ACADEMIC ARTICLES

An analysis of YouTube comments on BTS using text mining

◆ **HO KYOUNG KO**

Professor, Ajou University (South Korea)

ABSTRACT

In this study, I applied text mining methods to the analysis of comments to examine people's thoughts regarding BTS. I analysed approximately 100,000 comments left on YouTube videos of or about BTS. As indicated in the results of the TF–IDF analysis, the designation "people / person" is reflective of the foundations of commenters' perceptions of BTS. Through various contents, BTS has shown people their sincerity and cheerful side. People have familiarised themselves with BTS via such videos and thereby have come to perceive the group members as being the same kind of people as the viewers are. Of course, people have focused on different aspects of BTS. As shown in the topic modeling results, some have focused on the chemistry among the members, while others have concentrated on the appeal of their favourite members. However, regardless of their area of focus, people have shared a sense of togetherness with the group and have empathised with the lives of its members. Furthermore, they have been moved by the group's dedication and growth, expressing support for its members' hearts and thoughts. Also, network analysis results indicated that people have felt that BTS deserves their success based on the group's teamwork, character, and performances. People have also recognised that their music is an outgrowth of their dedication and passion.

KEYWORDS

BTS video, YouTube comments, text mining analysis

INTRODUCTION

Why do people love BTS? In the album Map of the Soul, the leader of the seven-member South Korean group, RM, states the group's love for its fans thus: "We would like you to know that our music is our fan letter to you. We are each other's fans and each other's idols" (Park, 2019). Greeting fans and expressing their appreciation for fans' support and love have long been part of BTS's daily routine. According to Lee (2019), their fandom ARMY (Adorable Representative MC for Youth) is unrivaled around the world in its organization, dedication, and pride. Therefore, BTS and ARMY are bound by an exceptionally strong spirit of togetherness (Yonhap News, 2019).

To keep in step with the recent intense interest in BTS, there have been attempts to analyse the factors behind the group's success. Amongst numerous such factors, the foremost is considered to be musical excellence. In the TV programme *Good Insight*, Big Hit co-CEO Bang Shi hyuk, founder of Big Hit and developer of BTS, comments on the quality of his company's contents, saying they aim "for a level of quality that we would not be ashamed of" and "to produce better-quality content than we have done previously" (Lim, 2018). After all, their success is based on their excellent musicality and high standard of performance. Many media, looking for other success factors, reported that their success is due to the use of SNS-based mobile networks (social media). However, Lee (2018) stated that their success cannot be attributed to the use of SNS alone, since many other singers and groups are also using SNS.

According to Lee (2018), another factor that has been mentioned is the mutual bond BTS has formed with fans through horizontal and mutually interactive social networks. Ultimately, BTS may be regarded as a talented group of musicians who have created and harnessed the "BTS-ARMY effect" via effective communication with their fans.

There is a noteworthy moment in the TV programme *Good Insight* (Lim, 2018) when a producer asks a foreign fan how they came to love BTS. Here, the fan responds, "I liked their choreography, and I came to see that they were truly good people. Eventually, that led me to love their music and everything about them." Some say they even fell in love with BTS's charm despite having no interest in popular music. Do many people, then, look for other elements of BTS outside of music? Or is there a basis for making people truly accept BTS's music or communication with fans, attracting people's interests? In any case, these interviews suggest that there may be other factors—besides BTS's music and how the group shares it—that draw people to the group.

What might constitute the context or significance of the aforementioned belief that BTS consists of "truly good people"? How do people think about them? That is, what are people's perceptions of them? To understand this, we must first survey people's perceptions of BTS. As such, in this study, I applied text mining methods to the analysis of internet comments to examine people's thoughts regarding BTS. Text mining enables researchers to use a large dataset to discover information that is of interest amongst users not only at the word level but also at the context level (Grimmer & Stewart, 2013). Moreover, Internet comment analysis has recently been used to examine people's perception of an event or object (Chumwatana, 2018). Therefore, the analysis of large numbers of comments can be used to assess people's thoughts or emotions in a meaningful way about the subject (Pudaruth, Moheeputh, Permessur, & Chamroo, 2018).

