

## Sentiment Analysis on Artist's Song and its Changes Over Time (1496 Words)

### Introduction

Recently, there has been great increase in importance of social media platforms when analyzing public's opinion. Twitter is one of the platforms that has become popular grounds for sentiment analysis. According to Giachanou and Crestani, "user-generated information" can be constructive documentation when comprehending general opinion on certain topics (Giachanou & Crestani, 2016). The analysis of social media became a substitute of doing individual surveys manually (Jain, 2013). The contents of tweets can become good indicator whether the audience has preference on certain music artist, product, or service (Jain, 2013). BTS is one of the leading K-pop boy bands who has been actively releasing albums until today. According to Billboard, their song "DNA" was the first song to be appeared on Billboard Hot 100 chart (Pascual, 2022), which can be a point of reference when indicating "bigness of hits" (Molanphy, 2013).

### Background & Related Study

The main purpose of user's activities on Twitter is to discuss their opinion freely and openly with others, which creates "user generated content" that fulfills four statuses of "wise crowds" – diversity, independence, decentralization, and aggregation (Vossen, 2013). The function of hashtags helps to aggregate information by topics, which can further contribute greatly as an adequate source for sentiment analysis.

With these advantages, Twitter has become one of the main sources that is used to predict on certain results. As one of the examples, the sentiment analysis on tweets has been used to predict on the popularity of the movie. Similarly, the tweet data has been used as a core feature in predicting music sales and chart performance. According to the research done by Vossen, the tweets are collected and employed to anticipate on the music sales. The main finding of the research is that daily collection of tweets and cumulative reach of tweets sent by unique users play as great features for predicting music sales (Vossen, 2013). However, the sentiment analysis had minute impact on the conjecture of the music sales.

Furthermore, Tsiara and Tjortjis explores on the relationship between number of tweets and Billboard chart position and how sentiment analysis of tweets effects the chart position of a song. The correlation coefficient was marginal between public's interest is whether positive or negative and the ranking position of artist's song (Tsiara & Tjortjis, 2020). The reasonable interrelationship between number of mentions of a song and the performance on the chart was discovered.

Through this paper, I analyzed on how public's opinion changed over time on artist's song, specifically "DNA" by BTS. The song was the first song to be appeared on Billboard's Hot 100 chart, which can be pointed out to be the beginning of popularity in USA. To investigate on how sentiments and interested topics change between before and after the release of the song, I intend to answer following research questions:

- *How does people's sentiments change over time on an artist's song? Do degrees of emotion change before and after the release of the song?*
- *How does interested topics among twitter users change before and after the release of the song? Does the change have relationship with the change of public's sentiment?*

## Method

### Data

To answer the research questions above, I selected four different timeframes: a week before the release of the song, a week of the release, a week after the release, and two weeks after the release. About 1000 tweets will be collected from each timeframe. The python package “snsraper” is used to scrape the tweet data to circumvent the problem of rate limits and time-period limits. About 4000 tweets were gathered in total. To implement TwitterSearchScraper module, following words were used as main keywords: BTS, DNA, #bts, and #dna. To perform the module with narrower search criteria, I used billboard and streaming as a subset keyword. However, “streaming” was not used when collecting a week before release tweet data. Also, only tweets written in English was collected for the analysis. The unrelated tweets were taken out manually during exploratory data analysis (EDA). The tweets posted by same username and identical tweets posted by different username was not filtered because the redundancy can show user’s level of interest.

### Analysis

#### Sentiment Analysis

The sentiment analysis of tweets was carried out by using Valence Aware Dictionary and sEntiment Reasoner (VADER) tool. VADER is one of the appropriate tools as it is a lexicon and rule-based tool that is specifically accustomed to social media. The tool quantifies the sentiment of the tweets by calculating compound score of positive, negative, and neutral ranging from 0 to 1. With the label given, the percentage of types of tweets will be calculated for the analysis.

The implementation was done for each collection of tweets from four different timeframes in the following steps:

1. Creating data frames for each timeframe:

To scrape the Twitter data, TwitterSearchScraper function was used, and unnecessary columns were removed from the dataset.

