Identifying Contexts of Distress in College Students' Reddit Posts: A Comparative Study of Classical NLP and Large Language Models

Carine Graff and Nikhil Krishnaswamy
Situated Ground and Natural Language (SIGNAL) Lab
Colorado State University
{carine.graff, nkrishna}@colostate.edu

Abstract

Mental health is a salient and growing societal concern among college students. Social media platforms such as Reddit offer a rich source of data regarding how students talk about their mental health, and NLP tools may potentially assist in identifying when a student is struggling. In this paper, we investigate how different NLP tools can be used to extract context surrounding college students expressions of distress. We construct a novel dataset from Reddit posts (*College Distress on Reddit*, or CDR), and examine the "classical NLP pipeline", and modern generative LLMs on this data. Our dataset exploration is conducted in parallel with, and contrasted against the Dreaddit dataset to examine cross-domain variation. Results show that standard or "classical" NLP tools extract a limited number of concrete entities, whereas generative models can infer more nuanced causes. However, LLMs struggle with knowledge extraction in specific content areas. Our work shows how important it is to be wary of LLMs, especially in mental health contexts.

Keywords: distress, mental health, generative AI

1. Introduction

Stress is often referred to as the "silent killer" (Balwan and Kour, 2021). It is particularly prevalent in academic environments, where it can trigger anxiety, depression, and abandonment of one's academic pursuits (Pascoe et al., 2019). According to a Gallup study, students report leaving higher education primarily to protect their wellbeing with 54% citing emotional stress and 43% citing mental health reasons.[1] Reported emotional stress doubled after the pandemic and remains high (Marken, 2024).

It is thus crucial to better understand what circumstances are detrimental to students' success during their college journey. Córdova et al. (2023) highlight that high levels of academic stress increase the probability of exhibiting signs of linguishing mental health, and the importance of identifying stressors to find solutions such as helping students develop coping mechanisms.

Amid these concerning trends, AI-powered tools have emerged as a potential coping mechanism among college students. Their effects can be positive or negative. Maples et al. (2024) report a participant whose Replika chatbot on several occasions prevented them from taking their own life, however, in February 2024, a teenager lost their life after conversing with a Character.ai chatbot (Yang, 2024). In Pesonen (2021)'s study, students trusted and were satisfied with the academic and non-academic support chatbot with a trust score of 71%. Young adults frequently use online forums both to cultivate a sense of community or belonging (Moore and Chuang, 2017), or for emotion regulation (Vermeulen et al., 2018), which makes them valuable platforms for observing expressions of distress. By understanding how students express stress and emotional strain on digital platforms, researchers can develop insights into how to better identify and respond to such cues in real time, and the extent to which automated tools can help.

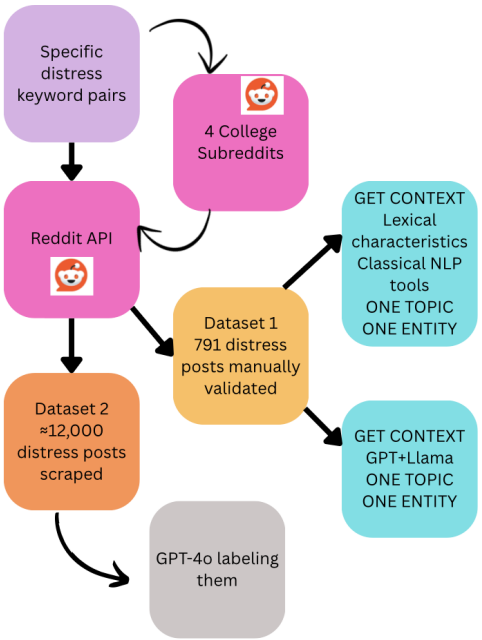

Figure 1: General Pipeline Diagram for CDR.

In this paper, we focus on detecting the context

[1] Gallup, 2024: State of Higher Education Report can be found on that page

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9657–9668
11–16 May 2026.

surrounding online expressions of distress among college students, who are a high-risk but underexamined population. We use Reddit posts as a data source to explore the context in which student expressions of stress occur. By comparing classical NLP methods, entity recognition via modern generative LLMs (GPT-4o and Llama 3) (Fig. 1), and examining the results of topics and entities overlap in GPT-4o and Llama 3, our work uniquely triangulates NLP paradigms to evaluate how different methods capture states of distress in student narratives (Fig. 2). This multi-method comparison invites reflection on the potential and limitations of standard NLP tools and LLMs in sensitive domains such as mental health.

The research questions are as follows:

- **RQ1**: What types of entities are most commonly associated with distress-related keywords in college students' Reddit posts and what patterns emerge in their surrounding contexts?
- **RQ2**: How do different types of NLP tools (e.g. spaCy NER vs. GPT-4o or Llama 3) differ in their ability to extract contextually relevant entities surrounding distress signals?
- **RQ3**: To what extent do different LLMs agree in the context they extract across datasets?

Our key contributions include:

- A novel dataset that concentrates on expressions of distress by college students specifically.
- Analysis of results bringing new insights into the feasibility and limitations of leveraging standard NLP tools and LLMs for mental health analysis.
- Insights into two LLMs (GPT-4o and Llama 3) and their capacity to extract knowledge from narratives compared to "classical" NLP techniques.

Our paper connects mental health and cognition to computational linguistics, NLP, and generative AI. It shows how different NLP paradigms interpret student distress and raises important considerations for using standard NLP tools and LLMs in mental health settings.

