Using Text Mining in Film Industry through Text Analytics on Reviews of the Movie Kingsman Series

Ji-Heon Song, Sung-Jun Kim

***Abstract*: The movie reviews are important to success in film industry. So, the objectives for text analytics are reviews and scores of the movies 'Kingsman: The Secret Agent' and 'Kingsman: The Golden Circle' (It is referred to as the following Kingsman series 1 and series2).**

These data were collected by crawling on the movie reviews page of Korea's largest portal site. Morphological analysis and Frequency analysis were used as an analysis method. Crawling and analysis were used in Python, an open source program.

Many studies using movie reviews focus on using complex and difficult analysis techniques or algorithms. However, this study confirms the important and objective reaction factors to movie viewers through frequency analysis using morphological analysis which is the most basic of text analytics. It shows that the viewer's response to the next series of films can be predicted. Therefore, even the basic analysis method can be useful for the film industry if it is appropriately used.

A more effective method can be found by comparing the results of this study with other methods of analysis. And, it is the basis for supporting the effectiveness of this study.

***Index Terms*: Bigdata, Movie review, Natural Language Processing(NLP), Text analytics**

I. INTRODUCTION

In the current society, unstructured data are exploding. These are the data we most easily access, and they are also diverse. Among them, text data are the most easily accessible data. There is a technique called text mining that analyzes it. In addition, the space where these texts show the strongest power is SNS and Internet, they affect many people acquiring information. One of them is about the movie. Viewers are more influenced by the subjective information of many people obtained through movie reviews than objective movie information before watching movies, it affects movie success and failure. Already there are studies that movie review affects sales, and various studies are being conducted [1]. However, most of these studies focus on how refined the text mining should be, and there is no research on how to use text mining as much as possible. I saw the news of the movie Kingsman Series 3 which will be released this winter. Therefore, this study examined the extent to which movie reviews could be utilized in the film industry, such as the predictions of the next series, through the analytics result of the most basic text mining method.

In section 2, we discussed previous studies on the film industry and text mining. In section 3, the methods used in this study are mentioned. For that, this study collected movie reviews and scores of Korean portal sites and analyzed them using Python. First, I analyzed the overall review and checked the reviews with less than 5 out of 10 star scores. because, this movie is a very popular movie in Korea, and the scores are very highly focused. Also, it is difficult to make emotional dictionaries personally for emotional analysis. However, since data on this is too small to be suitable for analyzing meaningful results. So, in this study, text mining was implemented and analyzed focusing on the whole review. In section 4, detailed and diverse analysis results were derived. In section 5, we rearrange the effectiveness of appropriately used text mining and conclude how these results can be used in the film industry.

II. RELATED WORKS

Through the studies on the film industry in Korea, I confirmed that the internet movie review of viewers has an important influence on the movie performance. Text mining will be studied about the definition, role, and the existence of latest studies through research on text mining. Through these, we reaffirm the necessity of this study. To do this, we define a practical analysis method for text mining for movie reviews.

A. *Film Industry*

The study of the importance of movie review to the film industry or its relationship with the box office has been preceded. In the case of movie evaluation through the traditional method, the research on the satisfaction of the movie selected through the portal site higher than the movie magazine indicates the importance of the portal site [2]. In addition, the amount and direction of reviews of these portal sites have been shown to have a greater impact on the box office [3]. In the study of general audience and critic's choice of movie, in case of general viewers, they give high rates when the movie is not Korean and have famous leading actor [4]. In addition, the major factors in the selection of the screening films in the Korean Movie Consumer Survey are contents / plot (86.9%), genre (77.4%),

Revised Manuscript Received on May 22, 2019.

Ji-Heon Song, Big Data Industry Security, NamSeoul University Graduate School, Cheonan, Korea, wlgjs218@gmail.com.