2. Preprocessing the contents into machine-friendly mode:

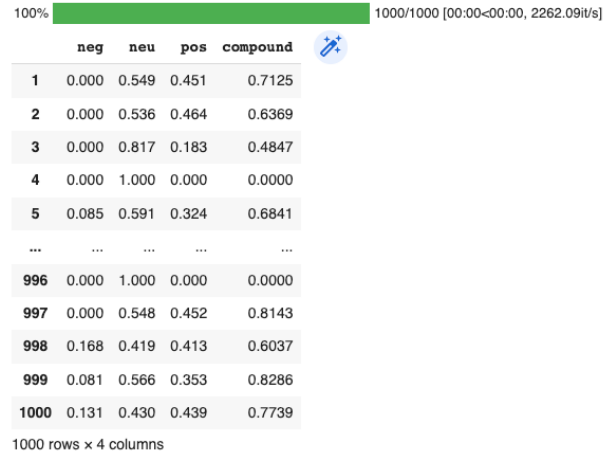
The preprocessing was done on the “content” column of the data frame, which shows the post of the tweets collected. The process of removing nonessential elements such as urls, @mentions, hashtag signs in front of the words, “\n”, and non-English words was done. Then the contents were tokenized and created new columns containing extracted tokens. The process of lemmatization was done also. The below image (Image 1) shows the example of the data frame:

id		url	date	content	username	text	tokens	hashtag	tokens_back_to_text	lemmas	lemmas_to_text	lemma_tokens
0	1	https://twitter.com/PsychicBoySuga/status/9092...	2017-09-16 23:53:57	Online Promotion for @BTS_twt #LOVE_YOURSELF DNA #...	PsychicBoySuga	Online Promotion for LOVE YOURSELF DNA BTS	[Online, Promotion, for, @BTS_twt, #LOVE_YOURS...	[LOVE_YOURSELF, DNA, BTS]	Online Promotion for LOVE YOURSELF DNA BTS ht...	[Online, Promotion, love, dna, BTS, https, t, ...	Online Promotion love dna BTS https t co vorcz...	[Online, Promotion, love, dna, BTS, https, t, ...
1	2	https://twitter.com/VickyJimin/status/90920291...	2017-09-16 23:50:42	#BTSALBUM1DAY \n #DNA \n/ are you ready \n \n	VickyJimin	BTSALBUM1DAY DNA are you ready lets party BTS	[#BTSALBUM1DAY, #DNA, /, are, you, ready, lets...	[BTSALBUM1DAY, DNA, BTS]	BTSALBUM1DAY DNA are you ready lets party BTS...	BTSALBUM1DAY, dna, ready, let, party, BTS, ...	BTSALBUM1DAY dna ready let party BTS https t...	[BTSALBUM, 1DAY, dna, ready, let, party, BTS, ...
2	3	https://twitter.com/king_kimyohan/status/90920...	2017-09-16 23:45:42	What's DNA ? \n BIGHIT: Do not answer \n Me: Die...	king_kimyohan	What DNA BIGHIT Do not answer Me Die now antis...	[What's, DNA, ?, BIGHIT, :, Do, not, answer, M...	[BTSALBUM1DAY, BTS, DNA]	What DNA BIGHIT Do not answer Me Die now antis...	[dna, bight, answer, die, artis, d, differenc...	dna bight answer die artis d difference btal...	[dna, bight, answer, die, artis, d, differenc...
3	4	https://twitter.com/jingyeokebts/status/909200...	2017-09-16 23:42:01	Two words, six letters, legendary and taking t...	jingyeokebts	Two words six letters legendary and taking the...	[Two, words, ,, six, letters, ,, legendary, an...	[BTS, DNA]	Two words six letters legendary and taking the...	[word, letter, legendary, take, world, storm, ...	word letter legendary take world storm bts dna	[word, letter, legendary, take, world, storm, ...
4	5	https://twitter.com/mikiyaa/status/9091995086...	2017-09-16 23:37:09	Sorry I forgot if there date 17 🍌 @BTS_twt #BT...	mikiyaa	Sorry I forgot if there date 17 BTS HER DNA LO...	[Sorry, I, forgot, if, there, date, 17, 🍌, @BT...	[BTS, HER, DNA, LOVE_YOURSELF]	Sorry I forgot if there date 17 BTS HER DNA LO...	[sorry, forgot, date, 17, BTS, dna, love]	sorry forgot date 17 BTS dna love	[sorry, forgot, date, 17, BTS, dna, love]