By integrating people's views on BTS and examining the group's contexts via internet comment analysis, we may understand how the group is perceived among people. Accordingly, this study aims to investigate people's perceptions towards BTS by analyzing comments on various BTS-related videos through text mining, one of the big data analysis methods.

In fact, while good looks, excellent performances, and well-made songs surely are factors that draw people to BTS, some have remarked that at the core of BTS's appeal is their message—and the empathy and healing that are thus effected (Jeong, 2019). In this study, I aimed

to explore the fundamental force that has driven people to empathize with and be healed by the words and messages of BTS, such that they were led to accept their music in earnest.

The purpose of the present study is therefore to provide statistical and quantitative illustration of ideas and reasons why ARMY loves BTS, and I will provide evidence and support from within the fandom for this reasoning.

RESEARCH METHODS

SUBJECTS OF ANALYSIS

DATA COLLECTION

The data analysed in this study are comments left on BTS's YouTube videos. Video clips were selected by searching for "BTS," filtered to include only those videos that were at least 20 minutes long and had a high number of views. These comments were collected using a crawler coded with Python. Data collection took place from January 10–15, 2019, during which time more than approximately 300,000 comments were crawled from 15 YouTube videos. While some of the YouTube videos were produced by Big Hit, such as "2018 BTSFESTA" and *Burn the Stage*, some of the videos were edited by fans.

Comments collected by crawling included those that are not written in letters (e.g., that were written using special characters and emotional characters) or syntactically meaningful sentences (e.g., a flock of words); these are excluded from the analysis. Through this process I analyzed approximately 20,000 comments written in Korean. The reason I excluded comments composed only of words (not written in sentences) is that while these comments can be used as data for frequency analysis, they cannot be applied to finding the relationships among words when used in network analysis within a sentence.

DATA CLEANING

Data cleaning was conducted using the natural language processing (NLP) method. Using the R-Studio software, this was done in

two stages—pre-processing of text data and morpheme analysis. In the pre-processing stage, words and phrases that are unnecessary for analysis were deleted from the collected text data. Stop words, such as various special symbols and punctuations included in text to which a meaning does not need to be given during the analysis, were eliminated. Then, normalization, which unifies the words that are similar to the words that appear incorrectly but are differently represented, had to be carried out repeatedly. The next step, morpheme analysis converts text into natural language for information extraction. In this study, "KoNLP," which is a Korean-language processing software package, as well as the "tm" and "stringr" packages, were used. By conducting part-of-speech analysis for pre-processed data, common nouns are extracted. Then there were words that needed to be filtered further after extracting the nouns. To get good analysis results, I repeated the normalization process along with the removal of stop words. In addition, since single-letter words appearing in noun extraction are often meaningless, they are removed and only words with two or more letters are extracted.

ANALYSIS METHOD

In this study I employed text mining techniques to extract high-frequency words from amongst the comments on videos related to BTS and to identify the contexts of those words. Text mining is based on the "statistical semantic hypothesis," which states that it is possible to extract meaningful information from text data. This hypothesis posits that "the intentions of people can be derived based on the statistical patterns of word usage in their utterances or writings" (Turney & Pantel, 2010). That is, what people write, which words they use, reveals meaningful thought patterns.

Topic modeling, which is one of the techniques of text mining, can analyze partial topics in the entire set through the linkage relationship between them by paying attention to the coexistence or the sharing of independent topics in closely connected studies. It is an efficient analysis method for identifying trends

over time for meaningful data (Mishne, 2007). The three characteristics of the text mining technique are:

1. objective analysis is possible, as cognitive ability is not used;

2. by visualizing the results of the analysis, perspectives from the study can be quickly and easily communicated; and

3. derived results can be illuminated from multiple angles depending on a researcher's point of view (Wang & Blei, 2011; Wei, Chang, Zhou, & Bao, 2015; Wong, Liu, & Mennamoun, 2008).

