A Text Mining Approach to the Analysis of BTS Fever

Research-in-Progress

Soobin Choi **Gayeon Park**

Hee-Woong Kim

Abstract

K-POP is steadily growing with global competitiveness. The rise of K-POP's popularity has continued to create Korean idol groups. However, many idol groups were dismantled and there is lack of measures for overseas advance and success. Therefore, this study aims to analyze the success factors of BTS by focusing on the text mining techniques. After collecting Twitter's online postings using crawling technique, we will analyze in three text mining techniques: topic modeling, keyword extraction, and term frequency analysis. By analyzing data with three text mining methods, we will derive how BTS could success globally and form a huge fandom. And with the derived key factors, we will suggest a success strategy based on the analysis results. In contrast to previous studies that were centered on case studies or interview, this study has implications in that the actual data was collected and analyzed through three text mining techniques.

Keywords: BTS, K-POP, Text Mining, Topic Modeling, Keyword Extraction, Term Frequency Analysis

Introduction

The term 'Korean Wave', a term referring to Korean popular culture preference phenomenon, is constantly expanding around the world. The popularity of K-POP (Korean pop) has been steadily evolving around the world for the last 20 years and is rapidly emerging as a new content of the Korean Wave. K-POP has a huge fan base not only in China, Japan, and Southeast Asia, but also in Eastern Europe (Jang and Song 2017; Fu and Shon 2018). Through the social network service (SNS), K-POP has formed a global fandom, increasing its awareness and proving its popularity (Choung 2010). BTS, Korea's representative idol group, have a lot of global fans and the number is continuously increasing.

After one year of the debut, BTS started touring in Seoul with performances in Japan, Philippines, Singapore, Thailand, etc., and quickly met with global fans. In 2015, BTS was selected as the Korean representative for the “World Wide Act”, which selected as the most active artist in MTV EMA 2015 (European Music Awards). MTV EMA said, "The next generation of large-scale K-POP artist group

that has shown remarkable growth rapidly since its debut in 2013. BTS is trying to write and work directly with the music, and to convey the thoughts and stories of the people through BTS music".

On the iTunes K-POP chart in the US in 2015, the BTS's album 'Hwayang-Yeonhwa PT. 2' and the title song 'RUN' won the top album chart and top song chart, and the best K-POP album 4th place and singing section '2015 best K-POP album and 2015 best K-POP song 20' at the same time. And "RUN," the album title track, were ranked fourth and third, respectively. When they released their second album 'WINGS' in 2016, it was ranked 26th in the US Billboard 200 chart. In addition, Billboard Album Chart and Billboard Digital Song Chart were ranked # 1. Also, for the first time as a Korean singer, they entered # 16 on the Indie album chart, including # 62 on the UK official album chart. In addition, it has ranked # 1 on the latest weekly charts for Japanese music store tower record comprehensive album, # 1 on the Chinese video site, and # 1 on the iTunes music charts in the iTunes 97 countries all over the world.

BTS entered the Billboard 200 with Billboard # 7 on the first week after releasing the 'LOVE YOURSELF' album in 2017, and the five albums released until that time were on the Billboard 200 charts in succession. "DNA" was also ranked on Billboard's 100 Best Songs of 2017 for # 49, and BTS was the only K-POP group. The BTS has once again proved its popularity as a Korean idol group for the second time in the history of the Korean singer in 2012 after the release of ABC's new year's show "New Year's Rocking Eve 2018" live on December 31, 2017 throughout the United States.

In 2017, Big Hit Entertainment, an entertainment agency of BTS, raised its sales and operating profit by 162.3% and 213.5% from 2016 to KRW 92.4 billion and KRW 32.5 billion, respectively. Hana Financial Group expects Big Hit Entertainment sales to reach KRW 140 billion and operating profit to reach KRW 50 billion in 2018. Big Hit Entertainment's 2017 operating profit surpassed the YG (25.2 billion won), JYP (19.5 billion won), and SM (10.9 billion won), respectively, which are classified as Big 3 for domestic entertainment agencies.

The popularity of K-POP continues to grow, and the range of activities is so wide that if a group has many fans profitability is easier. As the K-POP market becomes more active, various entertainment agencies in addition to Korea's three leading entertainment agencies, SM, YG and JYP will continue to compete for a new debut (Jang 2009). Many new groups continue to prepare for a debut, but most of the groups are disbanded in a short period of time without getting popular. At this point, research on strategies to grow into content that meets peoples' needs and expectations is needed (Park et al., 2013).

