Team 1.8
r/cuboulder Subreddit Topic Modeling and Sentiment Analysis

Chris Goswick
Aria Jia
Maddie Wallace

Leeds School of Business
UNIVERSITY OF COLORADO **BOULDER**

Agenda

- Research Question & The Data
- Topic Extraction
- General Sentiment Analysis
- Deep Dive into Some Topics
- Limitations and Future Study

00 Introduction

r/cuboulder

An unofficial place for people to discuss the University of Colorado Boulder, ask questions about the university, meet other Buffs, and stay informed about relevant campus issues.

01

Research Questions

1. What are the most common topics and themes discussed on the CU Boulder subreddit?

2. How do they relate to the overall sentiment of the community?

3. Do the topics differ before and after covid pandemic (2020)?

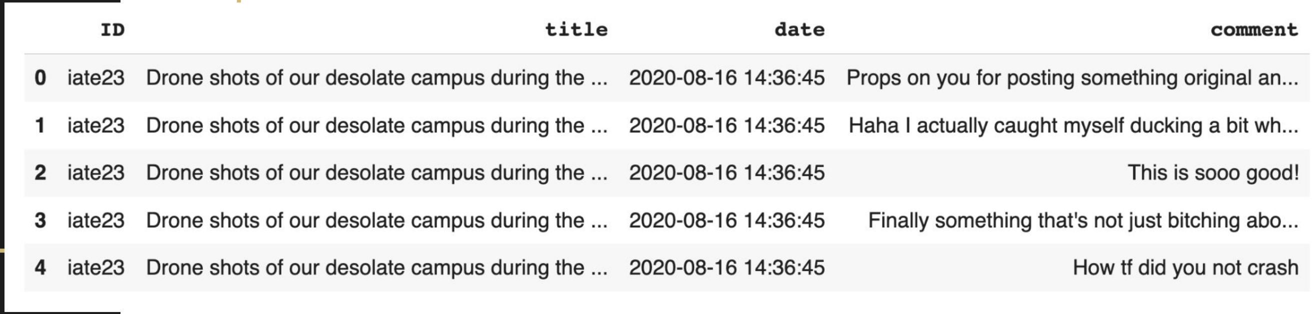
02

The Data

Data Details

Pull data through Python Reddit API Wrapper (PRAW)

1. Top 500 posts of r/cuboulder subreddit data (2016-2023)
   - A CSV file with columns: ID, title of the post, date, comments
   - Total rows: 21,313



	ID	title	date	comment
0	iate23	Drone shots of our desolate campus during the ...	2020-08-16 14:36:45	Props on you for posting something original an...
1	iate23	Drone shots of our desolate campus during the ...	2020-08-16 14:36:45	Haha I actually caught myself ducking a bit wh...
2	iate23	Drone shots of our desolate campus during the ...	2020-08-16 14:36:45	This is sooo good!
3	iate23	Drone shots of our desolate campus during the ...	2020-08-16 14:36:45	Finally something that's not just bitching abo...
4	iate23	Drone shots of our desolate campus during the ...	2020-08-16 14:36:45	How tf did you not crash

Data Cleaning

1. Remove all “[deleted]” or “[removed]” comments
2. Remove stop words
3. Remove URLs, mentions, non-ascii characters, extra spaces, and hashtags in the text
4. Split into two dataset of timeframe 2016-2020 and 2020-2023