Sung-Jun Kim, Big Data Industry Security, NamSeoul University Graduate School, Cheonan, Korea, jun1977@nsu.ac.kr, +82-10-4830-0007.

actors (69.3%), and the Internet (59.6%) is the highest source of reference information, and SNS is showing an upward trend [5]. In a study on the relationship between movie score and box office, the higher the movie score, the more successful the movie. Then, commerciality, non-Korean nationality movies, major distribution companies, and action / SF movies as genres were identified as success factors [6].

B. *Text Mining*

In the case of text mining, many studies have already discussed concepts and how they are used.

> Text mining is a technique and method for extracting valuable information from text data, which are unstructured data, and deriving insights. Unstructured data refer to data that are not measurable by number, which is exploding with the use of various media on the Internet. Among these, text data are data that can be easily accessed most frequently in various places. Roughly, text mining is also used as a concept of text analytics [7].

From the past, the explosive increase in unstructured data have been predicted. According to IDC forecasts, the unstructured data of 5 zettabytes in 2014 will increase to 40 zettabytes by 2020, and the methods for dealing with them will also become more important [8]. Natural Language Processing (NLP), one of the most important of these, is a technology that can help computers understand text as human beings understand it. In other words, the role of NLP is to organize the text information in a computerized manner and send it as input for analysis [9]. Studies using these text mining techniques are also actively underway in various fields. These are researches that a study to recognize the situation by web keywords scraping and grouping the problems of hygiene in the cruise industry [10], a study on the prediction of movie performance through emotional analysis of movie review data [11], a study to predict the entertainment elements for success through the sentimental analysis on the review in the Media [12], a study using sentimental analysis in reviews through comparison with competitors in ramen market [13], a study on the relationship between food and climate change using text frequency analysis [14], research to resolve the complaints of fashion companies by analyzing text elements of portal site comments [15].

III. Research Method

Where to gather data for text mining is also a very important part. In this study, movie reviews as data are scrapped from the movie review page of the largest portal site in Korea, which is the most popular Korean information gathering, with Python, an open source program, and conducted text analytics.

A. *Analysis Object*

The target of the analysis is review and score of the movie review page of the Korea's biggest portal site. It was selected as the most used website in Korea by the latest statistics of SimilarWeb. This portal site has a share of at least 65% in Korea although the specific figures vary from one statistic to another. Thus, this is the most suitable object to know the reaction of Koreans. On this page, a total of 39,000 data including 20,000 reviews and scores of Kingsman Series 1 and 19,000 reviews and scores of Series 2, were selected for analysis.

B. *Analysis Tool*

Before using tool, learn the NLP's most important and early ongoing Lexical Analysis procedures. This consists of large steps of a sentence splitting in the corpus, tokenization, which is the smallest unit that has meaning to analyze a document or sentence, morphological analysis (text normalization) that analyzes words in their general form, including lemmatization that changes to the basic form in which token's part of speech information is preserved, and stemming that changes the token into short form, and Part-Of-Speech-Tagging, which divides categories of tokens into like verbs and nouns [16].

Movie reviews and scores as analysis object were scraped with a target crawling using Python's beautiful soup library.

Frequency analysis was used as a method of text mining, using KoNLPy from Python. Korean-based force tags such as KoNLPy can be obtained by easily utilizing relatively high quality. Since sentence splitting, tokenize, lemmatization, and force tagging are all done at the same time to perform basic lexical analysis of text mining.

