University of Colorado, Boulder r/cuboulder Subreddit Topic Modeling and Sentiment Analysis

Chris Goswick
University of Colorado, Boulder
Boulder, Colorado
chgo2707@colorado.edu

Aria Jia
University of Colorado, Boulder
yiji5722@colorado.edu

Maddie Wallace
University of Colorado, Boulder
mawa9164@colorado.edu

Abstract

In this report, we have conducted topic modeling and sentiment analysis on the CU Boulder r/cuboulder subreddit. We have chosen to investigate this topic because of recent events surrounding the University of Colorado at Boulder. Understanding the public perception of an institution like CU Boulder can provide insights into the strengths and weaknesses of the university, which can, in turn, help in improving its reputation and competitiveness. Additionally, the findings of this project can be used by the university to better fine-tune its marketing and outreach efforts, as well as to identify areas where improvement is needed in its communication and engagement with students, future students, and stakeholders. Current students, potential future students, and school administration will be able to see and understand true student perceptions of the University. The perceptions of the university as seen through the r/cuboulder subreddit could also reflect the views of the general public. Our data has been sourced from Reddit.com, "an American social news aggregation, content rating, and discussion website" [1], where users can post freely with minimal moderation, allowing insights into true feelings to be attained. Administrators will be able to see what students and the community enjoy most as well as what they strongly dislike about the university, enabling them to conduct change where necessary.

1. Introduction

The University of Colorado, Boulder (CU Boulder) is a public research university located in Boulder, Colorado. Founded in 1876, the university currently offers nine schools with over 150 different academic programs, and as of January 2022, is home to over 35,000 students [2].

Throughout the last few years, CU Boulder has had plenty of events that have captured the attention of the community surrounding the university as well as national attention. We expect some of these events to be heavily present in our topic modeling analysis. Throughout the years since 2016, the pandemic of COVID-19 in 2020 has been recognized as the most life-changing and impactful event for all. Thus, in this project, we conducted topic extraction modeling and sentiment analysis for the CU Boulder subreddit community and made comparisons between before and after Covid (2016 vs 2020). We wanted to find out if there were any notable differences in the community outlook that could be attributed to the pandemic.

1.1. COVID-19

In March 2020, COVID-19 shocked the world and left no communities untouched, including CU Boulder. The administrative choices that followed due to COVID-19 caused strong opinions to be shared by almost everyone. Opinions were formed and shared around topics of all kinds, some of which included online classes, mask mandates, and vaccine requirements. Because the pandemic has impacted so many things in our lives, we chose to analyze its impact on the CU Boulder community. In this analysis, we will look at data from 2016, before the pandemic, and 2020, the start of the pandemic.

1.2. Reddit

Reddit.com is "an American social news aggregation, content rating, and discussion website" [1]. Users post content to the site which can then be up or downvoted by fellow users. "Posts are organized by subject into user-created boards called 'communities' or 'subreddits'". Each subreddit is moderated by either Reddit administrators or by community-specific moderators who are not Reddit employees and are chosen by the community or creator of the community to assume that responsibility.

2. Related work

2.1. "Understanding Fortnite's Reddit Community using Unsupervised Topic Modeling"

While researching ideas for our project we came across this article by Jerome Cohen, which inspired our project: "Understanding Fortnite's Reddit Community using Unsupervised Topic Modeling" [10]. This article talks about scraping data from Reddit using various APIs, as well as "The Business Case" [10], which helped us understand how analyzing the CU Boulder subreddit could benefit the CU Boulder community. Although we did not follow the steps Cohen took for his analysis, we agreed that we all found interest in the idea and liked the premise.

2.2. "Sentiment Analysis on Reddit News Headlines with Python's Natural Language Toolkit (NLTK)"

This article [11] by Brendan Martin and Nikos Koufos introduced us to the Reddit API scraper PRAW, as well as the concepts on NLTK and VADER. The article helped us to analyze our own data after we collected it and additionally, gave us ideas for future research.

