April 30, 2017
CMDA 3654 T/TH

Project Part 3

Team Name and Members
Our team name is Twitch Scrapers, because our project revolves around scraping information from the website www.twitch.tv. The members of our group are **Jacob Smith**, **Jordan Deang**, and **Kevin Jiang**.

Topic
Twitch is a website that hosts “channels” of users streaming themselves playing video games. Each stream is accompanied by a built-in chat, where viewers can communicate with the broadcaster as well as other viewers. Some streams are watched by thousands of people at any given time, resulting in a massive amount of data being produced in the form of chat messages. Twitch provides an open source API for developers to use to make calls against in order to retrieve information from the website using languages like Python.

Final Key Findings
All of these findings are based around three Twitch VODs, one stream by Legend of Zelda player ClintStevens, one stream by League of Legends (the most popular game on twitch) player imaqtpie, and one stream of a popular League of Legends competitive match. We decided to diversify which VODs we analyze so that we can make generalizations about the chat as a whole while also finding specific details on each stream.

Active unique chatters:

Number of unique chatters in the Clint VOD: 1459
Number of unique chatters in the NA LCS VOD: 23574
Number of unique chatters in the imaqtpie VOD: 9752

Clint’s unique chatters visualization, shown in Figure 1, shows the least amount of activity. His numbers are still surprising however, compared to his viewers. While Clint averages 2,000 - 3,000 viewers and had nearly half of them actively chatting, Imaqtpie had over 25,000 viewers and under 10,000 chatters.

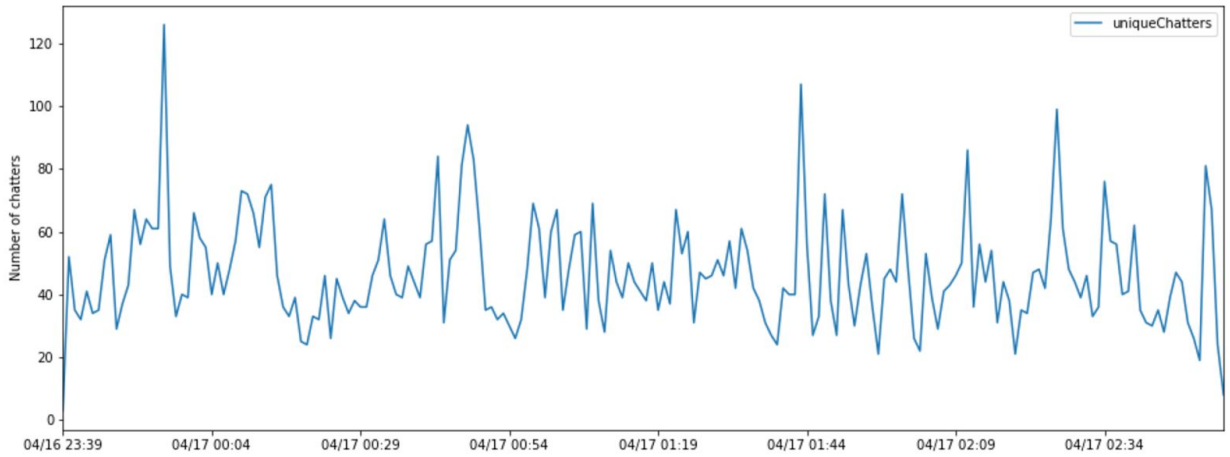

Figure 1: Active chatters for ClintSteven's stream

The number of active chatters in the NA LCS stream is shown below in Figure 2. The data starts to run flat at first but then begins to build up as the game progresses. We believe the small amount of active chatters during the first part of the stream may be caused by the game not having started at the time and with the commentators discussing various team roles and activities.

