Armchair Analysts

CS 5010 Final Report

Produced by: Aditi Rajagopal, Bradley Katcher, and Charlie Putnam
Computing IDs: ar5vt, bk5pu, cmp2cz

Note: Sources are cited as hyperlinks through the document.

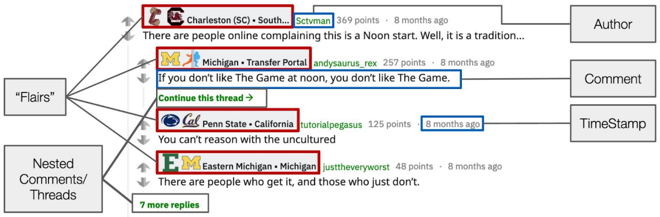
Introduction

We live in an age where commentary is abundant. A known home for commentary and discussion for any topic under the sun is reddit.com. Users (also known as Redditors) create discussion forums dedicated to certain topics called "subreddits." For the scope of this project, we would like to examine sentiment across a key popular college sports events via their respective subreddit posts, and visualize some of the trends we find. Specifically, we look at the subreddits for College Football (/r/CFB), focusing on the rivalry game of the 2018 matchup of Ohio State vs. Michigan football game stemming from personal interest, and because users can self-identify either program/fanbase affiliation via the Reddit "flair," or user-specified metadata (image, text).

Methodology:

We wrote a program using Reddit's "praw" API to scrape top-level comments from any given post, based on the input of any post's URL. In creating this web-scraper, our driving motivations were usability and interoperability. On the client-side, we wanted the user to be able to scrape and store data from any Reddit post with ease. With this in mind, we pulled data that would be useful to any reddit user. On the development side, we wanted to minimize code redundancy (i.e creating a flexible RedditPostParse class). By using the pandas python library, and exporting the parsed data into a csv, we give the user an open-ended, but an easy way to chart and analyze their favorite subreddit post.

For each comment, we obtained information regarding the comment's author, the user-declared "flair," the text of the comment, the timestamp (date and time) that the comment was posted, and the net "score" of the post (up-votes minus down-votes). Based on the collected text, we then run the data through a sentiment analysis package in order to gain a sense of the sentiment of the comments and gain insights from them.


Charleston (SC) • South... Sctvman 369 points · 8 months ago
There are people online complaining this is a Noon start. Well, it is a tradition...
Michigan • Transfer Portal andysaurus_rex 257 points · 8 months ago
If you don't like The Game at noon, you don't like The Game.
Continue this thread →
Penn State • California tutorialpegasus 125 points · 8 months ago
You can't reason with the uncultured
Eastern Michigan • Michigan justtheveryworst 48 points · 8 months ago
There are people who get it, and those who just don't.
7 more replies

Throughout this analysis, we decided to focus much of our effort on analyzing a “rivalry” game. We do this for a few reasons: 1) rivalries tend to be the most “charged,” and thus we’d expect to see strong (positive and negative) comments from fans on both sides, 2) these games are some of the most viewed, therefore, we expect them to have the most amount of comments, increasing our effective sample size. When commenting on posts, fans are able to declare themselves to be part of a given fan base by declaring a “flair,” i.e. giving themselves a special marker that is indicative of their team. We utilize this self-identification in order to separate users. While this could potentially lead to some bias (i.e. some fans may not declare a “flair”), we think in rivalry games, fans would be more inclined to “represent” their side by declaring a flair.

In order to obtain a sentiment score, we utilized the VADER (Valence Aware Dictionary and sEntiment Reader) package, which is a lexicon and rule-based sentiment analysis tool, designed for social media. This analysis tool has been trained on social media posts and has the capability to distinguish between a variety of inputs in order to obtain an accurate depiction of the sentiment of a given input. For example, VADER can process negation, punctuation, word-shape (e.g. ALL CAPS), degree modifiers, slang, emoticons, emojis, initialisms, acronyms the presence of links and other multimedia, etc. VADER provides a positive, negative, neutral, and compound score (a weighted average of the three prior scores) based on the text given. The majority of our analysis will focus on the compound score, where 1 is entirely positive and -1 is entirely negative.

In order to apply the VADER package, we first cleaned the comment text that was scraped from the post, in order to achieve the most accurate results possible. Based on our first run-through of the sentiment analysis scores, we determined that the presence of hyperlinks, brackets, parenthesis, and question marks all contributed to seemingly incorrect scores (e.g. a clearly negative post giving a neutral or positive sentiment score). In order to maximize the validity of the score, we removed these characteristics. This seemed to drastically improve the prediction accuracy. An example of this can be seen below:

Table One- Sentiment Scores Prior to Cleaning:

Post Text	Negative Score	Neutral Score	Positive Score	Composite Score
[fuck ohio](https://i.imgur.com/hyIMZmw.jpg)	0	1	0	0

Table Two- Sentiment Scores with Cleaned Post:

Post Text	Negative Score	Neutral Score	Positive Score	Composite Score
fuck ohio	0.788	0.222	0	-0.5423

A clearly inflammatory comment went from being considered as completely neutral, to significantly negative with our data cleaning procedure. This left us with a dataset that contained 897 top-level comments, with information regarding the poster, time of post, post content and sentiment, number of upvotes, and poster flair.