3. Data

Since there isn't an established dataset with up-to-date comments and activity data online, we utilized Python Reddit API Wrapper (PRAW) to pull all posts and comments of r/cuboulder directly. PRAW is a Python library that provides a simple interface for accessing the Reddit API. It allows users to extract a wealth of information from Reddit, such as posts, comments, subreddits, and more. In our project, we utilized the PRAW library to pull all posts and comments from the r/cuboulder subreddit directly. This allowed us to collect the most up-to-date data available for this specific community. We will explain more details about the data collection in later sections.

3.1. Gathering data - PRAW

To collect the data for our analysis, we utilized the Python Reddit API Wrapper (PRAW) package to scrape the top posts and their associated comments from the r/cuboulder subreddit. Due to Reddit's limit of 1000 queries, we were only able to collect data for the top posts of all time. We created a dictionary to store the data and looped through the posts to add them to the dictionary.

To collect the comments associated with each post, we first identified the top posts from the year 2016 and ran a loop through them by post ID. For each post, we used PRAW to retrieve its comments and added them to a separate dictionary that contained the post ID, title, date, and comment text. The final dataset was stored in a pandas DataFrame with 21,313 rows and seven columns: ID, title, date, post text, score, total comments, and post URL. We then converted the timestamp from epoch to GMT to facilitate our analysis. The resulting data allowed us to conduct topic modeling and sentiment analysis on the CU Boulder subreddit and gain insights into the public perception of the university.

Next, in order to conduct the topic comparison before and after the pandemic in 2020, we split the whole CSV file into two data frames: one with posts and comments submitted before 12/31/2019, and the other with posts and comments after that date. It resulted in the dataset before the Covid pandemic having 2,679 rows of records, and the dataset after 2020 having 18,302 rows.

3.2. Cleaning the data

To ensure the validity and accuracy of our analysis, we performed several data-cleaning steps on the dataset we obtained from the r/cuboulder subreddit. First, we dropped any rows of records that contained '[removed]' or '[deleted]' in the comments column. These were comments deleted by the Reddit content moderator bots or human moderators, and these entries did not contain useful information for our analysis. We also removed any comments that were less than five characters in length, as these were unlikely to contribute to the overall sentiment of the post.

To further facilitate our analysis, we removed any punctuation from the comments using the String and Re packages in Python, as well as any hyperlinks using a regular expression. We also created a list of profanity words and substituted them with a filler word to further clean up the data. Upon our original analysis, we discovered hateful language towards the LGBTQ community and removed associated words. When we first ran topic modeling, we discovered that this was a topic being clustered together. To remove hateful associations in our data we removed the words to avoid hateful bias in our topics. The original comment can still be seen in its original form in the 'comment' column.

We also applied the Natural Language Toolkit (NLTK) library to remove stop words, which are commonly used words in the English language such as 'and' and 'the', that do not carry much meaning in the context of the comments. We will talk more about NLTK in section 4.4. We also converted the phrase 'gon na' to 'gonna' for consistency and ease of analysis.

In addition, to lemmatize the words in the comments column based on their part of speech, we used the NLTK library to create a function that reduced words to their base form, which is important for our analysis as it allows us to group together variations of the same word. We first tokenized the comments into individual words and then applied part-of-speech tagging to assign each word a grammatical role, such as noun or verb. Based on the assigned role, we applied lemmatization using the WordNetLemmatizer module in NLTK to reduce words to their base form.

Finally, we removed any rows that contained an empty string in the cleaned_comment column, as these did not contain any meaningful information. These data-cleaning steps resulted in a clean dataset with 12,547 rows and the cleaned_comment column as the input for our topic modeling and sentiment analysis. These data-cleaning steps ensured the accuracy and consistency of our analysis and allowed us to gain meaningful insights into the public perception of the University of Colorado at Boulder as expressed on the r/cuboulder subreddit.

4. Methods

For our analysis we have chosen to use BERTopic to do topic modeling, followed by NLTK's VADER for sentiment analysis. We analyzed the r/cuboulder data for the years 2016 to 2020 and also the data from r/cuboulder for the years 2020-2023. We chose to analyze two different data sets because we wanted to see the effect of the pandemic on the CU Boulder community and campus.