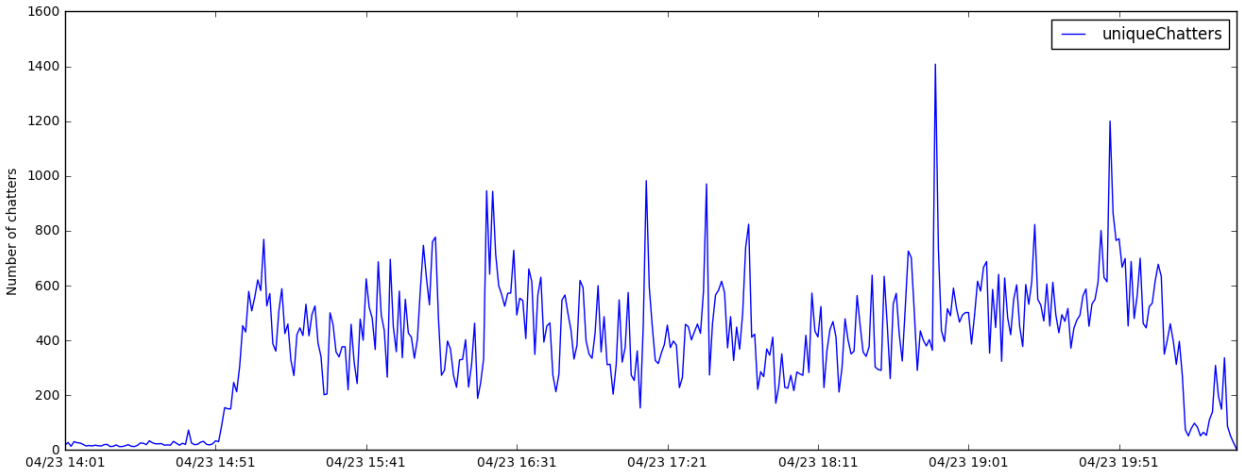

Figure 2: Active chatters for the NA LCS stream

Figure 3 shows the number of active unique chatters in imaqtpie's stream. The number of chatters begin to spike during various times probably because there was either an eventful or exciting play or some funny moments that happened to have occurred at those very moments.

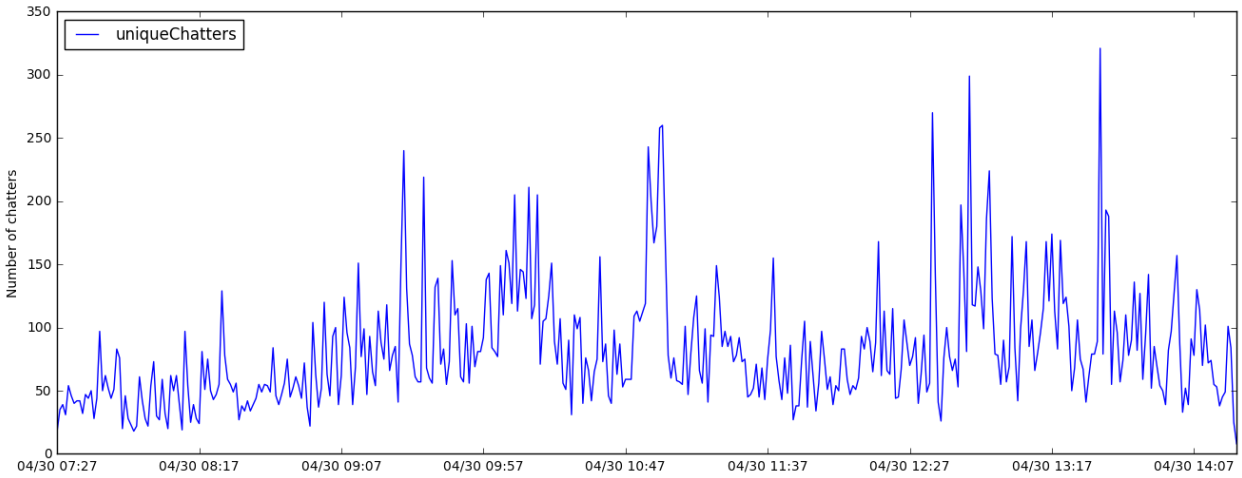

Figure 3: Active chatters for imaqtpie's stream

TD-IDF and # of unique chat messages:

