Analysis of Stress, Anxiety, and Depression among students using Reddit community data

Prahathish Kameswaran 115062056
Roopak Narayanasamy 114941190
Soundarya Venkatesh 114711711

Introduction

The mental health of university students is a major concern. Stress, depression, anxiety, and other mental health concerns can have a significant impact on a student's academic performance and overall well-being. This is why it is crucial to understand the factors that contribute to these problems and find ways to address them. A study shows that 8 out of 10 students get stressed, depressed, or feel anxiety during their time at university[1]. This could lead to insomnia, cognitive deficit, mood swings, and even physical illness. Some of the major stressors for students are examinations, peer competition, finance, or personal issues.

In terms of Sustainable Development Goals (SDGs), our project primarily focuses on SDG 3: Good Mental Health and Well-being. We are contributing to the overall goal by finding insights on mental health issues among students to help universities understand and come up with countermeasures. For example, universities can organize events and workshops during the time when students feel the most stressed during the academic year. However, our work also has implications for other SDGs, such as SDG 4: Quality Education. By finding insights on mental health issues in universities, we are helping to create a better learning environment for students, which could lead to improved academic performance and better career prospects.

Background

The domain of mental health has received a lot of attention over the years and continues to garner more. This is mainly due to the easy access to people's thoughts and ideas all over the internet. We have taken inspiration from some of these works to build our project. Bagroy et al's[4] "A Social Media Based Index of Mental Well-Being in College Campuses" uses a transfer learning based classification approach to analyze students' posts on Reddit, and develop a mental well-being index that tells meaningful patterns of mental health on campuses. The key takeaway from this work is that we try to use this as our reference in how we distinguish the training data sources as mental health-related or not.

Next, we looked into "Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses"[5] by Saha et al. This paper explores the use of social media data to model stress and emotional responses to incidents of gun violence on college campuses before and after the occurrence of a gun violence event. This gave us the idea to do a temporal analysis of stress over time to unearth any patterns or trends emerging over the course of a semester or through the pandemic.

Data

We scraped Reddit for all of our data needs. Specifically, we had to collect 3 sets of data.

- University Subreddit data - We collected the posts from 50 university subreddits to further analyze for indications of any mental health concerns.
- Mental Health data - We collected the posts from various Mental health communities like r/Stress, r/Anxiety, r/depression, etc to train our model on features resembling posts of users with mental health issues.

- Non-mental health data - We used the 2021 Reddit dataset from HuggingFace for our control data. We filtered out the posts that were from either the mental health or university subreddits.

We used a combination of PushShift API and the HuggingFace dataset for this task. We parsed the data and saved it in the same format for ease of use. We separated the fields using DELIM to avoid misidentifying a “,” in the text as a delimiter.

Data	Source	Size	No. of records
Mental health data	PushShift API	6 GB	7,833,425
Non-mental health data	HuggingFace Datasets	10 GB	12,733,772
University subreddit data	PushShift API	1.5 GB	2,718,329

The data had several fields, of which we were interested in - created_utc, subreddit, title, selftext, score, over_18, upvote_ratio and is_video. The title and selftext are the main fields that hold the content that we want for sentiment analysis. The other fields are used to maintain context or filter data.

Methods

Our pipeline has the following stages as we see in Fig 1: Collection, Preprocessing, Sentiment Analysis, Score analysis.

1) **Collection**: We collected the data using PushShift API and the HuggingFace Reddit dataset. We have described it in detail in the previous section.
2) **Preprocessing**: We then preprocessed the data in PySpark. We split the record with the DELIM and then concatenated the title and selftext, as the title by itself did not have context in most of the cases. Then we removed the URLs, punctuation and stop words, and lemmatized the words to reduce them to their root forms. We did this with the help of the nltk library. Since Reddit is a social media platform, there was also a wide usage of emojis. We converted the emojis to their text form using the emoji library and removed them. Finally, we converted the text to lowercase and returned the tokens.
3) **Sentiment Analysis**: For the sentiment analysis, we used a pre-trained DistilBert model[2][3] and added a linear layer(768, 1) on top of it to do transfer learning. We created a dataset with the posts from the Mental Health dataset labeled as 1 (indicative of mental health concern) and the posts from the Non-Mental Health Dataset labeled as 0. We trained the model with a 0.8 split of this data and we got the following correctness measures in the test dataset: **Accuracy: 0.9101, Precision: 0.8931, Recall: 0.9316, F1-Score: 0.9120.** We then used this model to predict the mental health sentiment scores of the university posts to do further analysis.