4.1. Topic modeling

Topic modeling is an unsupervised machine learning process in which an algorithm scans a series of documents (corpus) and is able to detect similar words and phrases within the documents and group them together. The number of groups, or clusters, is a parameter that can sometimes be tuned or chosen by the user. The resulting clusters represent the major topics within the series of documents.

4.2. BERTopic model

For our topic modeling purposes, we decided to use BERTopic. BERTopic "is a topic modeling technique that leverages BERT embeddings and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions" [5].

BERT stands for Bidirectional Encoder Representations from Transformers and was "designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context" [6]. BERT framework "was pre-trained using text from Wikipedia" [6]. Because BERT has been pre-trained on the entirety of English Wikipedia, it is meant to serve as a baseline on which users can build off of. This process is also known as transfer learning.

c-TF-IDF is "a TF-IDF formula adopted for multiple classes by joining all documents per class. Thus, each class is converted to a single document instead of a set of documents. The frequency of each word x is extracted for each class c and is L1 normalized. This constitutes the term frequency." [5].

BERTopic is an incredibly reliable tool to use in topic modeling. 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.

The algorithm behind BERTopic is simply six steps. First, it converts documents to numerical representations using sentence-transformers, similar to translating a book into a different language. Second, it reduces the dimensionality of these numerical representations using UMAP, similar to shrinking a giant balloon to a manageable size without losing its shape. Third, it clusters similar documents together using HDBSCAN, similar to sorting a pile of laundry into separate baskets based on their colors and fabrics. Fourth, it creates a bag-of-words representation by combining all documents in a cluster into a single document and counting how often each word appears in the cluster, similar to counting the frequency of different ingredients in a recipe. Finally, it assigns importance scores to words within each cluster using a modified TF-IDF algorithm, similar to ranking the ingredients in a recipe by how essential they are to the final dish. This generates easily interpretable topics with descriptive topic labels that contain important words, allowing for efficient topic modeling and exploratory data analysis in natural language processing.

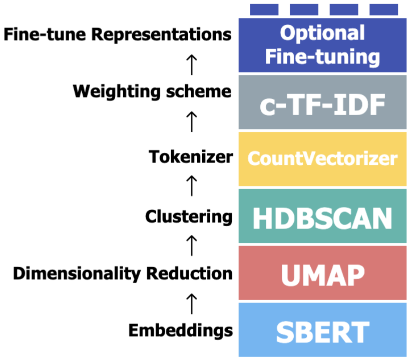


Figure 1: How BERTopic works.

4.3. Sentiment analysis

Sentiment analysis is a powerful tool in natural language processing (NLP) that allows us to measure the sentiment of a given text. Typically, sentiment analysis involves assigning a positive, negative, or neutral label to the text, along with a score that indicates the degree to which the text fits its assigned category. However, in order to conduct sentiment analysis, we typically require labeled data to train a supervised model.

In our project, we collected our own data from Reddit using PRAW, which meant that we did not have labeled data readily available. To overcome this challenge, we turned to unsupervised sentiment analysis using the Valence Aware Dictionary and sEntiment Reasoner (VADER), a module from the Natural Language Toolkit. While unsupervised sentiment analysis is less common in NLP, VADER is a robust tool that allows us to accurately measure sentiment in our unlabeled data. Although by using VADER there was still a chance that the final sentiment results could be biased and not that accurate, we were able to gain an overall understanding of the insights into the sentiment of the r/cuboulder community.

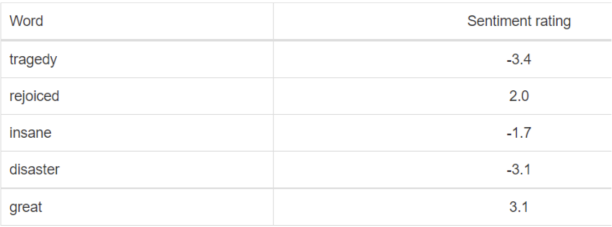
4.4. Natural Language ToolKit

The Natural Language Toolkit (NLTK) is a Python platform for creating platforms to work with human language data. It is a community-driven, free, open-sourced project. NLTK "provides easy-to-use interfaces to over 50 corpora and lexical resources…along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum" [7]. This platform would grant us access to VADER, which allowed us to conduct our unsupervised sentiment analysis.

