CMSC396H Presentation

Aniruddh, Shashank, Abhi, Tanay

The Problem

How information from related student subreddits (r/UMD) affect decision making

The What

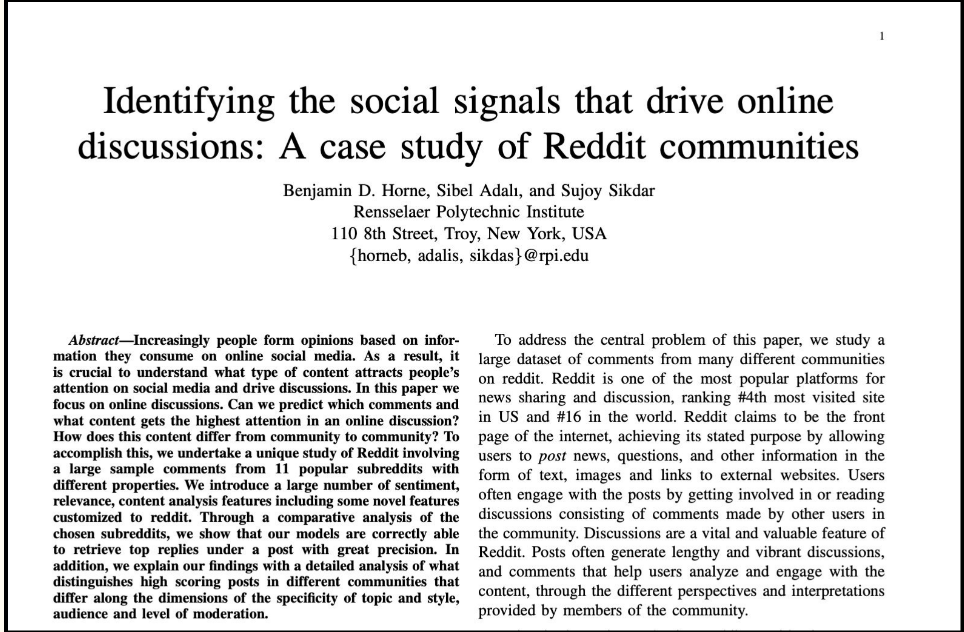
- Many students look for information from online resources, especially ones from their peers, and without a centralized location with all common questions and info, they look to common new hubs like Reddit
    - Reddit was chosen because if project went to completion, APIs exist to carry it out
- Our questions:
    - How does information obtained from Reddit comments, posts, and upvotes influence a student's decision-making process?
    - What is the impact of Reddit discussions on students' academic and housing decisions?
- Importance :
    - Highly upvoted comments and popular posts on Reddit often signal credibility and relevance to students, indicating that the information provided may be trustworthy and valuable. As such, students may be more inclined to consider and act upon recommendations or advice shared through these channels.
    - BUT! Can we trust this information?

Related Work

1. Social Signals that Drive Reddit Communities
2. Predict Post Commentary via Initial Commentary
3. Using ML To Predict Popularity of Reddit Comments

Their Work: Used sentiment, relevance, and content analysis specifically made for Reddit by looking at subreddit rules, characteristics, comments, upvotes, and users' karma to predict which Reddit comments and posts will be the most popular

Unknowns: The work does not cover how people themselves are influenced or why they chose to use Reddit. We are aiming to find out how the popularity affected people, but knowing why certain posts become popular are helpful for our project



Identifying the social signals that drive online discussions: A case study of Reddit communities

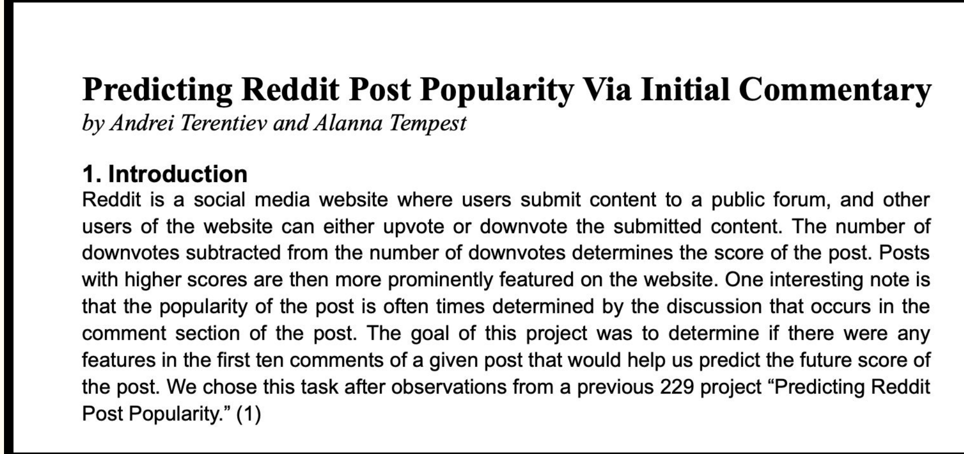
Benjamin D. Horne, Sibel Adalı, and Sujoy Sikdar
Rensselaer Polytechnic Institute
110 8th Street, Troy, New York, USA
{horneb, adalis, sikdas}@rpi.edu

