**BTS Sentiment Analysis**

Lisa Doan

Michigan State University  
735 E Shaw Lane  
East Lansing, MI 48825

doannhun@msu.edu

Crystal Lewis

Michigan State University  
603 Bogue Street, VanHoosen Hall  
East Lansing, MI 48825

lewiscry@msu.edu

Doyeon Kim

Michigan State University  
25 Beal St.  
East Lansing, MI 4882  
kimdoye2@msu.edu

**ABSTRACT**

1. The purpose of the project is to perform a sentiment analysis of the South Korean boy band BTS amongst Twitter users. With an increasing popularity worldwide and a rapidly rising Twitter follower count almost reaching 10 million followers, BTS is the perfect subject for analyzing which areas of the world this K-Pop group is generating positive, negative, and neutral responses. We want to find the changes of sentiment over time and how the impact of album releases, award shows, and any other events sway the opinions of Twitter users. One challenge that we face has to do with language barriers. BTS is a Korean based group with many fans worldwide so it is natural for many tweets to be written in a variety of foreign languages. This poses a challenge for sentiment analysis and to remedy this, we are filtering for tweets written in English.

**Keywords**

BTS, Twitter, json, Streaming, cluster, API, Python, tweet, retweeted

# **INTRODUCTION**

Twitter provides a unique platform for users to come together and discuss a variety of topics ranging from news, politics, comedy, pop culture, music, etc. BTS, a South Korean boy band, has seen an exponential increase in their number of followers and overall popularity from their interactions with fans on Twitter. In addition, they have released a new album in September and made significant advancements in America by appearing at the American Music Awards, releasing an English version of one of their songs, and performing on various talk shows such as the late night shows featuring James Corden and Jimmy Kimmel as well as the Ellen show. Because of these increased activities, it is clear that there will also be a spike in the number of tweets talking about BTS.

As BTS becomes a global boy band that is taking over the world, it is interesting to note the changes of people’s sentiment in the United States towards the group. It is undeniable that the group has garnered enough support from fans in the United States to have them invited to attend notable American shows but using the platform that they can best communicate with fans on and where trends and sentiments are best analyzed, Twitter is ideal for generating conclusions of opinions about BTS from Americans.

To complete our task, the group has completed exercises from the course, CSE 482 Big Data Analysis, which focused on retrieving and preprocessing tweets using Twitter’s Streaming API. This API allows us to collect tweets based that use the ‘BTS’ keyword over a certain time range. For the sentiment analysis portion of the project, we referenced a Python library that can determine the sentiment of a tweet by determining if a word has positive, negative, or neutral connotations.

The challenges of undertaking this project is how to filter the data and categorizing the overall sentiment of the tweets. Because Twitter is an inclusive platform that has users from all of the globe and BTS garners fans from numerous different countries, we had to address issues of retrieving tweets that were written in English. The sentiment analysis tool used has some weaknesses such as only working for words in English and its inability to process sarcasm. The majority of BTS’s fans are young millennials who exercise a variety of terms to express themselves which include a heavy use of sarcasm. Because of this, it is difficult to accurately gauge the true intent and sentiment of a tweet. Another issue is gathering tweets that aren’t retweeted (the same tweet but tweeted again by another person) and have geo-locations on. To ensure privacy, many people do not have their geo-locations on when they tweet, which made it difficult to gather accurate data.

Our project yielded results that were partially expected and unexpected. Because of the extremely filtered tweets, we were not able to get a definitively accurate sentiment reading of Americans who tweet about BTS. About 100,000 tweets were streamed but the relevant tweets equaled to numbers in the hundreds. However, the tweets we were able to analyze, yielded expected results: BTS garners a positive sentiment in the United States, with a particular emphasis in California.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*Conference’04*, Month 1–2, 2004, City, State, Country.

