Narrative

Dianne Kim

12/11/2020

Brief Substantive Background / Goal

The main question that inspired my project was whether or not K-pop is a viable soft power strategy for South Korea. "Soft power" is a term coined by Joseph Nye in 1990 to describe the strategy of using the powers of attraction through culture and ideology rather than coercion or "hard power." The South Korean economy has grown significantly since the country's independence in 1945, but it is still commonly believed that the country lacks the hard power necessary to become a global superpower like the United States and China. The popularity of K-pop has grown exponentially over even just the past few years, and K-pop artists are starting to break into the US music scene more with every passing day. I am curious as to whether the *Hallyu*, or Korean Wave, can be a strategic investment for South Korea to increase its soft power.

Admittedly, investigating this is a daunting task as soft power is tough to measure. In this project, I decided to use the resources I had available to conduct sentiment analysis and investigate word frequencies among both K-pop fans and non-K-pop fans. I carried out this project with the hopes that it would yield some interesting results, and provide a good approach for bringing the research further.

Collecting Data

I collected data from a variety of sources. First, I wanted to see if K-pop oriented news websites, which K-pop fans frequent to consume news not only about their favorite South Korean celebrities but also about what is happening in South Korea in general, reported differently than more general news websites. I chose to scrape data from AllKpop, which I knew to be one of, if not the most, popular K-pop-oriented news website. I filtered their news article feed to see only the "best" articles written by AllKpop staff instead of regular users. I assumed that the "best" category meant that these articles would have the most engagement from its readers. I wrote a function to first scrape all of the URLs of the articles that showed up on this page. Then, I wrote a function to scrape the title, author, body, number of comments, and upvote percentage of each article.

At this point, I ran into an issue where I did not know how to scrape information from a website with infinite scroll instead of numbered pages. After some research, I learned that the RSelenium package could help me simulate scrolling down the webpage. The toughest part, surprisingly, of using this package was actually getting started–I spent probably the majority of my time at this portion looking up what to install onto my computer so that the package could open a remote session. Thanks to some extremely intelligent and generous people on the Internet, I was able to figure it out and run the package. Once I got the remote session started and opened the website, all I needed to do was use a for loop to scroll down the page a certain number of times. I aimed for 1000 articles, but the website actually stopped loading articles after the 912nd. This was close enough, so I continued onward.

After AllKpop, I used an API that I found called News API. Like its name suggests, it is a very helpful API that provides lots of information about a huge variety of different news sources. I put together a get request to pull information about articles that mentioned K-pop. Due to my limited amount of foresight, I was only able to get information from 100 articles, as their rate limit is pretty strict. If I were able to re-do

this project I would definitely plan well in advance so I can use this API regularly for a week or so to collect more data.

Finally, I was curious to see the responses to these articles, or even just South Korea in general on Twitter. I used the Twitter API to pull the most recent 5000 tweets that mentioned @allkpop, AllKpop's twitter handle, and the same number of tweets that mentioned South Korea. The former would be my dataset for how K-pop fans react to topics related to South Korea, while the latter for how a more general group of people react.