***Abstract*—Increasingly people form opinions based on information they consume on online social media. As a result, it is crucial to understand what type of content attracts people's attention on social media and drive discussions. In this paper we focus on online discussions. Can we predict which comments and what content gets the highest attention in an online discussion? How does this content differ from community to community? To accomplish this, we undertake a unique study of Reddit involving a large sample comments from 11 popular subreddits with different properties. We introduce a large number of sentiment, relevance, content analysis features including some novel features customized to reddit. Through a comparative analysis of the chosen subreddits, we show that our models are correctly able to retrieve top replies under a post with great precision. In addition, we explain our findings with a detailed analysis of what distinguishes high scoring posts in different communities that differ along the dimensions of the specificity of topic and style, audience and level of moderation.**

To address the central problem of this paper, we study a large dataset of comments from many different communities on reddit. Reddit is one of the most popular platforms for news sharing and discussion, ranking #4th most visited site in US and #16 in the world. Reddit claims to be the front page of the internet, achieving its stated purpose by allowing users to *post* news, questions, and other information in the form of text, images and links to external websites. Users often engage with the posts by getting involved in or reading discussions consisting of comments made by other users in the community. Discussions are a vital and valuable feature of Reddit. Posts often generate lengthy and vibrant discussions, and comments that help users analyze and engage with the content, through the different perspectives and interpretations provided by members of the community.

Their Work: Used various ML models like K-Neighbors, Perceptron, SVM, and Logistic Regression that were trained on comment length, sentiment analysis, and “polarity” on posts with at least 10 comments

Unknowns: Which features were the most important and speculation as to why they are, both of which directly fit into our project



Predicting Reddit Post Popularity Via Initial Commentary

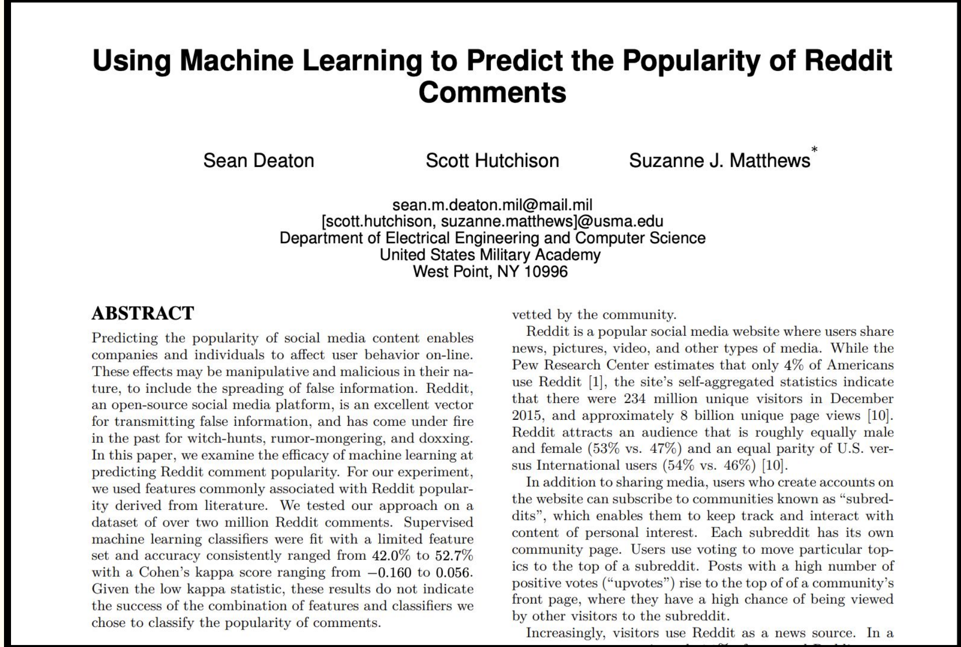
by Andrei Terentiev and Alanna Tempest

1. Introduction

Reddit is a social media website where users submit content to a public forum, and other users of the website can either upvote or downvote the submitted content. The number of downvotes subtracted from the number of downvotes determines the score of the post. Posts with higher scores are then more prominently featured on the website. One interesting note is that the popularity of the post is often times determined by the discussion that occurs in the comment section of the post. The goal of this project was to determine if there were any features in the first ten comments of a given post that would help us predict the future score of the post. We chose this task after observations from a previous 229 project “Predicting Reddit Post Popularity.” (1)

Their Work: Defined popularity as top 25% of comments within each *subreddit*, and used r/datasets (ironic); used ML models to predict how popular some comments would within 42-52% accuracy

Unknowns: How to predict popularity of other facets like posts based on upvotes since this study focused solely on comments. Does provide us with starting point on how to find datasets though.