Image 1. Example Data Frame of Scraped Tweets

### 3. VADER analysis:

The quantification of sentiment was calculated using VADER library and preprocessed texts were used as the input text. The percentage of positive, negative, and neutral tweets were calculated for each timeframe. The below image (Image 2) shows the example of data frame obtained from VADER analysis:



	neg	neu	pos	compound
1	0.000	0.549	0.451	0.7125
2	0.000	0.536	0.464	0.6369
3	0.000	0.817	0.183	0.4847
4	0.000	1.000	0.000	0.0000
5	0.085	0.591	0.324	0.6841
...	...	...	...	...
996	0.000	1.000	0.000	0.0000
997	0.000	0.548	0.452	0.8143
998	0.168	0.419	0.413	0.6037
999	0.081	0.566	0.353	0.8286
1000	0.131	0.430	0.439	0.7739

1000 rows x 4 columns

Image 2. Example Data Frame Obtained from VADER Analysis

### *Topic Modeling*

For the second part of my analysis, I implemented topic modeling by using Latent Dirichlet Allocation (LDA). Before, creating the model, I used word cloud to visualize the most used words using the preprocessed tweets and compare the difference in frequency and importance of keywords between different timeframes. Then, another preprocessing was done such as removing stop words and making bigram. Then the dictionary was created using the words. The base model was created using LDA function under gensim library. To evaluate the model, both perplexity score and coherence score were obtained. Using GridSearchCV, the best parameters were found. To find appropriate parameter value for number of topics, coherence values were computed for each topics created. With the parameters acquired, final model of LDA was made and executed topic modeling. To visualize the result, I used pyLDAvis under gensim library and created interactive visualization. The below image (Image 3) shows the example visualization of LDA model:

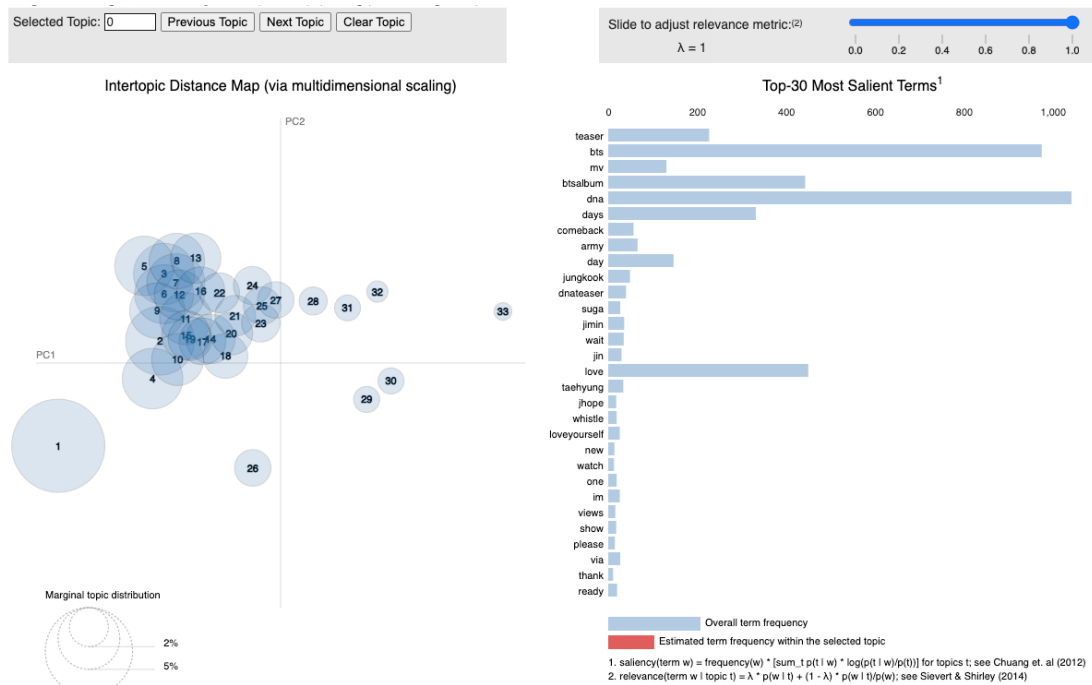
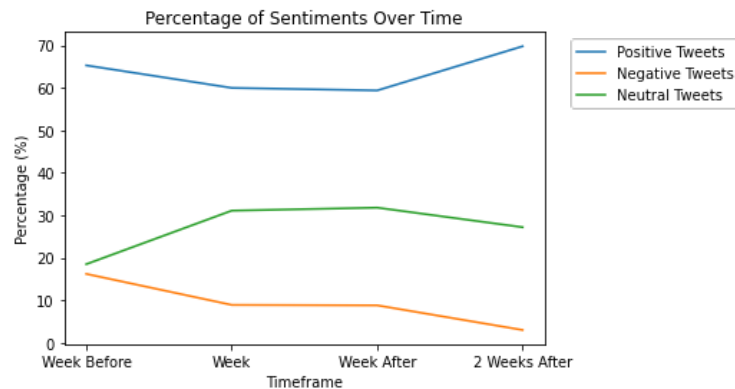