By using text mining techniques, complex and diverse language is processed as structured data and statistically analyzed, and accordingly, meaningful information about existing trends can be reorganized.

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY: TF–IDF

Term Frequency (TF), which is one of the most commonly used text mining techniques, indicates how many times a particular word appears in documents, and the larger its TF value, the more important the word in the document. However, if it appears frequently in all documents, it may mean that it is a commonly used word rather than being contextually important (Yook, 2017). For instance, "Bangtan" and "BTS," which appear a great deal in this study, become universal words that indicate a subject people want to discuss in the comments.

In this study, to filter out universal words occurring in all documents, I did not apply simple term frequency (TF) mining but the term frequency–inverse document frequency (TF–IDF) method. This method re-processes the frequency

$$TF - IDF = TF \times \log(\frac{N}{DF})$$

TF : frequency of a certain term within a comment

IDF : $\log(\frac{N}{DF})$

ID : number of comments that contain a certain term

N : total number of comments

of word occurrence based not on simple frequency but on the probability of frequency (Ramos, 2003). In this study, the TF–IDF values were calculated using the following equation:

TOPIC MODELING

Topic modeling is a useful method for topic analysis in that it statistically analyses the frequency of words within text data, thereby automatically extracting and categorising "topics"—or latent themes—from the total data (Teh, Jordan, Beal, & Blei, 2012). Topic modeling is a statistical text processing technique based on an algorithm for extracting certain topics from large volumes of textual data, wherein a matrix of documents and words is used to extract inference on the occurrence probabilities of latent topics within documents (Griffiths & Steyvers, 2004). In this study, we employed the Latent Dirichlet Allocation (LDA) model, presented by Blei et al. (2003), because, compared to other methods, LDA allows for easier interpretation of results (Blei, 2012) and addresses issues of over-fitting, thereby making it better suited for deriving various topics from large unstructured data (Griffiths & Steyvers, 2004).

In conducting LDA analysis, it is necessary to set the total number of iterations and suitable number of topics. In this study, the total number of iterations was set at 1,000, and pilot LDA analyses were conducted for a varying number of topics, from 5 to 12.

When setting a number of topics, there is a method for estimating topics by non-parametric statistics in an optimized probability model (Teh et al., 2012); a method of producing the final result by combining similar topics after creating many topics and conducting LDA (Song, 2010; Yu, 2014); and a method in which a topic analysis is performed in consideration of the number of cases, which yields a result whose term classification is meaningful or highly accurate (Griffiths & Steyvers, 2004).

KEYWORD NETWORK ANALYSIS

For network analysis, we first derived the characteristic vectors of each of the words in the comments and conducted correlation analysis

on those words. We then conducted network analysis using the derived correlation vectors. Also, a directed graph was generated by deriving conditionally independent and conditionally dependent nodes.

ANALYSIS RESULTS

TF–IDF ANALYSIS RESULTS

I extracted 100 words that were found to be meaningful in the full set of comments by using TF–IDF analysis. Here, we have presented only the top 30 with the highest priority in Table 1. Looking at the meanings of some of the words that were extracted, the word that had the most important effect was "video," followed by "people (person)," in terms of frequency. This was due to the fact that commenters referred to BTS as "people." Also, terms that reflected a sense of togetherness, such as "ARMY" and "us," followed in terms of frequency. The fact that BTS videos had a high occurrence of the affective terms "tears" and "moving" is also very significant.

Furthermore, commenters referred to individual BTS musicians as "members," meaning that they perceived each to be a part of the whole team. The next words that occurred frequently, such as "heart," "support," and "thoughts," reflected peoples' perceptions of BTS.