Using Machine Learning to Predict the Popularity of Reddit Comments

Sean Deaton Scott Hutchison Suzanne J. Matthews*

sean.m.deaton.mil@mail.mil
[scott.hutchison, suzanne.matthews]@usma.edu
Department of Electrical Engineering and Computer Science
United States Military Academy
West Point, NY 10996

ABSTRACT

Predicting the popularity of social media content enables companies and individuals to affect user behavior on-line. These effects may be manipulative and malicious in their nature, to include the spreading of false information. Reddit, an open-source social media platform, is an excellent vector for transmitting false information, and has come under fire in the past for witch-hunts, rumor-mongering, and doxxing. In this paper, we examine the efficacy of machine learning at predicting Reddit comment popularity. For our experiment, we used features commonly associated with Reddit popularity derived from literature. We tested our approach on a dataset of over two million Reddit comments. Supervised machine learning classifiers were fit with a limited feature set and accuracy consistently ranged from 42.0% to 52.7% with a Cohen's kappa score ranging from −0.160 to 0.056. Given the low kappa statistic, these results do not indicate the success of the combination of features and classifiers we chose to classify the popularity of comments.

vetted by the community.

Reddit is a popular social media website where users share news, pictures, video, and other types of media. While the Pew Research Center estimates that only 4% of Americans use Reddit [1], the site's self-aggregated statistics indicate that there were 234 million unique visitors in December 2015, and approximately 8 billion unique page views [10]. Reddit attracts an audience that is roughly equally male and female (53% vs. 47%) and an equal parity of U.S. versus International users (54% vs. 46%) [10].

In addition to sharing media, users who create accounts on the website can subscribe to communities known as "subreddits", which enables them to keep track and interact with content of personal interest. Each subreddit has its own community page. Users use voting to move particular topics to the top of a subreddit. Posts with a high number of positive votes ("upvotes") rise to the top of of a community's front page, where they have a high chance of being viewed by other visitors to the subreddit.

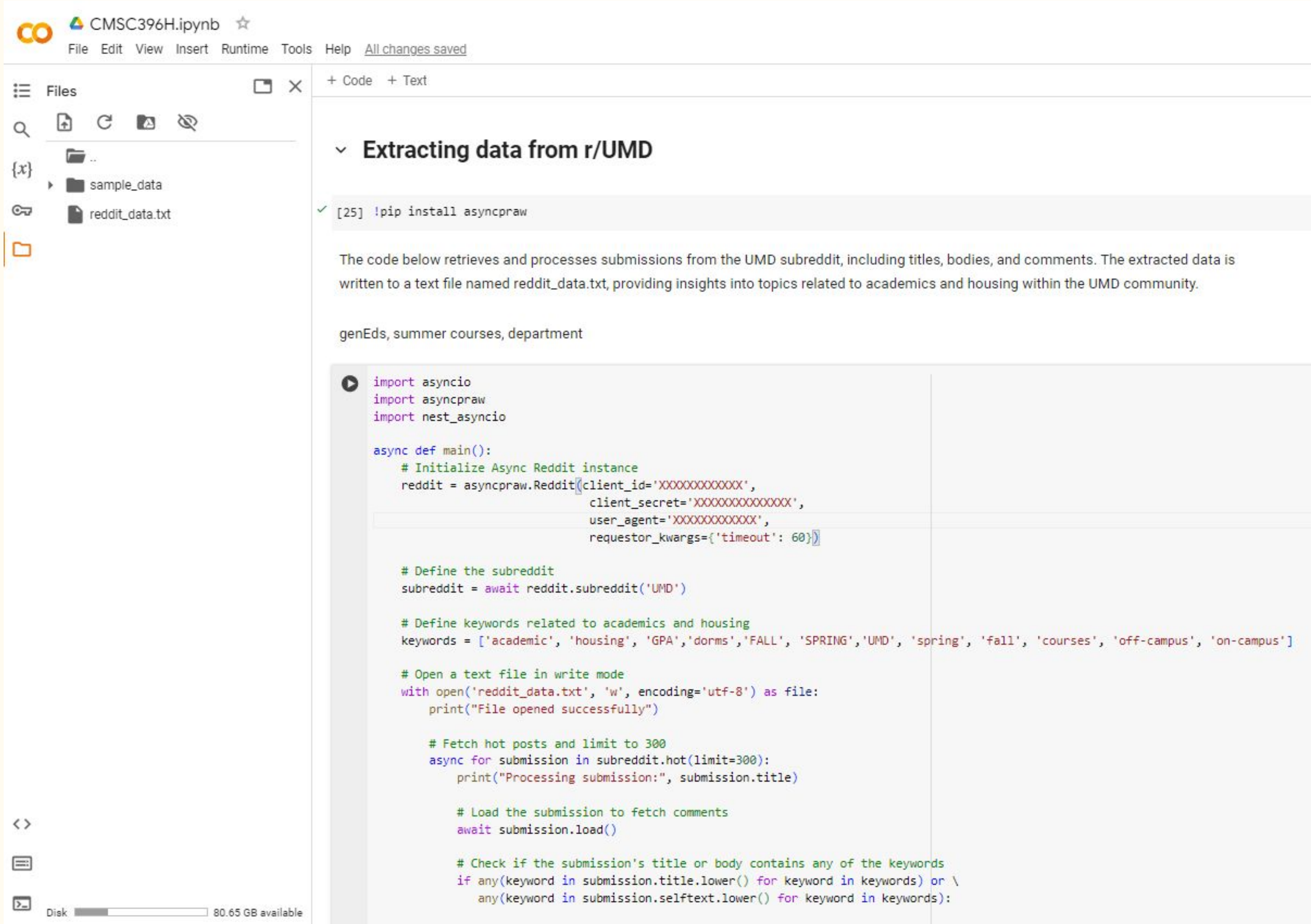
Increasingly, visitors use Reddit as a news source. In a