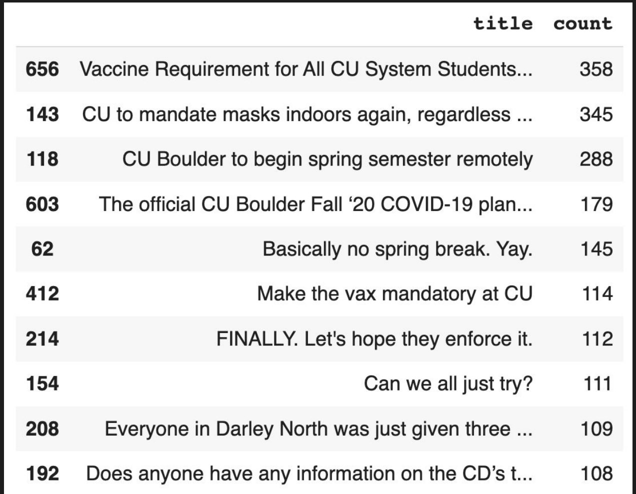
03

Initial Exploratory Analysis

Posts with the highest comments count

After 2020

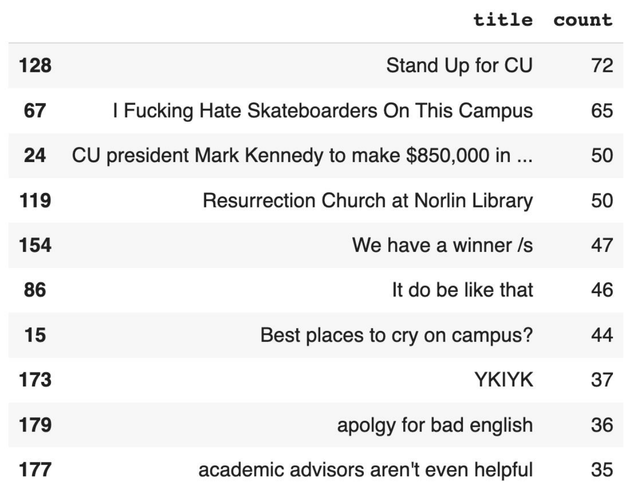
Covid, Vaccine, Holiday schedule…



	title	count
656	Vaccine Requirement for All CU System Students...	358
143	CU to mandate masks indoors again, regardless ...	345
118	CU Boulder to begin spring semester remotely	288
603	The official CU Boulder Fall ‘20 COVID-19 plan...	179
62	Basically no spring break. Yay.	145
412	Make the vax mandatory at CU	114
214	FINALLY. Let's hope they enforce it.	112
154	Can we all just try?	111
208	Everyone in Darley North was just given three ...	109
192	Does anyone have any information on the CD’s t...	108

Before 2020

Random topics related to the life at CU

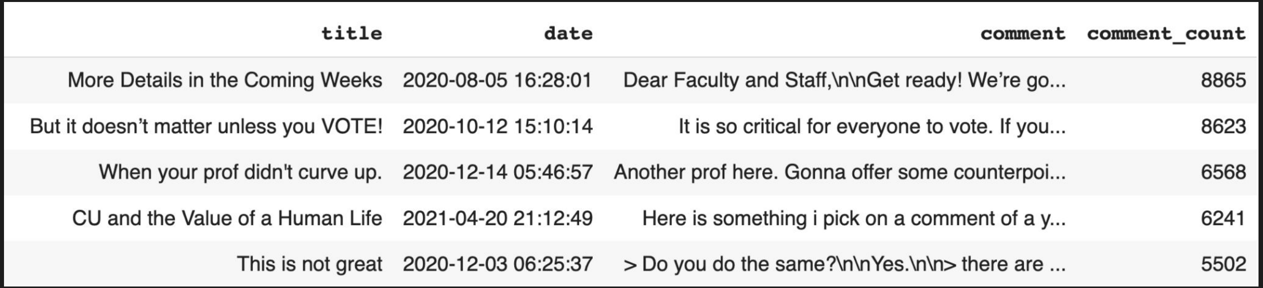


	title	count
128	Stand Up for CU	72
67	I Fucking Hate Skateboarders On This Campus	65
24	CU president Mark Kennedy to make $850,000 in ...	50
119	Resurrection Church at Norlin Library	50
154	We have a winner /s	47
86	It do be like that	46
15	Best places to cry on campus?	44
173	YKIYK	37
179	apolgy for bad english	36
177	academic advisors aren't even helpful	35

Posts with the longest comments

After 2020

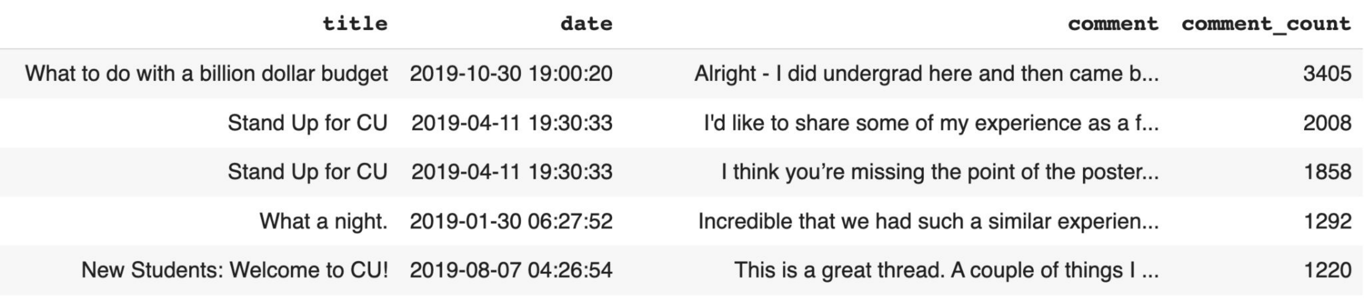
Students have most detailed express in voting, grading, and feelings