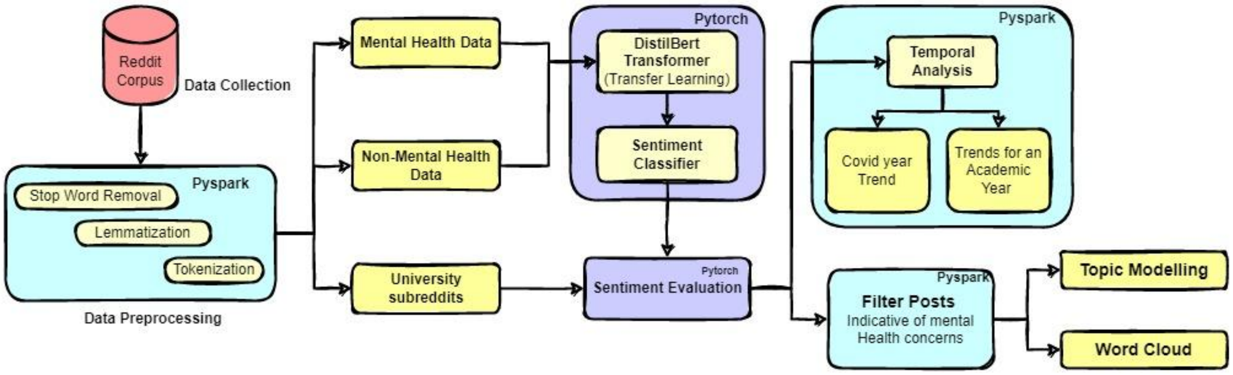


Fig 1: Model Implementation flowchart.

The model used the following hyperparameters to train the model:

batch_size = 64, max_length = 512, learning_rate = 16e-4, num_epochs = 1, dropout = 0.1

We used the Adam optimizer to only apply the gradient to the model's linear layer parameters. We used Binary Cross Entropy loss as the criterion and the StepLR scheduler with a gamma rate of 0.5 to decay the learning rate over time.

4) **Score Analysis**: This stage is of 3 steps.
   a) Temporal analysis - We grouped all the university post scores by their timestamp to find any temporal patterns in the data over the academic year or through the pandemic.
   b) Topic modeling: We then performed topic modeling using the Latent Dirichlet Allocation using the sklearn library to find recurring topics that are closely associated among posts that scored high in the sentiment analysis.
   c) WordCloud: Finally, to give a visual representation of potential causes for students' distress, we calculated the word frequencies among the posts that scored high in the sentiment analysis and generated a word cloud using the wordcloud library.

Evaluation/Results

Our analysis gave us the following results:

1. **Overall stress score for the universities:**

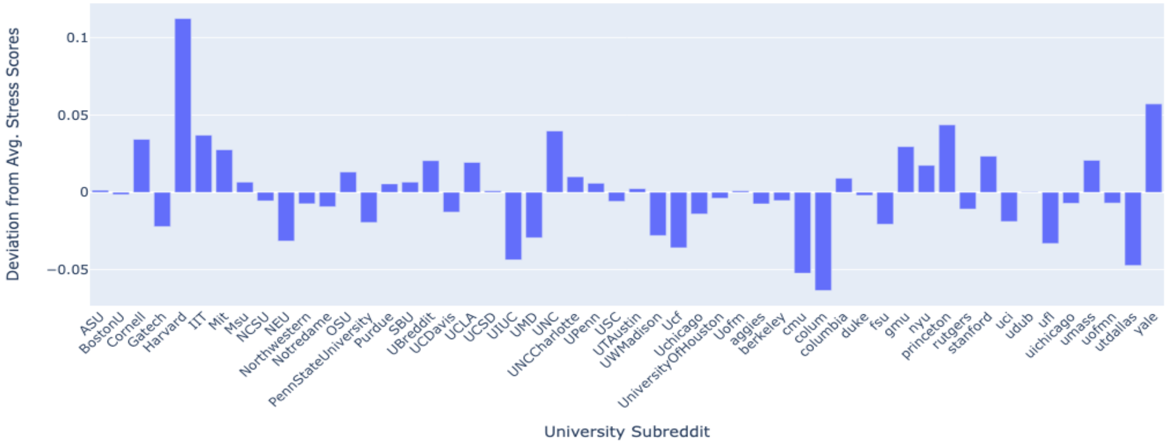


Fig 2: University level Average Stress Scores (avg = 0.39)

