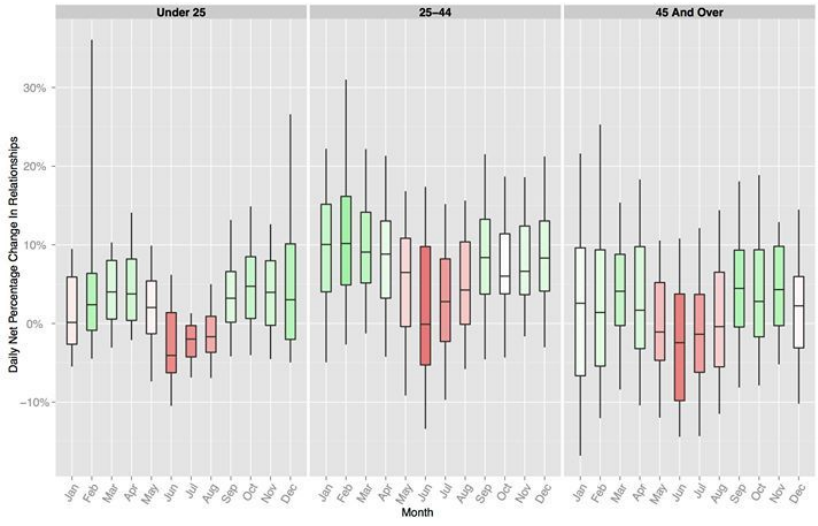
Relationships & Sentiment Analysis of the Boston Community via proxy of Social Media

By Zeke D'Ascoli, Seth Friman, Anika Rabenhorst, Nick Harper, Tommy Sarni

Simple, Sentimental Surveys

After much deliberation, we settled on a topic focusing on sentiment analysis of social media posts from throughout the Boston area. We had hoped to map time against sentiment analysis and see that events were actually having an effect on the overall Boston community, and wondered if relationships would be correlated. Long story short: what we found did not support this hypothesis. In fact, many of the visualizations that come from our data does not prove much. Nonetheless, the process was still interesting. And we ended up going down other paths.

Throughout this project, we faced many challenges in finding useful data. Last week, in our presentation, we spoke of our likely sources: a self-produced survey of Northeastern students, an already produced analysis of Facebook relationships, Twitter data, and the Boston subreddit. Unfortunately, we quickly found problems with accessing and utilizing some of these sources. Twitter, for instance, would only allow us to look deeply into things posted within the last week or so, the specifics of which are included below. Because our project dealt a lot with comparing over time, this was staggering news. The ability to track certain hashtags was now gone from us. Reddit also placed limits on what we could download but wasn't nearly as strict. Zeke was able to download the first 10,000 comments for each month of 2018 from the r/Boston subreddit. We ended up using this for a variety of visualizations. We didn't have any large complications with Facebook because we were already limited to the report prepared by data scientists Jackson Gorham and Andrew T. Fiore.



This chart is from the "Right Time for Love" report. Tracks changes in relationship status on Facebook by the month. Also separates for various age ranges. ***NOT OUR CHART***

The aforementioned report was an inspiration for this

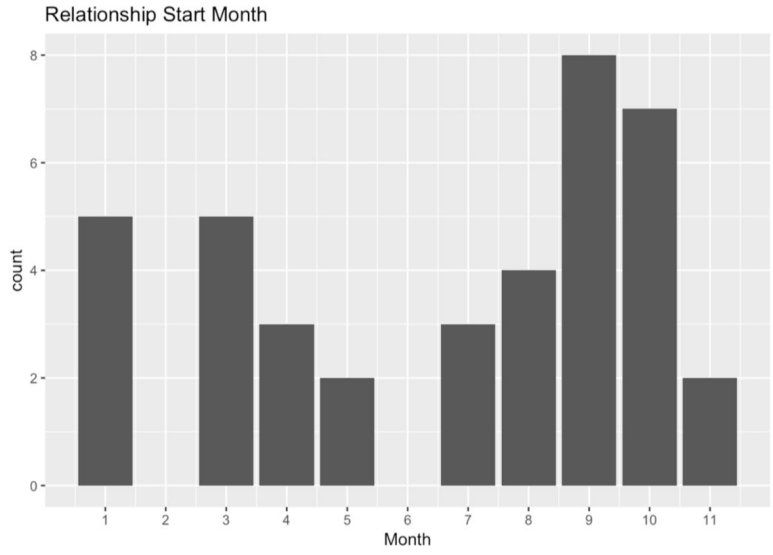
project. Using U.S. Facebook data from 2010 and 2011, Gorham and Fiore were able to look at the lining of up relationships start/end and times of the year. They thought a lot about net gain or loss or relationships over a certain time period, even narrowing down to the holidays that were most popular for coupling. Obvious contenders included Valentines Day and Christmas, along with the days surrounding these holidays. Their analysis also showed distinct differences in the seasons, with summer having the biggest net loss of Facebook relationship statuses being dating/married. We assume this has to do with the presence of summer flings versus winter "cuffing season." People often want someone to bring home with them for the holidays. The reason why I (Anika) want to emphasize the specificity of Facebook statuses is that they vary from actual relationships. In my personal experience, it would take a lot for me to make a relationship "Facebook official." This is a difference that comes from the data here being 8/9 years old. Additionally, I have noticed that our generation is not likely to be very active on Facebook, preferring apps like Instagram, or Snapchat. If one were to look at this same data 8 years later I would expect that the overall proportion of coupled statuses for users to be lower among the "Under 25" group than it was in 2010.

While we can learn a lot from other people's studies, it's even more important to do our own analyzing of data. With our focus centered on the Boston community, we dove deeper into Northeastern students. After careful consideration of what questions are appropriate to survey, we sent out the "Relationship Survey"; it gathered 52 responses. This was far less than we had hoped for, and once again, a setback. Even worse, only 12 responses came from women. Although the number of responses is lackluster, we still believe that this survey was worthwhile. Especially for the laughs coming from trying to assign sentiments to the three words responders would use to describe their relationships.