The final cleaning step before our analysis was to clean up the flairs. Reddit users can tag themselves in many different flairs in order to identify their fan base. We separated these flairs in order to determine what fanbase individuals self-identified with. If they indicated only Michigan or Michigan and any other flair than Ohio State, they were identified as Michigan fans. If they indicated only Ohio State or Ohio State and any other flair than Michigan, they were identified as Ohio State fans. Redditors who had flairs that were neither Ohio State nor Michigan or fans who lacked flairs were considered "neither." Redditors who posted both flairs (Ohio State and Michigan) were considered fans of whatever flair they posted first.

In an effort to go beyond the original specification of the project we: used the reddit API to create our dataset (rather than downloading the .json data from the post URL), created a mutable class so that you can pull and analyze comments from any Reddit post, and explored open-source tools to conduct semantic analysis via VADER sentiment, and visualizations via WordCloud

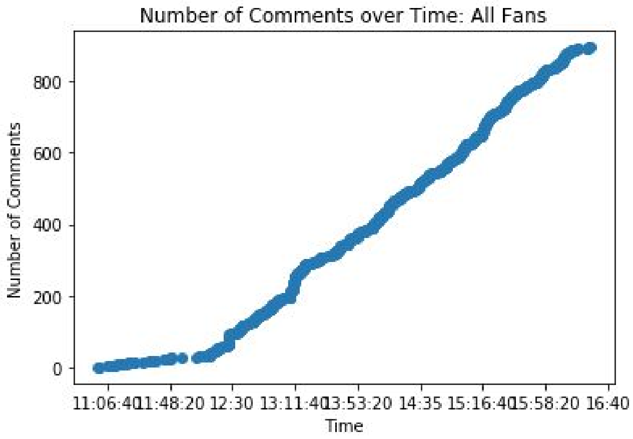
Results:

The Ohio State University vs. University of Michigan: Game Thread

Section One: Exploring Differences in Comment Rates

From this game thread, using flairs, we were able to identify 236 comments pertaining to University of Michigan Fans, 236 comments pertaining to Ohio State University Fans, and the remaining 425 identified neither, so-called "unaffiliated fans." Upon initial inspection, we noticed keyword similarities in flairs between University of Michigan and Michigan State University fans - so we made an effort to make sure we were not miscategorizing fans if the school names are too similar. The following chart shows the distribution of comments over time[1]:

Figure 1: Comments Over Time

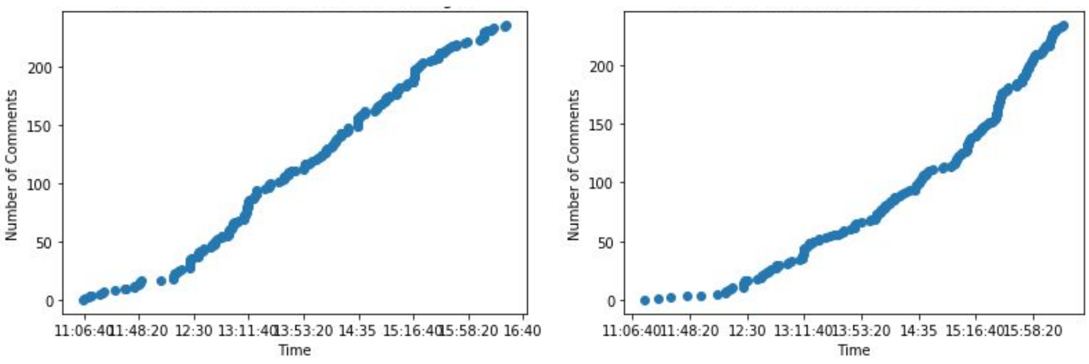


Two immediate takeaways from the graphs are that comments don't really "pick up" until the start of the game (scheduled at noon), and that around 1:15 pm (13:15), there seems to be a bit of a spike in comments (a sharp vertical increase in the graph, compared to a relatively monotonic slope). We hypothesize that this

[1] Note: For this graph, the three "latest" observations were excluded, as they were outliers in time and would cause the graph to be distorted.

occurs at the end of the first half/start of half-time, but are surprised that we don't see this same spike at the end of the game.

Figures 2 & 3: Comments Over Time by School Michigan (left), Ohio State (right)