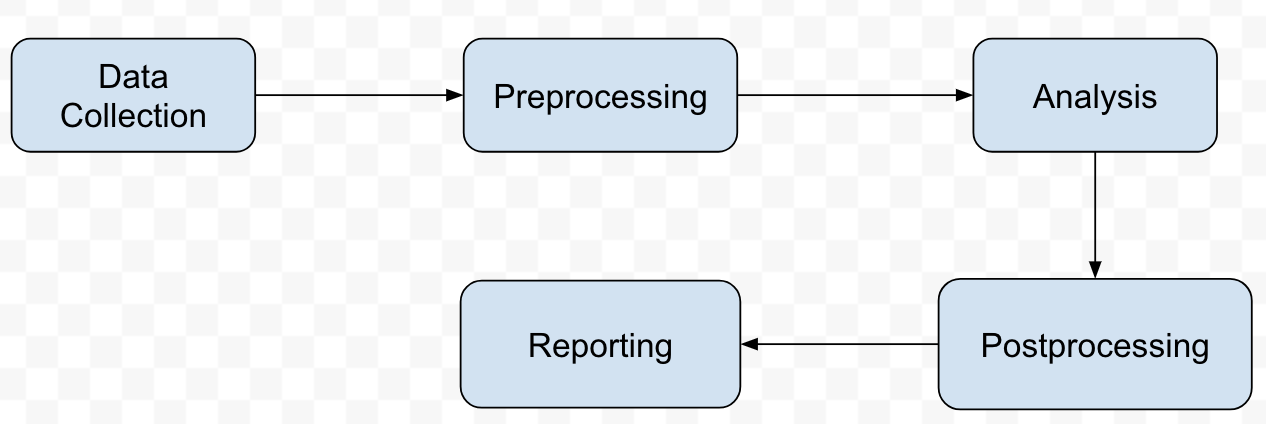
Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

# **PRELIMINARIES**

The problem domain is within the scope of Twitter’s web and mobile applications. The tweets are gathered using Twitter’s Streaming API which gathers all tweets written in English with the keyword, ‘BTS’. The streamed tweets are saved in the JavaScript Object Notation or JSON format. The analysis that we are performing on the tweets is determining the overall sentiment of a tweet by classifying and predicting if a word is positive, negative, or neutral. By summing up the positive, negative, and neutral terms in a tweet, we can classify what kind of sentiment the tweet expresses overall. The data that is generated is multiple time series for four days, with the first column representing the state the tweet comes from and the second column representing the net daily sentiment of the tweets from the corresponding state.

To further expand on the analysis being performed, the sentiment analysis method we used predicts whether a tweet is positive, negative, or neutral. By using the TextBlob library, we are able to provide it with the text of a tweet and it returns if the sentiment of the tweet. The prediction aspect of the sentiment analysis helps classify sarcasm if the text is repeated but does not completely and accurately handle it.

# **METHODOLOGY**

Figure 4.1 Schematic Diagram 

Data collection was the first module in the sentiment analysis process. This was conducted by using the Twitter API to stream tweets over the course a week. The stream was utilized for 30 minutes at a time to collect the resulting tweets into a json file. The tweets that were saved did have any specific parameters applied to them. This leads to the next step in the diagram which was preprocessing. Due to the massive amount of tweets, the data needed to be preprocessed so that only tweets that were relevant would be analyzed. For this project, the sentiment analysis would only use tweets that were only tweeted once and that geolocations. Hence, tweets that had the retweet attribute in the JSON set to null or false were excluded from the data set. Additional preprocessing steps were conducted during the streaming process too. Only English tweets were used for this analysis hence only English tweets were collected. This was also done by checking the language attribute in the JSON files, and if the language code was anything other than English it was excluded from the data set. The tweets were also filtered by US location, as well as State because the sentiment analysis was only relative to the 50 state of the US. Following completion of preprocessing, sentiment analysis was then conducted. Within the analysis portion, more processing of the data was required too. Newlines, spaces and other miscellaneous characters are extracted from the tweet. From here the text portion of a tweet is collected. The actual sentiment analysis was conducted using the TextBlob API sentiment. TextBlob is a python library used for natural language processing. It is based off another well known python library Natural Language Toolkit (NLTK). The sentiment method in this library can also calculate the polarity as well. To use this method, only the text of the tweet could be analyzed, hence why the preprocessing as mentioned previously was done. This leads to the post processing. In this step, the tweets are organized into categories based off on whether or not the sentiment is negative, positive or neutral. This set up the data for the last step, where reporting was conducted. Daily net sentiment was calculated, and a time series based off the state and daily net sentiment was created. This information is then presented in the format of a report. The Pandas python library was used to create the time series structure. All of this information was also presented in a jupyter notebook as well. One of the challenges with this project was the formatting of the JSON files. . For some of the files, there were extra newline characters. When the sentiment analysis code was ran on these files, it through an error because of this extra character. To ensure uniformity of these files, extra processing to unify the JSON file so that the rest of the code could work. Additionally, there were issues regarding the language of the tweets. Initially tweets were streamed without accounting for language. However upon researching TextBlob and realizing that the language for the method being used functioned in English, the the data needed to be streamed again. Due to the massive volume of tweets, it was easier to stream the data again as opposed to try and just filter it out.