Despite of the popularity of K-POP, previous research has focused on case studies or interviews (Ubonrat and Shin 2007; Cho and Sim 2013). There has been a lack of empirical analysis, especially about the critical success factor of K-POP. This study, therefore, aims to examine the critical success factors of K-POP by focusing on the BTS case. So in this research, we will analyze the success factors using text mining techniques and suggest success strategies that helps new idol groups can grow into sustainable content.

We will do the text mining approach to analyze social opinions about BTS by crawling social media data. Specifically, the main issues of the BTS will be derived through data analysis using three text mining techniques: topic modeling, keyword extraction, and term frequency analysis. Also, we will conduct interviews with BTS fans, entertainment officials, and industry experts for additional analysis. Finally, with the derived key factors, we will propose critical success factors based on the text mining of social opinions and the interview-based additional interpretation of those identified factors.

This study has academic significance in that it uses a mixed research method to derive major success factors through quantitative research analyzing actual data and qualitative research through interviews. Also the result of this study has a practical implications in that it suggests the success strategy and the overseas advance plan of the newborn idol group in the future. In addition, it provides management insight because it is closely related to the sales and operating profit of the entertainment agency.

Conceptual Background

Previous Research on K-POP and BTS

Most of previous research on K-POP was conducted through interviews. There were studies about K-POP culture and success factors, the spread of the Korean Wave, and the relationship between popularity and social media (Jin and Yoon 2016; Cho and Sim 2013; Leung 2012; Oh and Park 2012).

Table 1. Previous Research on K-POP

Author	Method	Field	Purpose	Results
Jin and Yoon (2016)	Interview	K-POP Culture	Study how the Korean Wave phenomenon is integrated into the culture of social media-led	Social media platform and social interaction of fans interplayed and Korean wave spread rapidly
Cho and Sim (2013)	Interview	K-POP Success Factors	Analyze K-POP's global success factors and explore the impact of smart media on K-pop's globalization	Changes in the smart environment and social media play a crucial role in spreading the K-POP craze
Leung (2012)	Interview	K-POP Wave	Study how K-POP is built, spread and accepted globally	An online K-pop fan culture builds a digital community that strengthens the bond between the fans and the connection between fans and music
Oh and Park (2012)	Interview	K-POP Sales	There is an empirical relationship between social media and K-POP popularity in Europe	Social media plays an important role in validating the K-POP business model.
Ubonrat and Shin (2007)	Interview	K-POP Consumption	How Thai youths became consumers of K-POP	Direct contact between artists and fans (concerts, fan meetings) and fandom activities play an important role in K-pop consumption

Jin and Yoon (2016) studied how the Korean Wave phenomenon is integrated into the social media-led culture and confirmed that social interaction with fans through the social media platform quickly spread the Korean Wave. Leung (2012) conducted a study on how K-POP is spreading and accepting Korean Wave worldwide, and online K-POP fan culture builds digital community to strengthen the bond between the fans and strengthen the bond between fans and music. The previous studies related to K-POP are summarized as Table 1.

As it can be seen from previous studies, most studies related to K-POP have been conducted through interview methods, and there are insufficient studies to identify the success factors by collecting and utilizing actual data. Therefore, this study will collect and utilize actual online social media data as well as interview data about BTS, and through this, we will derive the main success factors of BTS and suggest the success strategy of the K-POP in the future.

Research Methodology

The research process of this study is shown in Figure 1. First, we crawl the data and collect the latest data about BTS on Twitter. After data preprocessing, LDA topic modeling, keyword extraction, and term frequency analysis of text mining techniques are used to derive the success factors of BTS. Secondly, we will analyze the code with the data obtained from the interviews of the BTS fans, the entertainment officials, and the industry experts and derive the success factors of BTS. As a result, we will derive the key success factors of K-POP through quantitative data analysis of text mining and qualitative data analysis using interviews. And with the derived results, we will propose a success strategy for K-POP in the future.