Number of unique chat messages in the Clint VOD: 8491
Number of unique chat messages in the NA LCS VOD: 63072
Number of unique chat messages in imaqtpie's VOD: 23841

A broad generalization of the words used in the stream visualized as shown below for ClintSteven's stream:

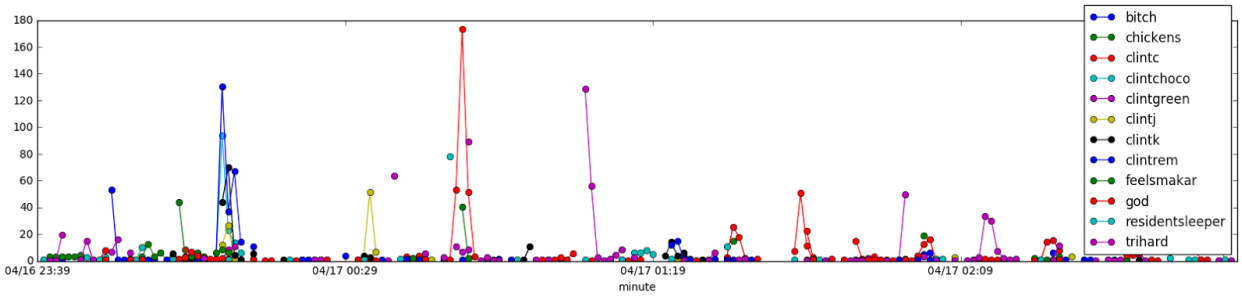

Figure 4: TF-IDF of ClintSteven's stream

Figure 4 depicts the importance and frequency of words that show up in ClintSteven's live stream. From the data that was analyzed, we can see from above that the most frequently used word is "clintc" which makes sense as Clint is the main host of the stream.

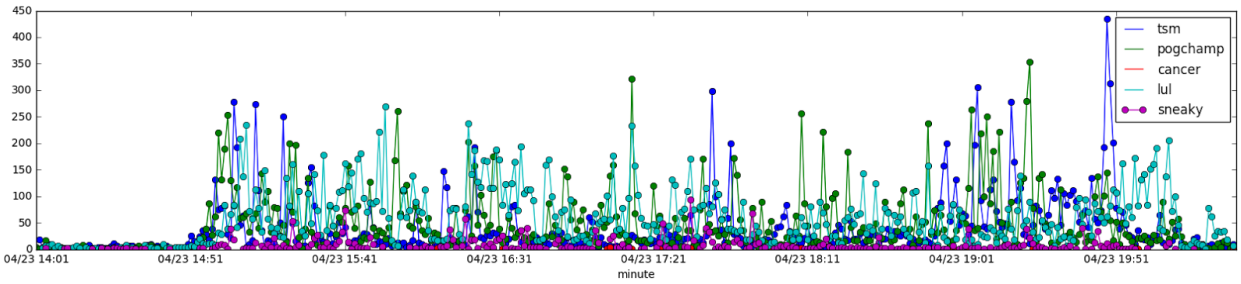

Figure 5: TF-IDF of a sample of words from the NA LCS stream

Because the stream for the NA LCS data was so large and widely popular on Twitch, the amount of data accrued from the stream is substantially more extensive than the other streams, as shown above in Figure 5. An important thing to note is that only a few words were selected to visualize the data of its significance and of course, the frequency of these words is very high as opposed to ClintSteven's stream. Figure 6, shown below, represents the the number of events triggered that produced boring times. Interestingly, there is a large spike at around 18:11 that caused a minor boring period to drastically decrease. This may be due to a notable play that was made during the stream that caused an uproar in the stream chat.