TOPIC MODELING ANALYSIS RESULTS

Based on this, an inter-topic distance map (IDM), which visualizes how the topics are related, was used to determine the final number of topics to equal five. If the number of topics was more than six, they were largely divided into three groups. Then, two groups contained one topic each and the remaining topic was included in the other group. When the number of topics was five, they were classified into four groups. Two topics appeared in only one group, and accordingly, that group was classified as the most significant. Based on this, the number of topics was set to five and potential topics for comment topic classification were extracted (*Figure* 1).

After extracting topics, words that make up each topic were compared and analyzed. Then, the core theme of the words that make up the topic was deduced by comprehensively considering the context of the comments. Lastly, the relationship between topics was effectively verified by visualizing the topic modeling result through IDM. For the words that made up the topic used in this study, words extracted through term frequency–inverse document frequency (TF–IDF) analysis were used.

After 5 rounds of topic modeling on the comments were collected, I was able to categorize the following five topics. The associated keywords and characteristics of each topic follow.

Word	TF–IDF	Word	TF–IDF	Word	TF–IDF
Video	522.73	Begging	110.35	Flowery road	75.70
People (person)	343.94	Thoughts	104.18	Each other	75.07
ARMY	303.15	Songs	103.89	Eldest brother	73.71
Us	261.21	Leader	102.88	Heartfelt	72.09
Tears	231.54	Purple	102.12	Love	71.41
Moving	198.73	Idol	97.00	Face	71.30
Image	173.52	Growth	90.54	Stage	70.70
Member	167.11	Best	88.71	Appeal	64.64
Heart	132.12	Fighting	87.32	Youngest brother	59.85
Support	125.46	Practice	77.96	Becoming a fan	57.27

TABLE 1. TOP 30 WORDS AND PHRASES EXTRACTED BY TF–IDF ANALYSIS

TOPIC 1: THE ARMY AND OUR SINGERS (UNITIZATION / IDENTIFICATION AS THE SAME GROUP)

The comments that correspond to the first topic fall under the context where fans have identified themselves with the artists, such that they see themselves unitised as the same group. Examples of the comments include:

"He is a leader of BTS as well as RM of the ARMY."

"May we look back and laugh at our times in the distant future . . . This toast gives us all a warm rest."

In other words, the most significant characteristic of Topic 1 is that people think of BTS by putting them in the category of "us."

TOPIC 2: MOVING AND FUN, SONGS AND LYRICS

The comments that correspond to the second topic fall under the context where fans are moved and derive entertainment from images of BTS, leading them to further empathise with, and be attracted to, the group's songs and lyrics.

"It was fun at first, and I laughed a lot, but later I do not know why I had tears in my eyes."

"Bangtan moved me and I laughed. I like how the song and the lyrics came out."

That is, in Topic 2, the most notable characteristic is that people are impressed with BTS, which makes fans or commentors accept BTS's songs more sincerely.

TOPIC 3: GROWTH AND THE STAGE, DEDICATION.