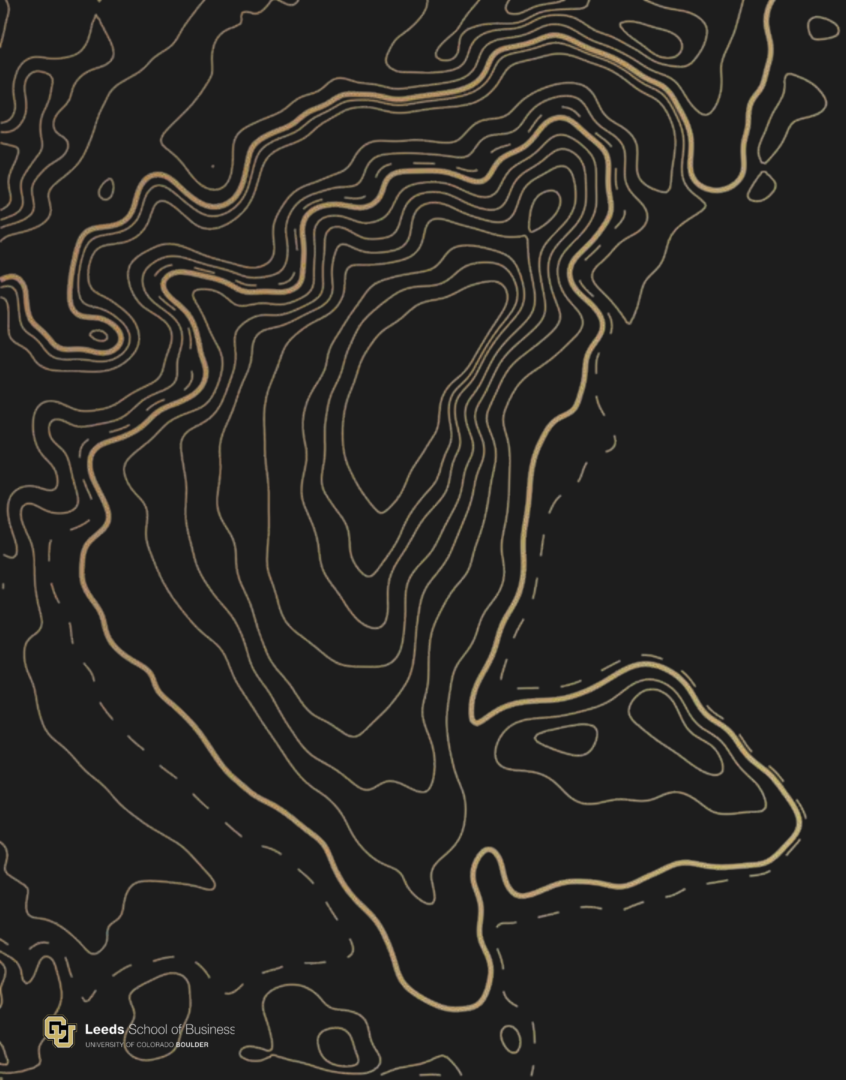
title	date	comment	comment_count
More Details in the Coming Weeks	2020-08-05 16:28:01	Dear Faculty and Staff,\n\nGet ready! We’re go...	8865
But it doesn’t matter unless you VOTE!	2020-10-12 15:10:14	It is so critical for everyone to vote. If you...	8623
When your prof didn't curve up.	2020-12-14 05:46:57	Another prof here. Gonna offer some counterpoi...	6568
CU and the Value of a Human Life	2021-04-20 21:12:49	Here is something i pick on a comment of a y...	6241
This is not great	2020-12-03 06:25:37	> Do you do the same?\n\nYes.\n\n> there are ...	5502