In Fig 2., we got the average stress score for each university subreddit. All of them averaged around **0.39** with slight differences, as shown above. This was expected and is a good indicator of the overall temper of the community. We also observed that most of the Ivy League universities were on the higher spectrum.

1) Temporal analysis

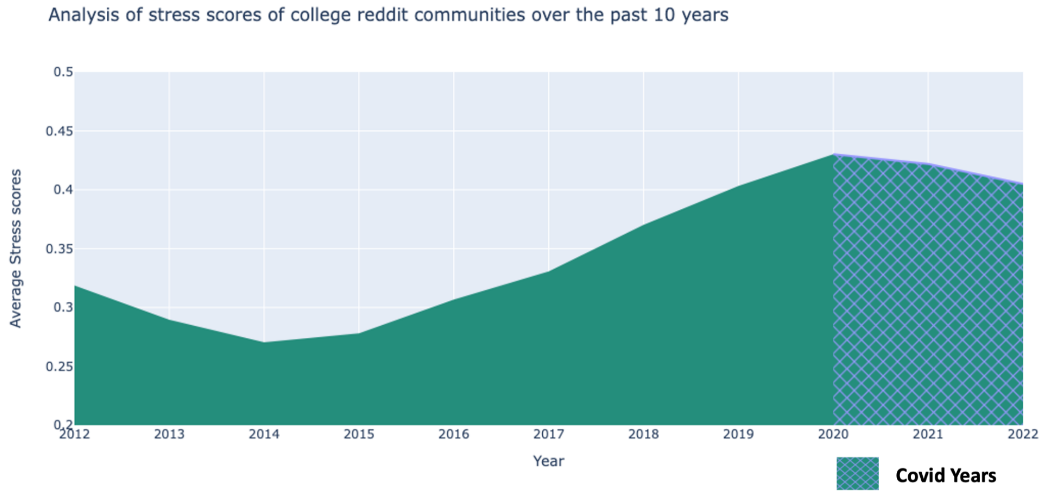


Fig 3: Analysis of stress scores of college Reddit communities over the past 10 years

In Fig 3., we grouped the posts based on the year to analyze how the stress score changed over the years. This graph indicates a slight peak in the stress score during covid years (cross-grid). This was what we had anticipated as the pandemic was a very stressful time for the students having to adapt to remote learning, not being able to find jobs, health scares, etc.

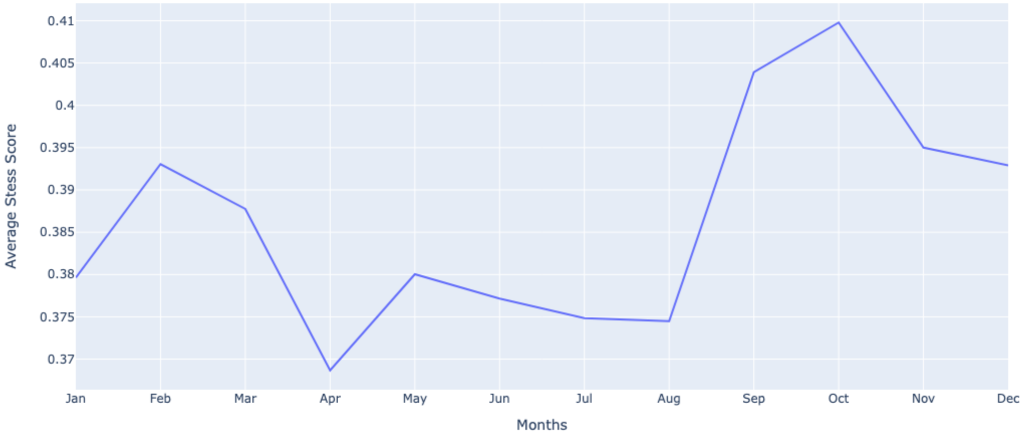


Fig 4: Temporal analysis over an academic year

In Fig 4., we grouped the scores for all the universities respective to the month and plotted a line graph to analyze the temporal pattern of students' stress levels in an academic year. We clearly see that during the semesters (January - April and August - December) students get too stressed. We could notice a peak in the stress score while the semester nears completion as students get stressed for their final exams. And during the summer break, we could see that the stress level among the students decreases.

