Topic Modeling-Based Analysis of YouTube Comments Related to Jeonse Scams

Yeji Lee

Data and Social Media Analysis
ye_ji@akane.waseda.jp

June 23, 2024

GitHub repository: https://github.com/2yeeji/Data-and-Social-Media-Analysis-51/blob/main/DSMA_Final_Project_Topic_Modeling_of_YouTube_comments_related_to_Jeonse_Scams.ipynb.

1 Introduction

Topic modeling allows for various tasks, including understanding public perceptions. In (Oh and Kim (2023)), the focus was on how urology-related news is delivered, revealing a distinction between articles providing medical information and those promoting services or products through LDA topic modeling of news article titles and content. In (Lee et al. (2023)), discourse on online remote education in Korea before and after the pandemic was explored by conducting LDA topic modeling of news articles from Korean media outlets. (Sun et al. (2023)) aimed to explore public reactions to the metaverse by performing BERT-based topic modeling on tweets, identifying the relationship between the expectations of the South Korean government and those of the public regarding the metaverse.
This study aims to explore public perceptions of Jeonse Scams through topic modeling using Latent Dirichlet Allocation model based on YouTube comments. According to (권경선 (2023)) (김성용 and 신광문 (2023)), large-scale Jeonse scam cases, known as 'Villa King' and 'Villa God', began to be uncovered in the latter half of 2022. While Jeonse scams have existed before, these incidents have led to increased public interest. Jeonse, a unique real estate system practiced in a few countries including Korea, Bolivia, and certain regions of India, has not been extensively researched. This report is significant as it sheds light on public perceptions of Jeonse scams in Korea for the first time. *According to Insight (2024), YouTube is a massive platform with 2.7 billion users, and people spend an average of 19 minutes and 39 seconds on YouTube per day. In Korea, YouTube penetration is as high as 89.9%, ranking 8th worldwide, indicating active use of YouTube. In 2019, according to 한국언론진흥재단(2023), only 12% of people consumed news through online video platforms, including YouTube. However, by 2023, this number surged to 25.1%, more than doubling in just four years. Among them, 98.8% utilized YouTube to watch news, indicating that YouTube accounted for the vast majority of news consumption through online video platforms. Furthermore, YouTube has the feature of commenting, allowing users to exchange opinions with each other. This report will utilize these characteristics to understand public perceptions of Jeonse scams.*
Topic modeling allows for various tasks, including understanding public perceptions. In (Oh and Kim (2023)), the focus was on how urology-related news is delivered, revealing a distinction between articles providing medical information and those promoting services or products through LDA topic modeling of news article titles and content. In (Lee et al. (2023)), discourse on online

remote education in Korea before and after the pandemic was explored by conducting LDA topic modeling of news articles from Korean media outlets. (Sun et al. (2023)) aimed to explore public reactions to the metaverse by performing BERT-based topic modeling on tweets, identifying the relationship between the expectations of the South Korean government and those of the public regarding the metaverse.
This study aims to explore public perceptions of Jeonse Scams through topic modeling using Latent Dirichlet Allocation model based on YouTube comments. According to (권경선 (2023)) (김성용 and 신광문 (2023)), large-scale Jeonse scam cases, known as 'Villa King' and 'Villa God', began to be uncovered in the latter half of 2022. While Jeonse scams have existed before, these incidents have led to increased public interest. Jeonse, a unique real estate system practiced in a few countries including Korea, Bolivia, and certain regions of India, has not been extensively researched. This report is significant as it sheds light on public perceptions of Jeonse scams in Korea for the first time.

2 Background

2.1 Jeonse Scams

According to (김성용 and 신광문 (2023)), the Jeonse system is a real estate system where tenants deposit a higher security deposit to the landlord instead of paying monthly rent, and they do not have to pay separate monthly fees. At the end of the contract period, the security deposit is returned to the tenant, making it a unique real estate system. (권경선 (2023)) describes Jeonse scams as acts perpetrated by landlords, property developers, intermediaries, or presale agents involved in Jeonse contracts, deceiving tenants and misappropriating Jeonse deposits. The significant problem with Jeonse scams, highlighted due to the large-scale incidents in the latter half of 2022, is that the majority of victims are young adults or newlywed couples.

2.2 Topic Modeling of Korean YouTube Comments

(김가은 (2023)) (최윤정 (2023)) (권경인 and 신성미 (2023)) (최재서 et al. (2023)) are all studies that utilized YouTube comments for LDA topic modeling. In Study 김가은 (2023), analysis was conducted using Python 3.5.2 and Origin Pro 16, with Okt used as the morphological analyzer for determining the number of topics based on the umass coherence score. Study (최윤정 (2023)) conducted web scraping using Python 3.10 and R for topic modeling, employing KoNLP for morphological analysis, and determining the number of topics through model comparison. (권경인 and 신성미 (2023)) utilized NetMiner 4.5 for topic modeling and determined the number of topics using the cv coherence score. Lastly, (최재서 et al. (2023)) conducted topic modeling analysis using R 4.1.1 with packages such as topicmodels and KoNLP, and extracted the optimal number of topics using findtopicnumbers().