These comments fall under the context where fans have closely followed the group's growth trajectory since debut, have been moved by the stage performances made possible through the group's efforts and growth, and have not only continued their support for the group, but have redoubled that support over time.
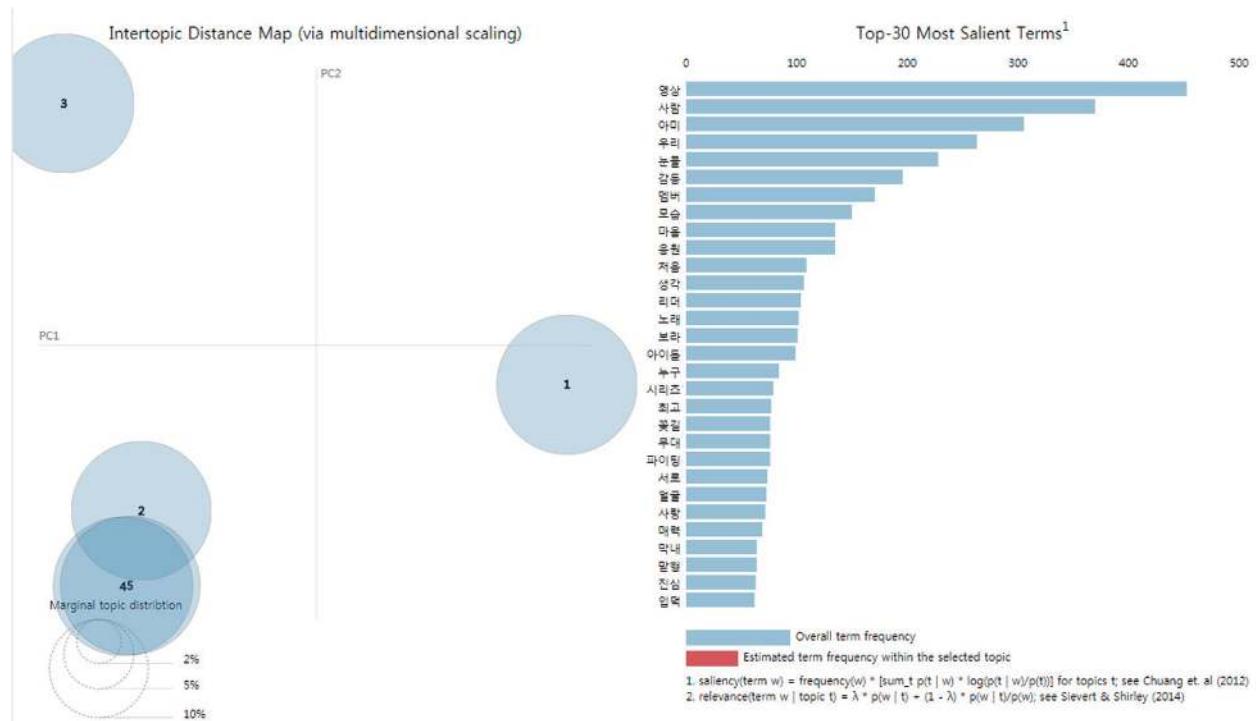

FIGURE 1. IDM FOR THE ENTIRE TOPIC

"I gave it a lot of thought. As the seven individuals who had lived their own lives are united in the name of Bangtan and always walk together, conflicts large and small can arise at any time. It is also a series of steps towards an outcome. Seeing that they go through the process well, I am happy that these people are my favorite people."

"There were a lot of trial and error, a lot of pain, and a lot of enemies. However, they have overcome all of these, and come to the stage where people can truly see them. They are stuck in the studio, working like crazy. They are full of passion, and as they mentioned in the interview, one of the three words that characterizes them best is work."

That is, the noteworthy characteristic of Topic 3 is that people are attracted by BTS's ability to overcome difficulties, grow, and show their enthusiasm.

TOPIC 4: CHEMISTRY BETWEEN MEMBERS, COMMUNAL LIFE

These comments focus on the chemistry between BTS members and their communal way of life, onto which fans project an ideal model of human society. Fans support the community within the group and, sometimes, reflect or project it onto the society they themselves belong to.

"Being able to value and care for each other and give happiness are great achievements by themselves."

"How good would it be if we could always be together like this?"

The most significant characteristic of Topic 4 is that fans agree with and respond to BTS's way of thinking.

TOPIC 5: THE APPEAL OF THE FAVOURITE MEMBER.

These comments focus on the individual appeal of the artists in the various contents produced by BTS and are centred on the favourite member of each fan.

"Taehyung is the most handsome man, almost like he is not human. I am curious of Taehyung's mother and father."

"When I see Jungkook, I just smile like a mother. It feels very healing."

"I like Jimin — he has a good heart, he's warm, bright, thoughtful, and he's always working hard."

"He is one of the handsomest men in the world, nice, cute, he's good at cooking and singing, he's humorous and witty. Where can you find this 27-year-old? He is my ultimate preference."

The last topic, Topic 5, shows fans' appreciation for BTS's appearance and style.