More contents on onboarding and Orientation for new students

Before 2020



title	date	comment	comment_count
What to do with a billion dollar budget	2019-10-30 19:00:20	Alright - I did undergrad here and then came b...	3405
Stand Up for CU	2019-04-11 19:30:33	I'd like to share some of my experience as a f...	2008
Stand Up for CU	2019-04-11 19:30:33	I think you’re missing the point of the poster...	1858
What a night.	2019-01-30 06:27:52	Incredible that we had such a similar experien...	1292
New Students: Welcome to CU!	2019-08-07 04:26:54	This is a great thread. A couple of things I ...	1220



04

Topic Extraction

BERTopic Model

! pip install BERTopic

Brief Introduction of BerTopic

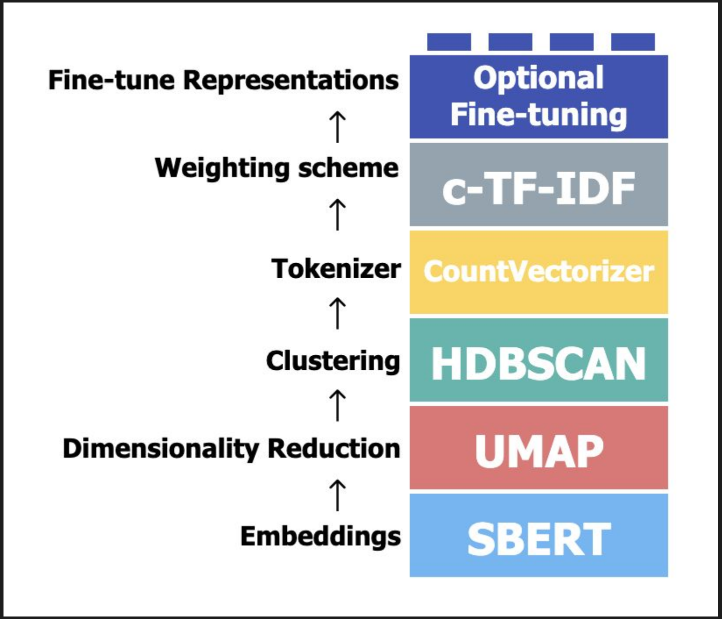
BERTopic

- ✔ BERTopic can automatically group similar documents together based on their content.
- ✔ It is based on **BERT** embeddings and **c-TF-IDF** to understand the context and meaning of the text.
- ✔ It is useful for tasks like **clustering** and **topic modeling**, and can be applied to a wide range of text data, such as news articles, social media posts, and customer feedback.
- ✔ It even offers built-in **visualizations** functions!
- ✔ It has great documentations: https://maartengr.github.io/BERTopic/index.html

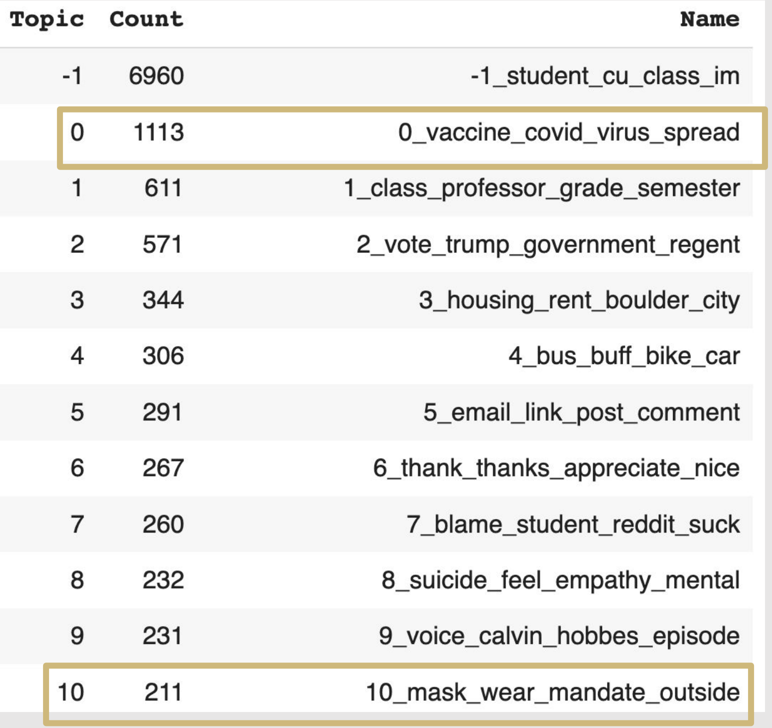
Behind Algorithm of BerTopic

Why is BERTopic Reliable?