2) Topic modeling:

We further filtered all the posts from the dataset that scored above the stress threshold (0.7) and ran the Latent Dirichlet Allocation algorithm on it to find the 10 most recurring topics or groups of closely associated words. We have tried to further interpret the groups as a topic. We think this can be avoided by using a more extensive stop-word list specialized for online datasets at the preprocessing stage.

Table 1: List of words grouped by LDA

Topic	Words
Housing	['looking', 'housing', 'need', 'advice', 'roommate', 'aid', 'student', 'financial', 'freshman', 'fall']
Social	['depression', 'campus', 'people', 'night', 'today', 'tonight', 'food', 'around', 'party', 'day']
Courses	['class', 'anyone', 'summer', 'taking', 'course', 'math', 'taken', 'online', 'take', 'know']
Advice	['anyone', 'like', 'transfer', 'hall', 'what', 'else', 'going', 'getting', 'chance', 'thought']
Admission	['student', 'campus', 'fuck', 'university', 'job', 'college', 'berkeley', 'state', 'purdue', 'ucla']
Dorm life	['dorm', 'campus', 'room', 'good', 'best', 'lost', 'place', 'know', 're', 'go']
Study	['major', 'study', 'schedule', 'science', 'engineering', 'program', 'course', 'exam', 'research', 'grade']
Help	['help', 'question', 'please', 'health', 'need', 'survey', 'hate', 'plan', 'graduation', 'mental']

3) Word Cloud

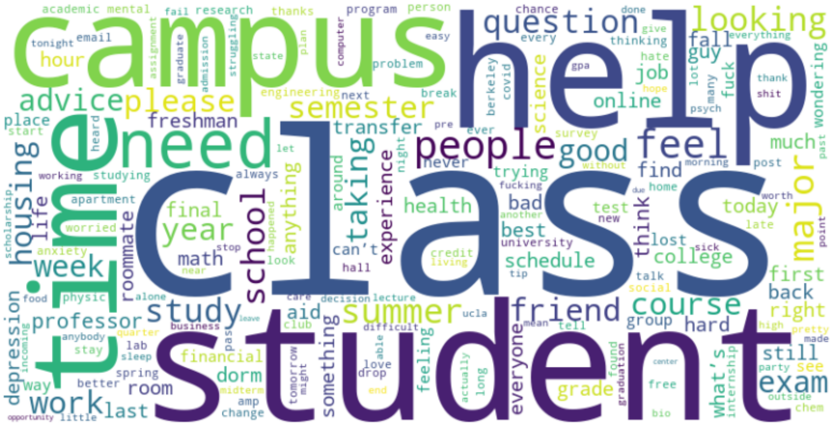


Fig 5: WordCloud for the most frequents words used in higher stress scored post in reddit

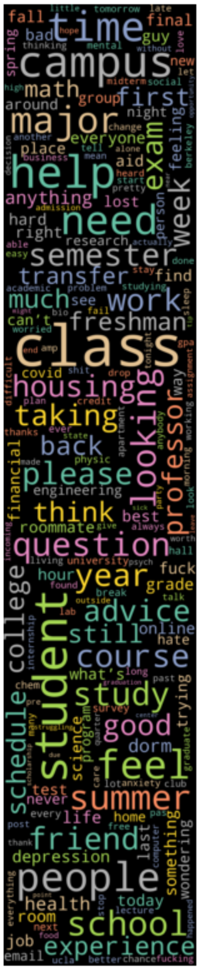
In Fig 5., it shows the most frequent words used in higher stress scored post in the reddit. From this WordCloud, we could clearly see that most things students get stressed about are campus, class, housing, exam, course, major, financial, final year, semester, advice, and many more.