This five-topic analysis shows that people are paying attention to BTS's various aspects. Further, the commonality of the topics reveals that rather than simply liking and paying attention to BTS just because their songs, looks, and shows are excellent quality, people feel empathetic and supportive toward them as human beings, not just as entertainers.
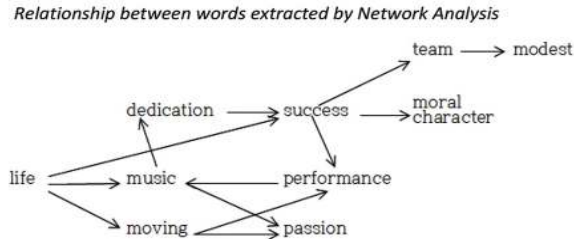
KEYWORD NETWORK ANALYSIS

In addition to using topic analysis to find out what people are paying attention to or are interested in, we can analyse the pattern of the words that are connected to explore the sense in which the words were working. Figure 3 displays the result of network analysis conducted using the correlation vectors as inputs.

Network analysis results indicated that some of the keywords representing BTS were "success,"

"life," "dedication," and "moral character." In particular, "success," "music," and "moving" were derived from "life," suggesting that BTS's music reflected their unique views on life and that the perceptions of fans converted the lives of BTS members into images of success and being emotionally connected.

Furthermore, words such as "performance," "music," "dedication," and "success" were interconnected via a strong network. Thus, in peoples' perceptions of BTS, their performances, music, dedication, and success were considered to be strongly associated with each other under a single category. Fans also believed that the music of BTS was imbued with the members' lives and performances, and that their music embodied their dedication and passion. As BTS are musicians, many elements surrounding music are connected by the network. However, what is special is that people think of their music in connection to "life" and "performance." Further, people believe their music represents their effort and passion.

FIGURE 2. RELATIONSHIP BETWEEN WORDS EXTRACTED BY NETWORK ANALYSIS

Lastly, the final several words representing them are "character," "teamwork," and "modesty." That is, the most basic words in the network of words are "moral character" and "modesty." Through the presence of these words, it can be interpreted that people's description of BTS— their perception of BTS's image—is based on BTS's character and modesty.

CONCLUSION

Recently, there have been various attempts to analyse the factors behind BTS's success (e.g., Kim, 2017; Kim, 2018). Amongst the factors analysed are how the members communicate with fans and how the group's content is produced and distributed in addition to various other factors, such as musical characteristics. However, the focus of this study has been on peoples' perceptions of BTS rather than BTS's success.

BTS members have stated that they have communicated and empathised through their content. Since BTS would hardly be the only group to create content and communicate with the public, it is difficult to attribute the extent of empathy the group has achieved solely to their content creation or use of social networks for communication. That is, all groups are involved in creating content and communicating with the public, but few achieve BTS's success. The more important factors to be considered are the aspects of BTS that people focus on and which of these elicit affective support amongst fans.

As indicated in the results of the TF–IDF analysis, the designation "people / person" is reflective of the foundations of commenters' perceptions of BTS. Through various contents, BTS has shown people their sincerity and cheerful side. People have familiarised themselves with BTS via a variety of videos, and thereby have come to perceive the group members as being the same kind of ordinary people as themselves. Of course, people have focused on different aspects of BTS. As shown in the topic modeling results, some have focused on the chemistry among the members, while others have concentrated on the appeal of their favourite members. However, regardless of their area of focus, people have shared a sense of togetherness with the group and have paid attention to the state of the lives of its members. Furthermore, people have been moved by the group's dedication and growth, expressing support for its members' hearts and thoughts.

According to Lee (2019), BTS has drawn people's empathy and ARMY have formed a strong rapport with them. From the comment analysis, the specific word "empathy" did not appear very often. However, regarding the characteristics of BTS, or the circumstances and events they

faced, people feel connected with the group by words of emotional support. Thus, although the word "empathy" is not high in frequency, one can assume that people empathize with BTS based on the words that express the underlying and nearly identical emotional response at many events, like "touching," "heart," "feeling," "consolation," and "comfort." Also, based on the group's teamwork, character, and performances, people have felt that BTS deserves their success. They have also recognised that the group's music is an outgrowth of its dedication and passion. Ultimately, such perceptions imbue a human vitality into the beautiful music and great performances of BTS, thereby taking on an organic synergy as well as taking the world by storm.