By leveraging pre-trained language models like BERT and density-based clustering techniques like HDBSCAN, BERTopic can efficiently create clusters of semantically similar documents and identify the most important words associated with each topic. Additionally, it creates easily interpretable topics with descriptive topic labels that contain important words, making it easier for users to understand the topics represented by each cluster.



Top 10 Topics
after 2020



Topic	Count	Name
-1	6960	-1_student_cu_class_im
0	1113	0_vaccine_covid_virus_spread
1	611	1_class_professor_grade_semester
2	571	2_vote_trump_government_regent
3	344	3_housing_rent_boulder_city
4	306	4_bus_buff_bike_car
5	291	5_email_link_post_comment
6	267	6_thank_thanks_appreciate_nice
7	260	7_blame_student_reddit_suck
8	232	8_suicide_feel_empathy_mental
9	231	9_voice_calvin_hobbes_episode
10	211	10_mask_wear_mandate_outside

Key Takeaway

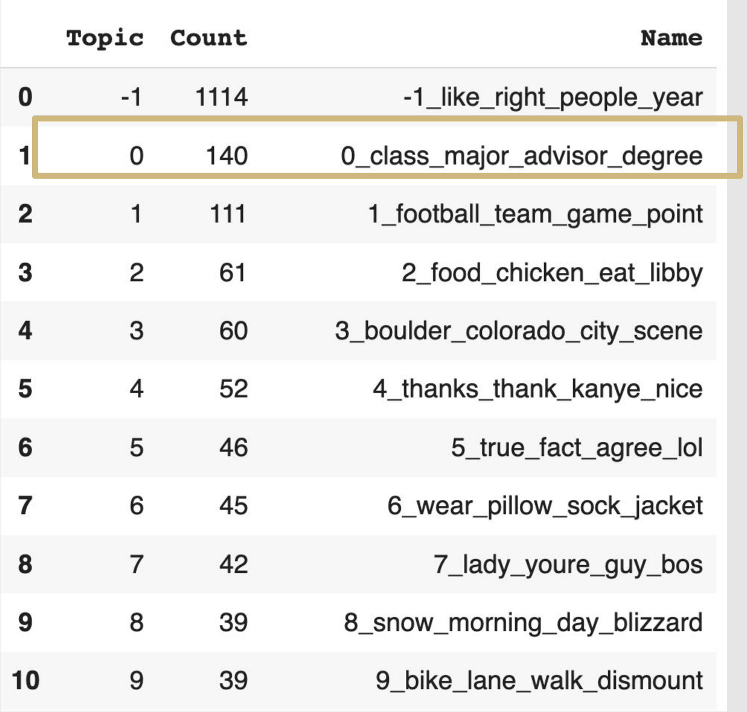
✔ Top1 topic is about **covid** and vaccination

✔ Other popular topics

Grading,
Life at CU (housing, commuting, etc.),
Politics and voting,
Mental health related topics

***Note**: -1 are all outliers contents*

Top 10 Topics before 2020



	Topic	Count	Name
0	-1	1114	-1_like_right_people_year
1	0	140	0_class_major_advisor_degree
2	1	111	1_football_team_game_point
3	2	61	2_food_chicken_eat_libby
4	3	60	3_boulder_colorado_city_scene
5	4	52	4_thanks_thank_kanye_nice
6	5	46	5_true_fact_agree_lol
7	6	45	6_wear_pillow_sock_jacket
8	7	42	7_lady_youre_guy_bos
9	8	39	8_snow_morning_day_blizzard
10	9	39	9_bike_lane_walk_dismount

Key Takeaway

✔ Top1 topic is about academic study

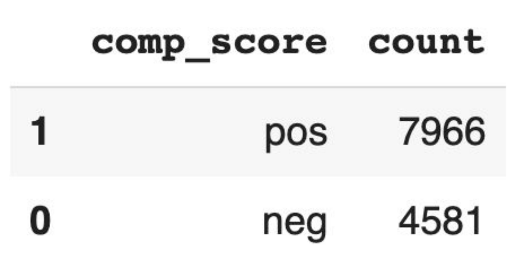
✔ Other popular topics

Sports,
Eating,
After school life and weather
(snow, biking, city scene, etc)