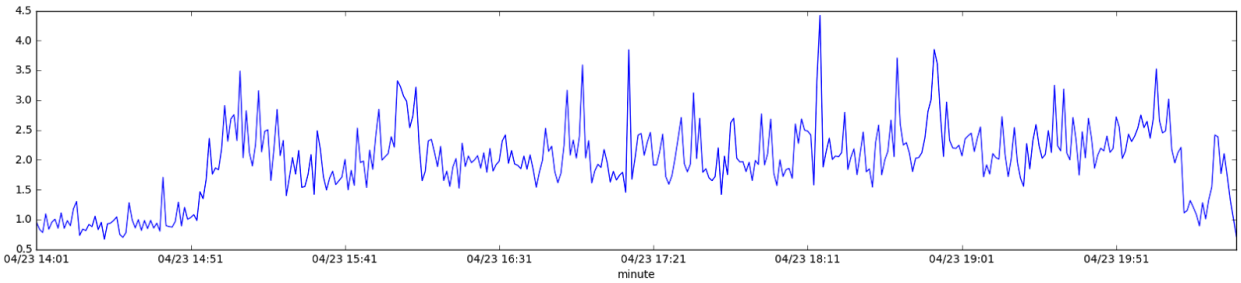

Figure 6: The boring periods generated throughout the NA LCS stream

Sentiment Analysis:

We conducted sentiment analysis on the processed messages of each minute of each VOD using the positive and negative lexicon used in the class. We also created our own lexicon of positive and negative emotes since emotes are a large part of Twitch chat. This showed that emotes were generally positive. This contradicted some of our results from the basic sentiment analysis.

Normal lexicon analysis: The visualizations and mean sentiment numbers showed that chat was mostly negative. Clint was overwhelmingly negative. Most negative sentiment came from use of the word "cancer," and much of the positive sentiment came from people saying they "love" Clint. Imaqtpie's stream was the most positive of the three while also having the highest positive spike of the three. However, an in-depth look at the messages showed that this was mostly because of one message being repeatedly copy and pasted that happened to have a positive

word. The NA LCS stream was also more negative, with many negative spikes happening when people began to talk badly about the North American teams. A time series of the sentiment for each VOD is shown in Figures 7, 8, and 9 below.

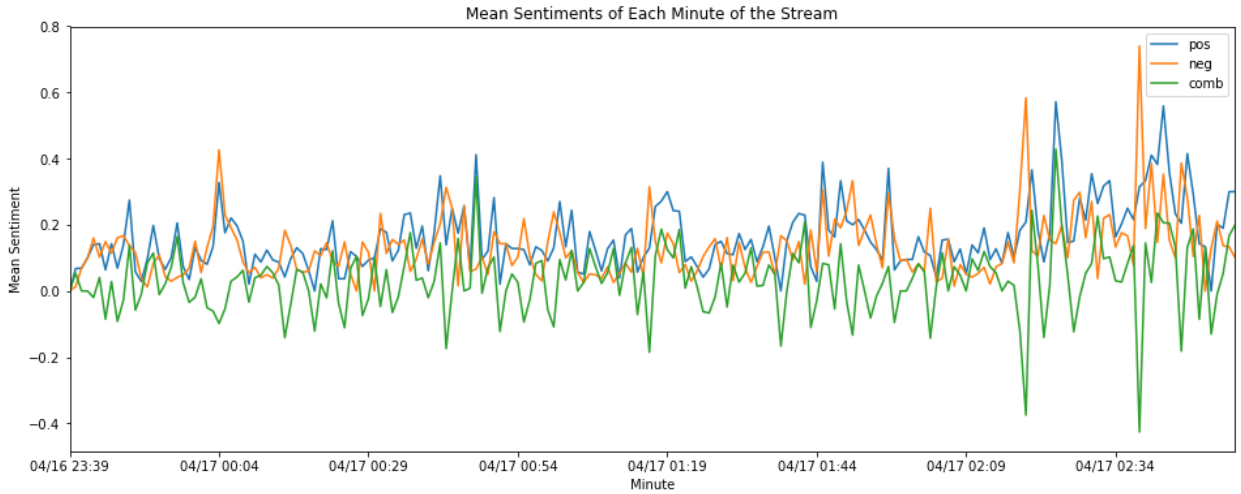

Figure 7: Time series of mean sentiment of each minute throughout ClintSteven's stream