2. Related Work

2.1. Mental health on social media

Most studies focus on mental health disorders such as schizophrenia, depression, or suicide risk. For instance, Kim et al. (2020) built a deep learning model for detection of schizophrenia and other mental disorders in Reddit posts. In this paper, we

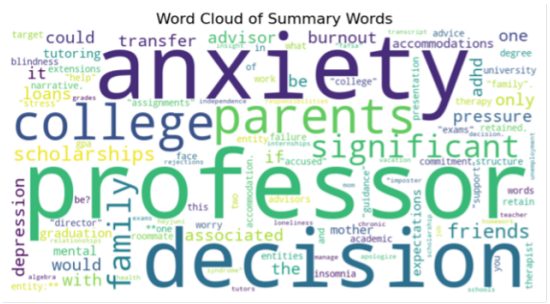

Figure 2: GPT-4o word cloud for one-word topic summary (CDR).

are interested in early detection, not necessarily of mental illness, but of states of distress that may deter students from their academic goals. In addition, the works related to this topic usually use one methodological approach, such as supervised classification or lexicon-based filtering. Li et al. (2020) and Cha et al. (2022) use symbolic methods to construct depression-specific lexicons, whereas Murarka et al. (2021) and Dinu and Moldovan (2021) apply transformer-based classifiers to Reddit posts labeled by subreddit or keyword. Others, such as Yang et al. (2023), explore the use of LLMs for mental health prediction and explanation. However, these studies rarely compare classical and generative approaches within the same pipeline. Moreover, few studies attempt to identify the causes of distress within students' social media narratives. One exception is the work by Dhanasekar et al. (2021), which introduces the Maxx chatbot to infer possible sources of distress based on conversational input, even though their evaluation remains limited in scope.

With all the social platforms available nowadays (Facebook, Tweeter, Reddit, online forums), more data is available to researchers especially in the domain of mental health. Nijhawan et al. (2022) analyzed tweets using Latent Dirichlet Allocation (LDA) to gather words and expressions that were common to those narratives and applied both ML and BERT models with strong results. Alhamed et al. (2024) introduced a novel dataset including Twitter posts from the same users before and after a diagnosis of depression to study linguistic changes when individuals undergo a change from one state to another. Those studies concentrate on chronic stress or mental disorders, such as depression, and use Tweeter as their main source. Other studies started to use Reddit, for instance Yates et al. (2017) created a dataset including self-reported depression: Reddit Self-reported Depression Diagnosis (RSDD).Turcan and McKeown (2019) with their Dreaddit dataset containing posts of several stress-related (PTSD, anxiety, etc.) subreddits cannot be ignored. Those datasets are available publicly and we will be using Dreaddit in this paper for

a comparison with our own dataset. Since 2017, CLEF eRisk has released datasets based on Reddit posts. Naderi et al. (2019) used the eRisk 2018 dataset on self-harm and anorexia with supervised learning based on bag of words and CNNs, as well as a hybrid model using all the methods, achieving low to moderate results. All these classifications are based on sentiment, linguistic characteristics, and use ML models.

2.2. Academic environment

Some works concentrate on academic settings, such as Oryngozha et al. (2024), who leveraged a Logistic Regression model based on bag of words trained with the Dreaddit dataset to explore stress in engineering and computer science academic communities in Kazakhstan. The overall stress level in academic texts was 29%, with the key emotions being sadness and fear. Moreover, the key factors contributing to this stress for students were studying, classes, professors, English skills, major IT companies, internships, and sleep. Other studies use surveys, not necessarily social media platforms. Our work diverges by applying automated NLP tools to student distress narratives on Reddit, aiming to determine their context.

2.3. Determining Context

Researchers have applied topic modeling to uncover themes included in mental health posts. For example, Franz et al. (2020) analyzed TeenHelp.org subforums entitled “Self-harm”, “Depression and Suicide”, and “Friends and Family” using both human annotators and LDA (from the "topicmodels" package in R. Their 15-topic model detected topics representing self-injurious thoughts and behavior (SITB) and related issues such as families, friends, and negative affect. Some other research incorporates symbolic knowledge and external information to improve context understanding. Gaur et al. (2018) mapped each subreddit to the Diagnostic and Statistical Manual of Mental Disorders 5th edition (DSM-5) category with a multiclass classifier. By leveraging medical ontologies, the model extracted meaningful context (symptoms, diagnoses) rather than treating posts as isolated text. Existing research highlights the importance of contextualizing mental health signals on social media platforms. Our work is inspired by existing approaches that use different methods, which we improve by applying a pipeline that uses classical and generative methods for a specific domain: college students’ states of distress on Reddit.

3. Method

3.1. Data Collection Strategy and Platform

The relevant social media platform used herein is Reddit. Reddit offers an anonymous space for individuals to express their concerns, as an alternative to confiding in a relative or sharing their thoughts in interviews. It includes texts that are longer than on most other social media platforms, such as Twitter, which would provide more context to students’ narratives than just one or two lines of Twitter posts (Turcan and McKeown, 2019). Additionally, according to Tejaswini et al. (2024), Reddit is the best social media platform to find data related to stress.

3.1.1. Datasets

Two Reddit-based datasets are used in this study:

- **Dreaddit:** The Dreaddit dataset from Turcan and McKeown (2019) includes ten subreddits: `ptsd`, `relationships`, `domesticviolence`, `homeless`, `almosthomeless`, `food_pantry`, `stress`, `anxiety`, `survivorsofabuse`, and `assistance`.
- **CDR:** CDR (*College Distress on Reddit*) is an original dataset created for this research, that focuses on expressions from distress from college students specifically. CDR enables us to examine how distress is articulated by college students in ways that may be different from the general population, such as in contexts that are unique to academic environments.

CDR Data Scope and Time Frame We focused on posts including specific keywords related to distress. The data containing the specific distress keywords ranges from April 2, 2003, to July 31, 2024. We concentrated on original posts (OPs) rather than comments, as that is where individuals usually post for help and hope for others to give their input. Several subreddits were used the titles of which pertained to college students’ experiences in higher education: `college`, `collegeRant`, `collegeadvice`, and `ApplyingToCollege`.