While the surveys do not have as many responses as we hoped, they still give insight into the college dating and relationships on campus. We first wanted to take a look at the demographics of who is answering the survey. We originally wanted to go into deeper with this including sex, and sexual preference along with other things, however Northeastern University policies had restricted us. So we are only left with the users age and gender. Though with minimal responses from women, that meant that we could no longer differentiate between men and women when analyzing and treated the survey population as one. So age was our biggest differentiating factor. This is very important for the survey because we can see that the results of the survey come from more new college students rather than some that have been here for a while, meaning that many of the relationships that are recorded are either coming from high school experiences or during freshman or sophomore year. This is important because those relationships would (hopefully) vary from relationships of juniors and seniors due to the maturity of the individuals and the relationship itself.

Our final survey used the following questions for demographics: (1) age, (2) gender, and (3) college within the university. Then we separated responders for deeper questioning based on the following question, “Have you been in a relationship in the past 4 years?” This broke off the students off into three groups: those who are currently in a relationship, those who have been in a relationship within the last 4 years, and those who have not been in a relationship within the last 4 years. When a user was filling out information about a past or current relationship they were asked the following questions: when did the relationship start and end (if finished), was the relationship long distance, and what three words would describe the relationship. For those who were currently single, only one more question appeared: “Reasons for being single.” There were two pre-written options and an option to input another reason. The responders overwhelmingly chose “I haven't met anyone I wanted to date,” with only one person choosing “I don't want to be in a relationship” and one responder choosing to add their own reason.

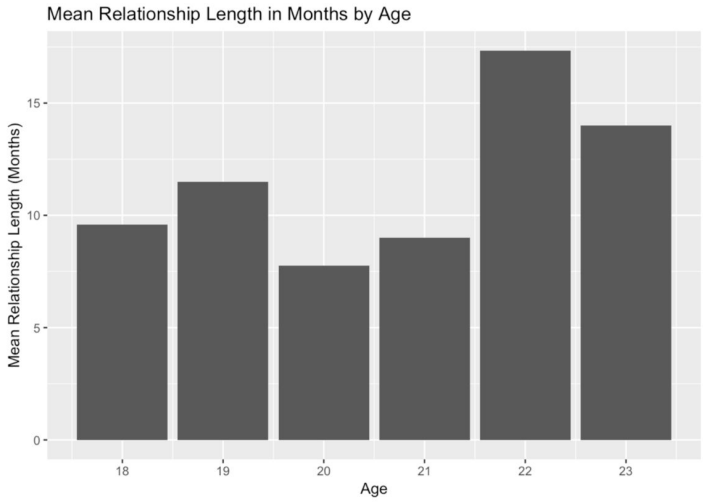
We also looked into the months when relationships start and end to see if we can compare to Facebook’s data about relationships. We did find that the dates people inputted may have had the wrong year, 2019. One example was a relationship supposedly lasting from October 2019 to November 2019. We were not surprised to have seen this problem, though it is frustrating. When looking at the figure depicting the starting months of relationships, there is a certain accuracy to what Facebook predicted.



September and October have the greatest amount of beginnings in the survey and they are some of the months where more people get together, however they are not as popular as some of the colder months according to Facebook. One guess for why this could be true for college students specifically is that the school year starts in September. For first-years, this means meeting a lot of new people within a short time period. Any one of the hundreds of people you meet within your first month of college- through classes, residence, and clubs- has the potential to spawn into a relationship. The breakup data we collected did not contain enough results to really tell us anything because there was at max three breakups for a

month, so there is not a big enough disparity between months to compare it to Facebook's data or to other months itself.

With both the start and end dates of the relationships, we were able to come up with how long each relationship last and sort them by the age of the individual. The results for this were very interesting because they confirm the idea that the older you are the longer your relationship will be. There is a decent jump between eighteen to twenty-one and twenty-two and up. This would make sense as these students get older they would be interested in more serious and longer relationships rather than flings that do not last long at all.



One of the most intriguing things from the analysis was when we ran sentiment analysis on the words submitted by responders about their relationships. Due to the nature of the survey and the way we collected data, typical sentiment analysis was not possible. Additionally, many people are not very good at following directions and submitted answers that were not in the format that we required. To compensate for this, we created our own sentiment analysis that came with a few ground rules. First, every word is either positive (+1), neutral (+0.5), or negative (+0). Second, every score is on a scale from one to four. One is the minimum score and means that this person had all negative sentiments. A score of four means that this person used only positive words to describe their relationship. But, since people did not always follow directions we had to establish what to do in these situations. For submissions where the only response was "Good", they were missing two of the words. To compensate for that we treated "Good" as a positive sentiment, and the remaining two words as neutral as there was no positive or negative. Consequently, this submission would receive a final score of two. Other responses contained phrases such as "Time consuming". We opted to treat these as one word rather than analyzing both "time" and "consuming" separately, as they are one sentiment. From this, the reduction of overall words in the submission meant that the remaining spot was considered neutral. While these ground rules were helpful and concise, it adds a level of subjectivity to our project. On group Facetime we sometimes had to discuss what categories best fit words. For example, one responder put the series "Intense sexual serious." If this was an already finished relationship, one might place "intense" and "serious" in the box of negative sentiment. However, as it's a current

relationship one might view “intense” as being similar to “passionate.” Though we did our best to account for this, this problem of subjectivity may have infiltrated our sentiment scale.

We proceeded with the process of visualizing this information to see if it produced anything interesting. We did not want to continue with the same scale of the x-axis being month of the year because we figured these variables would be completely independent. Rather, we needed a new measurement for what could affect sentiments. Ultimately, we went with relationship length. We wanted to see if there was any correlation between relationship length and feelings about the relationship. We took a mean sentiment for each length of the relationship and below is what we found:

We were surprised to find that an odd trend existed in the data. The shorter relationships, represented by the smaller x-values, had a larger amount of positive sentiments. This seems consistent with logic as there may have not been enough time to truly develop strong enough negative feelings about the relationship as it most likely did not become that serious. Additionally, there was a large spike in the longer relationships. Again, this is logical because for a relationship to have lasted a long time, there must be some fairly positive things to say about it. In the middle of these two positions, there was a dip in the sentiment. This is most likely due to relationships

having time to develop past the initial phase of being exciting and new, and sour before reaching long-term bliss. It should be acknowledged that we lacked a lot of relationship data for mid-length relationships. While we quickly drew this conclusion from the graph, it is quite possible that the visual representation would show up differently had there been more responses.