4.5. VADER

VADER, Valence Aware Dictionary, and sEntiment Reasoner, is a rule-based lexicon analysis tool that is especially sensitive to inferences communicated in web-based media, like Reddit posts. VADER utilizes a mix of words that are, for the most part, marked by their semantic direction as positive or negative. VADER not only tells us the polarity score but also tells us just how positive or negative an instance is.

VADER uses lexicons of sentiment-related words. Each word in the vocabulary is labeled as positive or negative and is evaluated on how positive or negative it is. The more positive a word is, the higher its evaluation score, and the more negative a word is, the more negative its evaluation score is. See the image below for examples of words and their evaluation scores.



Word	Sentiment rating
tragedy	-3.4
rejoiced	2.0
insane	-1.7
disaster	-3.1
great	3.1

Figure 2: Example words from VADER and their sentiment ratings.

In total, VADER will output four things: positive, negative, neutral, and compound. Positive is the degree to which the sentiment is positive, negative is the degree to which the sentiment is negative, neutral is the degree to which the sentiment is neutral, and compound is the total amount of lexicon scores that have then been normalized to be between -1 and 1.

VADER is an excellent choice for social media text analysis because it has been trained on social media text data. It is able to score things like emoticons, emojis, and punctuation, whereas other sentiment analysis tools are not able to score these things.

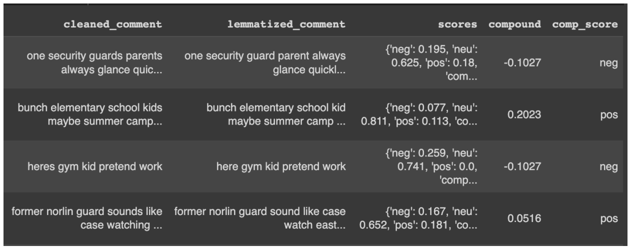
5. Empirical applications, experiments, and results

5.1. Empirical applications

When applying the methods discussed above we took multiple steps. First, we used the BERTopic model to do topic extraction. Within this step, we used UMAP to perform dimensionality reduction and to prevent any stochastic behavior in order to be able to reproduce the results. In our UMAP model, we set n_neighbors to 15, n_components to 5, and a random state of 43 for reproducibility. We added a vectorizer model of CountVectorizer to eliminate stop words. We then initialized and trained a BERTopic model with an embedding model of all-mpnet-base-v2, our vectorizer model, and our UMAP model. We chose our embedding model of all-mpnet-base-v2 because this model is a step up from the default model, and although it takes more computing time, it produces higher-quality results.

Because BERTopic uses HDBSCAN for clustering, the user is unable to specify the number of clusters wanted. This can be seen as beneficial because it requires the user to trust HBDSCAN to find the right number of topics that are in the corpus, versus a forced number. The user can then use topic reduction to reduce the number of topics. When this is done, HBDSCAN clusters related topics together and joins the clusters, keeping outliers in mind as well. In the implementation of our model for 2020, we used automatic topic reduction to reduce the number of topics to a maximum of 100 topics. When finished, our topics decreased from 221 topics to 100 topics. This ensured that we had the most true and relevant topics present. However, in our model for 2016, after running the model, we only had 47 topics, which was not enough to reduce. The small number of topics is due to the limited amount of data we received from 2016.