CDR Sampling Method and Filtering Critera The data collection focuses on posts containing words of distress. According to BetterHelp, distress includes a range of emotions and may present as mental anguish or physical symptoms.[2] Healthline defines mental distress as: “Any uncomfortable or unwanted emotions that come up when one experiences difficulties” (Tartakovsky, 2022).[3] While all

[2] BetterHelp definition of distress
[3] Healthline article on distress

distress (exclusively negative) is a form of stress, not all stress constitutes distress (Franks, 2023). To gather more evidence about which words are usually used in research on negative stress, we consulted the NRC Word-Emotion Association Lexicon: which contains 13,875 words classified into ten sentiments: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*, *positive*, *sadness*, *surprise*, and *trust*. That lexicon contains two files for each emotion, "sadness-NRC-EmoIntV1.txt" and "sadness-NRC-EmoIntV1-withZeroIntensityEntries.txt". We concentrated on the first listed file with non-zero intensity entries (on words with meaningful associations to specific emotions). For example, for the sadness emotion, "worried" has a sadness intensity of 0.62 and a fear score of 0.46. These show a strong association with emotional states related to distress. We used these scores to select terms such as "anxious," "cry," "stuck," and "overwhelmed" to ensure each word represents a measurable degree of emotional distress. Some words in the NRC Affect Intensity Lexicon such as "depressed" have a high score for example for sadness intensity (0.85) or 0.65 for "lonely", but we chose to include "worried" and "overwhelmed" because we not only took into account emotional salience, but also the context. "Worried" would be more often used in daily expressions compared to "depressed", which carries a clinical connotation. For instance also, "mournful" and "gloomy" are associated with sadness too, but they are less used in digital discourse. In addition, we checked if our keywords appeared often in social media discourse with the GoEmotions dataset that comprises Reddit comments labeled for twenty-seven emotions. We consulted the Feeling Words List as well from the book by Gilson et al. (2009). Yahya and Abdul Rahim (2023) also mention words that are usually found in texts that convey depression, such as "worried", "hurt", "anxious", "nervous", "cry", etc. Based on these considerations, we selected the following distress keywords: **help**, **cry**, **stress**, **overwhelmed**, **anxious**, **worried**, **stuck**, **scared**, **hard**, **struggling**, **sad**, and **unsure**. A simple search for indivdual words through the Reddit API was determined to likely overwhelm the results with false negatives, so we decided to search for *word pairs*: *help-stress*, *cry-stress*, *scared-stress*, *hard-stress*, *hard-struggling*, *help-sad*, *worried-unsure*, *worried-stuck*, *worried-stress*, *stuck-stress*, and *overwhelmed-anxious*.

CDR Data Collection - Reddit API Data was collected using the Reddit API web interface PullPush. We concentrated on original posts (not comments) in four specific subreddits: `college`, `collegeRant`, `collegeadvice`, and `ApplyingToCollege`. Looking for the specific word pairs gave us more chances of finding posts expressing distress. We scraped the Reddit API with Selenium and obtained around 12,436 posts total containing our keyword pairs. One annotator validated 791 posts containing the word pairs to make sure that the narratives were indeed conveying distress while trying to maintain a balance in the number of posts.[4] Given the size of the dataset and limited resources, we decided to use an LLM to classify the remaining data. To make sure the model could be used as a judge, we had two annotators (an expert and a non-expert) label fifty balanced samples. According to Calderon et al. (2025), fifty labeled samples represent a minimum requirement for assessing whether an LLM can be used as a judge. Among those fifty samples, annotators had a raw agreement of 92% on the 25 distress posts (coming from the initial 791 distress-annotated posts). The first annotator's labels were constant for these items, making Cohen's Kappa uninformative by design. This spot-check provides evidence of reasonable consistency between annotators and increases confidence in the distress labels assigned to the remaining posts.

We then used Calderon et al. (2025)'s approach (Calderon et al. (2025), Appendix C.2) to determine if an LLM is as good as a non-expert annotator by computing three metrics relative to the expert: - RMSE, which measures how far each annotator's (non-expert and the model) numeric prediction is from the expert's; ACC, which checks for exact categorical agreement; and SIM, which captures semantic or graded similarity when exact equality is not expected. These metrics were all computed for both, the LLM and the non-expert against the expert, allowing us to compare their mean scores. We then calculated the difference of means (Δmean) and tested its significance with a permutation test. The Δmean represents how much closer the LLM is to the expert than the non-expert:

- $\Delta_{\text{mean}} > 0 \Rightarrow$ LLM is closer to the expert
- $\Delta_{\text{mean}} \approx 0 \Rightarrow$ LLM and non-expert are equally close
- $\Delta_{\text{mean}} < 0 \Rightarrow$ Non-expert is closer to the expert

The results indicated that GPT-4o's performance (mean accuracy of 0.88, 95% CI 0.78-0.96) was comparable to the non-expert's (accuracy of 0.86, 95% CI 0.76-0.94). The mean difference between GPT and the non-expert was Δmean 0.02 with a paired-bootstrap 95% CI of -0.08-0.12). The winning rate showed that GPT performed as well as or better than the non-expert on 94% of items. Permutation tests on both the mean and the winning

[4] As a point of reference, in the Turcan and McKeown (2019) study, they annotated 3,500 posts, but there were five annotators for this task.

rate produced non-significant values ($p \approx 0.5$), indicating that there is no significant difference between GPT and the non-expert. The McNemar test, which examines paired disagreements on individual items was $n_{10} = 4$ (GPT right / NonExpert wrong), $n_{01} = 3$ (NonExpert right / GPT wrong), p=1.0, showing no difference in error patterns. All alignment metrics (Accuracy, SIM, -RMSE) and other tests such as permutation, bootstrap CI, and McNemar show that GPT's performance was statistically indistinguishable from the non-expert's. This supports treating GPT as a non-inferrior annotator, comparable to a non-expert. We thus decided to use GPT-4o for the rest of our data labeling.The LLM was given examples from the validated distress posts, as well as human-validated non-distress posts containing the distress keywords.