# **EXPERIMENTAL EVALUATION**

This section describes the experimental setup and results that were obtained.

## **Experimental Setup**

The data set before preprocessing was four JSON files containing the streamed tweets using the keyword ‘BTS’. The Twitter Streaming API was used in 30 minute increments to produce the four JSON files containing the tweets. After preprocessing, the files were organized in folders of the date they were streamed and separated in JSON files of the state that it was tweeted from.

Tweets that were streamed prior to November 6th were removed because they were not filtered based on the language of the tweet. The tweets on November 6th and afterwards are filtered for tweets written in English and using the keyword ‘BTS’. The total amount of streamed tweets were 28,359.

Our result is displayed using a time series where for each day, the net daily sentiment of tweets per state is shown. The results range from 0-5 net daily tweets per state. This is because most of the tweets that were streamed were retweeted and were missing geo-location information. After preprocessing and removing the irrelevant tweets, there were 1864 tweets. By further strictly filtering for tweets from the United States, the processed data set of tweets reduced even further. If we were able to filter for retweets, missing geolocations, languages, and tweets from United States early on in the data collection step and also by increasing the length of collection, we could have had more valuable data.

To validate the results, we incrementally tested the filtering methods (checking for retweets, tweets with no locations, and for tweets from users that live in the United States) on small versions of the streamed data. We began with JSON files containing about 20 tweets and then once confirmed that the preprocessing methods worked, generated the processed tweets JSON file, and ran the sentiment analysis code on the small processed file. After confirming the results by checking if they were indeed positive, negative, or neutral, we ran the processing and sentiment analysis methods on the larger data sets.

The software used to implement the project include Jupyter Notebook which allowed us to run Python 2 code and share documents that containing live Python code and visualizations. The Twitter Streaming API using a valid Twitter account was also utilized to stream and collect the relevant tweets. The library used to perform sentiment analysis is the TextBlob library. The Pandas library was used to create multiple time series to represent the granulated net daily tweets per state per day.