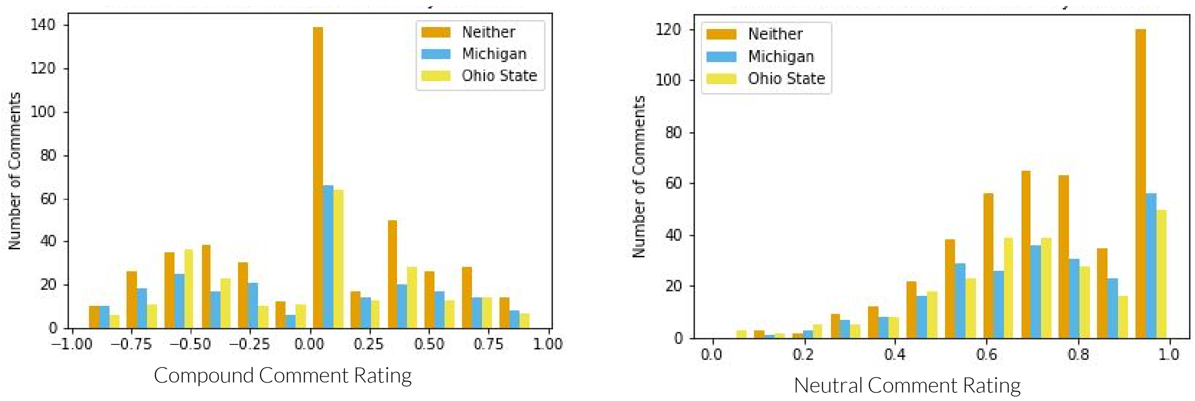


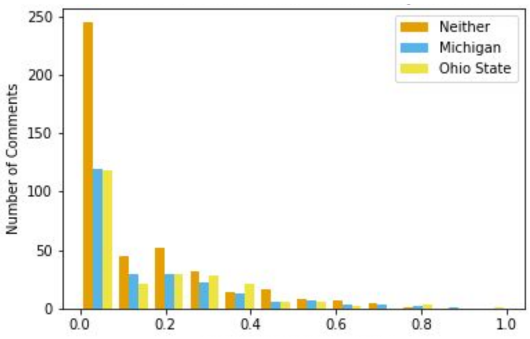
Based on the cumulative number of comments from Michigan and Ohio State fans above, we can see that in the case of this game thread, Michigan fans are overall more active commenters throughout the game.

Section Two: Distribution of Sentiment by Number of Comments

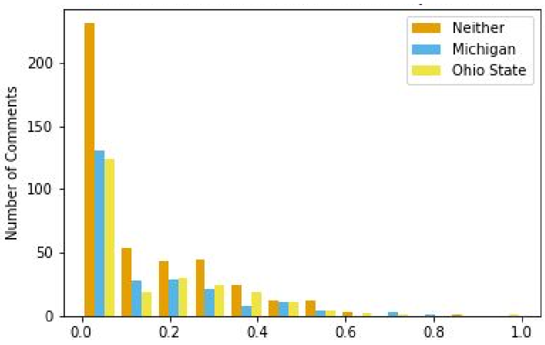
Below, we explore the distribution of comment sentiment by fan base for the compound sentiment score of comments. It appears that Michigan and Ohio State seem to have a similar distribution of comment sentiments. This can also be verified by exploring the component parts of the compound score (the portion of the compound score that is positive, negative, and neutral) with the graphs below:

Figures 4-7: VADER Sentiment Scores





Negative Comment Rating

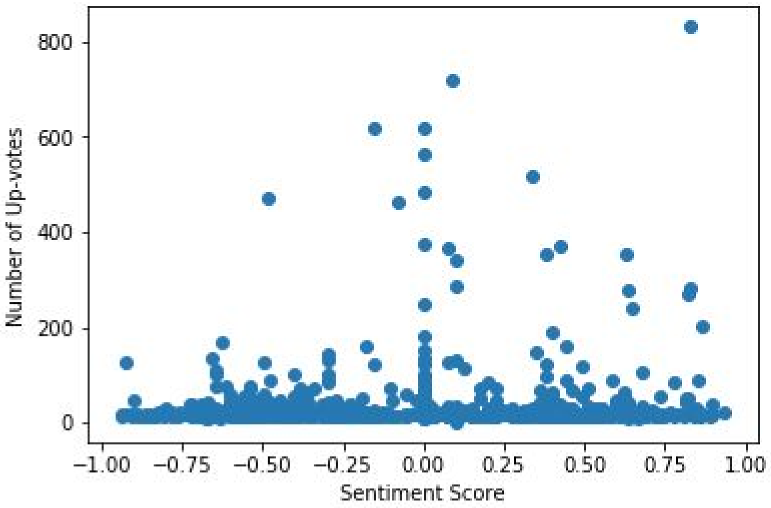


Positive Comment Rating

Section Three: Exploring the Relationship between Sentiment and Upvotes

We also sought to explore the correlation between the sentiment of the comments and the number of upvotes that a post received, combined among all commenters, as well as separated by fanbases:

Figure 8: Upvotes by Compound Sentiment Score -- All Redditors



Figures 9 & 10: Upvotes by Compound Sentiment Score by School Michigan (left), Ohio State (right)