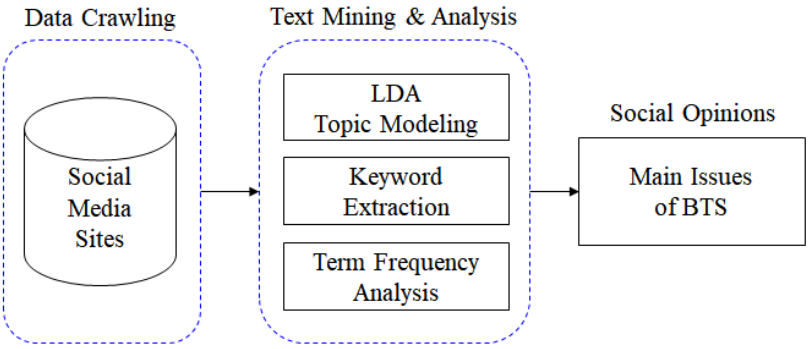

Figure 1. Research Procedure

Text Mining: LDA Topic Modeling

In this study, we will use topic modeling as the first technique of text mining. Topic modeling is a procedural probability distribution model that finds a certain pattern in a document or text and finds potentially meaningful topics. It is an algorithm that probabilistically calculates and extracts the result as a set of words likely to correspond to the topic on the assumption that the words constituting the paragraphs are not independent (Blei 2012; X. et al., 2014). After several equations, the probability that the i-th topic Z of the d-th document is assigned to the j-th is given. As a result, AB is the degree of relevance of the d-th document to the k-th topic multiplied by the relevance of the n-th word of the d-th document to the k-th topic (see Equation 1). Therefore, this study aims to analyze the database that derives the success factors of BTS through LDA topic modeling.

$$\mathbf{P}\left(\boldsymbol{Z_{d,i}} = \mathbf{j}\middle|\boldsymbol{Z_{-i}}, \mathbf{w}\right) = \frac{\boldsymbol{n_{d,k}} + \boldsymbol{a_k}}{\sum_{i=1}^{k}(\boldsymbol{n_{d,i}} + \boldsymbol{a_i})} \times \frac{\boldsymbol{v_{k,w_{d,n}}} + \boldsymbol{\beta_{w_{d,n}}}}{\sum_{j=1}^{V}(\boldsymbol{V_{k,j}} + \boldsymbol{\beta_j})} = \boldsymbol{AB} \quad \ldots\ldots \text{ Eq 1. LDA Probability Distribution}$$

Text Mining: Keyword Extraction

In this research, we will conduct keyword extraction method as the second technique of text mining. A keyword extraction method is defined as finding key keywords in the data field. By extracting keywords that are topics, it can create a basis for understanding trends (M. and A., 2015). The following TF-IDF weight model is used for keyword extraction (see Equation 2). The TF-IDF weight model evaluates the relative importance of words within a document and determines the ranking of the most similar documents. Therefore, a word that has a large TF-IDF value can be used as an important measure to extract key words (Lee and Kim 2008).

$$\mathbf{TF-IDF} = \mathbf{TF} \times \mathbf{IDF} = \boldsymbol{F_{t,d}} * \mathbf{log}\left(\frac{\boldsymbol{N}}{\boldsymbol{n_t}}\right) \quad \ldots\ldots \text{ Eq 2. TF-IDF Weight Model}$$

Text Mining: Term Frequency Analysis

In this study, we will use the term frequency analysis technique as the third text mining technique. The term frequency analysis corresponds to the TF in the TF-IDF model and TF is the abbreviation of the Term Frequency (see Equation 3). In other words, the number of occurrences of a particular word in text data can be used to determine which word is most frequently referred to (J Pastizzo and F Carbone 2007). Therefore, in this research, we try to identify the most frequent occurrence of term in Twitter data by using word frequency analysis method, and derive the success factors of BTS by using frequency.

$$\boldsymbol{tf_{i,j}} = \frac{\boldsymbol{n_{i,j}}}{\sum_k \boldsymbol{n_{k,j}}} \quad \ldots\ldots \text{ Eq 3. TF Equation}$$

Data Collection and Analysis Results

Data Collection: Twitter.com

The data for this study were collected from Twitter (Twitter.com) using crawling techniques for BTS related posts. Twitter is a social network service and a microblog service. It is a platform where people from all over the world can share and communicate their opinions and thoughts via tweets that are

shorter than 140 characters. The reason why we choose Twitter for data collection is that the BTS account, which is now called the best twitter star, has tweets that are selected as the most tweeted accounts in the world. Therefore, we can analyze and synthesize vast opinions. As a result, we have collected 20,716 of BTS related data on Twitter since January 1, 2017 through December 31, 2018. Not only the fans of BTS, but also the members of BTS actively participated and we were able to see many articles about BTS in public.