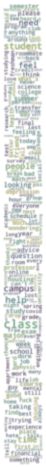
Conclusion

We analyzed students' stress levels over the years to find patterns and causes that might contribute to them. It can be seen that the stress levels increased over the pandemic. We also noticed that students' stress scores were the least over the holidays (especially summer). We also tried to find common topics that occur among high-scoring posts to find insights on what stresses students. Campuses can understand these factors that contribute to their students' stress levels and take measures to improve them. Universities can then use this information to develop targeted interventions, such as workshops or counseling services, that address the specific stressors identified by students. In addition to that, universities can also raise awareness about mental health and provide students with the resources they need to manage stress and anxiety. This can include offering free therapy sessions or support groups, providing access to mindfulness and meditation resources, or even implementing policies that promote a healthy work-life balance.

Reference

1. Asif S, Mudassar A, Shahzad TZ, Raouf M, Pervaiz T. Frequency of depression, anxiety and stress among university students. Pak J Med Sci. 2020 Jul-Aug;36(5):971-976. doi: 10.12669/pjms.36.5.1873. PMID: 32704273; PMCID: PMC7372668.
2. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter" by Sanh et al. (2019) for the DistilBERT model: https://arxiv.org/abs/1910.01108
3. "Attention is All You Need" by Vaswani et al. (2017) for the Transformer model: https://arxiv.org/abs/1706.03762
4. Bagroy, Shrey et al. "A Social Media Based Index of Mental Well-Being in College Campuses." Proceedings of the SIGCHI conference on human factors in computing systems. CHI Conference vol. 2017 (2017): 1634-1646. doi:10.1145/3025453.3025909.
5. Koustuv Saha and Munmun De Choudhury. 2017. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 92 (November 2017), 27 pages. https://doi.org/10.1145/3134727
6. Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.


little
tomorrow
late
fall
time
final
hope
bad
guy
spring
thinking
mental
without
love
campus
new
let
midterm
social
high
math
group
first
opportunity
around
night
major
change
feeling
another
everyone
berkeley
decision
place
tell
alone
business
mean
aid
help
exam
hear
heard
start
pretty
anything
lost
week
admission
person
hard
need
right
research
actually
able
easy
semester
done
stay
transfer
find
academic
problem
studying
much
see
work
sleep
might
bio
fail
can't
freshman
tip
worried
class
tonight
gpa
end
amp
difficult
covid
shit
drop
way
assignment
housing
looking
plan
credit
taking
professor
thanks
ever
apartment
working
state
anybody
back
financial
made
physic
morning
please
engineering
party
look
think
sick
best
incoming
roommate
give
always
leave
question
worth
hall
living
university
psych
fuck
college
hour
year
found
grade
break
talk
outside
student
lab
advice
internship
still
online
hate
course
what's
long
chem
past
graduation
study
trying
pre
survey
schedule
program
center
many
struggling
quarter
good
science
dorm
scholarship
due
feel
graduate
care
lot
anxiety
club
test
summer
never
every
life
home
pass
post
free
friend
last
computer
thank
something
wondering
depression
people
everything
point
stop
today
health
lecture
room
school
happened
next
food
job
experience
email
ucla
better
chance
fucking


semester
start
please
aid
see
heard
need
program
club
anything
look
around
student
studying
freshman
roommate
back
feel
credit
think
love
what's
never
half
science
college
summer
person
transfer
tomorrow
research
health
way
past
final
last
feeling
today
test
better
friend
change
amp
fall
chem
email
survey
people
social
break
bad
math
post
looking
depression
another
hour
group
academic
everyone
next
schedule
much
lot
wondering
everything
long
year
anxiety
right
found
plan
business
thanks
advice
question
room
every
professor
online
pretty
housing
university
can't
campus
problem
lost
help
might
spring
study
covid
drop
grade
class
free
exam
major
ever
worried
week
chance
school
engineering
dorm
apartment
hard
job
physic
work
many
life
lab
course
guy
mental
still
good
home
fuck
thinking
taking
always
find
best
trying
working
experience
place
hate
talk
stay
time
tell
first
financial
something