Data preprocessing and exploration To explore the context surrounding the keyword pairs in CDR, the human validated distress posts were used (hereafter we will refer to this as **CDR-D1**: 689 distress-human-validated posts—as opposed to non-distress-labeled—stripped from duplicates from 12,436 posts total). For Dreaddit, we randomly chose a balanced number of stress-labeled posts for each of the ten subreddits with a final number of 612 posts (hereafter referred to as **Dread612**).

Classical NLP tools We employed classical natural language processing (NLP) tools and generative methods to explore and extract knowledge from the collected data. The hypothesis would be that the main topics that stress students out are finances, housing, classes, or grades (Oryngozha et al., 2024). Before using classical NLP tools, the data was cleaned from special characters and stop words. We then applied the following tools over each keyword pair category in CDR-D1 and over each subreddit in Dread612:

- **YAKE!:** YAKE! (Campos et al., 2020, 2018a,b) is an unsupervised, domain-independent automatic keyword extractor designed to extract keywords from individual documents.[5]
- **NER:** Named Entity Recognition was performed on all the narratives per keyword pair / subreddit, using the spaCy `en_core_web_sm` model. Those entities could provide concrete anchors for distress.
- **Topic Modeling (LDA, SVD):** Latent Dirichlet Allocation (LDA) considers each post as a mix of topics and each topic is a distribution over words. It will learn these distributions while observing patterns of word co-occurrence across the dataset. It was applied to all the concatenated posts within a one keyword pair / subreddit with the Gensim `LdaModel`. Each topic was represented as a list of top-weighted words. We took the document formed by all the group narratives, computed TF-IDF and applied singular value decomposition to reduce the dimensionality using scikit-learn with `TruncatedSVD` and `TfidfVectorizer`. This gave us an approximation of semantic structure to identify dominant topics across the whole sample.
- **Parse Tree Analysis:** This analysis was performed for each keyword pair / subreddit with the NLTK CFG for defining the grammar and ChartParser, spaCy for pre-processing, POS tagging, and dependency information. We looked at context words and the top ten co-occurrences of the keywords and then proceeded with the word pairs (e.g., Fig. 3).

LLMs: GPT-4o and Llama 3 To examine differences between classical NLP tools and modern LLMs, we used two recent LLMs—GPT-4o and Llama 3—to extract the knowledge or context surrounding the distress target words in CDR-D1. GPT-4o was accessed via the OpenAI API and Llama 3[6] was accessed via HuggingFace and together.ai. Each LLM was prompted to extract the context surrounding the target words of distress for each narrative per keyword pair group. Then we asked it to extract topics and significant entities associated with these words. The LLM then had to give one and only entity that was predominant and one word that summarizes the topic related to the target words. The same procedure was then performed on Dread612, with the only change being that the target words were the subreddit titles (e.g. 'relationships'), maintaining consistency with our classical NLP experimental setup.

4. Results and Discussion

RQ1: To discover the types of entities that are most commonly associated with expressions of distress, we separated the extraction methods by topic or entity and compared them to each other. For topics, we computed the overall Jaccard similarity between the topic words extracted by LDA, SVD, YAKE!, GPT, and Llama (Fig. 4). Classical NLP tools, such as LDA, YAKE!, and SVD shared a modest overlap, and had little overlap with GPT and Llama. Meanwhile the LLMs shared a similar modest overlap between themselves, suggesting some shared semantic grounding but still with a notable and inconsistency in topic selection. Table 1

[5] https://pypi.org/project/yake/

[6] Specifically, `meta-llama/Llama-3.3-70B-Instruct`

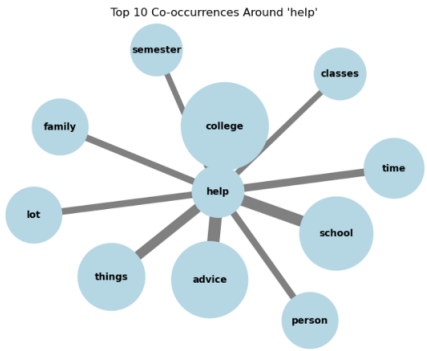

Figure 3: Top 10 co-occurrences around the word "help" (CDR-D1).

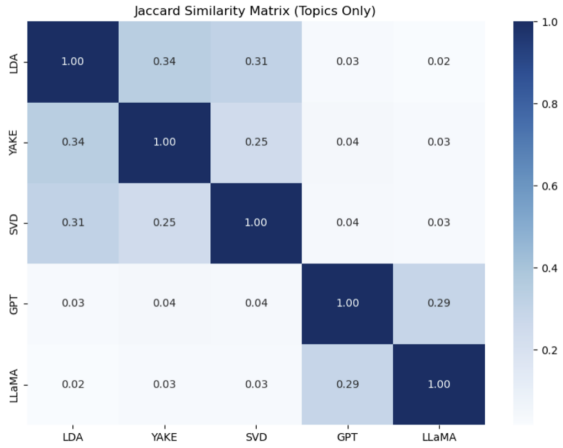

Figure 4: Jaccard Matrix for Topics (CDR-D1)

shows the top 10 topics for each method. Notably, "overwhelmed" appears in the top 10 selected topics by 4 of 5 methods. Classical methods appear more apt to extract "concrete" topics such as 'class', 'exam', and 'job' (i.e., distinct roles or events that cause stress) while LLMs more frequently extract abstract thematic or emotional context ('homesickness', 'burnout', 'regret').

Among the methods yielding entities (NER, Parse tree analysis, GPT, and Llama), NER shares almost nothing in common with LLMs (Fig. 5). While the top 10 co-occurrences from the parse-tree analysis show some overlap with LLMs (e.g., 'degree', 'failure'), GPT and Llama themselves share very few retrieved entities in common, further highlighting inconsistency across even modern techniques. Table 2 shows the top 10 entities for these methods.