After our topics had been created, we moved on to sentiment analysis within each topic. From NLTK VADER we imported SentimentIntensityAnalyzer and created an object for later use. We then used the sid polarity scores to extract scores. Then, we filtered the dictionary generated from the previous step to only include compound scores (as opposed to including the degree to which each instance was positive, negative, or neutral), as that is what we are most interested in. Finally, we added a column to the data frame which would tell us if a comment was negative or positive. The cutoff scores we determined for our project were positive equaling a positive compound score and negative equaling a negative compound score. This is a hyperparameter set by the user.



cleaned_comment	lemmatized_comment	scores	compound	comp_score
one security guards parents always glance quic...	one security guard parent always glance quickl...	{'neg': 0.195, 'neu': 0.625, 'pos': 0.18, 'com...	-0.1027	neg
bunch elementary school kids maybe summer camp...	bunch elementary school kid maybe summer camp ...	{'neg': 0.077, 'neu': 0.811, 'pos': 0.113, 'co...	0.2023	pos
heres gym kid pretend work	here gym kid pretend work	{'neg': 0.259, 'neu': 0.741, 'pos': 0.0, 'comp...	-0.1027	neg
former norlin guard sounds like case watching ...	former norlin guard sound like case watch east...	{'neg': 0.167, 'neu': 0.652, 'pos': 0.181, 'co...	0.0516	pos

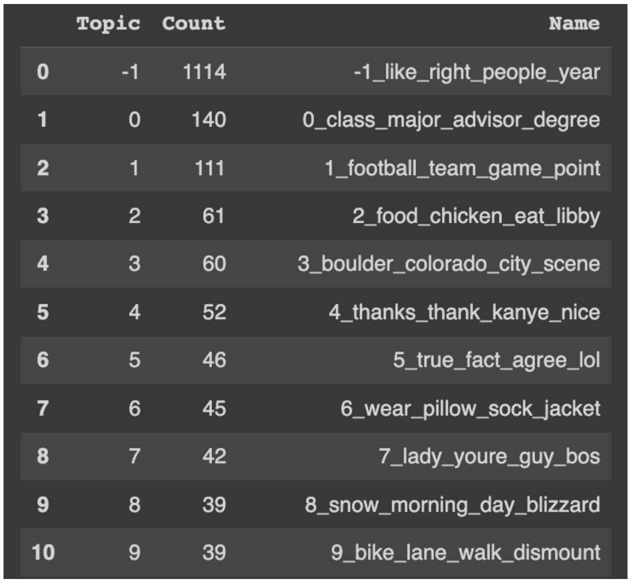
Figure 3: Image of the compound score as well as the overall sentiment.

5.2. Experiments

As mentioned in the previous section, BERTopic uses UMAP which has a stochastic nature, meaning it has uncertainty or randomness in its outcomes. Because of this, we were able to run our embeddings once and then run our model several times with different parameters until the topics detected were uniform and made sense.

5.3. Results pre-Covid

Our results from our data from 2016 contained topics revolving around general Boulder things, such as information about campus or classes. Please note that the topic labeled -1 is a topic consisting of outliers, which includes posts that do not fall into any of the topic categories created. Also, the topics are shown with the rank of their popularity, which means topic 0, class and grades, is the most common topic that people were talking about. We can also see a topic about the football team, which makes sense since the CU Boulder football team "finished the season 10–4, 8–1 in Pac-12 play to win their first Pac-12 South Division Title" and "it was their first winning season since 2005" [12]. This is all cause for conversation around the team. Additional topics include the weather, the general beauty of Boulder, and conversation around the bike lanes in Boulder, which tends to be a popular topic of conversation since many people use biking as a form of transportation to get around Boulder.



	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

Figure 4: Top 10 topics from 2016 data.

The visualization of the topic clustering is shown in figure 5. This distance map is an interactive distance map. If you wish to experiment with it, please see our code notebook. The visualization was accomplished using the built-in function in the BERTopic model. It clustered similar topics into groups and showed them in a 2D graph. When we interact with this chart, we can see what topics are similar from a high level. From the chart, we can see that there are 9 topic clusters with the 3 most concentrated clusters at the upper left side of the graph. These are topics similar to topics 0, 1, and 2.