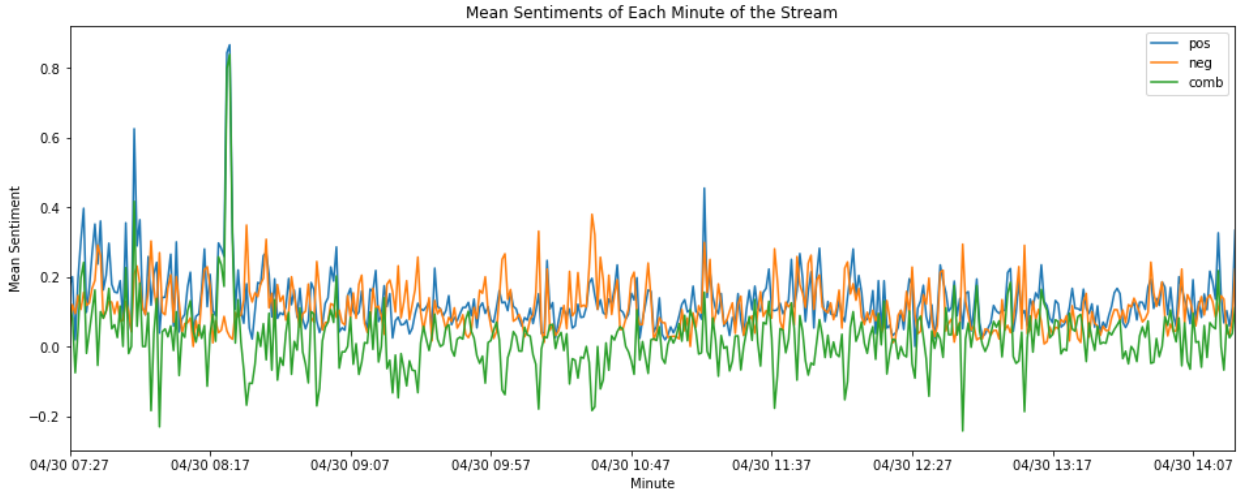

Figure 8: Time series of mean sentiment of each minute throughout imaqtpie's stream