Reddiment

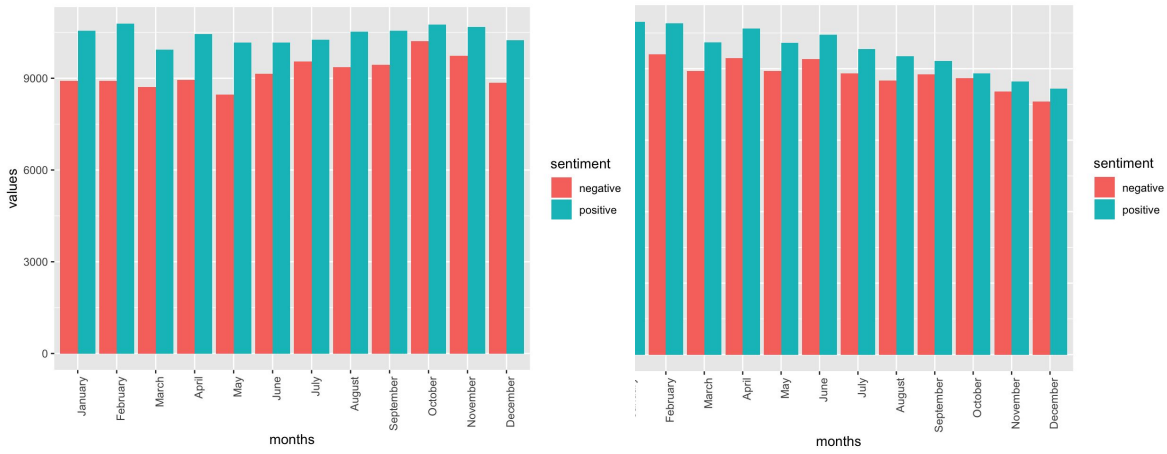
For another portion of this process, I (Zeke) wondered what knowledge we could ascertain about the Boston area through Reddit, and how these results stacked up with the other factors of our analysis. To best synthesize this with the relationship study, I decided to analyze the average sentiment of the Boston subreddit, or as I call it, the "reddiment". For any given time of year, I wanted to gauge the average reddiment of Boston, and hopefully use the comments as a proxy for general underlying moods of Boston. I was looking for any potential trends to compare with our other analysis.

Firstly, I needed a way to access the data I wanted: Reddit comments. Reddit is fully accessible via the public domain, so ideally I had access to unlimited historical data. Reddit also provides and maintains a useful Python API called PRAW (Python Reddit API Wrapper), and being comfortable in Python, I was excited to use. However, after roughly a week of forcing and hacking very basic functionality of what I wanted, I realized PRAW was not designed to sift through historical data. After a few more days of searching through public datasets, I finally found a cache of every single Reddit comment from 2018 saved into Google's BigQuery data collection. This can be found here. And even better, Google allowed up to a terabyte of free data. This dataset contained a lot of information, but what I cared most about was the text of the comment and what subreddit that comment came from. Using their built-in SQL Requester, I downloaded the first 10,000 comments of every month of 2018 for the Boston subreddit ("/r/boston"). I also thought it might be useful to compare my results for a different subreddit, "/r/relationship_advice", a community dedicated to giving advice for people struggling with relationship trouble. I also downloaded the same amount of data for this subreddit. I had a vague prediction of how much more polarized the "relationship_advice" sentiment data would be after I requested February's and March's comments. The first comment retrieved from February was "He wants [us] to move in together…". March's first comment was "When she stabbed me in the back and lied about the whole thing".

```
#standardSQL
select subreddit,body
from `fh-bigquery.reddit_comments.2018_12`
where subreddit = 'boston' LIMIT 10000
```

Now, with the comments, we were ready to analyze the data! I began by stripping all the comments down and running sentiment analysis with the R tidy* packages. The first lexicon I wanted to analyze was the "bing" lexicon, where dictionary words are

grouped binarily into either “positive” or “negative” sentiment. This allows for a fairly quick, but very brief, insight. Using techniques learned in class this semester, I created a program to count the number of positive and negative words out of each month’s 10,000 comments. Displayed below are the results.



Results from counting the number of positive and negative words per 10,000 comments per month. Boston sentiments on the left, Relationship_Advice on the right.