C. *Analysis Method*

The analysis process consisted of basic text mining procedures such as data extraction, extracted data preprocessing using konlp and stop words, information extraction, and interpretation [17]. In this study, we first extract the data by web scraping using a movie review site as a target. Second, the extracted data were used in Kind Korean Morpheme Analyzer (kkma) class of KoNLPy package. Kkma is a morphological analyzer and natural language processing system written in Java, developed by the Intelligent Data Systems (IDS) Laboratory at Seoul National University (SNU) [18]. After separating the sentences, nouns were extracted as preprocessing steps include tokenize, stop word removal, and part-of-speech (POS) tagging. Third, the Python code in Figure1, shows process of morphological analysis and frequency analysis. Frequency analysis was performed on nouns that were pre-processed and finally extracted. Frequency analysis was performed by adding +1 every time nouns appeared. Fourth, the frequency analysis was visualized as a word cloud, and the top 30 nouns were extracted for various insights. Finally, the final analysis was based on related works.

```
word_frequency = {}
noun_list = []

stop_list = [ "check", "now", "going", "moving",
"ing",
                "hello", "I", "tomorrow", "movie",
"score", "time",
              "been a while", "exactly",
           “completely” ]
line_number = 0
for line in kingc_lines[:20000]:

   line_number += 1
   print(str(line_number) + "/" +
str(len(kingc_lines)), end="\r")
   sentences = kkma.sentences(line)
   for sentence in sentences:
      nouns = kkma.nouns(sentence)
      for n in nouns:
         if n not in stop_list and len(n) > 1:
            if n == "Colin":
               n = "Colin Firth"
            elif n == "Firth":
               n = "Colin Firth"
            noun_list.append(n)
            if n not in word_frequency.keys():
               word_frequency[n] = 1
            else:
               word_frequency[n] += 1
```

Figure 1: Python Code for Morphological, Frequency analysis

IV. Results and Discussion

Word cloud shows the result of visualization of frequency analysis of movie Kingsman 1 and 2. The most significant feature of the word cloud is that words with many frequencies of analysis result are expressed in large letters. This method is more effective when the gap between a word with a low frequency and a word with a high frequency is large. According to Word Cloud in Kingsman 1, there are words ‘Action’ and ‘Great’ that stand out. According to word cloud on Kingsman 2, the most obvious word is ‘Action’ and ‘Series 1’, and the ‘Expectation’ is next big. However, the words that can be distinguished from the current visualization data are too few to analyze, and cannot be assured that only these are important words. That is, the frequency of words is similar in order to notice the detailed relationship between words.

Therefore, the results of the frequency analysis were ranked in order of frequency, and the top 30 words were printed at once. These are translated into English and replaced with the following Table 1. When the top 10 words of Kingsman1 (left) and Kingman2 (right) are compared, they mostly overlap with 7 out of 10, which are ‘Action’, ‘Colin Firth’, ‘Cruelty’, ‘Fun’, ‘Expectation’, ‘Scene’, and ‘Think’. Among them, words that are considered important can be narrowed down to Action, Colin Firth, Cruelty, and Fun. It was based on the main elements of a hit movie studied in Related Works.

According to preliminary studies, the main elements of the hit movies are content / plot, genre, actor foreign film. Among them, labeling the main words of the movie Kingsman series, Action can be classified as Genre, Colin Firth as leading actor, Cruel as characteristic word, Fun as positive word. Since the movie Kingman has information that it is a well-hit movie, it can be said that these factors are the characteristics of the movie and the reason it was popular. There was concern and interest in the famous director Matthew Vaughn through other top keywords in Kingsman1 of Table1. In addition, the great feature of Kingsman1 was the use of new words or abbreviations such as ‘Dae-bak’, ‘Byeong-mat’, and ‘Kang-chu’. These words are the words most commonly used by the 90's generation in Korea. Among the words, the word ‘Dae-bak' appeared in the Korean word dictionary of the National Korean Language Institute in 2002 and was designated as a standard word in 2008 [19]. That is, it can be seen that this movie, which has similar characteristics to the generations that pursue 'simple' and 'fun' features of the Korean 90's generation, has been effectively appealed as B movie [20]. Therefore, in addition to the fact that the film ratings in Korea is set to 19, the review of this movie disproves that the young generation audience of the 90 '. It can be used to extract the target age range from the review text. Through the other top keywords that ‘Series1’, ‘Expectation’, and ‘Previous Series’ of Kingsman2, it was found that the movie audience had a lot of expectation for the first movie when viewing the second movie. With this information, you can say that the audience of movie Kingsman3 will have interest and expectation on the movie that has already been watched. It can also be expected that the movie will be a hit if it keeps the concept of the same leading actor and fun and cruel.