Image 3. Example of LDA Visualization

## Result

The following plot shows the sentiment changes over the time:



Looking at the graph, I was able to analyze that there were apparent number of differences between positive, negative, and neutral tweets for each time frame. It was interesting to see how percentage of positive tweets show increasing trend despite the percentage goes down during the week and the week after. However, for negative tweets, the decreasing trends can be seen explicitly. From this we can observe that public tends to have higher tendency to have optimistic reactions toward “DNA” by BTS over time. Also, it was interesting to see how the percentage of positive tweets was highest two weeks after the release of the song.

In addition, analysis on relationship of number of unique users and the size of dictionary was done. Dictionary is bag of unique tokens from the text data. Looking at the table (Table 1) below, both number of unique users and number of unique tokens or the size of dictionary increases greatly on the week of the release. Also, it is engaging to see how number of unique of users are similar during week before and two weeks after but the size of dictionary differs greatly. From this we can infer that the variety in content of the texts decreases over the time.

Number of Unique Users		Size of Dictionary	
Week Before	604	Week Before	1615
Week	682	Week	1863
Week After	464	Week After	1616
2 Weeks After	608	2 Weeks After	1125

Table 1. Number of Unique Users vs. Size of Dictionary

This phenomenon can be also shown through visualization through word clouds. Looking at the below image (Image 4), we can perceive that the volume of keywords increases a lot during the week of the release. However, the most presented keywords do not change over time.