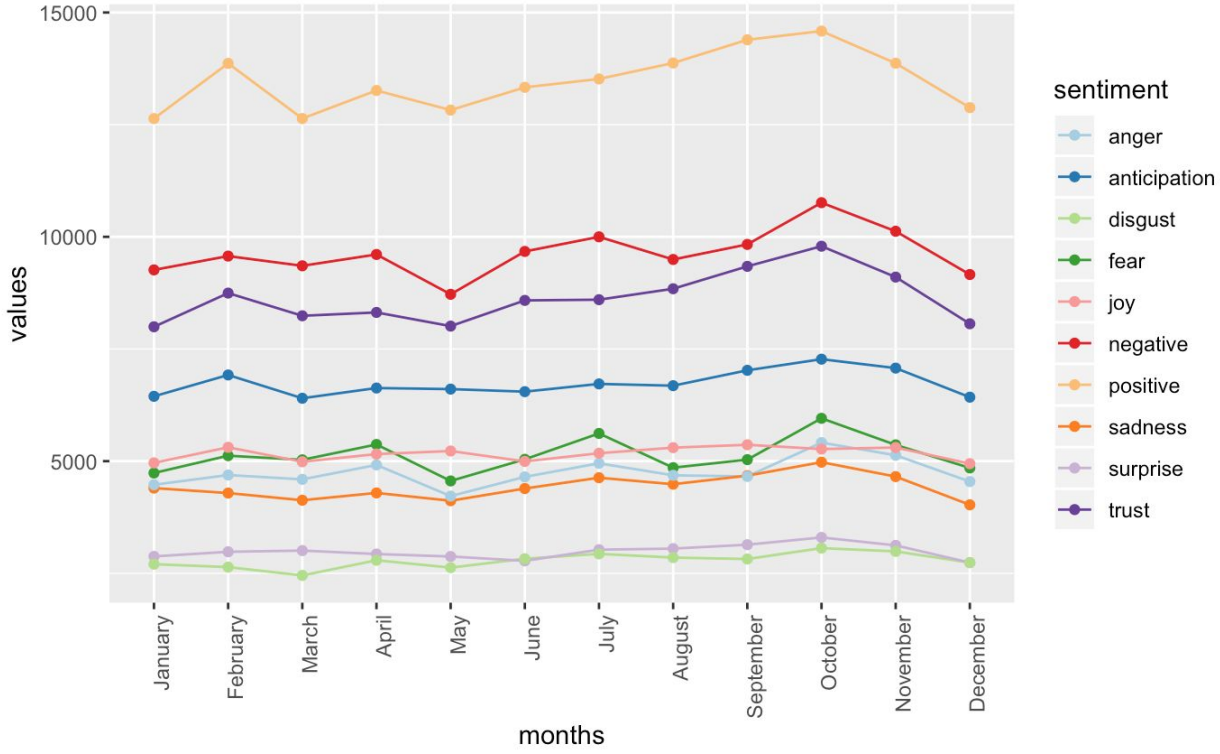
From the beginning, there is an apparent, but slight, sinusoidal trend of the Boston reddiment, while a strictly downward trend in the Relationship Advice reddiment. This may be because the holiday season can be an especially stressful time, bringing home new partners to your family or dealing with finding the perfect gift for the one you love, leading not necessarily to more negativity, but clearly to less positivity.

Also, the number of sentimental words per month approximately doubled. After further investigation, average length of each comment for /r/boston (of the 120,000 comments pulled) was 170.8174 characters. The average length of each comment for /r/relationship_advice was 291.9423 characters, significantly higher than that of /r/boston. I also graphed each mean character count per 10,000 comments per month for both subreddits, but there was very little change, so including the graphs would have been superfluous. This mean count implies that the standard commenter on /r/relationship_advice types more than the standard /r/boston commenter. This analysis was a good start, but I wanted to know more about the overall reddiments, not just positivity/negativity and average response length.

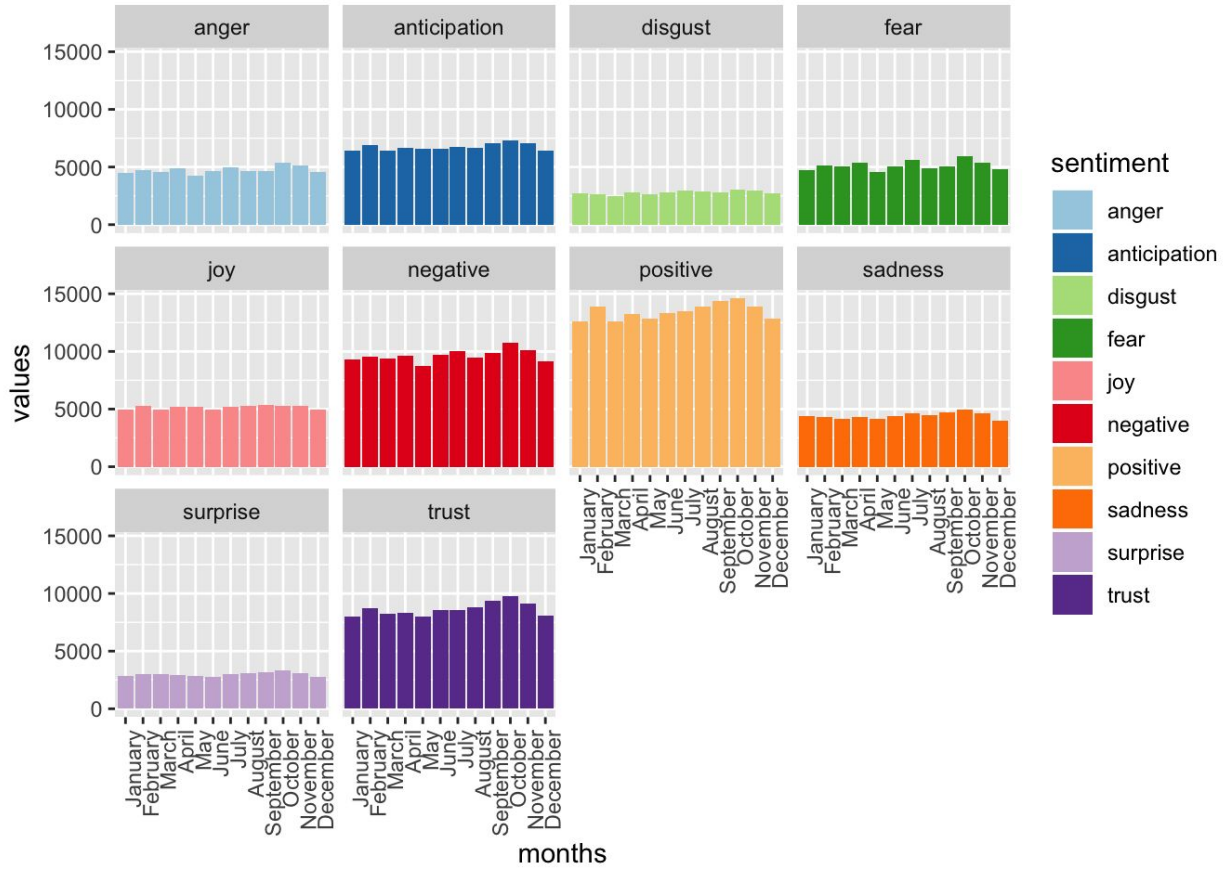
Following this, we chose to look deeper at ten different types of sentiment using a different grouping lexicon (NRC), which provided us with: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and the classics- negative and positive. We are unsure of whether or not words would be double-counted; for example a word being categorized under “positive” and “trust” at the same time. For this reason, we have

chosen to represent these sentiments both as a line graph together and facet wrapped bar charts. First, looking at the charts about r/Boston:

This chart was originally displayed with bars for each sentiment for each month, but we quickly realized that would be too crowded. We do want to stress that the lines

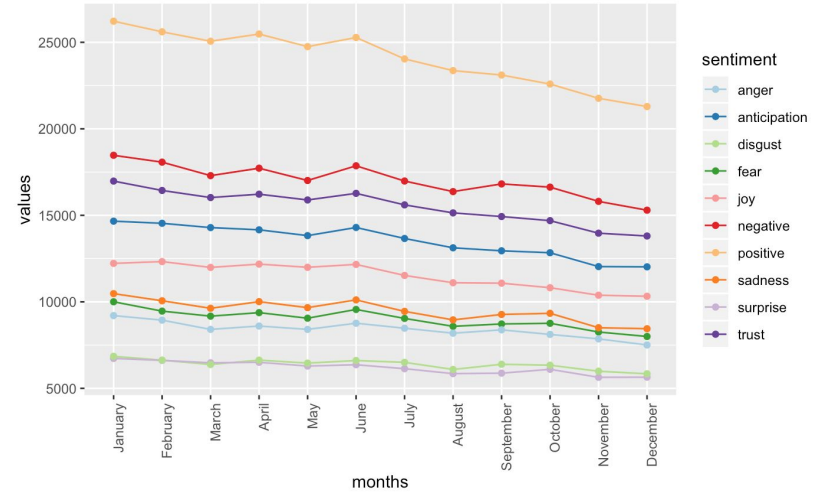


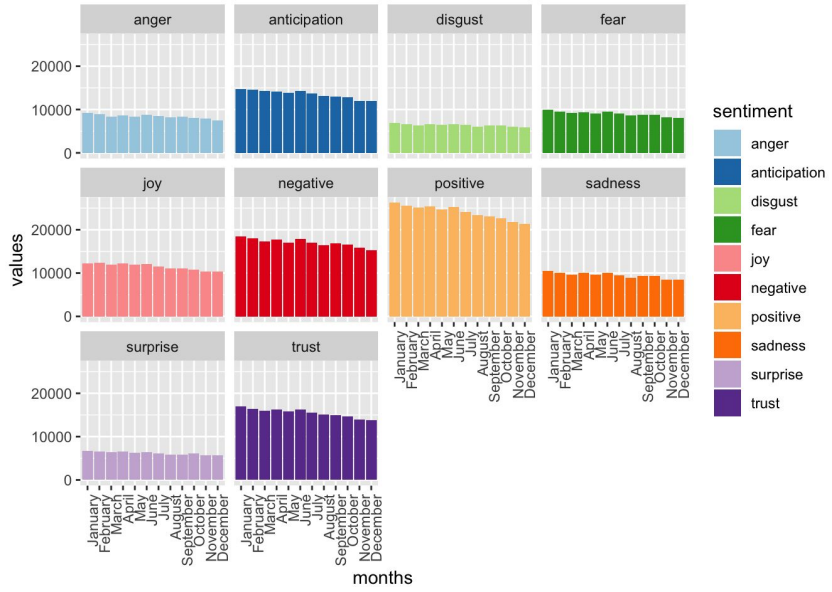
are really only connecting points for months, it should not be interpreted as a continuous variable. We can see that positive and negative are the most common sentiments, which supports the idea that words can have multiple sentiments assigned to them. We noticed that there is a general spike of emotional activity around February, a dip in March, and another emotional upswing in the fall.



The above graph is the facet-wrapped version of the previous data from r/Boston. Nothing huge jumped out at us besides the overall increase in most of the sentiments around September and October. This affected most of the sentiments tracked; this is surprising considering we have the same number of comments each month. This early-fall emotional upswing appeared to affect positivity and trust the most drastically. A possible reason for this is because in September and October, most college students move into school and consequently become more active in the Boston subreddit, sharing their thoughts on the new world. Or perhaps more beautiful nature gives rise to Reddit users commenting on the novel experience of a “New-England autumn”. It may even be combination of the two: new students enjoying a stunning fall!

Now, the same charts, but using the r/Relationship_Advice subreddit. These can be seen on the following page. In these graphs, we noticed that there was once again decreasing positive sentiment. One would think that means negativity would increase, but we didn't find that either. It seems that the advice given in the beginning of the year may be more positive, hopeful, and supportive. This is perhaps due to an idea of a "fresh start" at the beginning of each year, with more people desperately holding onto their New Year's resolutions. Unfortunately, this data did not line up well with the Facebook data or our survey. By "line up", we assume that positive, joy, and trust would be most associated with the seasons that had the most people getting together. Perhaps, however, the downward trend of the data means that less people need advice with their relationship, and are instead experiencing healthy relationships. Or, alternatively, the commenters providing advice on this subreddit are becoming more logically minded, giving less emotional or sensationalized advice and instead sticking to cold, hard facts.