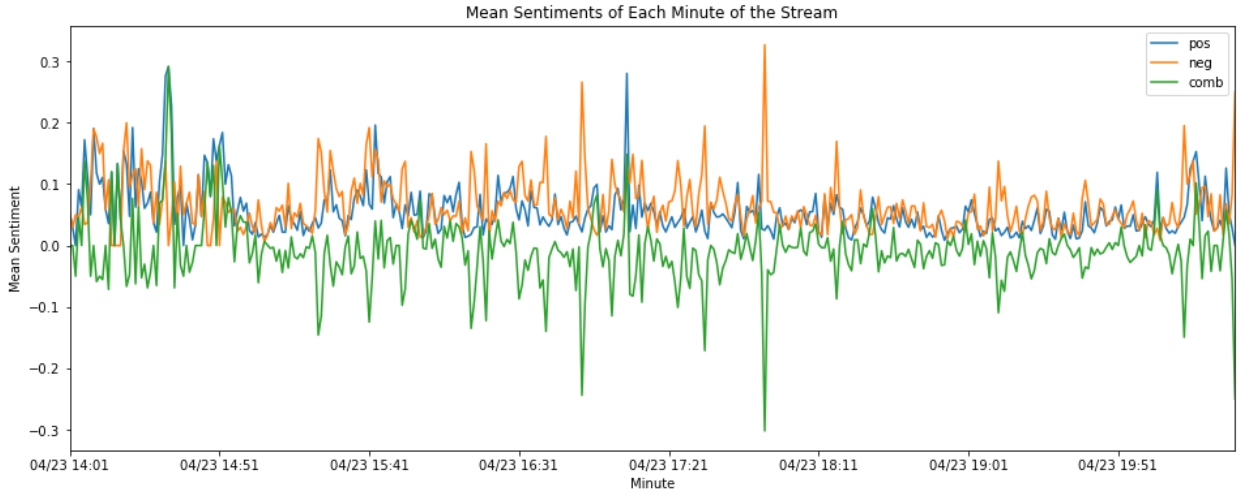

Figure 9: Time series of mean sentiment of each minute throughout the NA LCS stream

Emote sentiment analysis: The numbers were mostly positive, with “PogChamp” and “LUL” being the most positively used emotes and “ResidentSleeper” being the most negatively used emote. The NA LCS stream was overwhelmingly positive due to the repeated use of the “LUL” emote throughout the entire stream. Time series visualizations of the emote sentiment of imaqtpie and NA LCS are shown in Figures 10 and 11.

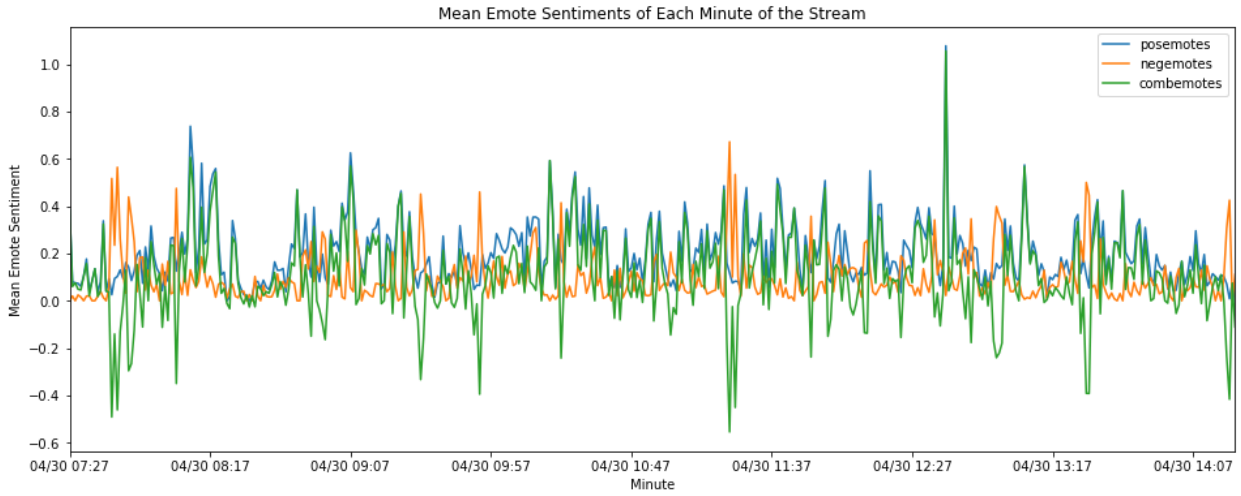

Figure 10: Time series of mean emote sentiment of each minute throughout imaqtpie’s stream