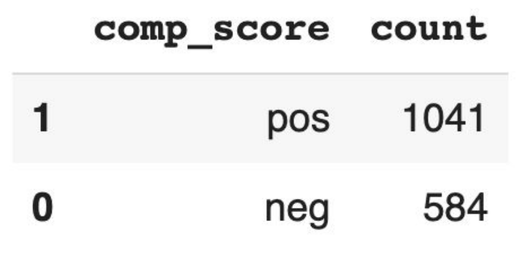
***Note**: -1 are all outliers contents*

Sentiment Analysis - NLTK VADER

After 2020



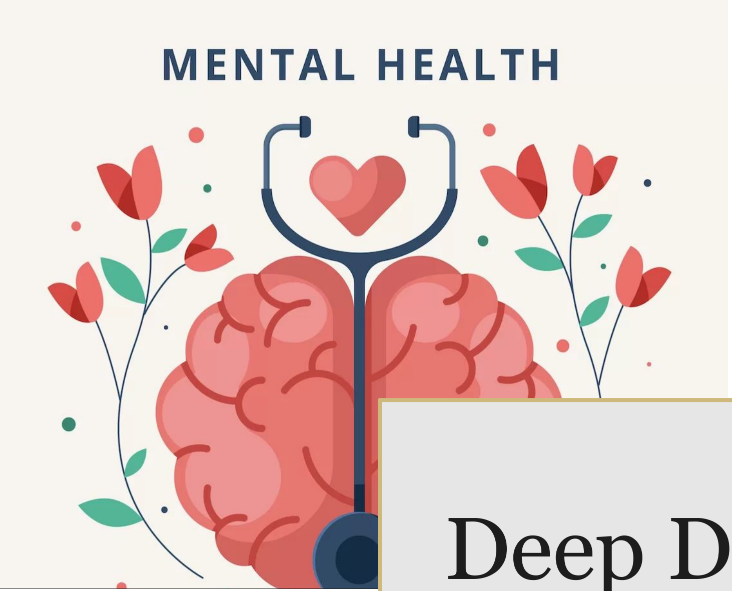
	comp_score	count
1	pos	7966
0	neg	4581



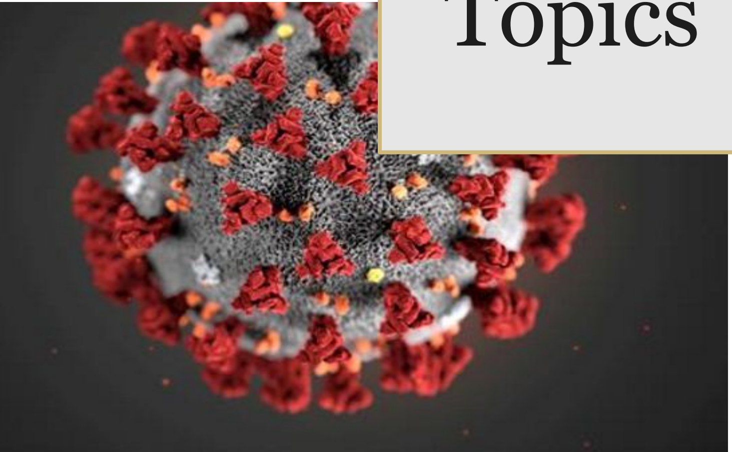
	comp_score	count
1	pos	1041
0	neg	584

Before 2020

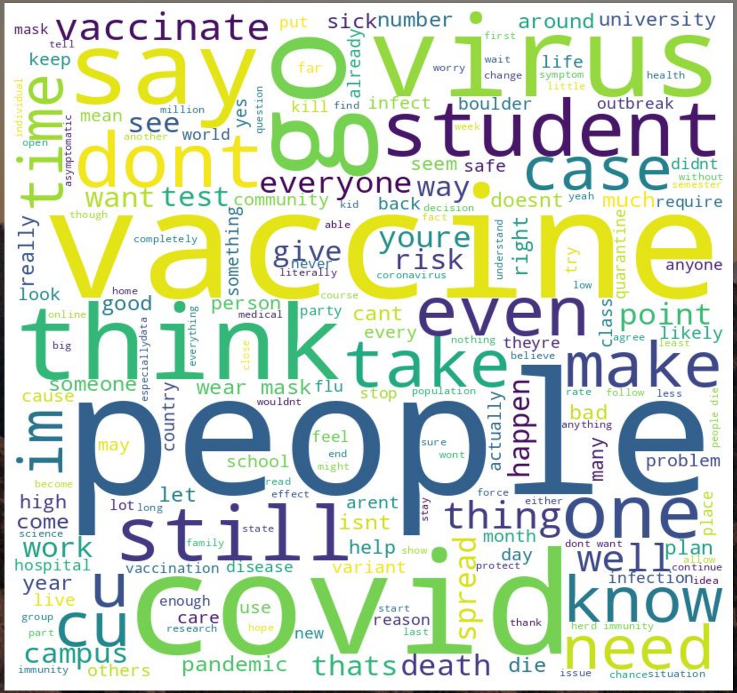
- ✔ Based on the VADER lexicon, the overall sentiment of the comments are all more positive.
- ✔ But the percentage of negative comments is also very high, especially after the covid year.
- ✔ We will deep dive into the sentiment for some specific topics