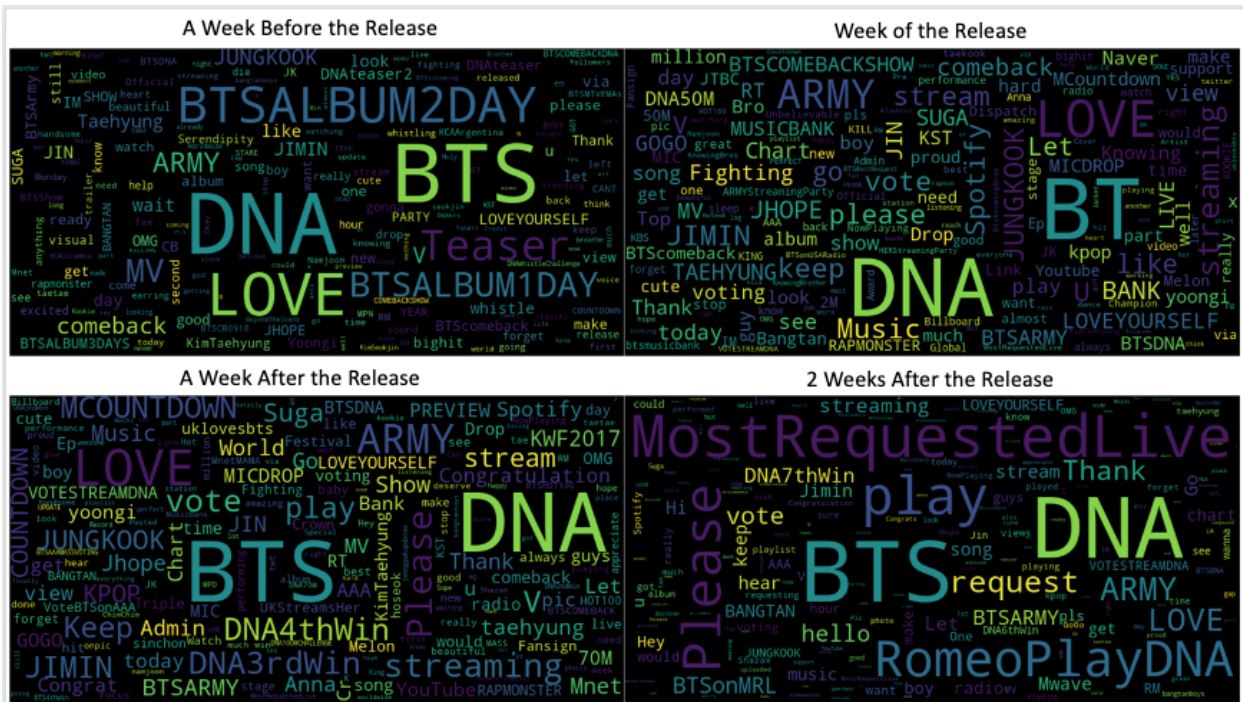


Image 4. Word Cloud for Each Timeframe

Also, looking at the visualization of LDA for each timeframe, I was able to analyze that discussed topics among the users have high similarities between each other, which infers that tweets among the users on “DNA” have high level of similarity.

## Conclusion and Limitation

In conclusion, public's sentiment changes over time on artist's song. However, the interested topics tend to not change over the time. From the analysis above, we can conclude that degree of positiveness increases as the time proceeds. However, the variety of topics or content was not apparent, which gives insight that the correlation between change of interested topics and sentiment cannot be defined.

For further and thorough research, the analysis can be done over longer term with higher number of collected tweets. The time lag can be considered, and public's response "honest" reactions can be revealed circumspectly. Also, more keywords can be used to possibly add more variety into collected tweets and reduce the noise in data.

## References

- Giachanou, A., & Crestani, F. (2016). Like it or not. *ACM Computing Surveys*, 49(2), 1–41. <https://doi.org/10.1145/2938640>
- Jain, V. (2013). Prediction of Movie Success using Sentiment Analysis of Tweets. *International Journal of Soft Computing and Software Engineering [JSCSE]*, 3(3), 308–313. <https://doi.org/10.7321/jscse.v3.n3.46>
- Molanphy, C. (2013, August 1). *How the hot 100 became America's hit Barometer*. NPR. Retrieved October 14, 2022, from <https://www.npr.org/sections/therecord/2013/08/16/207879695/how-the-hot-100-became-americas-hit-barometer>
- Pascual, D. (2022, May 2). *BTS' 10 top songs on the Billboard hot 100*. Billboard. Retrieved October 14, 2022, from <https://www.billboard.com/lists/bts-top-songs-billboard-hot-100/bts-feat-designer-mic-drop/>
- Tsiara, E., & Tjortjis, C. (2020). Using Twitter to predict chart position for songs. *IFIP Advances in Information and Communication Technology*, 62–72. [https://doi.org/10.1007/978-3-030-49161-1\\_6](https://doi.org/10.1007/978-3-030-49161-1_6)
- Vossen, R. (2013). Does chatter matter? Predicting music sales with social media.