The Data

Using data from Reddit

We aimed to extract data from the University of Maryland (UMD) subreddit using the Reddit API and the Async PRAW (Python Reddit API Wrapper) library. The goal was to gather information related to academics and housing by analyzing posts and comments within the subreddit.

Keywords included 'Housing', 'Dormitory', 'Residence hall', 'Apartment', 'Off-campus housing', 'Housing options', 'Roommate', 'Lease', 'Rent', 'Landlord', 'Housing cost', 'Housing amenities', 'Commute', 'Housing application', 'Housing lottery', 'Academics', 'Classes', 'Courses', 'Majors', 'Minors', 'Degree requirements', 'Class schedule', 'Professors', 'Lectures', 'Assignments', 'Exams', 'GPA', 'Academic advising', 'Study groups', 'Academic resources', 'Research opportunities', 'Internships'



Extracting data from r/UMD

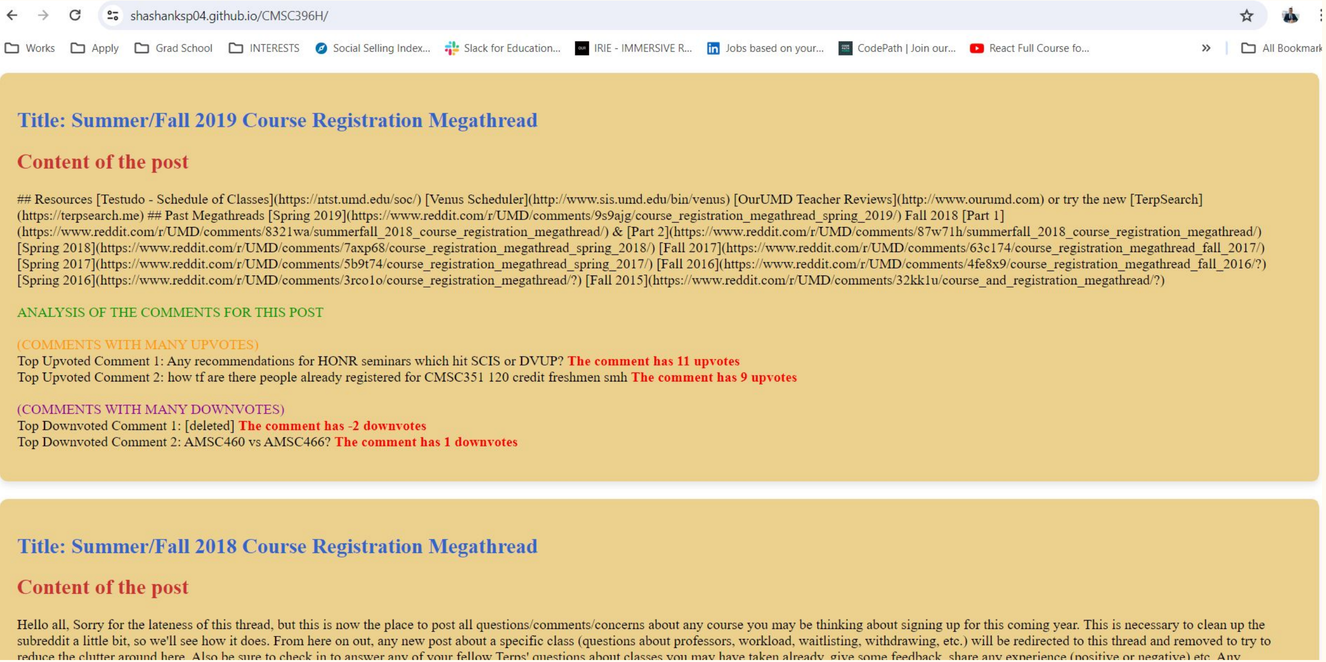
```
[25] !pip install asyncpraw
```

The code below retrieves and processes submissions from the UMD subreddit, including titles, bodies, and comments. The extracted data is written to a text file named reddit_data.txt, providing insights into topics related to academics and housing within the UMD community.

genEds, summer courses, department

```
import asyncio
import asyncpraw
import nest_asyncio

async def main():
    # Initialize Async Reddit instance
    reddit = asyncpraw.Reddit(client_id='XXXXXXXXXXXX',
                               client_secret='XXXXXXXXXXXXXX',
                               user_agent='XXXXXXXXXXXX',
                               requestor_kwargs={'timeout': 60})

    # Define the subreddit
    subreddit = await reddit.subreddit('UMD')

    # Define keywords related to academics and housing
    keywords = ['academic', 'housing', 'GPA','dorms','FALL', 'SPRING','UMD', 'spring', 'fall', 'courses', 'off-campus', 'on-campus']

    # Open a text file in write mode