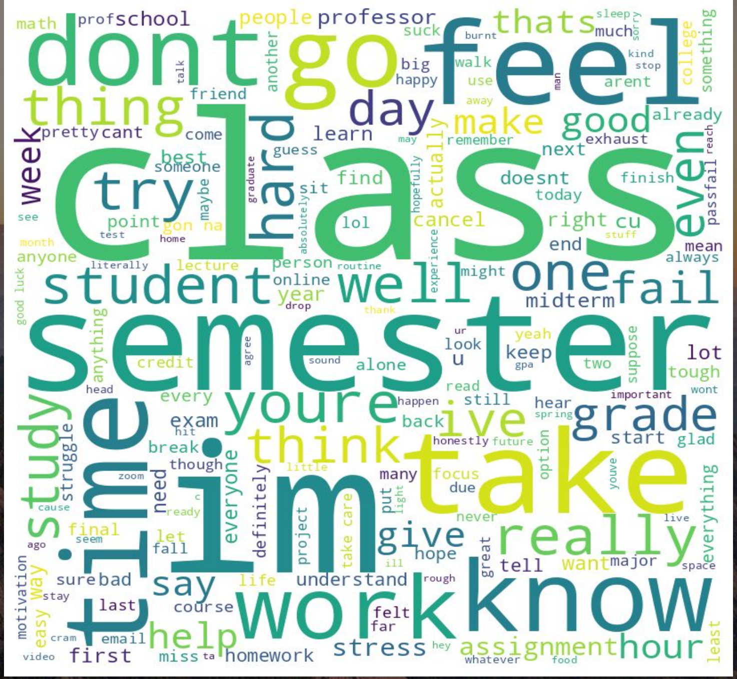
Deep Dive into Topics



Topic 0 - Covid



Topic 8 - Mental Health



Topic 0 - Covid

Top Words:

1. Vaccine
2. Covid
3. Virus
4. Spread
5. People

Leeds School of Business
UNIVERSITY OF COLORADO BOULDER

Comments of Interest

- “If anyone here needs help finding a vaccine, I volunteer with [Colorado Vaccine Hunters](https://www.facebook.com/groups/covaccinehunters) and VaccineFairy.org. (I'm also CU staff.) Feel free to DM for tips, tricks or even help booking an appointment, both in CO and your state if you're not here yet! Happy to help make it easy and stress-free :)“

- “Herd immunity requires a significant portion of the population gets vaccinated...thus lots of people not getting vaccinated is a threat to the worldwide community’s well being“

Topic 8 - Mental Health

Top Words:

1. Suicide
2. Feel
3. Empathy
4. Mental
5. Day

Comments of Interest

- “Very true. Life is pretty intense right now, and as we're approaching mid-semester, everyone's been far more stressed. I hope you're doing well, and thank you for the friendly reminder to take care of ourselves and each other.“
- “... All in all, I feel like CU screwed students over by taking away spring break, and it was all done just to try and get people back in the dorms/on campus. To add insult to injury, they didn't even have the decency to have the wellness days on a Friday/Monday so students could have a three day weekend…“

Conclusion and The Future

- **Sentiment Analysis**
  - We conducted unsupervised sentiment analysis for this project, which could be very biased depending on the dataset the pre-trained sentiment models used. A more accurate sentiment results could be obtained by spending time manually labeling the text data we have.
- **Topic Extraction**
  - Bertopic is quite modular and has many hyperparameters that can be tuned which can allow you to build your own topic model throughout a variety of sub-models. More experimenting and tuning of the model could be conducted for a more comprehensive and accurate topic clustering.

Thank you