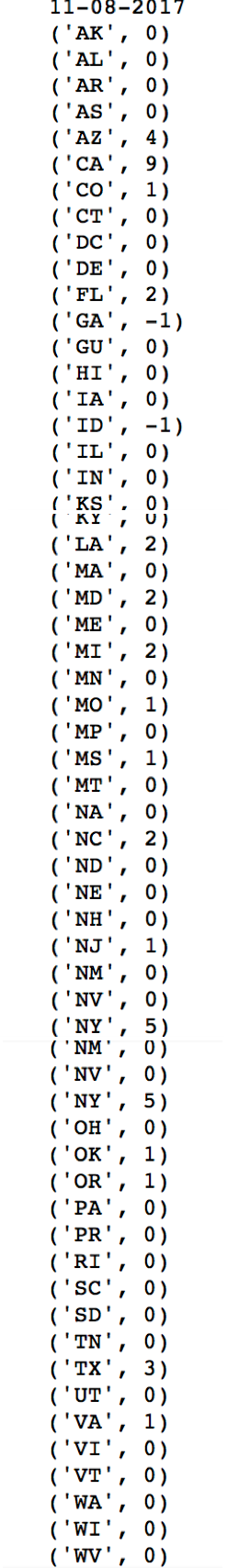
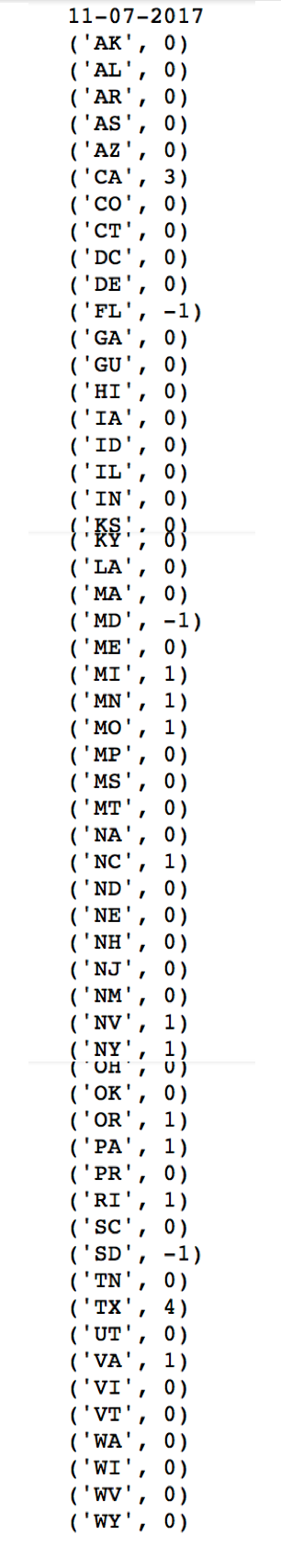
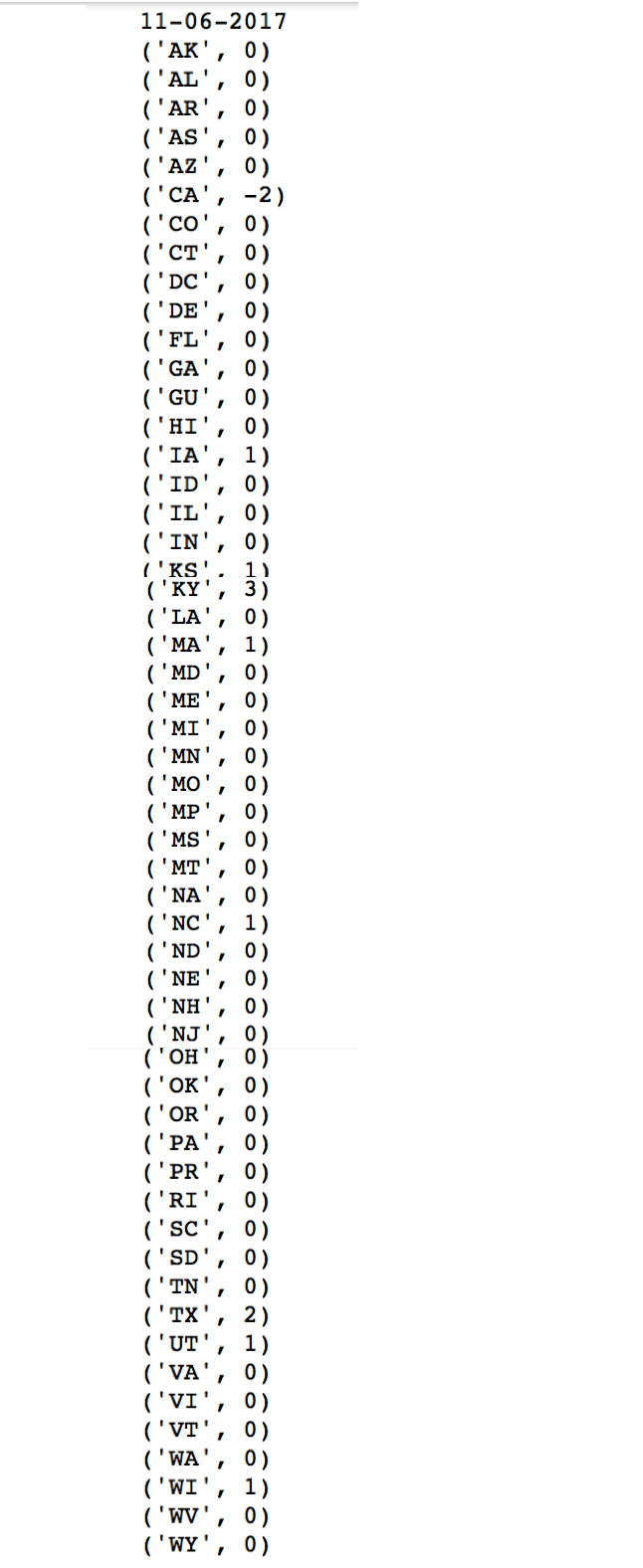
|  |  |  |
| --- | --- | --- |
| **Task** | **Completed Date** | **Completed By** |
| Project proposal | 10/12 | All members |
| Data collection | 10/14 - 11/8 | Lisa |
| Data preprocessing | 11/3, 12/4 | Doyun and Lisa |
| Intermediate report | 11/7 | All members |
| Develop sentiment analysis method | 11/26 | Crystal |
| Generate net daily sentiment per state for all processed data | 12/4-12/5 | Lisa and Crystal |
| Code documentation | 12/5 | Lisa |
| Final report | 12/7 | All members |