    with open('reddit_data.txt', 'w', encoding='utf-8') as file:
        print("File opened successfully")

        # Fetch hot posts and limit to 300
        async for submission in subreddit.hot(limit=300):
            print("Processing submission:", submission.title)

            # Load the submission to fetch comments
            await submission.load()

            # Check if the submission's title or body contains any of the keywords
            if any(keyword in submission.title.lower() for keyword in keywords) or \
               any(keyword in submission.selftext.lower() for keyword in keywords):
```

Results of extracting the information using PRAW



Title: Summer/Fall 2019 Course Registration Megathread

Content of the post

Resources [Testudo - Schedule of Classes](https://ntst.umd.edu/soc/) [Venus Scheduler](http://www.sis.umd.edu/bin/venus) [OurUMD Teacher Reviews](http://www.ourumd.com) or try the new [TerpSearch](https://terpsearch.me) ## Past Megathreads [Spring 2019](https://www.reddit.com/r/UMD/comments/9s9ajg/course_registration_megathread_spring_2019/) Fall 2018 [Part 1](https://www.reddit.com/r/UMD/comments/8321wa/summerfall_2018_course_registration_megathread/) & [Part 2](https://www.reddit.com/r/UMD/comments/87w71h/summerfall_2018_course_registration_megathread/) [Spring 2018](https://www.reddit.com/r/UMD/comments/7axp68/course_registration_megathread_spring_2018/) [Fall 2017](https://www.reddit.com/r/UMD/comments/63c174/course_registration_megathread_fall_2017/) [Spring 2017](https://www.reddit.com/r/UMD/comments/5b9t74/course_registration_megathread_spring_2017/) [Fall 2016](https://www.reddit.com/r/UMD/comments/4fe8x9/course_registration_megathread_fall_2016/?) [Spring 2016](https://www.reddit.com/r/UMD/comments/3rco1o/course_registration_megathread/?) [Fall 2015](https://www.reddit.com/r/UMD/comments/32kk1u/course_and_registration_megathread/?)