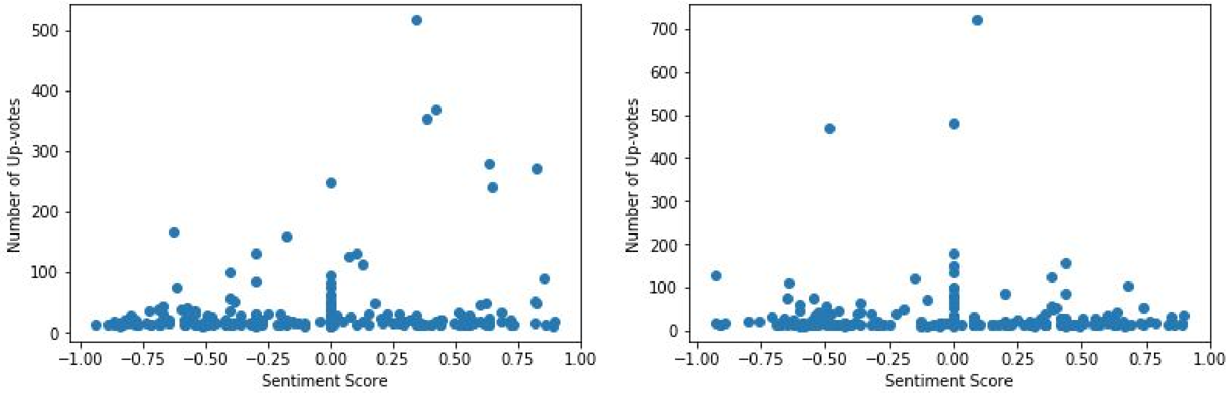


Table Three: Average Votes and Average Sentiment Ratings by Fanbase

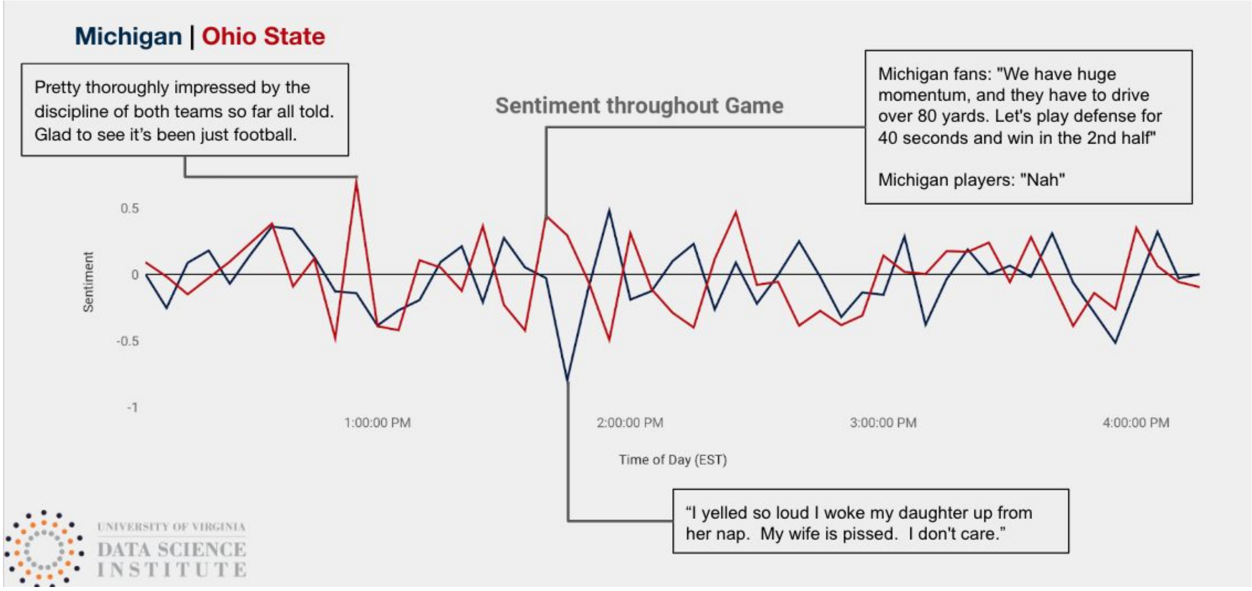
	Votes	Negative	Neutral	Positive	Compound
Michigan	34.88	0.148	0.72	0.126	-0.041
Neither	37.64	0.122	0.753	0.125	0.003
Ohio State	34.79	0.153	0.699	0.143	-0.031

From the above graphs, it appears that there is very little correlation between the compound sentiment score and number of upvotes received by a comment. This is verified by a computation of correlation coefficients between the number of votes and compound score by the fanbase. The strongest correlation is for Michigan fans, with a correlation of 0.156, followed by 0.081 for Redditors who are not part of either fan base, and -0.017 for Ohio State fans. In summary, there doesn't seem to be any significant correlation between average up-votes and comment sentiment. The average compound score seems to be almost exactly the same across fan bases, as well as through the components.

Section Four: Exploring Comment Sentiment throughout the Game

Given that we have time as a component of our data, we seek to explore the relationship of comments over time, in order to do that, we utilize the timestamp data collected from the scrape. First, we convert that data from UTC to EST for easier comprehension, since the game was in EST. Next, in order to capture a relatively smooth plot of sentiment over time (eliminate some of the noise associated with commenting), we resampled on time in order to create five-minute bins and computed an average sentiment score from there. This also allowed us to eliminate most of the "lag" that may occur between comments and game time, i.e. most comments are reactionary to what is happening during the game and we want to allow time for Redditors to comment and react to what was happening during the game.