For topics in Dread612, the highest agreement (0.35) is also between GPT and Llama. The ten topics common to all methods were ['abuse', 'anxiety', 'family', 'help', 'homeless', 'panic', 'relationship', 'shelter', 'stress']. Regarding entities, the aggregated Jaccard index between GPT and Llama reached 0.52, whereas other agreements are very low (e.g., 0.01 between NER/GPT and NER/Llama, 0.09 between Parse Tree Analysis/GPT and Parse Tree Analysis/Llama. Both datasets reveal human entities contributing to stress: 'friends', 'parents', or 'dad' and 'family'. While the LLMs show a moderate degree of overlap, Jaccard of 0.52 still signals substantial inconsistency and divergence in the entities they retrieve.

RQ2: To explore how different types of NLP methods differ in their ability to extract contextually relevant entities surrounding distress signals, we compared them to manually-labeled ground truth. The same expert annotator as mentioned in Sec. 3.1 annotated three distress samples from each keyword pair / subreddit in the respective dataset (33 samples total) for both Topics and Entities, and compared them to the topics and entities extracted by the different methods.

LDA	YAKE	SVD	GPT	Llama
also	stuck	much	fear	balance
going	unsure	also	balance	application
school	overwhelmed	go	application	overwhelmed
anxious	school	help	overwhelmed	regret
class	class	stuck	architecture	burnout
even	anxious	exam	burnout	homesickness
semester	years	overwhelmed	homesickness	overload
know	job	going	transition	depression
year	major	degree	regret	isolation
college	scared	got	overload	stress

Table 1: Top extracted topics across different methods (LDA, YAKE, SVD, GPT, Llama) for CDR-D1.

Metrics We converted all topics and entities into binary vectors and computed a Kappa (κ) score with a relaxed constraint—using WordNet to catch synonyms—between those binary vectors. This was computed per narrative, given one topic and one entity per narrative. Table 3 shows the results. We also employ Jaccard index to quantify overlap between sets, and cosine similarity (computed using SentenceTransformer `all-MiniLM-L6-v2` as a semantic similarity measure).

Discussion For Llama in CDR-D1 (Table 3), the average κ between all keyword pair groups was 0.79 for entities and 0.65 for topics, indicating substantial agreement between the human and LLM annotations. For GPT, average κ was 0.30 for entities and 0.61 for topics, indicating that GPT displays substantial agreement with humans on topics, but struggles with entities. When compared to the manual annotations, all methods provide very low relaxed Jaccard indices with the manual annotations.

The semantic content of the gold-standard and the SVD, YAKE!, and LDA sets are nearly completely disjunct, indicating that these methods capture different conceptual themes. SVD alone or LDA alone are not suitable for meaningful topic de-

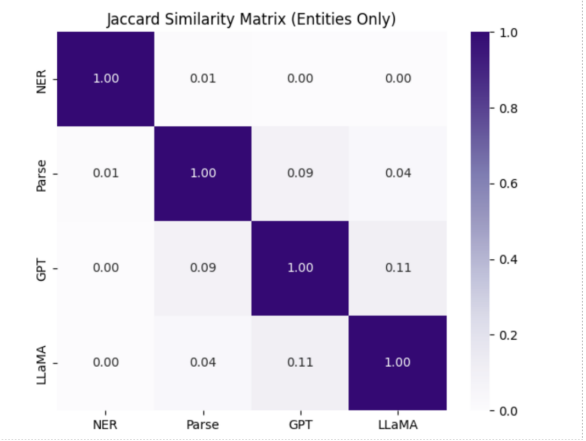

Figure 5: Jaccard Matrix for Entities (CDR-D1)

NER	Parse	GPT	LLaMA
spanish	class	degree	boyfriend
21f	rest	debt	roommate
second	degree	explosion	nyu
year	math	enrollment	teenager
week	life	insane	student
first	side	adhd	journalist
english	planet	medication	daughter
day	failure	future	mom
tomorrow	reason	financial	ucla
years	chest	failure	california

Table 2: Top extracted entities across methods (NER, Parse, GPT, LLaMA) for CDR-D1.

CDR-D1	κ	Jaccard	Cosine	Dread612	κ	Jaccard	Cosine
Topics				**Topics**			
SVD	–	0.05	0.48	SVD	–	0.05	0.31
LDA	–	0.05	0.45	LDA	–	0.04	0.28
YAKE!	–	0.04	0.51	YAKE!	–	0.04	0.31
GPT	0.61	0.10	0.64	GPT	0.50	0.68	0.83
Llama	0.65	0.14	0.70	Llama	0.16	0.28	0.61
Entities				**Entities**			
NER	–	0.01	0.31	NER	–	0.00	0.18
Parse	–	0.03	0.63	Parse	–	0.04	0.28
GPT	0.30	0.04	0.56	GPT	1.00	0.47	0.75
Llama	0.79	0.20	0.44	Llama	0.55	0.34	0.70

Table 3: Comparison of methods across Relaxed Kappa, Relaxed Jaccard, and Cosine metrics for CDR-D1 [L] and Dread612 [R].

tection in emotionally rich and implicit contexts like mental health posts. Standard NER taggers like spaCy are usually trained to recognize named entities, not abstract concepts. The parse tree analysis may extract more general terms and is not capturing the semantically important concepts that humans labeled as entities. GPT has a very low overlap and Llama's score indicates some semantic alignment, but the low scores show a gap between model reasoning and human understanding. Even though the relaxed Jaccard overlap scores were low, some cosine similarity values (≈ 0.51–0.63) reveal moderate-to-strong alignment in semantic space, close to the LLMs' scores, although it should be noted that most pretrained transformers have a tendency toward anisotropy in the embedding space (Ethayarajh, 2019) and so positive cosine similarity values may not be indicative of true similarity.