ANALYSIS OF THE COMMENTS FOR THIS POST

(COMMENTS WITH MANY UPVOTES)
Top Upvoted Comment 1: Any recommendations for HONR seminars which hit SCIS or DVUP? **The comment has 11 upvotes**
Top Upvoted Comment 2: how tf are there people already registered for CMSC351 120 credit freshmen smh **The comment has 9 upvotes**

(COMMENTS WITH MANY DOWNVOTES)
Top Downvoted Comment 1: [deleted] **The comment has -2 downvotes**
Top Downvoted Comment 2: AMSC460 vs AMSC466? **The comment has 1 downvotes**

Title: Summer/Fall 2018 Course Registration Megathread

Content of the post

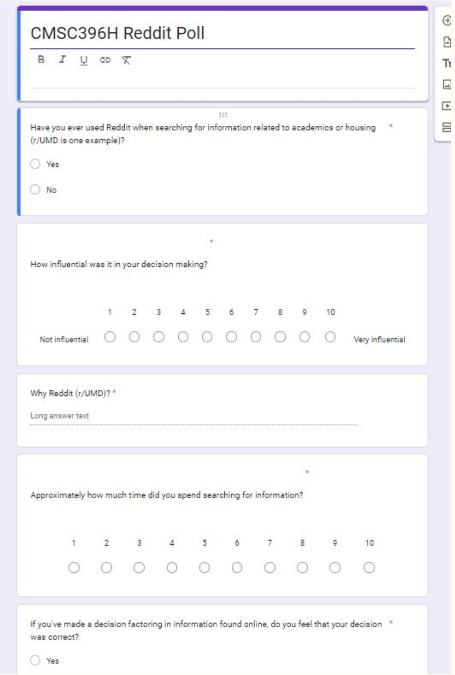
Hello all, Sorry for the lateness of this thread, but this is now the place to post all questions/comments/concerns about any course you may be thinking about signing up for this coming year. This is necessary to clean up the subreddit a little bit, so we'll see how it does. From here on out, any new post about a specific class (questions about professors, workload, waitlisting, withdrawing, etc.) will be redirected to this thread and removed to try to reduce the clutter around here. Also be sure to check in to answer any of your fellow Terps' questions about classes you may have taken already, give some feedback, share any experience (positive or negative) etc. Any

Findings

GOOGLE FORM SURVEYS

For our research, we opted to employ surveys to confirm the findings of our data. Specifically, we conducted surveys targeting University of Maryland (UMD) students. To ensure the inclusivity and representativeness of our sample, we made deliberate efforts to encompass students from diverse backgrounds. This involved reaching out to participants across a spectrum of majors, encompassing various college years (including freshmen, sophomores, and juniors), and students residing both off-campus and on-campus.

By specifically targeting UMD students, we aimed to capture firsthand accounts of their experiences with Reddit, shedding light on their usage patterns, preferences, and decision-making processes influenced by the platform.



CMSC396H Reddit Poll

Have you ever used Reddit when searching for information related to academics or housing (r/UMD is one example)? *

○ Yes

○ No

How influential was it in your decision making? *

	1	2	3	4	5	6	7	8	9	10	
Not influential	○	○	○	○	○	○	○	○	○	○	Very influential

Why Reddit (r/UMD)? *

Long answer text

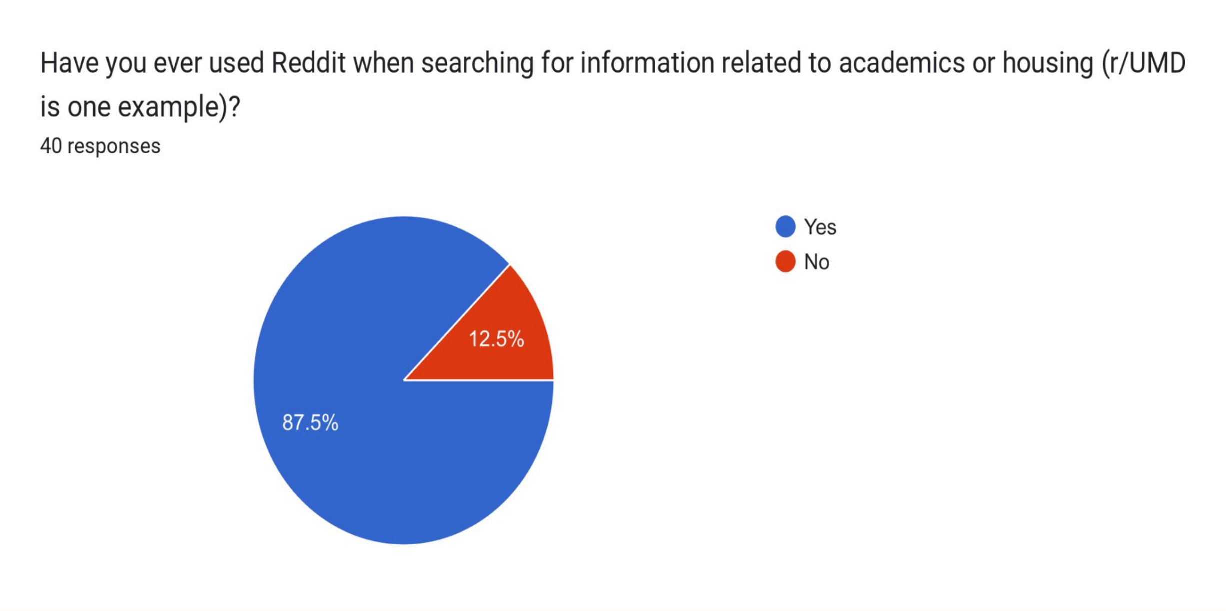
Approximately how much time did you spend searching for information? *