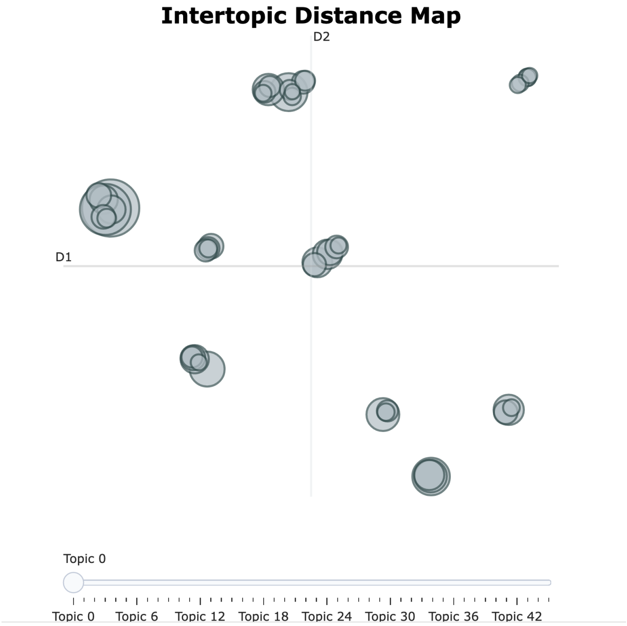
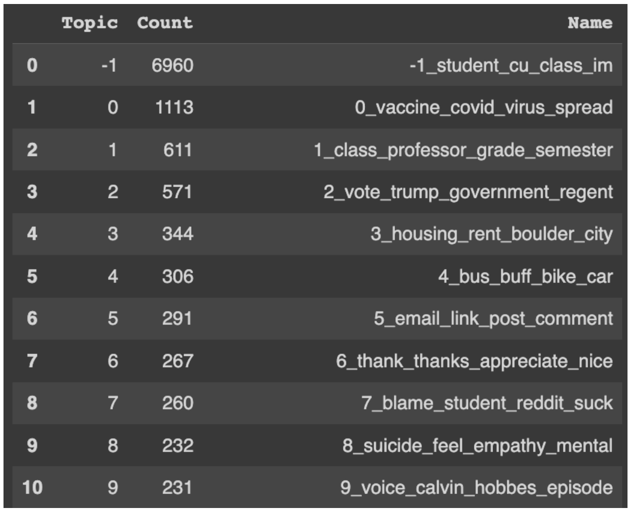


Figure 5: Topic clustering from 2016 data.

5.4. Results during and after Covid

Our results from 2020-2023 showcase what has been happening in the CU Boulder community since 2020. 2020 was the beginning of the COVID-19 pandemic. This resulted in lots of discussion from students, faculty, and the community about choices made by the university regarding in-person classes, the vaccine, and the spread of the virus. This is reflected in our findings since the topic of COVID-19 is the largest topic. You can also see a topic about the presidential elections in 2020, between Donald Trump and Joe Biden. This election was incredibly divisive, as the two candidates were very questionable in the eyes of the American people. Another topic we'd like to draw attention to is topic number 8. This topic emphasizes mental health and feelings people were tackling during the pandemic and have since been managing since. There has also been a more open discussion about mental health and getting proper help since the pandemic began. Lastly, there are still topics such as 1, 3, and 4 which cover general Boulder and CU Boulder topics, like classes and professors, housing, and methods of transportation.



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

Figure 5: Top 10 topics from 2020 data.

We have created the same visualization as the 2016 data for the 2020 topic clustering, as shown in figure 6. The interactive chart can be found in our code notebook. The cluster around the middle of the chart is topics similar to topic 1.

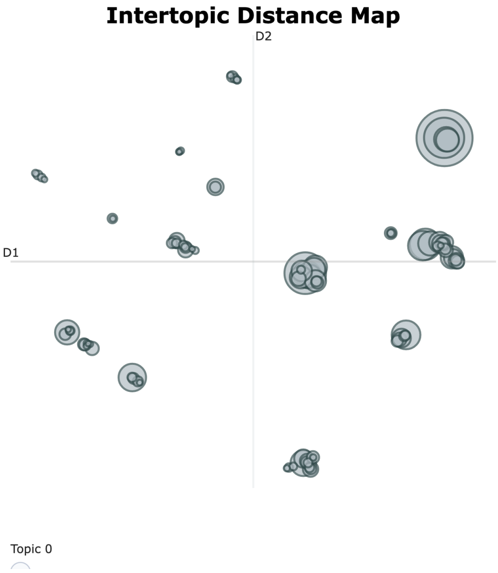


Figure 6: Topic clustering from 2020 data.