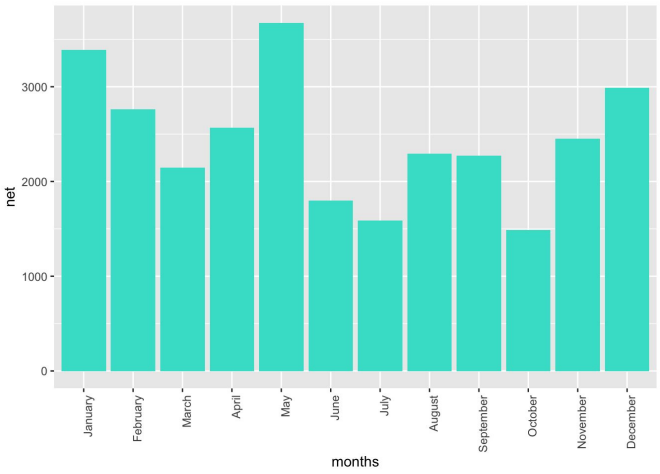




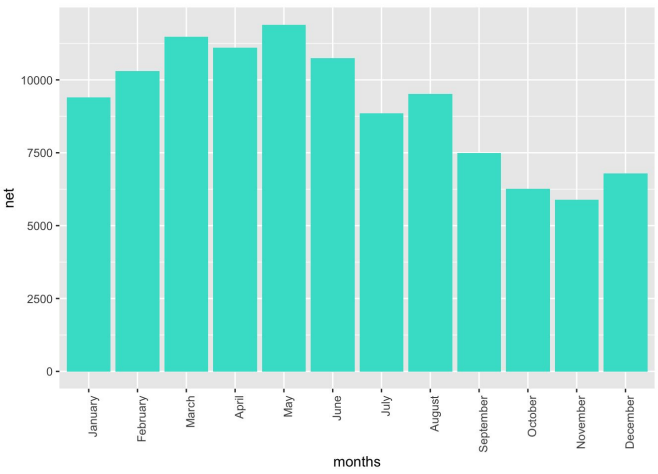
There was, however, another sentiment methodology I had not yet tried. As of right now, the lexicons I had used ("bing" and NRC) had both grouped words into binary categories of sentiment. The words in these lexicons were either positive or negative, but there was no ranking of *how negative or positive* a word was. This is good for quick categorization. However, under these constraints, the following two sentences would be considered emotionally equivalent: "I hate and despise him" and "He is a bad bad man". Clearly, though, these comments are not emotionally equivalent. Since both "comments" each contain two negative words ("hate" and "despise", "bad" and "bad"), simply counting the occurences of negative/positive words would not accurately determine the polarity of sentiment.

There was instead a third way of categorizing the comments. Under a lexicon titled "AFINN", each word in the dictionary is given a weighted score between -5 and +5 (-5 being most negative, +5 being most positive). This, I believed, would give a much better estimation of the "true" sentiment for each subreddit. I decided to refactor my code to work with this new "ranking" system, and then proceeded to compute the net-weight of all 10,000 comments for each month (the sum of all comments' scores).

/r/Boston data (left)



/r/Relationship_Advice data (right)



This was extremely unexpected. It was almost the opposite of the data we thought we had before. Boston's reddiment had a much different spike, however there was still a peak (albeit much smaller) between September and October, though there was now also one in May, followed by a rapid falloff. And on the other side, Relationship_Advice's reddiment score was almost perfectly sinusoidal. Both graphs also were a net-positive, leading us to believe either there were more words associated with positivity in the wordlist or people are just generally more positive. Optimistically the latter.

I wondered why the peak occurred in Boston around May, so I looked online for events that happened in May of 2018 in Boston. There were no overwhelmingly obvious

events, so I had to keep thinking. Then I realized that both most graduations for colleges and National Commitment Day are in May. Boston (or at least the subreddit) seems to be heavily influenced by students. Who would have guessed?

This new graph of relationship_advice made much more sense than a constantly downward sentiment plot. Here there is a definitive cyclical trend of when emotions are more positive within this subreddit. And, this fit much more clearly with the Facebook study, where the relationship curve had a maximum around springtime and a minimum around late-fall/early-winter. This was exactly the type of result we had been looking for when we started this project. This demonstrates that there is in fact a clear correlation between time of year and the sentiment within the subreddit /r/relationship_advice. To generalize this conclusion, people are more positive (beginning of relationships) towards springtime/summer and more negative (ending of relationships) towards fall/winter in regards to their relationships.

Twitter

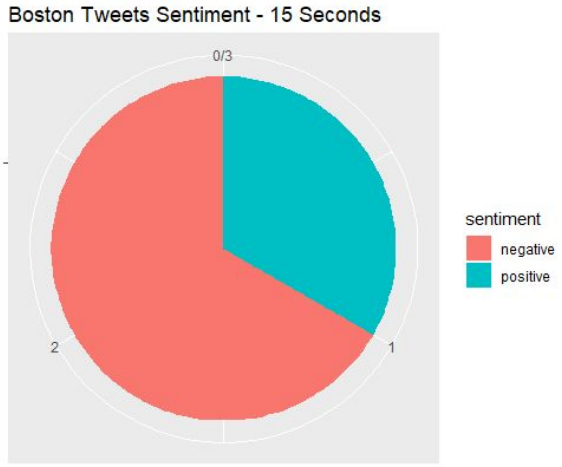
My (Nick) main focus for our project has been working with Twitter data in order to see what it may reveal about Boston and relationships, which had been our original goal. I had planned to see what kind of trends in hashtag use, phrases and emojis might reveal about when relationships start and end. One issue I expected to encounter was if there are even any hashtags or phrases or emojis that even correlate with aspects of relationships, such as when people get together or break up. However, before I got the chance to explore the different possibilities related to that, I encountered a bigger issue.

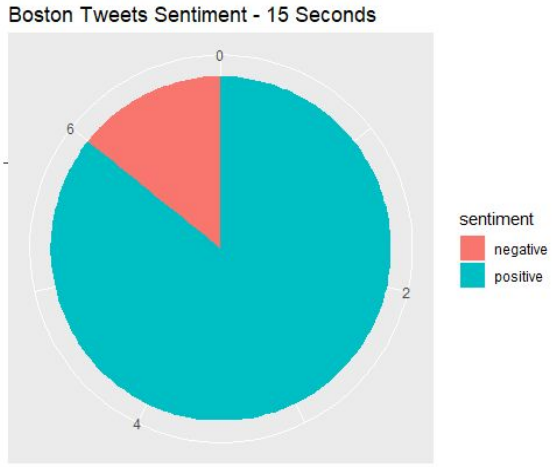
The data the Twitter allows to be accessed is provided on several different tiers, including paid and free options. Through the Twitter API, the free options limits the amount of tweets you can access and from specific periods of time. As a result, the frequency of tweets containing those items that I was looking for over the course of a year to see how it changed was simply not possible. I can not access that kind of historical data, and for the free option that does allow historical data retrieval, it severely limits the amount that can be retrieved. I looked for other potential sources that may contain large amounts of twitter data online, and did find a few that had been compiled for academic studies. The issue with these ended up being that the studies were specific to certain topics, and as a result the tweets they had compiled were unrelated to both Boston and relationships. At this point, I came to the conclusion that my hopes for a large Twitter dataset to analyze relationships with were a bust.

