Sentiment Analysis of the UMass Subreddit

Rebecca Lee

Stephen Harris | English 491DS-01

12 May 2021

Abstract: This project uses a Python program to scrape data from the social media platform Reddit, particularly from the UMass subreddit, and performs a sentiment analysis on the collected posts' titles and body text. It looks for language (particularly in words and sentences) that convey positive, neutral, and negative sentiments. In addition, the program calculates the average sentence length of each posts' body text. The aim of this project is to investigate subreddit members' general sentiment about the topic of housing and the specific kinds of content they post and talk about. The project also aims to see whether there is a correlation between the length of a posts' text and certain sentiments. One important conclusion of this project is that many users of the UMass subreddit tend to post significantly more neutral content than sentiment-laden content.

Website: The project is hosted on Google Colab at this link.

Specs: I used Python 3.10.4, the Natural Language Toolkit, TextBlob, and Praw (the Python Reddit API Wrapper). The project was coded on Google's Colab Research site.

Contact: For more information, please contact Rebecca Lee at rebylee@umass.edu

Introduction

Reddit is a social media platform that essentially functions as a forum for people to post comments, images, videos, news links, etc. to share with other users of the app. Within Reddit, there are "subreddits," which are merely subforums with a specific topic. For instance, the subreddit "askreddit" is a forum for people to post questions of all types for other users to answer. Likewise, the UMass subreddit is a forum for members to post thoughts, questions, opinions, etc. that have to do with the school.

The UMass subreddit has long served as a place for (prospective) students and parents to voice their thoughts, opinions, and questions regarding certain topics or issues related to the school. As a frequent user of the UMass subreddit myself, I have noticed that the topic of housing is typically a popular subject to talk about during specific times of the year. For example, during the spring semester, specifically the month of March, many students talk about room selection times when preparing for selecting their housing arrangements for the following fall semester. In addition, many prospective students during the summer or right before the start of a semester often ask questions about housing, such as an ideal spot on campus to live at. However, while these are just a few kinds of content posted on the subreddit, there are a range of others as well. Since the topic of housing is typically a recurring one and encompasses many kinds of content, this project aims to perform a sentiment analysis on the first 100 posts on the results page given the search term "housing." By conducting a sentiment analysis, we can conclude what users' thoughts and opinions are about housing in general and whether they are mostly positive, negative, or neutral. We can also infer what kinds of housing topics users are talking about on the subreddit by looking at sentiment analysis data. Lastly, this project also aims to see whether or not there is a correlation between positive or negative sentiment and the average sentence length in posts.

Method

In order to collect data, the project utilized Google Colab as an integrated development environment and the coding language Python to scrape data from Reddit. Since the UMass subreddit has been around for over a decade, the scope of the project was limited to the first 100 posts that appeared on the results page under the search term “housing.”

Part 1.1 of the project entailed importing packages to extract data from the UMass subreddit. For convenience and easy access to built-in methods, the project utilized Praw, the Python Reddit API Wrapper to scrape data from the platform. The “datetime” module was also imported in order to get the timestamp of each Reddit post (also called a “submission”).

Part 1.2 of the project involved the bulk of the data extraction. Using Python code and Praw, the subreddit name, subreddit title, subreddit description, the title and body text of the first 100 “housing” posts, and the timestamps of each post were all extracted and printed to the console. Once the data was scraped from the platform, each posts’ title and body text were stored in a dictionary variable called post_dictionary as key-and-value pairs. However, since the data needed to be reformatted to omit emojis, paragraph breaks, extra spaces, and account of missing punctuation marks, the title and text of each post were separated into two different string list variables. For instance, title_list was used to store the titles of each post while text_list was created to store the body text of every post. Separating the data also allowed for easier access, better organization, and better formatting when printing the data.

In Part 2.1 of the project, the scraped Reddit data was analyzed for sentiments using a data set of emotion words from the National Research Council of Canada. First, a csv file was uploaded and imported into the project file, which contained a list of emotion words in English. Once the csv file's data was read into the project file and formatted correctly, a list of English stopwords were stored into a list variable as strings. Then, a set of integer variables were created to keep count of the number of positive and negative words, as well as the number of emotion words encountered in the Reddit data. A separate integer variable was also created to keep count of the total number of words in the 100 sample posts collected. Next, NLTK and TextBlob were imported into the project file since some built-in methods were later used in the sentiment analysis.

When performing the sentiment analysis, there were two methods used. For the first one, data from the NRC was imported and used as previously stated. For each of the 100 posts collected, the code looped through every title and body text and separated the data into sentences. The data was then printed to the console for the purpose of viewing the data easier. Then, the code looped through each title and text in the two lists title_list and text_list, split the data into words, and stored them as a list of strings. The total number of words in each title and text was updated in two separate variables (one for titles and one for texts). After, each word in every title and text was checked to see if it was included in the list of stopwords; if it was not, the code continued to check if it was included in the list of words in the NRC data. If the word also belonged to the list of NRC emotion words, the code checked which emotion it correlated to, and whether it was a positive or negative emotion word. All of this data was later printed to the console. Additionally, TextBlob's built-in sentiment analysis tool was utilized to provide more insight to the data. Using a similar process as before in terms of looping through the data, the code performed a sentiment analysis on each posts' title and text content broken down in two ways: by post (general sentiment analysis) and by sentence of each post (more detailed sentiment analysis).