Another interesting note we'd like to make is the increase in the count. The count is the count of documents that fall under that topic. This increased count is most likely due to an increase in the use of Reddit because of the pandemic when people were looking for ways to connect, or simply an increase in the popularity of Reddit. It is hard to pinpoint exactly which reason is behind this, but a Google Trends search reveals that Reddit has steadily been increasing in popularity from the beginning of 2016 to the present.

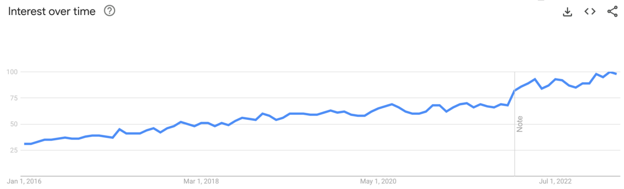


Figure 7: Increase in popularity of Reddit from 1/1/2016 to 1/1/2021.The note on the graph states: “An improvement to our data collection system was applied 1/1/2022.” [13]

5.5. Deep dive into topics sentiments

Since most of our data comes from 2020 and later, we decided to focus our exploration of topics on recent years. This is where we find the most helpful information about how students adjusted to the pandemic, as well as their mental health during and after. The two topics we chose were topic 0, which is about Covid, and topic 8, which has to do with mental health.

Covid forced public universities across the nation to become online full-time, which was unprecedented. This led to students having to make large adjustments to how they learn and live in the middle of the semester in the spring of 2020 through the spring of 2021. When classes did transition back to in-person in the fall of 2021, there was a mandated vaccine requirement for all students attending class in person. We labeled topic 0 as Covid because of its classification of words making up the topic. The top 5 words as far as BERT classification for this topic are vaccine, Covid, virus, spread, and people.

Diving deeper into this category required reading the comments for this topic to get a feel for what users were actually posting. We separated topics in this category by their positive or negative sentiment classification from VADER and started reading through the comments. What became clear from the positive comments is that Reddit can be used as a valuable news source. There were comments directly from faculty giving updates about the timeline and requirements for the return to the classroom. Another comment was from a community member that was involved in a program helping people find vaccines in the area:

> *“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 :)”*

Another comment from a user shows the value that the city of Boulder provides outside of academics:

> *“Stay positive, it’s amazing here. Being 5 minutes drive from a million trails is one of the best things about being here, and covid can’t stop that!“*

Not only was this encouraging others to stay positive during a difficult time, but it also encouraged people to continue living their lives safely.

On the negative side of Covid, there was a lot of discussion about the efficacy and safety of the newly created vaccines. Many discussions about Covid involved debating whether or not the new vaccine mandate would endanger students and if it was safe to return to campus. One comment that shows pushback to vaccine skeptics states:

> *“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”*

For CU, this could show that not enough information was distributed about the safety of the vaccine and could show a potential area for improvement if we ever find ourselves in a pandemic again.

The second topic we explored further was topic 8, which we called ‘mental health,’ due to its top words being suicide, feel, empathy, mental, and day. This topic was full of comments talking about how hard it was being remote students and criticizing some of CU’s decisions around remote instruction. In the spring of 2021, CU decided to cancel spring break in an effort to prevent people from traveling and spreading the virus further. This was replaced with a ‘wellness day’ and a reduced load from professors during what would have traditionally been spring break [14]. Many students took to Reddit to voice their frustrations with this decision:

> “ *... 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... ”*

It appears that the overwhelming majority of comments coming in stemming from this decision point to students feeling burnout with the lack of break. Another student succinctly commented: *“Big mood. motivation low. behind in classes. no spring break to help catch up :(“*.

However, not all the comments were negative for this topic. There were some commenters leaving some encouraging words for people feeling this burnout and frustration:

> *“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.”*

This shows that Reddit can be an important community for those who are in isolation. Despite being separated from the world, there is still a place to go seek connection with others who are sharing a similar experience to yours.