Data Analysis Method

This chapter describes how to analyze the collected data. As it mentioned earlier in the research methodology, we will analyze Twitter data by using three text mining techniques. Before analyzing the data, we first preprocess the data, and then we use the topic modeling method to derive topics from BTS on Twitter. Second, we will seize the keywords extracted from the collected data by using the keyword extraction technique. Finally, through the term frequency analysis, we will identify which words are most frequently mentioned through the frequency of occurrences of words in text data. As a result, the main success factors of BTS will be derived based on the results obtained through the three techniques.

Analysis Results

First of all, with the result of topic modeling on the total tweets (20,716), we could identify that the topics related to Loyalty to BTS, Voting for BTS, Loyalty to each BTS Member, ARMY (Fan Club), BTS Concert and Performance, Awards, Best Idol Group, BTS Music and Performance, BTS Collaboration, New Challenges of BTS are mentioned. Derived topics and topic-related keywords are listed in Table 2.

Table 2. Topic Modeling Results

Topic	Description	
	Label	Keywords
01	Loyalty to BTS	angel, proud, love, beautiful, thank you, best, talented
02	Voting for BTS	bts_twt, retweet, likes, replay, vote, awards, rates
03	Loyalty to each BTS Members	방탄소년단, 뷔, V, bts_twt, bts, jimin, 정호석, JIN, 태형,
04	ARMY (Fan Club)	ARMYs, btsarmy, 아미, mostrequestlive, bts_twt, minniejoons
05	BTS Concert and Tour	btsloveyourselftour, btsincitifield, concert, newyork, btsworldtour
06	Awards	billboard, pcas, disney, Europe, mpn, bts, epiphany
07	Best Idol Group	kpop, idol, thegroup, bts_bighit, US, universe
08	BTS Music and Performance	뷔, 방탄소년단, rap, music, jimmy, thesong
09	BTS Collaboration	bts_collab, vt_cosmetics, moment, making, g8, g6
10	New Challenges of BTS	rts, game, kpop, wanted, movie, tweet, heart,

Secondly, the analyze results through the keyword extraction as the second analysis technique, the top 30 key keywords are listed in order of the weight value which indicates the TF-IDF value. The top 30 extracted keywords are listed in Table 3 and it can be visually identified which words are important. Next to each keyword, weights are displayed, and the keywords are interpreted as words. As a result, it can be seen that not only keywords related to the BTS Member (8 keywords) are mentioned most frequently but also weights are also highly distributed. Followed by BTS Member related keywords, there are 6 keywords related to BTS Concert in the top 30, and weights are also generally higher. In conclusion, it can be understood that the core keyword in the BTS related Twitter data are the member

name of BTS and the concert of BTS. It means fans like each member of the BTS, and they also like BTS's songs and performances at the concert.

Table 3. Keyword Extraction Results

Keyword		Weight	Interpretation	Keyword		Weight	Interpretation
01	태형_뷔	1	BTS Member	16	boxy_smile	0.455	Performance
02	milktae_jpg	0.771	BTS Member	17	concert_venue	0.445	BTS Concert
03	shoot_dance	0.583	Performance	18	security_guard	0.433	BTS Concert
04	뷔_	0.534	BTS Member	19	family_members	0.433	BTS Member
05	eng_vers	0.516	BTS Music	20	rap_duo	0.412	BTS Music
06	btschartdata	0.498	SNS Voting	21	finger_heart	0.391	Performance
07	btsloveyourselftour	0.498	BTS Concert	22	concerts	0.385	BTS Concert
08	fanart	0.498	Fandom	23	슈가	0.368	BTS Member
09	lip_bite	0.498	Performance	24	fan_account	0.355	Fandom
10	universe	0.498	Idol Group	25	music_nominations	0.326	Awards
11	boys	0.498	Idol Group	26	monster_army	0.279	Fandom
12	taehyung	0.497	BTS Member	27	bts_collab	0.270	Collaboration
13	attendance_record	0.464	BTS Concert	28	prince	0.266	BTS Member
14	solo_spotlight	0.458	BTS Concert	29	heart_attack	0.262	BTS Member
15	btsvotingteam01	0.385	SNS Voting	30	idol	0.257	Idol Group