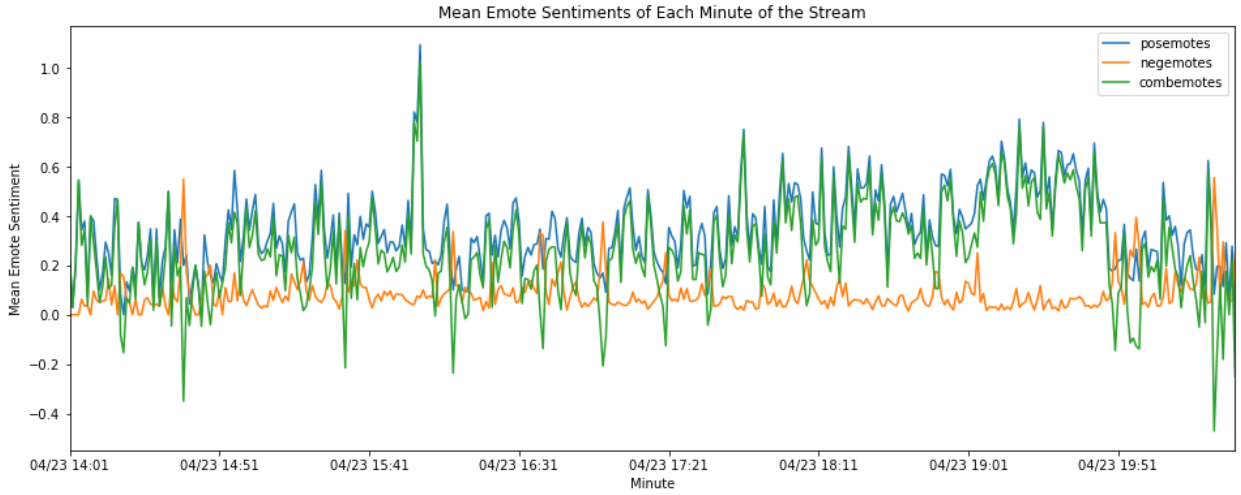

Figure 11: Time series of mean emote sentiment of each minute throughout the NA LCS stream

Computing the correlation between the basic sentiment analysis and our emote analysis shows very little correlation.

Pearson correlation for Clint: -0.016678942946537524
Pearson correlation for imaqtpie: -0.030186594007190898
Pearson correlation for NA LCS: 0.014037068941554232

This leads us to believe that the vocabulary used by Twitch chat is very complex and may require a combination of a normal lexicon and emotes to determine sentiment. However, the moments of high emote sentiment seemed more closely correlated to exciting moments in the VOD rather than the normal lexicon since exciting moments involved more use of positive emote “PogChamp.” A deeper lexical analysis of emotes by linguistics experts may be required before proper emote sentiment analysis can be conducted. Figure 12 below shows the time series difference between the combined normal lexicon analysis and the combined emote lexicon analysis of the NA LCS stream, the data likely most skewed by repeated use of emotes.

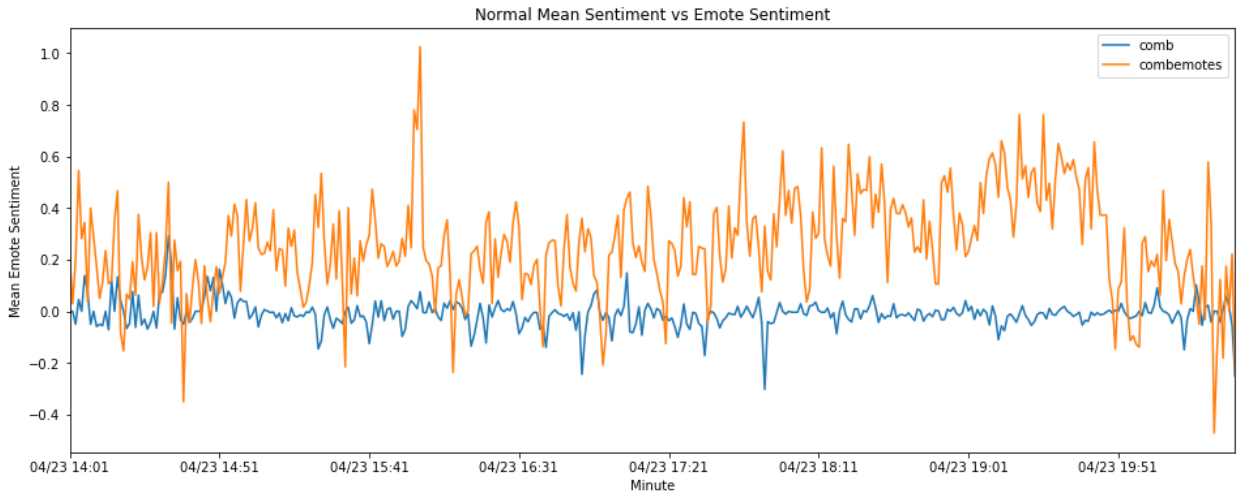

Figure 12: Time series of mean emote sentiment vs normal mean sentiment of each minute throughout the NA LCS stream