An identical analytical pipeline was applied to the Dreaddit subdataset (Table 3[R]). Overall, the numbers for all methods are lower for the CDR-D1 dataset compared to the Dread612 dataset (Table 3[R]). When looking at the Kappa score, LLMs seem to have moderate to strong agreement with humans for both datasets, except for GPT's CDR-D1 entities. However, while κ is computed separately for each model-human pair, the gap in values for each model (e.g., GPT's κ (0.30) and Llama's κ (0.79) for CDR-D1 entities) highlights inconsistent alignment with human annotations across models, reflecting disagreement in their interpretive patterns.

Summary Given the relaxed Jaccard overlaps, the topic and entity extraction methods—including LLMs—have more difficulties extracting knowledge in the CDR-D1 dataset, specific to college students and containing specific distress keyword pairs. The gap in κ values suggests inconsistency in how different models capture topics and entities.

RQ3: To determine to what extent different LLMs agree in the context they extract across datasets, we:

- analyze the affective content of both datasets
- compute the Jaccard overlap across outputs and assess whether these affective characteristics help explain differences in LLM alignment, as measured by Jaccard overlap across outputs

This approach allows us to test whether emotionally charged or neutral texts yield different agreement patterns between models.

Emotional language in CDR vs. Dreaddit Before digging deeper into the LLMs' knowledge extraction, we wanted to see how the Dreaddit dataset compared to CDR. Dreaddit already came with metrics from the Linguistic Inquiry and Word Count (LIWC), which is a well-known psycholinguistic dictionary and commercial product accessible via purchased license, such as `lex_liwc_affect`, `lex_liwc_social`, `lex_liwc_anx`, `lex_liwc_sad`, and `lex_liwc_family`, etc. `lex_liwc_affect` measures the proportion of words related to emotional processes and includes both positive and negative emotions. Examples would be "happy", "sad", "love", or "angry". `lex_liwc_social`

assesses the frequency of words pertaining to social processes and interactions, such as "friend", "talk", or "we". It reflects engagement in social contexts. `lex_liwc_anx` refers to expressions of anxiety or apprehension such as "worried", "nervous", or "afraid". It quantifies the use of words associated with emotions linked to anxiety. `lex_liwc_sad` captures language indicating sadness ("cry", "grief", or "depressed"). `lex_liwc_family` counts words related to family members or family relationships ("mother", "father", or "sister"). [7]

However, since we did not have access to LIWC, we reconstructed affect and social categories based on emotional words relevant to students (stress, overwhelmed, etc.) and social actors in academic settings (family, professor, roommate, professor, etc.) For the sentiment, we used VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a lexicon and rule-based sentiment analysis tool for social media and short informal texts. VADER uses a dictionary of about 7,500 words and phrases that have been labeled with a sentiment valence (positive or negative intensity). It will adjust sentiment using negation, intensifiers, punctuation, and capitalization. It returns a compound sentiment score between –1 and +1. We use that score as the sentiment metric for each post in our dataset. Posts with high affect and negative sentiment have a strong distress signal. Posts with neutral or positive sentiment even with high affect might be reflective or coping narratives. Posts with low affect and low social ratio may be more dispassionate factual narratives. We then computed the affect ratio and social ratio:

$$\text{Affect ratio} = \left(\frac{\text{number of affect words}}{\text{total words}}\right) \times 100 \quad (1)$$

$$\text{Social ratio} = \left(\frac{\text{number of social words}}{\text{total words}}\right) \times 100 \quad (2)$$

Lexicon-based methods such as VADER are known to be sensitive to domain and register. We thus interpret the sentiment scores as indicative signals rather than definitive measures of emotional content. Fig. 6 shows the results of this analysis. It seems that Dreaddit posts appear to have more emotional vocabulary. For instance, the narratives include PTSD, abuse, and trauma contexts and mention social actors more often such as families, partners, or friends. The posts in our dataset have a less negative emotional tone on average. That may indicate that our dataset is more topic-specific such as grades, time management, which is stressful, but not always described in deep emotional language. Dreaddit includes severe life situations, which leads to higher affect and social word ratios. As expected, BERT-based models achieve higher F1 scores when trained and evaluated within the same domain, while cross-domain evaluation exhibits degraded performance, indicating domain shift between datasets.

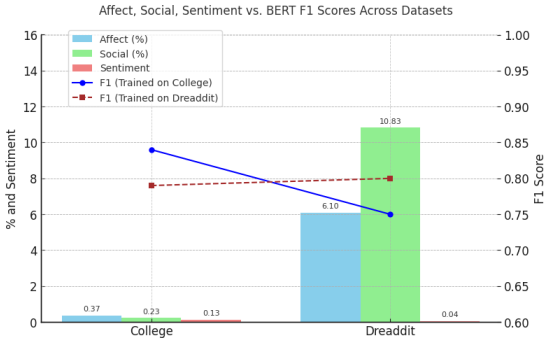

Figure 6: Affect and Social Percentages, Sentiment score, and F1 scores for the CDR and the Dreaddit datasets.

Jaccard scores To what extent do LLMs agree when extracting content across datasets? To explore GPT-4o and Llama 3's overlap in extracting topics and entities, we computed the average Jaccard scores per keyword pair for CDR-D1 and per subreddit for Dread612. For the topics in CDR-D1, the results show that the LLMs had the lowest overlaps in the 'worried-stuck' and 'scared-stress' keyword pair groups (0.18). For 'stuck-stress' the overlap between GPT and Llama was zero. 'overwhelmed-anxious' only yielded a negligible overlap of 0.09. The highest overlaps in topics were for 'hard-struggling' and for 'stuck-stress' with a score of 0.43. This moderate overlap indicates that similar circumstances may be expressed (i.e., experiences that evoke emotions may also contribute to struggles). For entities, the overlap for 'worried-stuck' was negligible (0.06). It was low for 'hard-stress' and 'cry-stress' (0.18 and 0.12), and 0.00 for 'stuck-stress'. The highest score was for 'help-stress' with 0.25, which is considered as low. For Dreaddit's topics, the overlaps were low for the 'relationships' and 'anxiety' subreddits (0.11). The highest overlaps were for 'stress' and 'ptsd' with 0.42 (moderate). The lowest entity overlap was also on the 'anxiety' subreddit (0.17). The highest overlap was for 'survivorsofabuse' with 0.81.