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

If you've made a decision factoring in information found online, do you feel that your decision was correct? *

○ Yes

Analysis

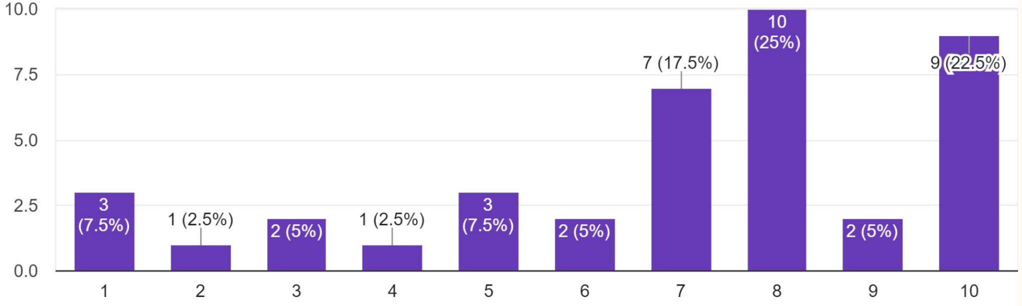


Have you ever used Reddit when searching for information related to academics or housing (r/UMD is one example)?

40 responses

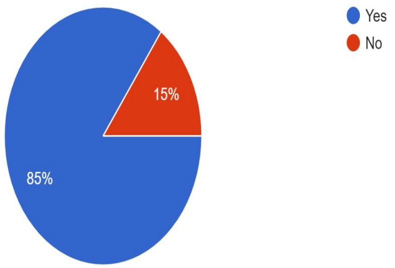
How influential was it in your decision making?

40 responses



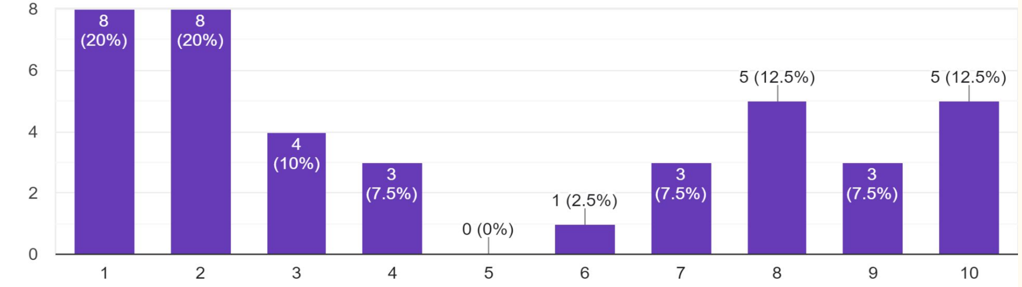
If you've made a decision factoring in information found online, do you feel that your decision was correct?

40 responses



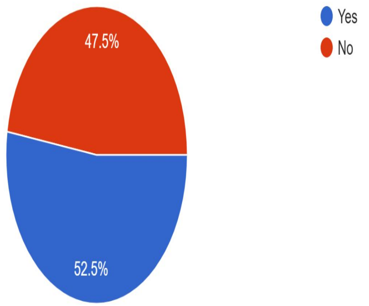
Approximately how much time did you spend searching for information (in hours) ?