Lastly, Part 2.2 of the project calculated the average sentence length for each posts' text content. Since each posts' title content was relatively short (usually less than one sentence long and only a few words), the project focused on calculating the average sentence length of every posts' body text. This was achieved by looping through the list of text (text_list) and separating the text into sentences. Each time a sentence was encountered, an integer variable that kept track of the number of sentences in a particular posts' text was incremented by one. In addition, each sentence was then broken down into words; the number of words in a specific text was later stored in a word counter variable. To calculate the average sentence length for a particular posts' body text, the word count was divided by the number of sentences. In order to account for zero division errors when running the code, if-else statements were used; if a post's text was not empty and contained content (i.e. the word counter was not equal to 0), then the average sentence length for that post could be calculated using the formula previously stated. Else, the number 0 would simply be printed to the console.

Results and Data

Based on the sample data collected from the first sentiment analysis (utilizing the NRC's data), the UMass subreddit appears to have more positive sentiment in the content related to the topic of housing. According to the data, there were a total of 7,220 words from the 100 posts collected, including words from both the titles and body texts. Of these 7,220 words, 682 were found to appear in the NRC data and were associated with positive sentiment; this means that about 0.09% of the sample words collected had positive sentiment. Meanwhile, 453 words of the 100 posts had a negative sentiment, calculating to about 0.06% of the total sample words collected. Based on the words accounted for, there were 90 anger words, 367 anticipation words, 78 disgust words, 150 fear words, 210 joy words, 322 sadness words, 143 surprise words, and 421 trust words. Therefore, ranking the word types in order based on frequency, the word types would appear as follows: trust, anticipation, sadness, joy, fear, surprise, disgust, and anger. These findings are interesting because it was originally assumed that the UMass subreddit would include more negative sentiments on the topic of housing, particularly with the housing arrangements and selections prior to the start of a semester. In fact, many subreddit members have voiced their complaints in the past about the school over-admitting students with little available housing facilities. However, the data clearly proves this assumption to be false, as the ratio of positive words to negative words is much larger than expected. In fact, the data also gives insight into the specific housing topics talked about in the subreddit. While many of the negative words point to topics of complaint or concern about finding housing arrangements with a large student population at the school, many of of the positive words point to other housing topics, such as looking for potential housing options (i.e. asking questions about housing), giving advice and resources about housing, etc. For example, the word "interested" (which contains positive sentiment) was used in a post where a user invited other users to directly message them about possibly becoming roommates since housing might be cheaper that way. However, positive and negative sentiment do not always predict the type of content that gets posted, as there are some negative words that appear in posts about asking for advice, where the negative word might be used to describe an individual's past experience before making a decision. For instance, the word "regret," which has negative sentiment attached to it, appears in a lengthy post from a user that asks the subreddit for advice about where they should live for housing arrangements. The user describes their previous experience going to a community college but "instantly regrets it" since they would like the true college experience. Therefore, the sentiment of a post can be sometimes be correlated to the type of content posted, but not always (e.g. the post is about someone's complaint about housing, the post is merely asking a question about housing, the post is sharing resources and advice about finding housing with a positive tone, etc.).

Looking at the sentiment analysis tool provided by TextBlob, the data shows that many of the post titles are neutral (as they have a polarity and subjectivity values of 0.0 or a value close to that). Taking a closer look at the data conveys that the majority of the titles are short and simple ones that indicate the post is about asking a question related to housing. For instance, the titles "Housing" or "Housing Removal" both have polarity and subjectivity values of 0.0, meaning both are neutral in sentiment and objective. These types of posts simply ask a question about housing and are easy to categorize. The second most common type of post titles are those that have a negative polarity with a range of subjectivity values (some posts have negative sentiment indicated by the negative polarity value but can have a low or high subjectivity value). For example, the title "Why it's so difficult to find housing in Amherst area" has a polarity value of -0.5 and a subjectivity value of 1.0; the post includes a news article link that talks about the housing crisis in the Amherst area with the rise in rental prices. Meanwhile, the title "Housing Appointment on the 28th, chances of a single?" has a polarity value of -0.07 and subjectivity value of 0.21. This post's title might be categorized in this way since it sounds as if the user is concerned with their housing appointment time (i.e. it appears later in comparison to the housing appointments of others) and is therefore afraid of not getting an ideal housing arrangement (i.e. a single dormitory room). Compared to the first example, however, this particular post title has a much lower polarity and subjectivity value since their topics and intention behind the post are different; while the first post aims to bring attention to and somewhat criticize the housing crisis in Amherst, especially for students looking for off-campus housing, the second post merely voices their subtle concern over getting preferred housing arrangements. Likewise, the sentiment analysis performed by TextBlob also shows similar findings about each posts' text data, or body data within each post. In other words, the title of each post seems to somewhat mirror the type of sentiment and subjectivity value exhibited in the text data. Using an example from before, the post titled "Housing Appointment on the 28th, chances of a single?" includes the body text: "Going to be a junior next year and my priority is 900. I have my apt tomorrow and I want a single, but I have no idea if there's even any chance. Any one who might know?" The text of this post has a polarity value of 0.18 and a subjectivity value of 0.4; as mentioned before, the title of this post has a polarity value of -0.07 and a subjectivity value of 0.21. At a first glance at this data, the subjectivity values of the title and text remain relatively close and low. However, what is interesting is the disparity in polarity values. Although the polarity values between the title and text are somewhat similar in the sense that they are close to the neutral value 0.0, it is clear that TextBlob's sentiment analysis tool reads the title as negative and the text as positive. One might disagree with this interpretation of the post's text due to the user's sense of doubt and worry despite their good priority number (for context: a lower priority number is better than a higher one when choosing housing at UMass; out of the tens of thousands of students admitted, this reddit user's was 900). Overall, while the text data seems to include more positive polarity values, they still remain relatively close to 0.0 which means there is more neutrality involved in the text data's sentiments. Therefore, within the text data, neutrality seems to be the most common sentiment (although subjectivity values have a wide range) while the second most common sentiment appears to be negative.