Summary Overall, the overlap results are lower in the CDR-D1 dataset. Given those results and as CDR is more specific than Dreaddit, (it is geared towards college students, it contains specific distress words, and it contains less overt emotion), LLMs are struggling to agree on the topics and entities surrounding our distress keywords. Dreaddit is more general and may be more rich emotionally, as previously discussed, making it perhaps easier for LLMs to extract knowledge around states

[7] See LIWC-22 documentation

of distress in that context, because the overt emotional language serves as a stronger indicator of important information.

5. Conclusion and Future Work

In this paper, we used standard NLP tools and LLMs to extract topics and entities surrounding keyword pairs of distress or stress subreddits in Reddit posts to explore how these methods perform in a specific dataset pertaining to college students (CDR) and a more general stress dataset (Dreaddit). The common entities found by both classical NLP tools and generative techniques yield similar results: entities extracted from college students' Reddit distress narratives reveal that students' human relationships are stressful for students. Being accepted, feeling like they belong, relationships with family members, pressure from relatives to succeed, etc. These are all stressors for college students. It also seems that similar stressors exist for the Dreaddit dataset. Classical NLP methods such as SVD, LDA, YAKE!, NER, or parse-tree analysis are not able to extract contextually rich topics or entities related to keywords of distress, unlike LLMs. However, even though the Kappa score was generally higher for CDR, low Jaccard overlap between GPT and LLama shows they disagree on what co-occurs contextually with the extracted entities or topics. One novel contribution of this paper is our unique CDR dataset, which enabled to show that compared to a more general stress dataset such as Dreaddit, LLMs' negligeable to low overlap on certain posts containing specific keywords of distress indicate they should be used with caution in mental health settings. Other methods for topic extraction could also be tested and perhaps a combination of methods could strengthen topic and entity extraction in distress narratives containing specific pairs of distress keywords. Our project contributes to the development of ethically aware and emotionally intelligent AI tools by raising awareness about the risks of overreliance on LLMs especially in the mental health domain. The code and data for this project can be found at https://github.com/SignalLab-CG/CDR_Reddit.git.

Limitations

Maintaining a good balance in the number of posts collected depending on the word pairs chosen was not guaranteed, as some word pairs were more present in some subreddits than others.

Temporal limitations Reddit's API limits to ten queries per minute for free when one is not logged in to a Reddit account. Some subreddits also only started existing at a certain date, and limiting the post search to a month, for example, did not yield enough results. We kept the range from April 2, 2003, to July 31, 2024, as looking at a more recent time only would not produce enough posts.

Sampling bias The Reddit posts in CDR are not representative of the general population, especially for the data collected here, it only represents a certain type of users from a certain age group (college students). As a result, overreliance on models trained in this domain could lead to misinterpretation of distress signals when applied to other populations. As people posting on Reddit are willing to share their thoughts publicly, that targets a certain type of personality, favoring more vocal or extreme viewpoints. In addition, some posts might get deleted by moderators or by users, which means some data might be missing from the API in the first place. This study focuses only on English data and does not concentrate or explore differences between cultures, who may express themselves less directly about their state of distress.

Prompt Sensitivity An important limitation of using LLMs is prompt phrasing. Small changes in wording, formatting, or example order can lead to differences in model output. This introduces challenges in tasks involving subjective interpretation such as entity extraction. The inherent variability in LLM behavior may affect the stability of results. Also, results obtained from API-based LLMs are subject to temporal variability. Even when using the same model names, outputs can differ across runs due to stochastic sampling, backend infrastructure changes, and silent model updates introduced by the provider. Consequently, exact replication of results at a later date may not be guaranteed, although trends and aggregate patterns are expected to remain stable.

Ethics Statement

Reddit is an anonymous public platform, but scraping data on Reddit has caused some issues in the past. Large-scale scraping, especially by companies intending to use the data to train LLMs often violated Reddit's terms of use—so much so that the API was no longer available for some time. Reddit then took action and changed some of the accesses to their API. Since June 2023, mainly for heavy usage, the API has been monetized. It is still free of charge for ten queries a minute without using authentication and one hundred queries a minute when using authentication (spez, 2023). The Reddit API page rules mentions that individuals have to respect Reddit's usage limits and its terms of use. When reading the API's terms of use, it is clear that API users are not allowed to train an AI model

without the consent of the "rightsholders in the applicable User Content" (Reddit, 2023). And further on: "If you are interested in using the Data APIs for commercial purposes, research in excess of rate limits, or for any use that is not expressly permitted under the Data API Terms, then you will need to enter into a separate agreement with Reddit." To ensure that we could use Reddit's data, we wrote to Reddit, who answered that for the specific case of a survey, one should contact the moderators of the subreddit. As the current project did not contain any surveys and as the Reddit API was available to query posts, we went ahead with our study. Even if users post on Reddit without disclosing their real names, for the current project, to avoid any issues, Reddit usernames were erased. In addition, we consulted with our local institutional research ethics board who determined that no human subjects research protocol was necessary for this study and that we could proceed with the research.

In general terms, AI handling of mental health and related issues should be treated with great care. There are an increasing number of reports of interaction with AI adversely affecting the mental states of people undergoing stressful situations or who have underlying mental health issues.[8] Our study further reinforces that NLP techniques, including classical methods and LLMs, handle distress contexts inconsistently, and are sensitive to changes in the specific of the data and domain (viz. college students represented by CDR and general internet posters represented by Dreaddit), and should therefore not be assumed to have the capability to act as automatic filter or detector.