Cleaning/Pre-processing Data

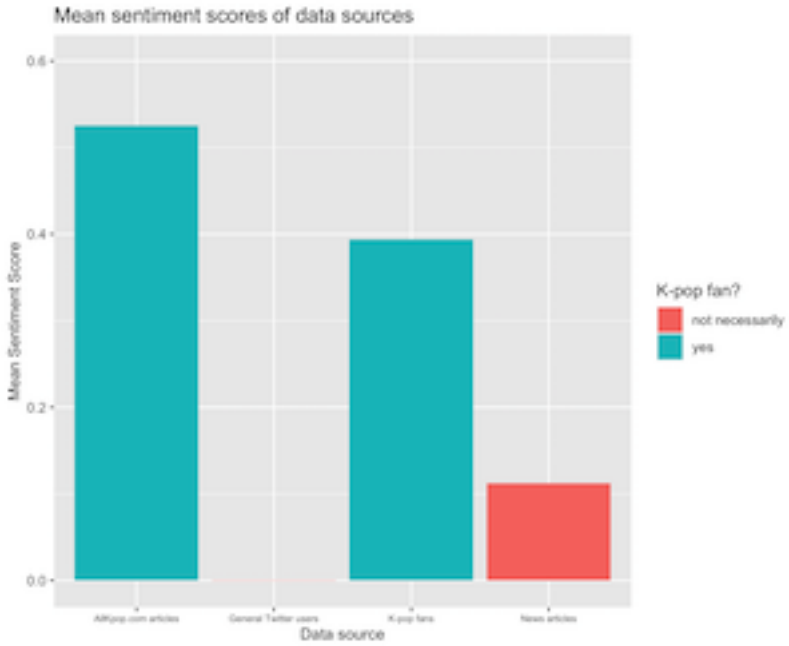
I had a lot of data, but thankfully it was not too complicated to pre-process my data for text analysis. I created a corpus of the content in the articles along with the content of the tweets to prepare for analysis. I removed stop words, converted all the letters to lowercase, removed numbers, and removed punctuation. I withheld from stemming as it would be difficult for sentiment analysis later on. For content that I collected with specific search terms, I removed those search terms from the corpora so that they would not show up for frequency analysis. If I collected my data properly, I assumed those terms would be one of the most frequent ones anyway.

Analysis and visualization

Once I had my corpora ready for analysis, I first calculated the frequencies of each term. I sorted the rows by decreasing frequency to easily see which terms were repeated the most. Then, I created a word cloud of a maximum of the 50 most frequent terms.

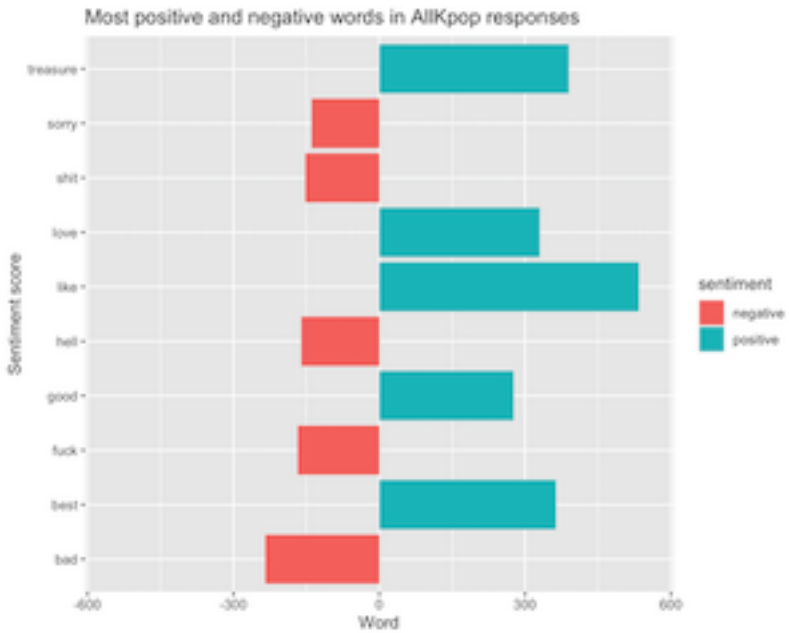
Next, it was time for sentiment analysis. I chose to use the AFINN sentiment lexicon because each word in the lexicon is rated with an integer between -5 to 5 depending on how negative or positive the word is. I thought having a more specific rating for positive and negative words would be useful for my sentiment analysis. I joined the AFINN sentiment lexicon with my frequency tables to add a column indicating each term's respective sentiment score. What surprised me at this step especially with the K-pop-fan-specific content was that so many words had an NA value for sentiment score. These terms were internet slang or K-pop-specific terms (jargon?) that would not be included in a standard English lexicon. Though it was a bummer, I replaced the NAs with 0s and moved on.

In order to account for frequency, I added another column multiplying the frequency of the term by its AFINN score to get its overall sentiment score. Then, I grabbed the top 5 and bottom 5 rows of each table to get the 5 most positive and most negative words. I visualized this data with a bar graph. Out of curiosity, I also calculated the mean sentiment scores for each dataframe and plotted them. It seems like the AllKpop articles have the highest sentiment score while general Twitter users had a sentiment score almost close to 0, leaning only slightly to the negative. There were no heavily negative sentiment averages.