**Fig. 1**. Timeline of Project

## **Experimental Results**

For filter and mapping tweets system, you should report the number of positive and negative comment about BTS with each regions in the States, which means you have to map the data set such as comments and geolocation. Then passing the analysis path which is figure it out that the comment is positive comment or negative. If the comment is positive then giving value as positive number and counting it, if it is negative number it will be negative number will be counting it. So in the mapper function, key is geolocation and value is numbers depend on the positive and negative comments.

Reducer input are list of the data set that mapped from mapper function. Reducer function will start to add up all positive number and negative number so show how much that location is favorable to BTS.



Following the results, It shows that CA is the most activative place for talking about BTS and next following is TX and NY. The most hostile reaction came from SD, GA and MD. Also the most friendly reaction can see from CA and NY.

## **Discussion**

During this processing there is little minor problem that we did not expected that which is people who use twitter, they turn off their location for twitter. So when we stream and collect data then mapping the data set with geolocation, but there is so many null value, so after when we filtered out the null value number of data set decrease a lot. So overall output is not reliable base on the original streaming data numbers. However our goal was try to figure out how each States and people think about BTS and relationship between the location and tweet character. So even we could not able to get many tweets, geolocation value was main key value. Following the results, States where immigrant people live a lot, that place shows positive and more activities tweets actions such as CA, NJ, and NY. But States where immigrant people are not living a lot such as SD, those place comments is negative or not actively tweets.

# **CONCLUSIONS**

The overall result is positive sentiment towards BTS given the daily net sentiment shown in the report. However, it should be noted that despite the vast number of tweets that were initially collected, the actual number used in the data set was far smaller due to filtering constraints based off the parameters of our project. Additionally, there are no specific states that tends to be more correlated with an overall positive, negative or neutral sentiment. Other things to account for these results could be the filtering of language. Given that BTS is a korean based group, and a good portion of their fanbase tweet in korean too, the results given are not wholly representative of the true sentiment regarding BTS. For future improvements and more conclusive results, the locations looked at, as well as the preprocessing of the data could be improved by including more regions and languages.

**REFERENCES**

1. <https://developer.twitter.com/en/docs>
2. <http://textblob.readthedocs.io/en/dev/index.html>
3. <http://code.activestate.com/recipes/577305-python-dictionary-of-us-states-and-territories/>
4. http://cse.msu.edu/~mayao4/cse482/project/Topics.pdf