Lastly, the term frequency analysis was used as the third analysis technique. It is a part of corresponding to the TF (Term Frequency) of the TF-IDF model, and it can grasp how frequently a specific word appears in the data. The results are listed in Table 4, top 30 frequency were selected. For each word, the interpretation is displayed along with the frequency of occurrences in the table. As you can see from the results table, we could find that 'bts_twt' has the highest TF value (1871), which is the official account of BTS on Twitter. This means BTS members and fans are constantly communicating globally through online official accounts.

As it can be seen on top 10 high frequencies, the frequency of terms about BTS Group is significantly higher, and the terms about the BTS Concert and fan club (ARMY) also have a high frequency. As a result of the data analysis, it can be confirmed that the fans mention a lot about the concert and words indicating BTS themselves through the official SNS account. Also, since the fan-related words such as the fan club (btsarmy) show high frequency, it can be seen that the fandom activity is very active.

Table 4. Term Frequency Results

Term		Freq.	Interpretation	Term		Freq.	Interpretation
01	bts_twt	1871	BTS Group	16	jimmyfallon	104	Fandom
02	btsxcitifield	1606	BTS Concert	17	btsvotingteam	100	SNS Voting
03	bts	1076	BTS Group	18	army	87	Fandom
04	뷔	516	BTS Member	19	제이홉	83	BTS Member
05	방탄소년단	447	BTS Group	20	태형	81	BTS Member
06	btsincitifield	349	BTS Concert	21	thesong	78	BTS Music
07	pcas	274	Awards	22	taehyung	78	BTS Member
08	citifield	252	BTS Concert	23	rm	76	BTS Member
09	btsarmy	243	Fandom	24	mostrequestliv	74	BTS Music

					e		
10	minniejoons	203	Fandom	25	Jungkook	73	BTS Member
11	btsloveyourselftour	201	BTS Concert	26	bts_bighits	71	Kpop Idol
12	idol	195	Kpop Idol	27	concert	71	BTS Concert
13	지민	160	BTS Member	28	정호석	71	BTS Member
14	thegroup	142	Kpop Idol	29	themusicvideo	47	BTS Music
15	newyork	113	Fandom	30	mpn	42	Awards

Using the three text mining techniques, we analyzed 20,716 Tweets related to BTS and found the results. The result of keyword extraction and term frequency analysis were mapped on 10 topics derived from the topic modeling. The 10 topics are Loyalty to BTS, Voting for BTS, Loyalty to each BTS Members, ARMY(Fan Club), BTS Concert and Tour, Awards, Best Idol Group, BTS Music and Performance, BTS Collaboration, and New Challenges of BTS. So the main success factors of BTS, which were commonly found through three text mining techniques, are Voting for BTS, Loyalty to each BTS Member, ARMY (Fan Club), BTS Concert and Tour, BTS Music and Performance, and Best Idol Group (7 factors). Therefore, the overall interpretation of mapped result is that we could find that ARMY, which is a fan club name of BTS, is actively promoting its presence by using twitter. Which means the official online fan account is active all over the world. Also, we could identify the interest and enthusiasm for the music, performance, and concert of the BTS and loyalty to each member of the BTS. It also proves that BTS is the most successful idol group, and it shows that BTS has formed worldwide fandom. The popularity is proved once again through the awards.

As a result, the main issues of BTS are: most of all, the global fandom (ARMY) is formed and it proves they are the most powerful and dedicated fandom in the world by active activity. Second, ARMYs (BTS fan club) plays the important role on twitter which promotes to people for voting when music awards voting are ongoing. Third, fans of BTS have great loyalty to each BTS members. So they keep mentioning on twitter about each member's name and this gives an effect of spreading the name of BTS. Fourth, BTS keep communicating face-to-face with fans through concert and a world tour. Fifth, BTS shows good music and stage that captured the hearts of fans. Sixth, BTS is recognized worldwide that they are the best idol group not only they are good at dancing and singing, but also they have good manners and shows kindness to everyone. Lastly, BTS was awarded at many global music awards, such as in 'Billboard' and 'PCAs' and so on, it proves they are successful globally.