CONFLICT OF INTEREST

The creator has no relevant conflicts of interest to disclose.

AFFILIATIONS

Graduate School of Education, Ajou University

REFERENCES

Chumwatana, T. (2018). Comment analysis for product and service satisfaction from Thai customers' review in social network. *Journal of Information and Communication Technology*, 17 (2), 271-289. doi:10.32890/jict2018.17.2.8254

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 1-31. doi:10.1093/pan/mps028

Jeong, H. (2019, January 23). *"Being to comfort my life" BTS from ARMY*. Sports Seoul. http://www.sportsseoul.com/news/read/722316

Kim, N. K. (2018). BTS *Insight Well done and sincere: Z-generation management strategy learned from BTS*. Secret book.

Kim, S. (2017). *This is Bangtan DNA*. Seoul: Dokseogwang.

Kim, Y. (2019). BTS: *the review*. Seoul: RHK.

Lee, J. (2019). *BTS and ARMY culture*. Seoul: CommunicationBooks

Lee, J. (2018). BTS, *art revolution*. Seoul: Parrhesia.

Lim, K. (2018, February 23). The future of BTS and K-pop (Season 2, Episode 67) [TV series episode]. In *Good Insight*. KBS. http://vod.kbs.co.kr/index.html?source=episode&sname=vod&stype=vod&program_code=T2014-0867&program_id=PS-2017122081-01-000&section_code=05&broadcast_complete_yn=Y&local_station_code=00. 2018.09.15

Mishne, G. A. (2007). *Applied text analytics for blogs* [Unpublished doctoral dissertation]. Universiteitvan Amsterdam

Park, H. (2019, September 16). *Our music is a fan letter to you*. Sports Seoul. http://www.sportsseoul.com/news/read/720836

Pudaruth, S., Moheeputh, S., Permessur, N., & Chamroo, A. (2018). Sentiment analysis from Facebook comments using automatic coding in NVivo 11. ADCAIJ : *Advances in Distributed Computing and Artificial Intelligence Journal*, 7(1), 41-48. doi:10.14201/adcaij2018714148

Ramos, J. (2003). Using TF−IDF to determine word

relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, ICML 2003.

Song, Z. (2010). *Research on text categorization based on LDA*. [Unpublished master's degree dissertation]. Xi'an University of Technology, Xi'an, China.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141-188. doi:10.1613/jair.2934

Yee W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:476), 1566-1581. doi: 10.1198/016214506000000302

Wang, C. & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 448-456, ACM. doi:10.1145/2020408.2020480

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4), 2264-2275. doi:10.1016/j.eswa.2014.10.023

Wong, W., Liu, W., & Mennamoun, M. (2008). Determination of unithood and termhood for term recognition. In M. Song & Y. Wu. (eds) *Handbook of Research on Text and Web Mining Technologies*, IGI Global. doi:10.4018/978-1-59904-990-8.ch030

Yonhap News Agency (2019, September 15). *Why are they enthusiastic for BTS*. https://www.yna.co.kr/view/MYH20190828014500704

Yook, D. (2017). Text mining-based analysis for research trends in vocational studies. *Journal of Korea Academia-Industrial cooperation Society*, 18(3), 586-599. doi.org/10.5762/KAIS.2017.18.3.586

Yu, S. Y. (2014). Exploratory study of developing a synchronization-based approach for multi-step discovery of knowledge structures. *Journal of Information Science Theory and Practice*, 2(2), 16-32. doi:10.1633/jistap.2014.2.2.2

SUGGESTED CITATION

APA CITATION

MLA CITATION