Using this approach, we were able to identify instances where "tensions ran high" or where we believe there was significant over or underperformance by a given team, based on the absolute difference between sentiment scores of the two fanbases. Below is a depiction of some comments that occurred when the average sentiment was at a relative extremum for a given fan base:



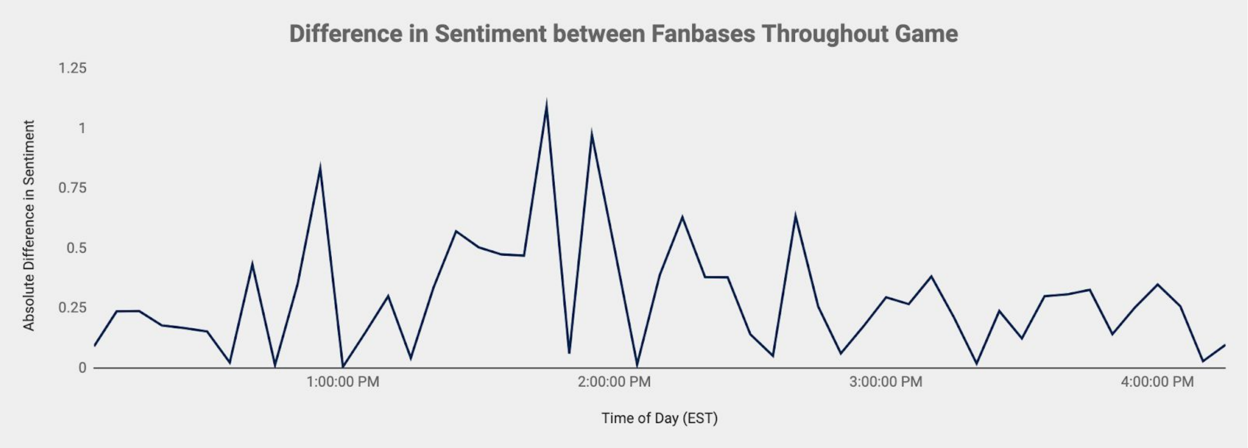
Pretty thoroughly impressed by the discipline of both teams so far all told. Glad to see it's been just football.

Michigan fans: "We have huge momentum, and they have to drive over 80 yards. Let's play defense for 40 seconds and win in the 2nd half"

Michigan players: "Nah"

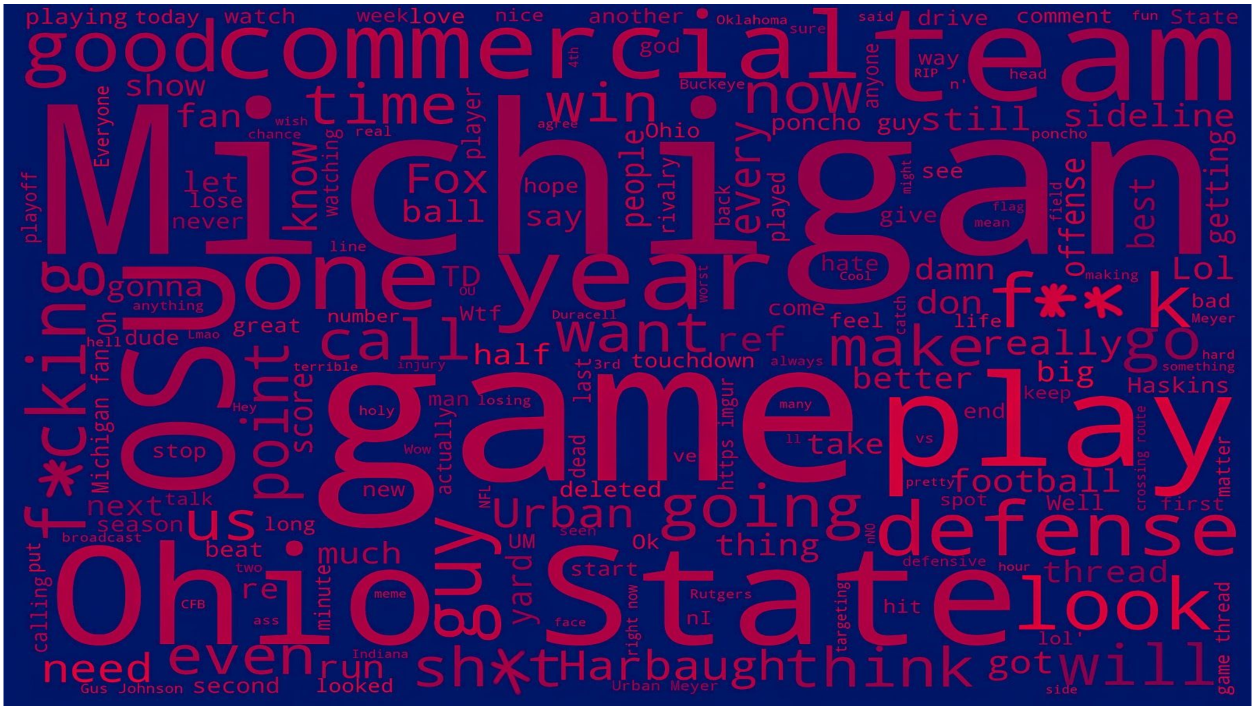
"I yelled so loud I woke my daughter up from her nap. My wife is pissed. I don't care."