Table 1: Result of Frequency analysis about movie Kingsman1 (left), Kingsman2 (right)

N	NOUN	COUNT	N	NOUN	COUNT
1	Action	2507	1	Series 1	2818
2	Colin Firth	2354	2	Action	2400
3	Cruelty	1379	3	Expectation	1619
4	Great	1364	4	Fun	1047
5	Expectation	939	5	Cruelty	943
6	Scene	862	6	Plot	763
7	Fun	843	7	Think	728
8	Man	827	8	Previous Series	723
9	Think	752	9	Colin Firth	689

10	Director	626	10	Scene	654
11	Suit	602	11	Great	629
12	Plot	579	12	Series 2	625
13	Manner	575	13	As much as	565
14	Dae-bak (a big hit)	443	14	Feeling	547
15	Spy	423	15	Previous Movie	542
16	The last	401	16	Man	519
17	Really	382	17	Kingsman	475
18	Byeong-mat (insane, using in B contents)	381	18	Disappointment	460
19	Head	368	19	Worth Seeing	380
20	Stranger	353	20	Time	369
21	Comic	345	21	Series 3	355
22	Feeling	339	22	Like	335
23	Kang-chu (Strongly recommend)	338	23	Story	334
24	Recommend	334	24	Harry	310
25	Action Movie	327	25	Part	300
26	First	320	26	Season	285
27	10	313	27	Entertainment	283
28	Bolman (Worth seeing)	310	28	Personal	264
29	People	306	29	Recommend	258
30	Time	302	30	Manner	256

V. CONCLUSION

Although it is a simple frequency analysis of movie reviews, the information available within it varies according to the analyst. In most review analysis, only the presence or absence of the word itself is judged to confirm the preference, or the review of various movies is analyzed at once to predict the successful elements. However, in this text analytics, we focus on one movie with a series, analyzing the characteristics of the movie, and what are the successful factors of the movie, which combines existing research.

As a result, basically, the films in this series appeared to be the main feature with the action genre, starring main actor Colin Firth, and interesting and brutal story. In addition, the age range of the consumers who actually leave a review through the words that were used mainly were found. So, this could be applied to catching the target age group of the movie. In other words, the Kingsman Series 3 to be released next is expected to be enough for existing audiences if they have the same interests, concepts and actors like existing movies, as in section 3, because the expectation of existing audiences is sufficient. It means, this information reveals which elements to focus on for success of the next series of movies Kigsman3 to be released. It is not only feedback on the film, but also If they are studied before the production of the actual film, the film makers can be provided good guidelines for the production direction. Although it is a simple text mining, there is a limit to other interpretations depending on what background and knowledge analyst is, but this is a part that can be overcome through collaboration of various experts. This text mining approach through reviews is expected to be more effective if applied to a variety of other industries or products.