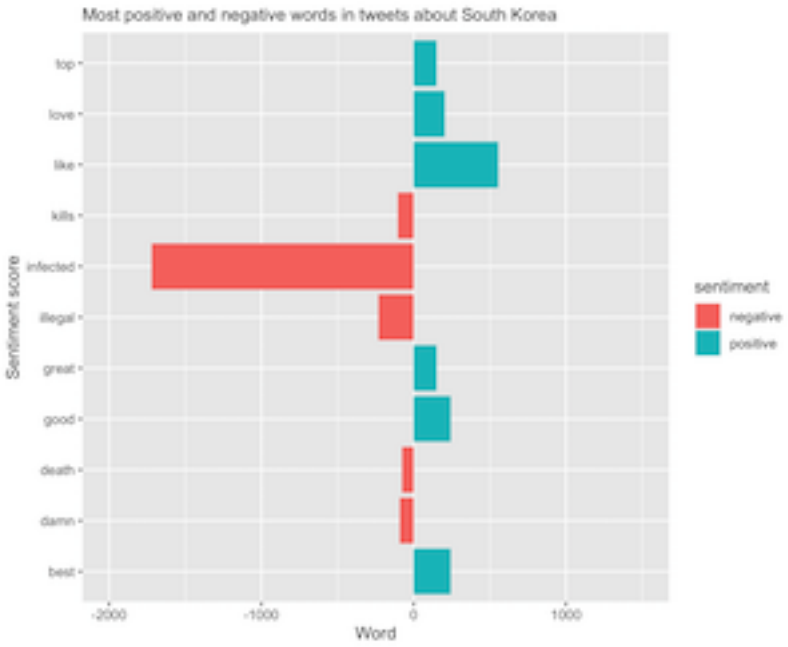


Conclusions/Future work

While my research is nowhere near sufficient to make claims about soft power as a whole, I do think that my data still shows that being a K-pop fan contributes to more positive sentiments regarding South Korean news overall. While both AllKpop articles and general news articles had overall positive sentiments, AllKpop was more positive. Assuming that the AllKpop staff and writers are K-pop fans themselves, this observation would support the idea that K-pop may work as a soft power strategy. However, judging from the top 5 negative (and vulgar!) words from K-pop fans on Twitter, one may argue that being invested in K-pop may actual lead to more polarized reactions–just as strongly negative as positive. Perhaps the result would be different, though, if Internet lingo were more accounted for.



Tweets about South Korea yielded a more uncertain result due to the fact that terms related to the COVID-19 crisis dominated the top negative words, such as “infected” and “death.” Collecting data from tweets published before 2020 may show more interesting information.



What stuck out to me most among the word clouds was that there were way more proper nouns in the word cloud of K-pop fan responses. There were more artist names, and even individual member names from these groups. The word “BTS” was more prominent in the general news word cloud than in AllKpop’s word cloud, insinuating more coverage of a super-popular artist such as BTS from general news websites while AllKpop covers a variety of different artists. The ‘expertise level’ of these different datasets is not necessarily surprising, but still interesting.

Future work could expand the project over a longer period of time and larger amounts of data. It would also

be interesting to create sentiment lexicons that contain Internet and fandom lingo especially to be used for social media (ex. lmao, go off king, skskskskk, and whatever else the kids say these days). These lexicons could provide a more accurate sentiment analysis of tweets or Facebook posts. Given more time and resources, I would also be interested in investigating the sentiment of Japanese K-pop fans and non-K-pop fans towards South Korea, as the two countries have a historically strained relationship.

Perhaps an investment into K-pop as a soft power strategy may be a high-risk, high-reward scenario in which those who are more invested in South Korean culture react not only more positively to news about South Korea, but also more negatively than those who are not. From my data, it seems that the general sentiment towards South Korea is neutral, while K-pop fans hold a more positive sentiment. I suppose is good news that none of my results shows a strikingly negative sentiment towards South Korea. The South Korean government is actively investing in the K-pop industry, and recently they created a *Hallyu* department in their Ministry of Culture, Sports, and Tourism. My research hints that this change may show some positive results and I am interested to see whether these investments pay off.