6. Limitations and future study

In most cases, sentiment analysis is supervised, meaning past data is labeled and used to construct a model for predicting future data. However, our project collected data from Reddit using the Python Reddit API Wrapper (PRAW), and lacked the resources and time required to label the data for supervised sentiment analysis. As a result, we had to rely on unsupervised sentiment analysis, which is not commonly used in NLP practices.

While our unsupervised sentiment analysis yielded valuable insights, it is important to note that it could be biased depending on the dataset used to train the pre-existing sentiment models. Therefore, to obtain more accurate sentiment results, it would be beneficial to allocate time and resources toward manually labeling the Reddit data.

In terms of topic extraction, while BERTopic is a highly modular tool that allows for fine-tuning of hyperparameters to construct a customized topic model through various sub-models, it is possible that more experimentation and tuning of the model could be conducted to produce a more comprehensive and accurate topic clustering.

7. Conclusion

In summary, our study successfully utilized BERTopic for topic extraction of the CU Boulder subreddit. The model has the potential to provide valuable insights into what topics and themes students are discussing, allowing the school and data science team to better understand the student community. Additionally, combining this model with sentiment analysis can help to assess student perceptions of various aspects of the university, including services, environment, policies, and more.

Though there's still room for refinement, the suggested future studies could benefit from further investigation into more accurate sentiment analysis techniques and additional experimentation with BERTopic to achieve more comprehensive topic clustering. Overall, this study offers a promising foundation for future research in the field of natural language processing and its applications to understanding student perspectives and opinions.

References

[1] Wikimedia Foundation. (2023, April 24). Reddit. Wikipedia. Retrieved April 2023, from https://en.wikipedia.org/wiki/Reddit

[2] *Academics Programs & Resources*. University of Colorado Boulder. (2022, March 2). Retrieved April 2023, from https://www.colorado.edu/academics

[3] Greenberg, D. (2022, December 20). *Deion Sanders hiring causes huge spike in Colorado merch sales*. Front Office Sports. Retrieved April 2023, from https://frontofficesports.com/deion-sanders-colorado-merch-sales-fanatics/

[4] Howell, B. (2023, April 18). *Football season tickets sold out for CU Buffs' 2023 season*. BuffZone. Retrieved April 2023, from https://www.buffzone.com/2023/04/17/football-season-tickets-sold-out-for-cu-buffs-2023-season/

[5] Grootendorst, M. P. (n.d.). *Bertopic - Bertopic*. BERTopic - BERTopic. Retrieved March 2023, from https://maartengr.github.io/BERTopic/api/bertopic.html

[6] Lutkevich, B. (2020, January 27). *What is Bert (language model) and how does it work?* Enterprise AI. Retrieved March 2023, from https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

[7] NLTK. (n.d.). Retrieved March 2023, from https://www.nltk.org/#natural-language-toolkit

[8] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[9] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[10] Cohen, J. (2019, September 27). *Understanding fortnite's reddit community using unsupervised topic modeling*. Medium. Retrieved March 2023, from https://towardsdatascience.com/understanding-fortnites-reddit-community-using-unsupervised-topic-modeling-30f984f58129

[11] Martin, B., & Koufos, N. (n.d.). *Sentiment analysis on Reddit news headlines with Python's Natural Language Toolkit (NLTK)*. Learn Data Science - Tutorials, Books, Courses, and More. Retrieved April 2023, from https://www.learndatasci.com/tutorials/sentiment-analysis-reddit-headlines-pythons-nltk/

[12] Wikimedia Foundation. (2023, March 28). *2016 Colorado Buffaloes football team*. Wikipedia. Retrieved April 2023, from

https://en.wikipedia.org/wiki/2016_Colorado_Buffaloes_football_team

[13] Google. (n.d.). *Google Trends*. Google trends. Retrieved May 3, 2023, from https://trends.google.com/home

[14] Oravetz, J. (2020, October 22). *CU Boulder Cancels Spring Break to curb covid-19 spread*. KUSA.com. Retrieved May 3, 2023, from https://www.9news.com/article/news/health/coronavirus/cu-no-spring-break-coronavirus/73-2ae2f497-7f64-41a1-b211-341ddc25cb86