REFERENCES

1. Y. Liu. (2006, January). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*. [Online]. *70(3)*. pp. 74-89. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1949819
2. E. A. Jeon, "A Study on the Correlation between Film Reviews and Audience Choices: With an Emphasis of Comparing Film Reviews of Portal Sites and Film Magazines," unpublished.
3. H. K. Kim, "The Impact of Time for Writing Movie Reviews on the Box Office Success: Focusing on the Volume and Valence of the Reviews," unpublished.
4. E. H. Oh, B. S. Chon. (2008, November). Determinants of Film Critics between Online Users and Professional Reviewers, *Korean Journal of Broadcasting and Telecommunication Studies*. [Online]. *22(6)*, pp. 267-289. Available: http://www.dbpia.co.kr/Article/NODE01094608
5. Korean Film Council. (2017, April). 2016 Theater Movie Consumer Survey [Online]. Available: http://www.kofic.kr/kofic/business/rsch/findPolicyDetail.
6. G. I. Koo, "A Study on Correlation between Film Rating and Film Performance - Focused on Portal Site Naver and Movie Journal Cine 21," unpublished.
7. L. Kaushik. (2013). Text Mining - Scope and Applications, *Journal of Computer Science and Applications*. [Online]. *5(2)*. pp. 51-55. Available: http://www.irphouse.com/
8. E. Simoudis. (2016, February 2). Insightful applications: The next inflection in big data [Online]. Available: https://www.oreilly.com/ideas/insightful-applications-the-next-inflection-in-big-data
9. L. Kumar, P. K. Bhatia. (2013, March). TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS. *Journal of Global Research in Computer Science. 4(3)*, pp. 36-39. Available: www.jgrcs.info
10. T. Shuting, B. Kang, H. S. Kim. (2018). Understanding the Food Hygiene of Cruise through the Big Data Analytics using the Web Crawling and Text Mining. *Culinary Science & Hospitality Research*. [Online]. *24(2)*, pp. 34-43. Available: 10.20878/cshr.2018.24.2.005
11. S. Lee, J. Cho, C. Kang, S. Choi. (2015, November). Study on prediction for a film success using text mining. *Journal of the Korean Data & Information Science Society*. [Online]. *26(6)*, pp. 1259-1269. Available: https://doi.org/10.7465/jkdi.2015.26.6.1259
12. A.J. Oh, S. H. Ahn, J. M. Byun. (2016, April). A Big Data Study on Viewers' Response and Success Factors in the D2C Era Focused on tvN's Web-real Variety 'SinSeoYuGi' and Naver TV Cast Programming. *International Journal of Advanced Culture Technology*. *4(2)*, pp. 7-18. Available: 10.17703/IJACT.2016.4.2.7
13. Y. S. Kim, S. R. Jeong. (2018, February). Competitive intelligence in Korean Ramen Market using Text Mining and Sentiment Analysis. *Journal of Internet Computing and Services. 19(1)*, pp. 155-166. Available: http://dx.doi.org/10.7472/jksii.2018.19.1.155
14. K. Y. Bae, J. H. Park, J. S. Kim, Y. S. Lee. (2013, November). Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm. *Journal of the Korean Data & Information Science Society*. 24(6), pp. 1429-1437. Available: 10.7465/jkdi.2013.24.6.1429
15. H. S. Oh, S. K. Cho, C. W. Kang, D. S. Lim. (2010, February). Fashion Company's Claim Data Analysis Using Text Mining. *Journal of the Korean Data Analysis Society*. *12(1)*, pp. 297-305. Available:

https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001422865
16. H. Wachsmuth. *Text Analysis Pipelines - Towards Ad-hoc Large-Scale Text Mining.* 1st ed. New York: Springer International Publishing, 2015, Ch2, pp. 16-53.
17. K. Nielbo, R. Nichols. *How to Work with Unstructured Data* [Online]. The University of British Columbia. Available: https://hecc.ubc.ca/quantitative-textual-analysis/qta-theory/how-to-work-with-unstructured-data
18. KoNLPy. (2014). KoNLPy: Korean NLP in Python[Online]. Available: http://konlpy.org/en/latest/
19. National Institute of Korean Language. (2002). New Words in 2002 [Online]. Available: https://www.korean.go.kr/attachFile/viewer/201903/e424bfc6-e6d8-4ada-8d45-bd62a7029488_0.pdf.htm
20. H. T. Lim, *90nyeonsaeng-i onda*. Seoul (Korea): Whale Books, 2018.

AUTHORS PROFILE

Ji-Heon Song Graduate School Student at Namseoul University in Master degree course. Major is Bigdata Industry Security. Main studies are Bigdata Analytics in variety fields. A research paper that Proposal Study on Public Data-Centered Crime Prevention Through Environmental Design (CPTED) in Seoul City Utilizing Classical Music was published. Also, recent projects were Tourism Bigdata Analytics. Meanwhile, new analytics studies started using the deep-learning algorithm.

Sung-Jun Kim Professor at Dept of Bigdata Industry Security, Namseoul University Graduate School. He studied about Information security and law. A research paper he participated as an assistant professor that Proposal Study on Public Data-Centered Crime Prevention Through Environmental Design (CPTED) in Seoul City Utilizing Classical Music was published. Also, recent projects were Tourism Bigdata Analytics. He is an advisor of the Government Committee, and an advisor of Business Agencies.