Through this chart, and the one below, we are able to clearly identify spikes throughout the game and explore at what points throughout the game that comment sentiments tend to trend together (towards the beginning and end), as well as when there are significant deviations (mostly towards the middle of the game, likely during the "heat of the moment") when most of the action was occurring. The following graph shows the absolute value of the difference between OSU and Michigan Fan's sentiment over time:



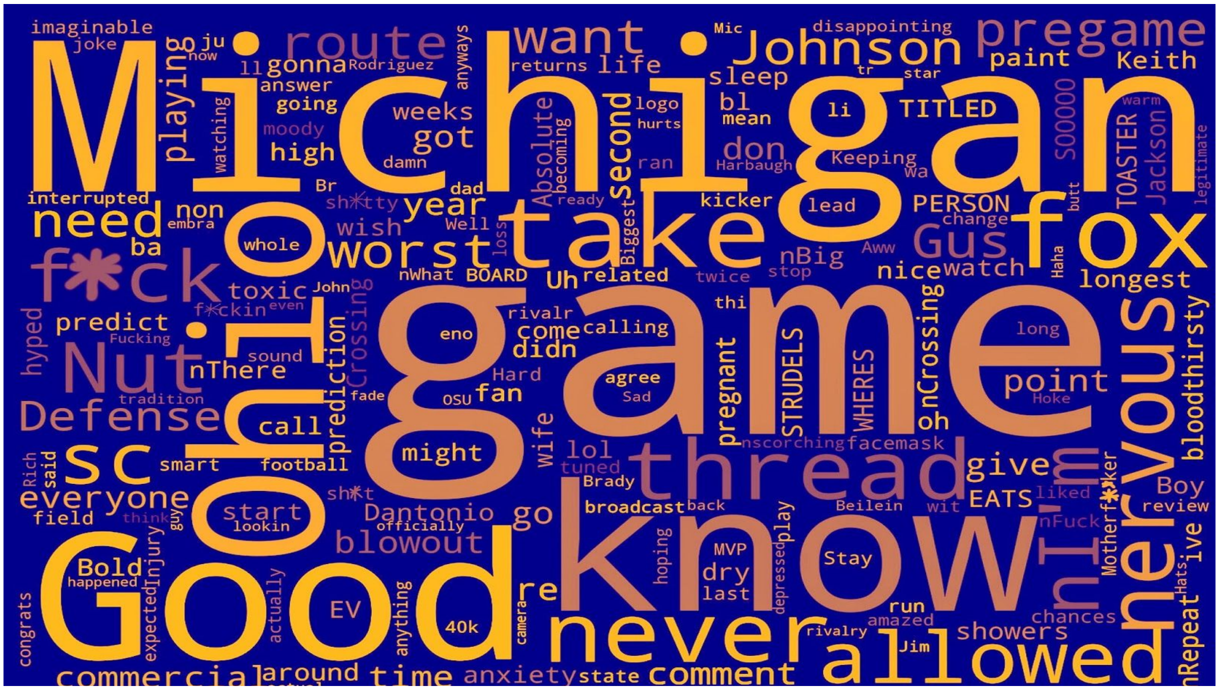
Section Five: Exploring Frequent Usage of Words:

In addition to understanding just the sentiment of what people were saying in their comments, we sought to further get a glimpse of the actual content that was being talked about. In order to do so, we constructed word clouds of the most frequently utilized words in comments. To further explore the differences between the fanbases, we looked at the top words used by each fan base as well. The following chart shows the most frequently used words by all Redditors. The relative size indicates the usage frequency of that word:

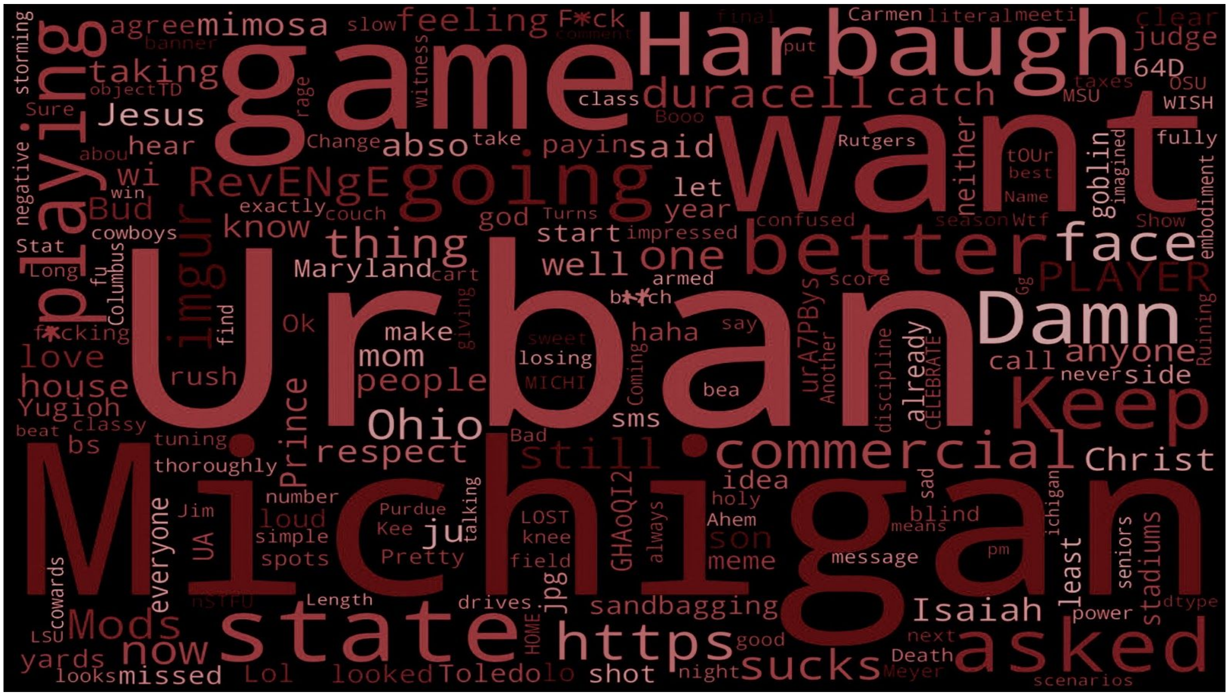


As you would expect, there were a lot of comments regarding the two schools, the use of the word “game” and “team,” as well as quite a few expletives, and commenting on commercials. The most utilized words by Michigan fans seem to be relatively similar as the general ones, although they have increased usage of the words “Good,” “know,” “never,” “allowed,” “nervous,” and “fox,” as can be seen from their word cloud below:

Michigan:



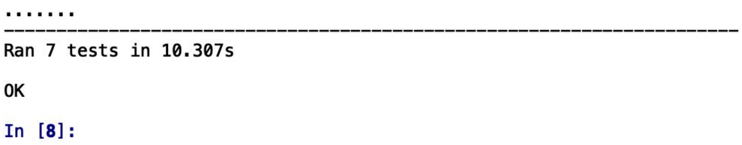
The Ohio State University:



Ohio State Redditors seem to discuss the coaches more often (both Harbaugh and Urban are used quite frequently), and it looks like they posted links more frequently with the presence of “https,” as well as many common words (such as want, playing, asked, etc.

Testing:

In terms of testing, we took a test driven development approach. As we built aspects of our project, we kept in mind testing scenarios, and proceeded to do incremental testing. We focused on ensuring that we were able to crawl the correct information in terms of object type, and we also wanted to ensure that we were storing the crawled and cleaned data correctly.



```
.......
----------------------------------------------------------------------
Ran 7 tests in 10.307s

OK

In [8]:
```

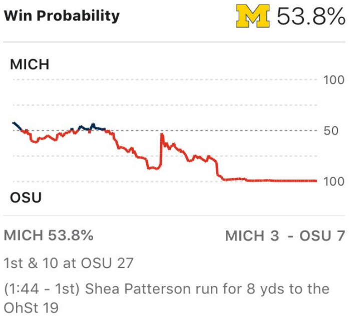
Applications and Reusability:

We took an agnostic approach when creating our project - we wanted to give the user the ability to scrape any reddit post, and gain usable data for visualizations and more. In looking at sports-related reddit posts, we believe that the data has a multitude of applications including, but not limited to: providing context and decision analysis for sports gambling, creating a “snapshot in time” for nostalgia purposes (reliving a live event etc), and providing sentiment feedback on broadcasting (reactions to commercials, songs being played during the broadcast, etc). In terms of any reddit post, we believe that our project may be helpful to those who are active participants in subreddit communities, those who are interested in joining a subreddit, and those who moderate subreddits. In terms of content moderation, our project would give moderators insight to flag keywords, identify abusive users, and identify active users. In terms of general subreddit users, using our project would enable them to analyze subreddit activity by crawling multiple posts, and also provide insight on content generation (what posts gain the most traction/activity).

Areas for Improvement

For the scope of this project, we made a few sacrifices up front in order to deliver our product in a timely manner. If we were given more time and resources to expand on this project we would have expanded our dataset to incorporate nested comments. If we had more computing power through UVA’s HPC cluster, Rivanna, we would have been able to use PRAW to pull all 21.8k+ comments from The Ohio State University vs. University of Michigan game thread. We would have also liked to have provided a few more case studies to further show the applicability of our work to other college sports games including the NCAA Basketball Championship game featuring University of Virginia vs Texas Tech University, and even pro-sports games including the SuperBowl.

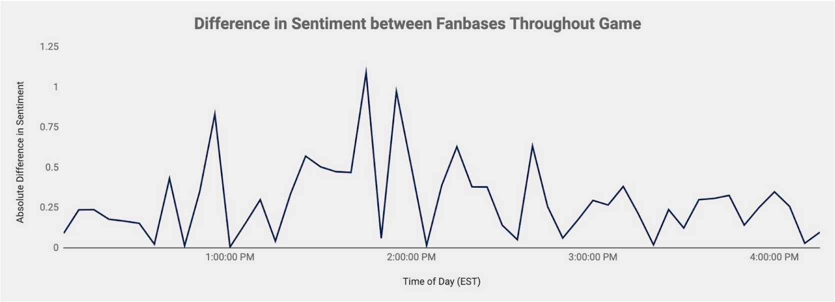
For this project, we wanted to map the play-by-play events of the game to the sentiment scores to see whether the events mirrored the sentiment expressed by the corresponding teams. We actually created a web crawler to pull the “Win Probability,” and the “Play-by-Play” data from ESPN.com. As per the graph on the left below, the probability of OSU winning the game is highlighted in red, and the probability of Michigan winning is highlighted in dark blow. Early on, we noticed a limitation of this data - the play-by-play details are not stored as a time-zone timestamp (i.e. UTC/EST etc), but as duration left in the quarter. Taking into consideration timeouts and commercial breaks, the data as is would not have mapped 100%, and we would have had to find a live recording of the game to verify how the timestamps overlapped.