Lastly, the project analyzed each posts' text data for average sentence length. The title for each post was ignored since the majority were short and therefore did not hold significant value to the study. Based on the data shown by average sentence length, there is no correlation between the average sentence length of a text and its sentiment. Prior to the study, it was assumed that longer sentences and text would indicate stronger (negative) sentiment, which is typically seen on product review pages. However, since Reddit is a different platform with different uses, this assumption cannot be applied here. As an example, one particular post's text reads: "Does anyone know if the admin has a plan to address the overcrowding or no? It gets worse every year and this year it's really bad." The polarity value of this text is -0.55 while its subjectivity value is 0.63; meanwhile, the text has an average sentence length of 13.5 words. While it may seem that the average sentence length is high given that there are 27 words in the text, there are only two sentences, which means that the text is not as long as one may think. Additionally, another post's text reads:

> It's now February 4th, and I was cleaning out my emails, only to see an email from 2 weeks ago saying that I must check in for the spring 2022 housing assignment…and I never did. It says 'If you do not properly check in, your assignment will be cancelled as a no show and you will be subject to the cancellation fee as of Thursday, January 27th at 3pm.' It says this however, I've receive no emails about being kicked out, and I still have full access to the residential hall, my room (still have key), and I even go to my RSD to pick up packages that are addressed to my room. Should I be worried? I've sent an email, but should I still be scared that at some point i'm going to have to leave my room?

This text has a polarity value of 0.04 and a subjectivity value of 0.60 with an average sentence length of 37.25. Compared to the previous example, this text is clearly longer and includes four sentences with a total of 149 words. However, contrary to previous assumptions, the data yet again shows that sentiments do not necessarily predict or correlate to the length of a post and its average sentence length. It is true that this particular post is longer, but its polarity value is relatively close to the neutral value 0.0. On the other hand, this text's subjectivity value is more subjective, which might explain the increase in average sentence length. Based on the text in the post, the user asks if they "should be worried" which indicates possible worry and concern over their housing situation. In general, however, the text data indicates that average sentence length does not necessarily correlate to sentiment.

Conclusions

Overall, this project researched the UMass subreddit using Python and Praw to see what kinds of sentiment members have surrounding the topic of housing, the types of housing content posted in relation to sentiment, and to determine whether average sentence length is correlated to certain types of sentiment. Based on the sample data, the UMass subreddit consists mostly of content with neutral sentiment with a range of subjectivity values, indicating posts that ask questions, ask for others' thoughts and opinions about a certain issue, etc. Meanwhile, content with negative sentiment ranked second in terms of frequency and typically included content indicating a user's negative thoughts, emotions, and criticism over an issue (e.g. the housing crisis). Lastly, the project discovered that average sentence length does not necessarily correlate to a certain type of sentiment (i.e. negative sentiment indicates higher average sentence length), which proved the previous assumption as incorrect.

For future studies, there are limitations to consider, especially when working with computer code. It is clear that humans are much more complex than computers and have a better method at gauging human sentiment when reading text. While conducting this project, there were some titles that were ranked as neutral even though they originally appeared as negative in sentiment. For example, one post was titled “All that’s left on campus” and ended with a skull emoji to communicate negative feelings about the housing crisis at UMass. This particular user wanted to convey their frustration about not getting their ideal housing arrangements for the fall semester of 2022, but TextBlob assigned it a value of 0.0 for its polarity and subjectivity. In addition, the original content of the body text of this post included a screenshot of a small number of housing options available on campus. However, because it is difficult to use sentiment analysis on emojis and images, the study omitted anything non-text based and solely focused on analyzing texts. Another limitation that must be considered is the NRC data; many of the words scraped from Reddit were not found in the NRC database and did not match that source’s emotion words. For future studies, it might be helpful to use a database that includes a wider range of words to ensure more accuracy and increase the scope of the project. Therefore, these are some limitations to consider when using such tools and technology to import and analyze digital data.