**Analysis of Tweets Containing #okboomer**

**Executive Summary**

The analysis discussed here is an exploration of recent tweets with the “okboomer” hashtag. According to the investigation, the dataset points to largely objective and neutral tweets with a clear connection to age and generational topics.

**Introduction**

Of all of the hashtags being used today, one of the most interesting to me is #okboomer. While this response is likely one of the least aggressive possible, it is full of passive aggressive dismissal. This millennial-driven come-back to being called “snowflake” is prohibitive to meaningful discussion about modern barriers to success, but it is also super entertaining. My goal in choosing this dataset is to see what attributes tweets containing this hashtag might carry.

**Data**

The sample of tweets I pulled from the Twitter API contains 2,780 tweets that all include the hashtag “okboomer.” These tweets were created between 5:05 am on 12/3/19 and 12:50 am on 12/8/19. Table 1 shows that the largest number (43% of total tweets) were located in the USA, the second largest group (41% of total tweets) was created without a specific location, and the rest (16% of total tweets) were listed as in another country. The attached appendix contains a table of the 50 most common tokens from the gathered tweets. Of interest are the words old, generation, age, and millennials. These tweets combined have a lexical diversity of 3.39.

Table : Tweet Count by Location

|  |  |
| --- | --- |
| **Location** | **Count of Tweets** |
| N/A | 1146 |
| Other | 448 |
| USA | 1186 |
| **Grand Total** | **2780** |

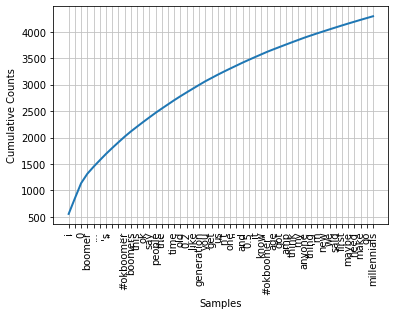


Figure : Token frequency cumulative counts

**Methods**

To collect these tweets, I used the Twitter API and tweepy. My code found a sample of tweets with #okboomer and cleaned up the stopwords, emoticons, and punctuation. In my function to write the texts to a csv file, I passed the filtered tweet through to the Textblob Sentiment function to determine the polarity and subjectivity of each tweet. With these two measures, I can get an idea of how this hashtag is being used on Twitter.

**Results**

There are a heck of a lot of tweets with #okboomer streaming into the internet all day, every day. In a five-day timespan, there were over 2,700 tweets. Unsurprisingly, this dataset suggests that these tweets are often associated in some way with age and generational issues. The sentiment analysis results (Figure 2) show that the more objective the tweet (closer to 0), the less polar it is (closer to 0). The more subjective the tweet (closer to 1), the wider the range of polarity.

Figure : Sentiment analysis - each datapoint represents a tweet. Polarity varies from negative (-1) to positive (+1). Subjectivity ranges from objective (0) to subjective (1).

Figures 3 and 4 below show the distribution of tweets as they fall into subjectivity and polarity ranges. Overall, the majority of tweets appear to be objective and not polarized.

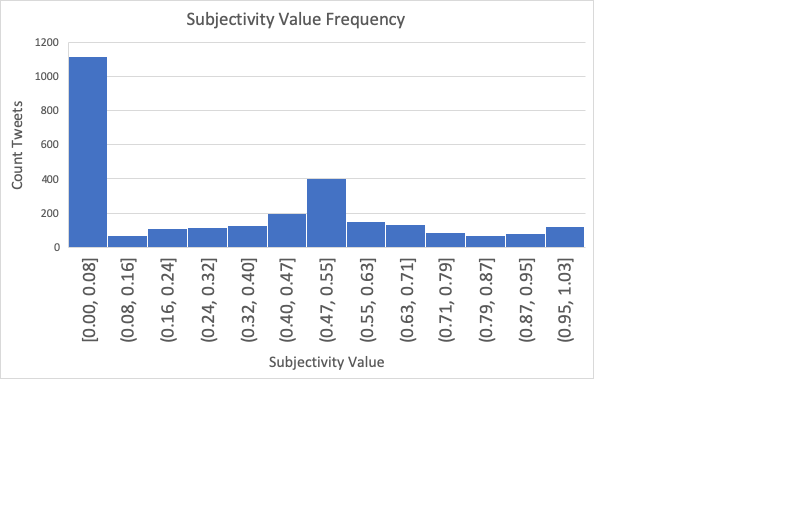


Figure : Frequency of tweet subjectivity – these tweets tend to be more frequently more objective (closer to zero).