Acknowledgments

We thank Dr. Labiba Jahan and Dr. Frank Coyle from SMU for their input and discussions on this topic, as well as our annotator for their assistance with the data annotation and the anonymous reviewers whose feedback helped improve this paper.

Bibliographical References

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260, Torino, Italia. ELRA and ICCL.

Wahied Khawar Balwan and Sachdeep Kour. 2021. A systematic review of hypertension and stress—the silent killers. *Scholars Academic Journal of Biosciences*, 6:150–154.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. European Language Resources Association (ELRA).

Ricardo Campos, Vítor Mangaravite, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018a. A text feature based automatic keyword extraction method for single documents. In *Advances in Information Retrieval (ECIR 2018)*, volume 10772 of *Lecture Notes in Computer Science*, pages 684–691, Grenoble, France. Springer.

Ricardo Campos, Vítor Mangaravite, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018b. YAKE! collection-independent automatic keyword extractor. In *Advances in Information Retrieval (ECIR 2018)*, volume 10772 of *Lecture Notes in Computer Science*, pages 806–810, Grenoble, France. Springer.

Ricardo Campos, Vítor Mangaravite, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Junyeop Cha, Seoyun Kim, and Eunil Park. 2022. A lexicon-based approach to examine depression detection in social media: The case of twitter and university community. *Humanities and Social Sciences Communications*, 9(1):1–10.

Pamela Olivera Córdova, Patricia Gordillo Gasser, Hernán Mejía Naranjo, Isabel Taborga La Fuente, Alberto Chacón Grajeda, and Alberto Unzueta Sanjinés. 2023. Academic stress as a predictor of mental health in university students. *Cogent Education*, 10(2):2232686.

Varshaa Dhanasekar, Yenugu Preethi, S Vishali, I R Praveen Joe, and Booma Poolan. 2021. A chatbot to promote students' mental health through emotion recognition. In *Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1412–1416. IEEE.

Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic detection and classification of mental illnesses from general social media texts. In *Proceedings of the International Conference on*

[8] https://www.bbc.com/news/articles/cgerwp7rdlvo

Recent Advances in Natural Language Processing (RANLP 2021), pages 358–366, Held Online. INCOMA Ltd.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Sharon Franks. 2023. Distress vs eustress. American Psychological Association. Accessed June 2025.

Peter J. Franz, Erik C. Nook, Patrick Mair, and Matthew K. Nock. 2020. Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide and Life-Threatening Behavior*, 50(1):5–18.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "Let me tell you about your mental health!": Contextualized classification of reddit posts to DSM-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, pages 753–762, Torino, Italy. Association for Computing Machinery.

Mark Gilson, Arthur Freeman, Marisa J. Yates, and Sarah M. Freeman. 2009. *Overcoming depression: Workbook*. Oxford University Press. Oxford University Press eBooks.

Jiyoung Kim, Jaewook Lee, Eunil Park, et al. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10:11846.

Genghao Li, Bing Li, Langlin Huang, and Sibing Hou. 2020. Automatic construction of a depression-domain lexicon based on microblogs. *JMIR Medical Informatics*, 8(6):e17333.

Brian Maples, Merve Cerit, Arun Vishwanath, et al. 2024. Loneliness and suicide mitigation for students using GPT-3-enabled chatbots. *NPJ Mental Health Research*, 3(4).

Stephanie Marken. 2024. Mental health, stress top reasons students consider leaving. Gallup News. Accessed June 2025.

Carrie Moore and Lisa Chuang. 2017. Redditors revealed: Motivational factors of the reddit community. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, pages 1923–1932. University of Hawai'i at Mānoa. Accessed: 2025-06-05.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (LOUHI 2021)*, pages 59–68. Association for Computational Linguistics.

Nona Naderi, Julien Gobeill, Douglas Teodoro, Emilie Pasche, and Patrick Ruch. 2019. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum*, Lugano, Switzerland. CEUR Workshop Proceedings.

Tanmay Nijhawan, Gopal Attigeri, and Thippeswamy Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9:33.

Nazzere Oryngozha, Pakizar Shamoi, and Ayan Igali. 2024. Detection and analysis of stress-related posts in reddit's academic communities. *IEEE Access*, 12:14932–14948.

Michaela C. Pascoe, Sarah E. Hetrick, and Alexandra G. Parker. 2019. The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth*, 25(1):104–112.

Joonas A. Pesonen. 2021. "Are you OK?" Students' trust in a chatbot providing support opportunities. In *Proceedings of the International Conference on Learning and Collaboration Technologies: Games and Virtual Environments for Learning*, pages 199–215, Cham. Springer International Publishing.

Reddit. 2023. Data API Terms. Accessed 2025-06-07.

spez. 2023. Addressing the community about changes to our API. [Online forum post]. Reddit. Posted June 9.

Margarita Tartakovsky. 2022. How to recognize emotional distress, plus 5 tips to help you cope. `https://www.healthline.com/health/mental-health/emotional-distress`. Healthline, May 13.

Vankayala Tejaswini, Sathya Babu, and Bibhudatta Sahoo. 2024. Depression detection from social media text analysis using natural language processing techniques and a hybrid deep learning model. *ACM Transactions on Asian and*

Low-Resource Language Information Processing, 23(1):1–20.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Anne Vermeulen, Heidi Vandebosch, and Wannes Heirman. 2018. Smiling, Venting, or both? adolescents' social sharing of emotions on social media. *Computers in Human Behavior*, 84:211–219.

Nur Hidayah Yahya and Hazlina Abdul Rahim. 2023. Linguistic markers of depression: Insights from english-language tweets before and during the covid-19 pandemic. *Language and Health*, 1(2):36–50.

Angela Yang. 2024. Character.ai sued after florida teen's suicide, as parents blame chatbot for death. NBC News article. Published May 3, 2024.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.