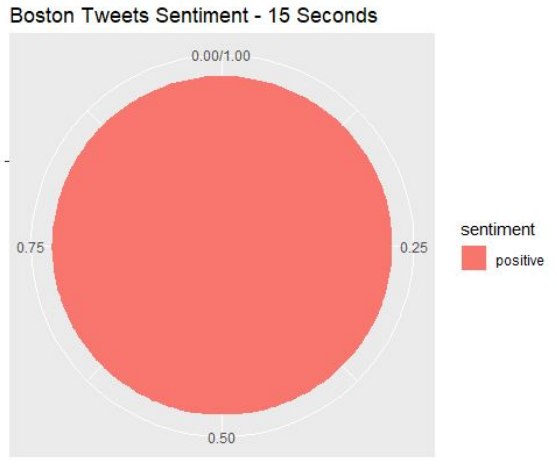
I did not want to give up on using Twitter to explore potential Boston trends, but I knew I would need to alter my focus. In doing so, I knew that I only would be able to worked with the severely limited data that Twitter would give me access to for free. This led me to RTweet, which would allow me to retrieve and use data with R, where I could visualize and analyze it. RTweet has several ways to get data, you can search for tweets containing hashtags, phrases. The limit is 18,000 tweets every 15 minutes, and there is no way to change when the tweets are from other than the default, which I assume is the latest tweets containing them. There is also the ability to search for tweets by geolocation and randomly stream about one percent of all tweets for a specified amount of time. There are plenty of other features included that relate to followers, trending, timelines and other Twitter features, however I determined they would not be useful for the kind of information I had hoped to find.

I decided to focus on randomly streaming tweets for various amounts of time, set to Boston on the geolocation, and do sentiment analysis to see if perhaps tweets from Boston tended to be more negative or positive. I also wanted to explore if the data I could get from random Boston tweets would even be able to provide me with any sort of valid information, considering that it is a small and random sampling of tweets from Boston. I was skeptical going into it, with little hope for consistent data, considering the nature of the source I was using and the small size of it. However I did still hope for

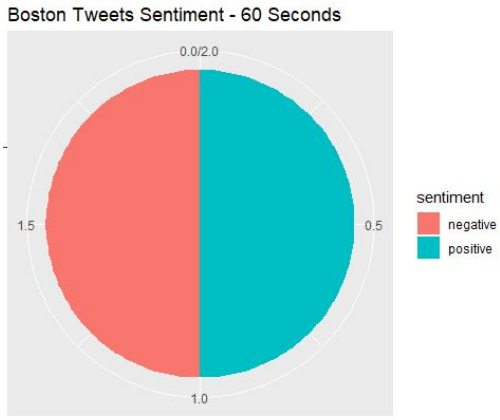
to find something interesting, even if it was not what I had originally set out to find. Using the geolocation feature of RTweet included using the Google API, just for the Google API key to include in RTweet. The most challenging and time consuming part of the project has been trying to figure out how to work with the Twitter data, that ended up being very underwhelming, and figuring out how to use R effectively. The conclusion that I came to about the data and changing focus ended up being similar to the experiences of my fellow group members, and Boston area sentiment analysis on social media ended up being more interesting and better to work with. Below are some visualizations of what I found out about sentiment analysis of Tweets from Boston over varying spans of random streaming times.

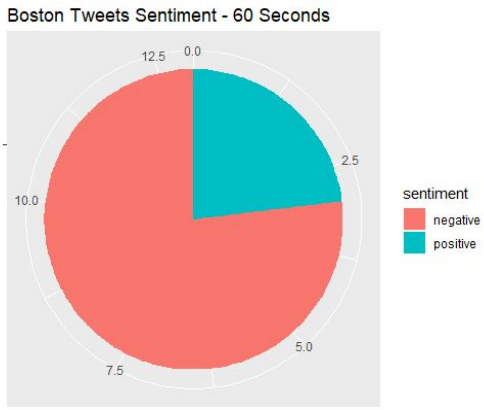


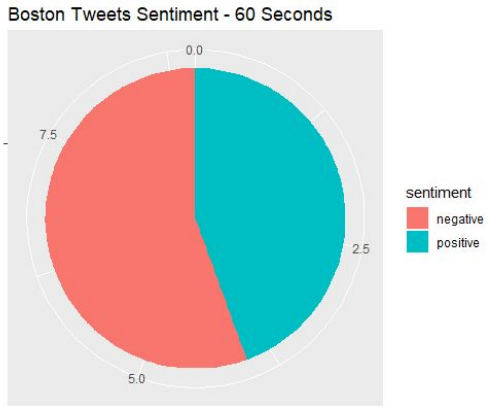


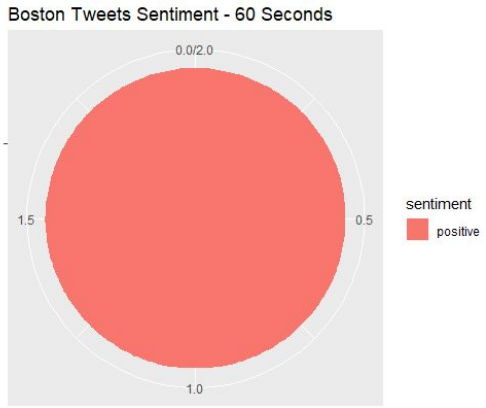


I started with 15 seconds of tweet streaming, which I quickly determined to yield far too few results. These would include only a few tweets, and from these tweets there would only be a few words that would even be determined to have positive or negative words related to them, having used the bing lexicon and summing up the amount of positive and negative words of all the tweets included. Some did not even include and positive or negative words to even visualize, so I have determined that this information is fairly useless and inconsistent.