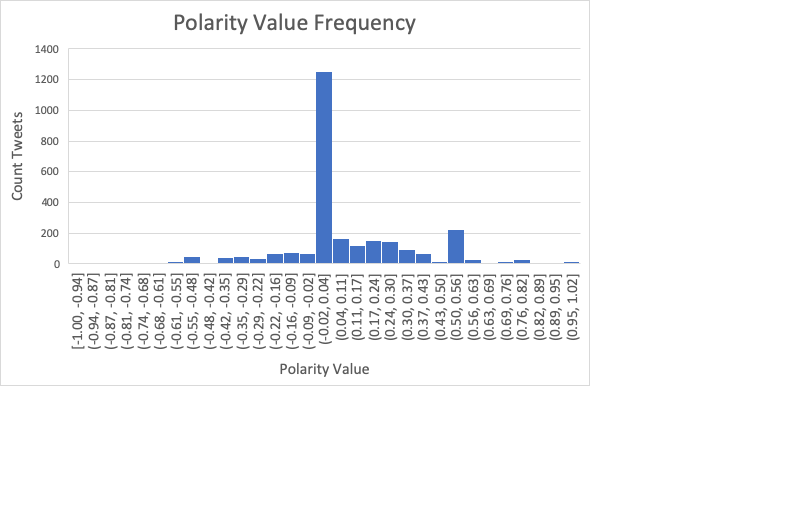


Figure : Frequency of tweet polarity - these tweets tend to be more frequently not polarized (closer to zero).

**Conclusion**

Congruent with the abundant memes I’ve seen on social media, this dataset suggests that the “okboomer” hashtag is very popular. It is likely that the trend is more specific to the US, but it is difficult to be sure since a significant portion of the tweets do not have a labeled location. Our sentiment analysis shows us that most of these tweets tend to be not opinionated and are overall neither positive nor negative in tone. However, there are two issues that could make this analysis misleading. First, I removed all stopwords, which means that negating words, like “not,” are not being considered. Second, the way this hashtag is used is most often when employing sarcasm, or after a quote from an older individual (e.g. “snowflake”, or “When a boomer says, ya just gotta save more money...”). Thus, the sentiment of the author may not be detectable by this current method. Another alternative is that a lot of the tweets in the data were actually re-tweets. Frequently, users will re-tweet someone else’s tweet without changing much, or by adding very few words. Personally, if I share something on social media that someone else posted, I rarely add my own comments other than something like “check out this interesting article.” The low lexical diversity can also be contributed by how brief many tweets and re-tweets are.

If I were to continue with this project, I would like to refine the cleaning segment. There were a lot of things that seemed to slip through the clean-up I did for stopwords and punctuation. Some of this is likely due to the high number of errors and odd things that people type into their tweets. I would also like to figure out a way to remove tokens that aren’t true words, without sacrificing any meaningful abbreviations or texting acronyms (cyberslang). It might also be interesting to compare original tweets to re-tweets to see how often people are agreeing with the #okboomer tweets, or if they are disagreeing. Overall, I think that I barely scratched the surface on this dataset, and there is a lot more to learn.

Appendix

|  |  |
| --- | --- |
| **Item** | **Count** |
| i | 556 |
| ' | 294 |
| 0 | 286 |
| boomer | 178 |
| . . . | 132 |
| " | 123 |
| 's | 120 |
| " | 112 |
| … | 108 |
| #okboomer | 106 |
| boomers | 98 |
| this | 91 |
| ok | 90 |
| say | 87 |
| people | 86 |
| the | 81 |
| `` | 79 |
| time | 79 |
| old | 75 |
| 0.2 | 72 |
| like | 71 |
| generation | 70 |
| you | 69 |
| get | 62 |
| us | 62 |
| n't | 59 |
| one | 57 |
| '' | 56 |
| and | 56 |
| 0.5 | 53 |
| item | 52 |
| know | 51 |
| #okboomer" | 51 |
| age | 46 |
| got | 45 |
| amp | 45 |
| think | 44 |
| my | 44 |
| anyone | 43 |
| thing | 30 |
| 'm | 39 |
| new | 38 |
| we | 38 |
| said | 37 |
| first | 37 |
| maybe | 36 |
| need | 35 |
| make | 35 |
| go | 34 |
| millenials | 34 |