Subscriber Sentiment Analysis: Although this was only relevant to Clint and imaqtpie's streams, the subscriber analysis had higher peaks and more extreme, sporadic data. Looking closely at the peaks revealed that this was again because of repeatedly copy pasted messages.

So, after performing repeated sentiment analysis on these three VODs, we have concluded that much of the sentiment is skewed because of copy and pasted messages both within one message and amongst users. If one user prints a negative word 15 times and it is copy and pasted by 50 other users, a massive negative outlier is created. This leads us to believe that twitch chat may require a more fine processing tool that can distinguish copy and pasted messages with legitimate messages while performing sentiment analysis. Some kind of combination of our unique message analysis and sentiment analysis is required to get a good idea of the chat's sentiment.

Analysis Process

For our analysis of the chat data, we used many of the text analytics methods we learned in class such as TF-IDF and sentiment analysis. Every text analysis method first required processing the data by cleaning it and generating the bag of words of each message. This allowed us to analyze the individual words. We also had to determine a way to bin these messages for analysis. We decided to bin the messages into every individual minute the 60 seconds is typically large enough to capture a moment that may happen in the stream and the subsequent chat reaction.

To determine the number of unique chatters in a particular minute, we created a list using the "from" column of the csv, which contains the username of the person sending each message. We then converted the list to a set, removing any duplicate names from the data set and found

the length of this set. We also created a column that stored the minute of each message, then grouped by this value and graphed the number of unique chatters in each minute of the VOD.

For the unique number of chat messages, we first read in the respective csv data into the program and then removed the irrelevant punctuations. Next, we removed any duplicate chat messages that were not unique to the data set by using the set data structure to perform this operation. And lastly, we outputted the number of unique messages which resulted in the amount displayed above.

All of our visualizations are time series graphs because the messages all come with a timestamp, and this lets us analyze connections between the text and the video itself.

For the sentiment analysis, we first used only the lexicon of positive and negative words provided to us in the class. However, this proved inadequate because it did not capture the wide vocabulary of emotes used in Twitch chat. So, we created our own emote lexicon of positive and negative emotes. Although imperfect, it did capture the moments in the VOD where many users would react in a short period of time with just one short emote such as “PogChamp,” an emote expressing excitement, or “LUL,” an emote expressing laughter. We also computed the Pearson correlation between the normal lexicon and our emote lexicon to see how effective our emote lexicon may be.

Changes from Part 2

Our only real change from Part 2 involved rerunning our analysis on a variety of Twitch VODs and comparing the results. This allowed us to refine our findings and categorize them into things that are specific to a Twitch channel and things that are general to Twitch stream chats.

In addition, we decided to include a separate sentiment analysis on the chat of subscribers only. Subscribers have access to special channel-unique emotes that regular chatters do not have, so we expect the data of the two parties to be slightly more positive.

Contributions

- Jacob Smith: Ran the active-chatters files that analyzes chat user activity on each of the VODs. Assisted with the report write-up.

- Jordan Deang: Performed sentiment analysis on all three VOD chat data sets. Created a emote lexicon to analyze a correlation between emotes and sentiment. Compared sentiment of subscribers vs. regular viewers. Wrote up the changes from Part 2.

- Kevin Jiang: Implemented the unique-messages file that analyzes the number of unique chat messages in the three recorded streams; created the TF-IDF matrix that represents the relationship between the words used for each messages with the corresponding dates, thereby determining its importance; analyzed miscellaneous activities such as boring periods; wrote up changes from Part 2