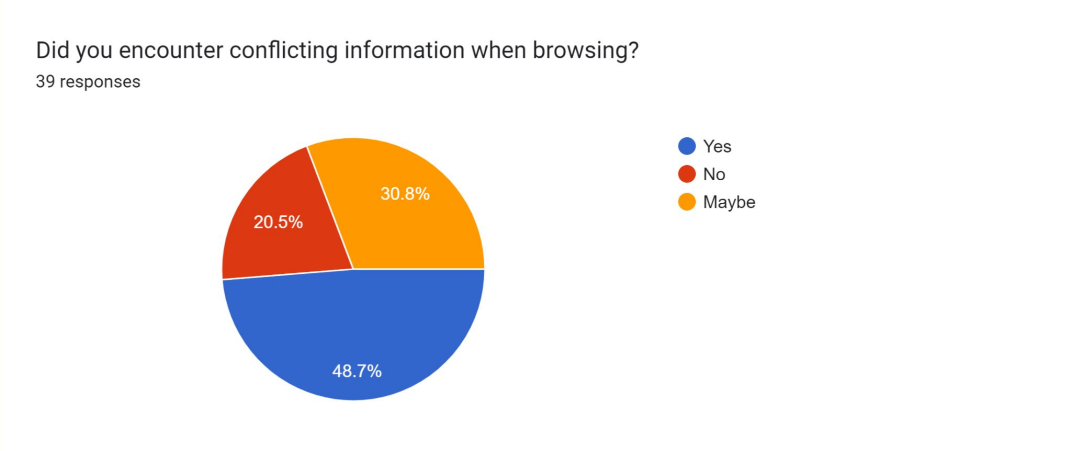
40 responses



Did you use any other sources besides Reddit?

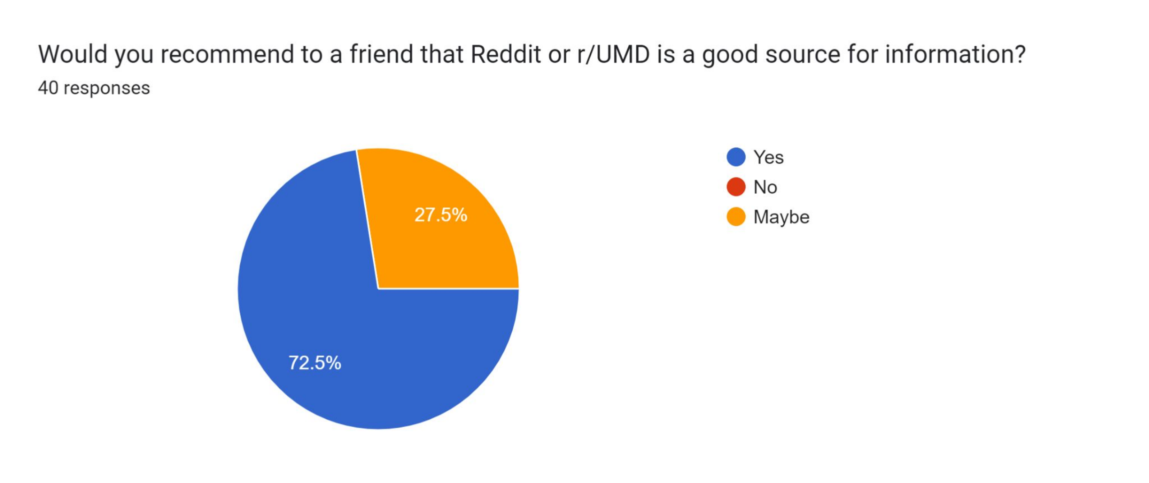
40 responses





Did you encounter conflicting information when browsing?

39 responses



Would you recommend to a friend that Reddit or r/UMD is a good source for information?

40 responses

Conclusion

- **Reddit’s Influence on Decision-Making**: Our analysis of the r/UMD subreddit has shown that Reddit is a pivotal source for UMD students, particularly in gathering insights on academic and housing decisions. This reliance is likely because highly upvoted comments and popular posts are perceived as more credible and informative.

- **Assessing Information Reliability**: Despite Reddit’s popularity among students, the variability in content trustworthiness necessitates a critical evaluation approach. This is crucial considering Reddit’s user-generated content where anonymity and varied user expertise can influence information accuracy.

- **Social Signals and Post Popularity**: Reflecting on related works like "Social Signals that Drive Reddit Communities," it's evident that user karma, sentiment, and content relevance play significant roles in influencing the visibility and credibility of posts. These factors are instrumental in shaping students' perceptions and their subsequent decision-making processes.

- **Comprehensive Data Extraction and Survey Analysis:** Using the Reddit API, we targeted and analyzed posts from the *r/UMD* subreddit related to student queries about academics and housing. The survey we conducted further substantiates our data extraction, indicating a strong correlation between Reddit discussions and student decisions in these areas.
- **Future Research and Tools for Credibility Assessment:** Based on our findings and the gaps identified in "Predicting Reddit Post Popularity Via Initial Commentary," future research should explore the mechanisms through which social signals affect user behavior on Reddit. Additionally, we should aim on developing educational tools to assist students in critically assessing online information's reliability. Lastly, machine learning techniques specified in previous papers would allow for predictions to be made about certain posts relating to the likelihood of their popularity.