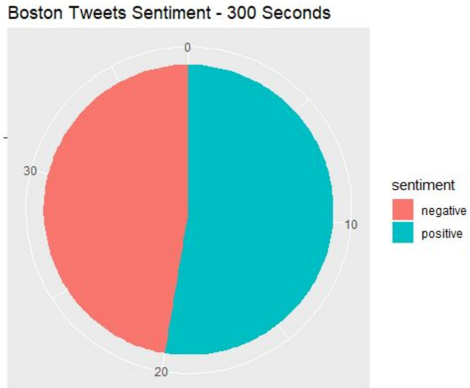


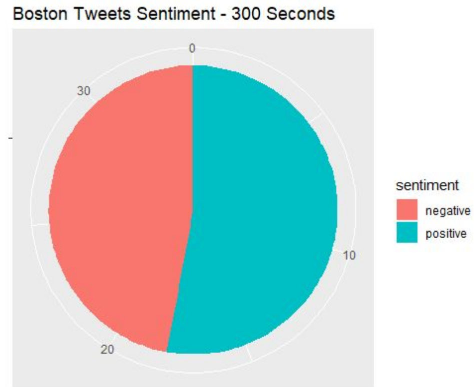


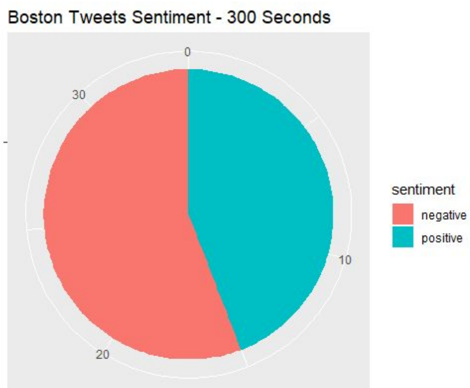


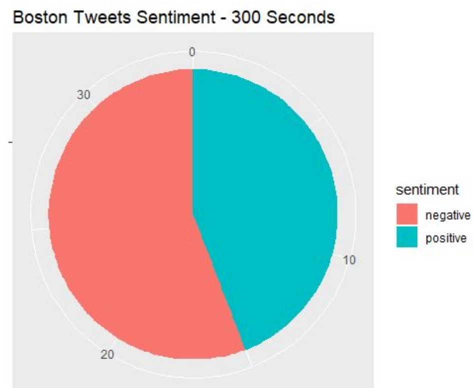


Above are some results from one minute of streaming. Once again, these yielded wildly inconsistent results. The sheer count of positive and negative words that were counted were lower than I was expecting. Even the amount of tweets was lower than I was expecting some of the time, yet other times was at a reasonable amount. This could be a result of the way the Twitter API works, or possibly just the amount of tweets in that time frame is less but I am not quite sure. Either way, it undermines any kind of trends that I was hoping for.









At five minutes of streaming, I have finally begun to see a more consistent pattern, albeit an expected one. While there is some swing back and forth, the average of each five minute stream ends up being just about 50/50. This seems like it would be the expected ratio. The general trend is the only thing I am confident in even suggesting with this type of data. Even with the five minute periods, there are only around thirty or so tweets that data is drawn from.

The Twitter data has ended up being an underwhelming and disappointing experience. It was a challenge to find anything useful to work with, and without streaming Tweets for an extended period of time, which I was unable to do, there are pretty limited conclusions to be drawn about Boston through a distant look at its tweets in the form that I have attempted to undertake, besides that the tweets appear to have a closely split trend of positive to negative sentiment.

Basic Conclusions

If this paper proved too dense or verbose, this section contains a "tl;dr" of our conclusions for each section.

Facebook:

- Facebook found that the most most positive net change in relationships status was in the winter; largest net loss in the summer
- Most popular dates for starting relationships surrounded major holidays, ie: Valentine's Day and Christmas

Survey:

- Sentiment analysis showed that responders with relationships on either ends of the spectrum in length spoke using more positively than mid length relationships
- Many problems with having a small sample, but found that relationships tended to begin in the Fall, around the start of the school year

Reddit:

- The Boston subreddit's sentiment lines up remarkably well with the presence of college students (more positive during student arrival, more negative during student departure)
- Within the Relationship_Advice subreddit, there is a clear association between the level of positivity and time of year, following a sinusoidal curve peaking in springtime and plunging during late fall/early winter

Twitter:

- Drew most recent tweets in 15 second, 60 seconds, and 5 minute intervals- found that the resulting ratios of negative versus positive were wildly inconsistent
- The five minute data pull was most consistent but showed close to 50-50 positive vs negative sentiment

Group Roles

With a lot of different components to this project, there was an easy division of tasks between the group in order to accomplish our end goal. We each took one aspect of this project to be our own The different tasks that we took on were Reddit data, Twitter data, the baseline Facebook data, the Relationship Survey, and overall writing and analysis.

Seth was in charge of the graphs and analytics for the Relationship Survey. With this, he had to create additional columns for the csv and write statements in Excel to extract more data that what the survey provided. Some of these included the length of a relationship, the sentiment score, and simplifying dates into just months. From here this data was put through RStudio, and, using ggplot2, we were able to visualize this data using bar charts. The survey offered some complications with visualization at times so Seth had to figure out what information was the most important and what was possible to visualize given our information.

Anika was primarily responsible for the writing process of this final submission. This involved gathering the data analysis from group members and talking to each of the others in depth about their process and what was found. Her goal was to create a story of our process through a more personal writing style. She thought this was the right way of presenting our final project because we didn't end up doing exactly as we set out to do. Of course, everyone else also was involved with the final writing. Anika also worked on pulling together our interpretation of the Facebook analysis and survey.

Zeke undertook collecting historical Reddit data, analyzing it, and visualizing it using data techniques learned this year in Bostonography. All data was obtained using Google's BigQuery public data collection of 2018 Reddit comments. Zeke then used R to program functions that could effectively parse the data and create useable data frames for visualization. Zeke also transcribed his process in this document and provided possible conclusions from the data.

Nick took charge of working with the Twitter data. This included using RTweet and RStudio to acquire, analyze and visualize Twitter data. A huge chunk of time spent on the project was wrangling with R and the Twitter and Google APIs in an attempt to get data that we could work with. He also wrote up the information about the shortcomings of the data, and just generally working with Twitter and what we concluded about it.

Tommy worked mainly on the analysis and comparison of the survey data and its relationship with Facebook's data. This included analyzing each data set separately and verifying its validity and relevance to the project, as well as coming up with some explanations that came when looking at the data. He also aided in the writing process and formatting of the final submission.