Future Research Plan

In the future research, we will further interpret and validate those main issued identified from the text mining approach. For this purpose, we are going to apply interview-based qualitative approach: fans, agency, and industry experts. We will conduct in-depth interviews with actual BTS fans about the reasons why fans like BTS and why fans keep going on fan activities. As a result of text mining methods shows one of the main factors of success of BTS is that ARMYs' (fan club) dedicated activities and powerful reactions, this interview will be an important part to derive key success factors of BTS. We will also interview the representatives of BTS 's agency and related people to understand the strategy that they have developed as a global idol group. We can then find the strategies how BTS could acknowledge their faces to the world. Lastly, we will interview experts related to the entertainment industry to find out the main factors of success. This interview part could give us another side of opinions of success of BTS. Based on the combination of the text mining analysis results and the interview analysis results, we will propose meta-inferences regarding the social opinions about K-POP (i.e., BTS), which may suggest the critical success factors. This study thus contributes to the literature by proposing the mixed methods approach and identifying the critical success factors for K-POP and digital contents business in general.

References

Blei, D. M. 2012. "Probabilistic topic models," *Communication of the ACM* (55:4), pp. 77-84.

Cho, Byung-Chul and Sim, Hichul. 2013. "Success Factor Analysis of K-POP and A Study on sustainable Korean Wave - Focus on Smart Media based on Realistic Contents," *The Journal of the Korea Contents Association* (13:5), pp. 90-102.

Choung, Jae-Min. 2010. "An Analysis of Youth Fandom Based on the Ground Theory," *Studies on Korean Youth* (21:3), pp. 91-119.

Fu, Kang and Shon, Seunghye. 2018. "A Comparative Analysis of Composition and Activities of Idol Group Fandom among Chinese Fans- Grounded theory analysis of TFBOYS and BTS fan cafe member interviews," *Media, Gender&Culture* (33:2), pp. 5-43.

J Pastizzo, Matrhew and F Carbone, Robert. 2007. "Spoken word frequency counts based on 1.6 million words in American English," *Behavior Research Methods* (39:4), pp. 1025-1028.

Jang, Gyu-Soo. 2009. "Overseas advance instance and strategic study of Korean pop music," *The Journal of Global Contents* (9:6), pp. 217-238.

Jang, Won ho and Song, Jung Eun. 2017. "The Influences of K-pop Fandom on Increasing Cultural Contact: With the Case of Philippine Kpop Convention, Inc.," *Journal of the Korea Entertainment Industry Association* (18:2), pp. 29-56.

Jin, D. Y., and Yoon, K. 2016. "The social mediascape of transnational Korean pop culture: Hallyu 2.0 as spreadable media practice," *New Media & Society* (18:7), pp. 1277–1292.

Lee, S. and Kim, H. 2008. "News Keyword Extraction for Topic Tracking," *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, pp. 554-559.

Leung, Sarah. 2012. "Catching the K-Pop Wave: Globality in the Production, Distribution, and Consumption of South Korean Popular Music," *Senior Capstone Projects*, pp. 149.

M. Habibi and A. Popescu-Belis. 2015. "Keyword Extraction and Clustering for Document Recommendation in Conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (23:4), pp. 746-759.

Oh, Ingyu and Park, Gil-Sung. 2012. "From B2C to B2B: Selling Korean Pop Music in the Age of New Social Media," *KOREA OBSERVER* (43:3), pp. 365-397.

Park, Cheol., Kim, In Kiw., and Jeon, Youn. 2013. "Policy Research: The Development Strategy of K-POP through the Improvement of Digital Music Distribution in Korea," *Korea Business Review* (17:3), pp. 251-265.

Ubonrat, Siriyuvasak and Shin, Hyunjoon. 2007. "Asianizing K-pop: production, consumption and identification patterns among Thai youth," *Inter-Asia Cultural Studies* (8:1), pp. 109-136.

X. Cheng, X. Yan, Y. Lan and J. Guo. 2014. "BTM: Topic Modeling over Short Texts," *IEEE Transactions on Knowledge and Data Engineering* (26:12), pp. 2928-2941.