Connor Elfers

Twitter Squid Games Triangulation Analysis

March 30th, 2022

**Frequent Terms**

**Chart, funnel chart

Description automatically generated**

**Process:**

Here are the 15 most frequently mentioned terms from the given twitter dataset. To achieve this result the Comment’s from the twitter data were tokenized, cleaned, counted, and reordered so that the top mentioned terms appeared first. I used the standard stop\_words vector provided by tidytext so that unimportant words were discarded from the analysis. I also personally removed a few custom stop words as the default stop\_words vector is generic and did not included terms specific to my data. These terms include http, twd, t.co, de, 13, 16, and im. I removed these as they would not provide any significant analytical results. After all unwanted terms were removed, I simply used the ggplot package to produce the graph shown.

**Analysis:**

The end result of this shows the top 15 most frequently mentioned terms. Within the top 5, we can assume that people love this show from the words mentioned such as “love” “fan” and “life”. I have personally seen the show can attest to holding the same feelings after viewing. So, I am fairly confident that most people are big fans of the show. However, to play devils advocate here, I could argue that the terms by themselves don’t mean anything because there is no context. The only context I hold is my own personal feelings. So, the term “fan” could be referring to literal fans that blow air. To get a deeper insight I then took this tokenized data and applied a sentiment analysis to it.

**Sentiment:**

Chart, bar chart, funnel chart

Description automatically generated

**Process:**

To produce this graph, I took the already tokenized data I produced from the first graph and by applying the default “nrc” sentiment data, from the tidytext package to it I was able to then filter only between terms that were considered positive and negative. After determining all negative and positive terms, each of the categories were counted and arranged to show the top 10 most frequent within each one. I once again used the ggplot package to produce the graph shown.

**Analysis:**

As with the previous graph the issues are still present regarding the lack of context. However, using sentiment analysis it gives us a slightly deeper look into the lack of context. Shown we can see two separate graphs, on one side terms that are considered negative are displayed on the other terms considered positive. From a visual perspective it may look like there are more negative terms than positive, however, when we take a deeper look we can see that actually there seems to be more positive terms. Looking at the x-axis (frequency) the lowest positive terms seems to be equal in frequency to the highest negative terms. So, we can assume that the show has an overall positive sentiment about it. At the same time however, just because there are more positive words than negative words does not mean the overall sentiment of Squid Games is positive, the same goes for negative terms. Although I mentioned that words by themselves contain no inherent context, through the context of the show, the negative words that are displayed align with what the show itself is about. That is, about killing and dying. So it can be assumed that the negative terms displayed are not actual negative sentiment towards the show. To note however, this is also a condensed group of terms that were categorized from a very broad field of sentiments. Example, the word “producer” was within the positive category. Now I for one do not consider the word “producer” a positive or negative term, neutral if anything. Through this graph and the previous we can safely assume that the show Squid Games is a rather popular show and has an overall positive sentiment towards it.

**How Retweets, Likes, and Engagement Effect One Another**

**Bubble Graph**

**Chart, scatter chart

Description automatically generated**

**Process:**

In order to achieve the results of this graph I needed to take Impressions, Twitter\_Activity, and Engagement and plot them against themselves. My first step was to remove the largest outlier from each of the three data types as in the end the skewed with the visuals of the graph and them being an outlier does not produce much relevant insight for the purpose of this graph. Afterwards I once again used the ggplot package to help me graph.

**Bar Graph**

**Chart, bar chart

Description automatically generated**

**Process:**

To make the presented graph I simply took the total sum of each of the three data types and plotted them against one another. I used ggplot packaged for the graph.

**Line Graph**

**Chart, histogram

Description automatically generated**

**Process:**

For the creation of this graph, I once again took the data from each of the three data types and plotted them against one another. I scaled the graph down to show relevant data in order to complete the analysis. I used ggplot package again for the creation of the graph.

**Analysis:**

* **Bubble Graph**

Here we get a great visualization of how each of the three data types effect one another. We can see that most of the data is concentrated on the left-hand side of the graph, under 500 Twitter\_Activity (Retweets). On that same side we see that engagement is also at its highest as well as impressions. This tells me that as the amount of retweets increase the level of engagement decreases along impressions appearing to decrease as well. We could come to the conclusion of that maybe the amount of retweets affects engagement and impressions in a negative way.

* **Bar Graph**

With the bar graph we can see just how much of a difference retweets has to impressions and engagement. Retweets is actually over 30 million thanks to about 129ish outliers, but in order to see both impressions and engagement I had to scale it down to 3 million. Having this much more retweets tells me that after a certain point in a tweet’s popularity lifecycle people seem to stop liking and engaging with thee tweet itself. Taking into context how twitter works, its very easy to retweet something as well as liking something, the difference is that when you like something it is saved to the ‘liked’ section on your twitter page. Perhaps people simply do not want to like everything as they save that area as a sort of storage and do not want it cluttered. Engagement is based off both retweets, likes, and comments thus having one side low and another side high with outlier will make the engagement closer to impressions than retweets.

* **Line Graph**

The line graph only confirms the observations made from the previous two graphs. Most engagement happens when retweets is under 1,000 and impressions under 1,000. This is totally in line with what would be considered a normal situation. If you say no and a thousand people say yes, your no won’t make much a difference. This applies to anything on the internet that is super popular, the expectation of your comment on a post of 1,000 comments is that it’ll probably not be noticed.

The conclusion I’ve made from these three graphs is that higher amounts of retweets a tweet has received the less likely it’ll be liked, thus making the engagement lower. Also, I’ve concluded that tweets with a modest amount of retweets and likes will have the highest engagement level as it still has room for people to talk with one another. I call this the silent majority as this is were most of the engagement happens, in areas, or in this case tweets, that are not mainstream or widely visible to everyone. Perhaps the tweets with a lot of retweets were from celebrities or controversial people, as they would be less likely to engage with the comments as they either A.) are a celebrity and do not want to bother to try to talk to everyone or B.) they tweeted something controversial and do not want to argue about it in the comments.

**Correlogram**

**Chart, treemap chart

Description automatically generated**

**Analysis:**

As per my other graphs you can see that 1.) there is extreme outliers with each category. This proves my idea of a 'silent majority' as each category plotted by itself you can see the area portrayed does not go very much out at all. Thus, the larger grouping has relatively small counts.