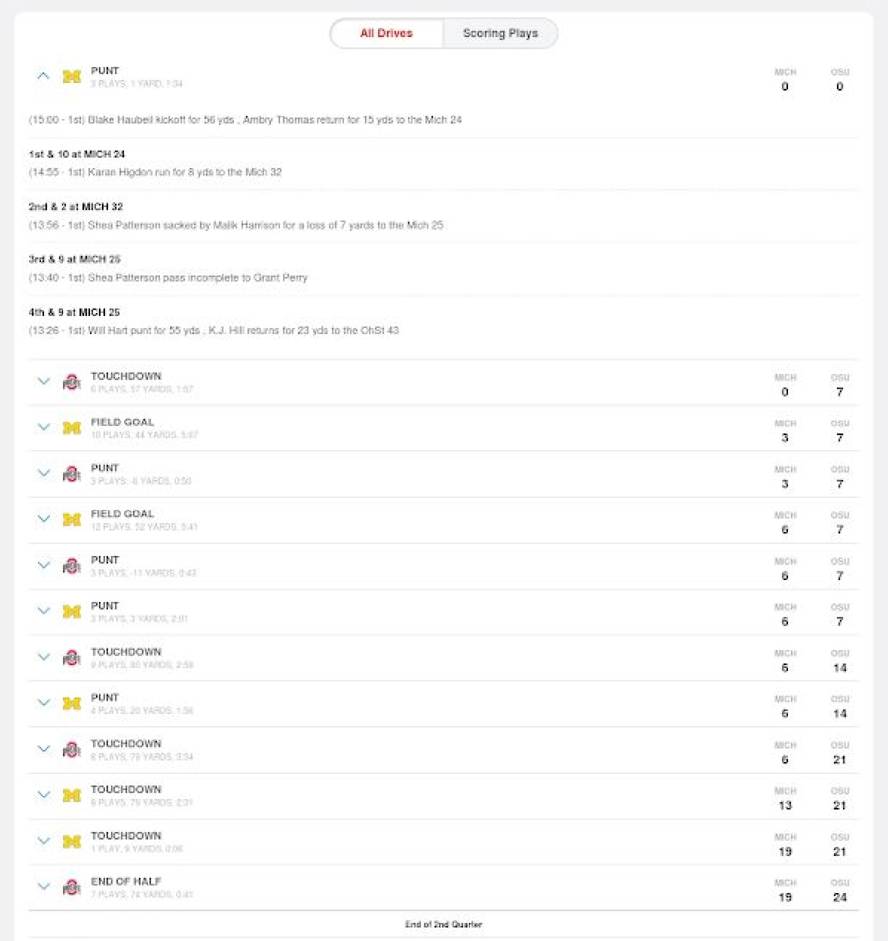


MICH 53.8% MICH 3 - OSU 7

1st & 10 at OSU 27

(1:44 - 1st) Shea Patterson run for 8 yds to the OhSt 19





PUNT
3 PLAYS, 1 YARD, 1:34
MICH 0 OSU 0

(15:00 - 1st) Blake Haubeil kickoff for 56 yds , Ambry Thomas return for 15 yds to the Mich 24

1st & 10 at MICH 24
(14:55 - 1st) Karan Higdon run for 8 yds to the Mich 32

2nd & 2 at MICH 32
(13:56 - 1st) Shea Patterson sacked by Malik Harrison for a loss of 7 yards to the Mich 25

3rd & 9 at MICH 25
(13:40 - 1st) Shea Patterson pass incomplete to Grant Perry

4th & 9 at MICH 25
(13:26 - 1st) Will Hart punt for 55 yds , K.J. Hill returns for 23 yds to the OhSt 43

	MICH	OSU
TOUCHDOWN 6 PLAYS, 57 YARDS, 1:07	0	7
FIELD GOAL 10 PLAYS, 44 YARDS, 5:07	3	7
PUNT 3 PLAYS, -8 YARDS, 0:50	3	7
FIELD GOAL 12 PLAYS, 52 YARDS, 5:41	6	7
PUNT 3 PLAYS, -11 YARDS, 0:43	6	7
PUNT 3 PLAYS, 3 YARDS, 2:01	6	7
TOUCHDOWN 9 PLAYS, 80 YARDS, 2:59	6	14
PUNT 4 PLAYS, 20 YARDS, 1:56	6	14
TOUCHDOWN 8 PLAYS, 79 YARDS, 3:54	6	21
TOUCHDOWN 8 PLAYS, 79 YARDS, 2:31	13	21
TOUCHDOWN 1 PLAY, 9 YARDS, 0:06	19	21
END OF HALF 7 PLAYS, 74 YARDS, 0:41	19	24

End of 2nd Quarter

Citations:

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.