Research	Data Used	Program Used	Preprocessing Method	Model Used	Choosing the Number of Topics
김가은 (2023)	YouTube Comments	Python	Special Characters Removal Korean Text Only Retention Synonym Processing Morphological Analysis using KoNLPy's Okt	Latent Dirichlet Allocation	Umass Coherence Score
최윤정 (2023)	YouTube Comments	Python(Crawling), R	Special Characters Removal Korean Text Only Retention Morphological Analysis using KoNLP	Latent Dirichlet Allocation	Compare Results
권경인 and 신성미 (2023)	YouTube Comments	NetMiner 4.5	Special Characters Removal Synonym Processing Homonym Processing	Latent Dirichlet Allocation	Cv Coherence Score
최재서 et al. (2023)	YouTube Contents and Comments	Webopmetric Analyst(Crawling), R	Special Characters Removal Synonym Processing Morphological Analysis using KoNLP	Latent Dirichlet Allocation	findtopicnumbers() function

Table 1: Literature Study of 2.2

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the most commonly used method in topic modeling, a generative probabilistic model for collections of discrete data such as text corpora. According to Blei et al. (2003), LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, LDA generates topics within a document based on the TF-IDF values of the text corpus. It is characterized by its ease of interpretation compared to other models and the requirement for morphological analysis.

3 Data and Method

3.1 Data Collection

The data crawling was conducted using the YouTube Data API. As of January 23, 2024, the top 20 videos based on 'views' were selected when searching for "Villa King Jeonse Scams KBS" or similar queries in Korean, appending the name of terrestrial TV channels. Only videos uploaded by news accounts of the three terrestrial TV channels in Korea were used. Shorts, streams, and videos other than news content from these channels were excluded. The video IDs were pre-entered for the desired videos from which comments were collected. The extracted data includes the comment content, commenter, comment timestamp, and number of 'likes', saved in Excel format.

3.2 Preprocessing

During the data collection phase, rows with empty comment content were deleted, and comments from the three channels were all consolidated into a single dataframe. For stopword processing, nouns, adjectives, and verbs were extracted from the entire document, and the top 500 words based on frequency were saved to a text file, leaving only unimportant words in the file as stopwords. This process was conducted by one native Korean speaker and reviewed by another. Text preprocessing was carried out as follows: website addresses were replaced with "URL", email addresses with "EMAIL", and only Korean characters and numbers were retained for analysis. Special symbols and characters were removed, and multiple consecutive spaces were replaced with a single space.

3.3 Latent Dirichlet Allocation

First, only nouns, adjectives, and verbs were extracted from the documents for tokenization. The tokenizer used for this purpose was the Korean model provided by spaCy. Using these words, corpora, dictionary, and id2word were created. Both alpha and beta values were increased from 0.01 to 0.02 with an interval of 0.001, and the coherence (u_mass) was measured for topic numbers ranging from 5 to 15 to select the best model. Each document underwent 100 iterations, and the random_state was set to 123 to ensure consistent results. Among the combinations of alpha and beta values, only the one with the highest coherence was selected, and the coherence was visualized. Through Figure 2, it was evident that a significant portion of the document was comprised of stopwords. To conduct a more precise analysis, not only stopwords but also words occurring less than 30 times or having a document frequency over 0.5 were filtered out.
When the coherence score was highest for both (alpha = 0.017, beta = 0.010) and (alpha = 0.017, eta = 0.011), the average coherence scores were compared. Both cases had a topic number of

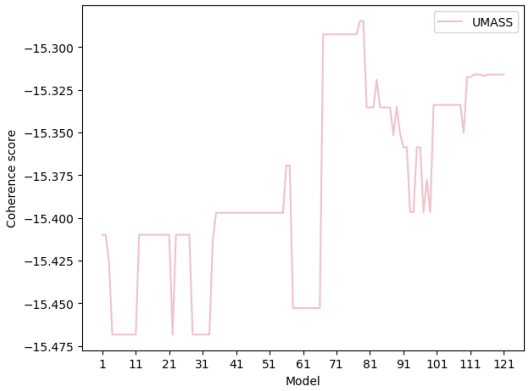

Figure 1: LDA Coherence

7. The average beta01 value was -16.858926329733343 for alpha = 0.017, beta = 0.010, and -16.858643695185204 for alpha = 0.017, beta = 0.011. Therefore, the case with alpha = 0.017, beta = 0.011 was selected as the final best model.

4 Results and Conclusions

4.1 Crawling

The data consists of 9509 entries for KBS, 7426 for SBS, and 8711 for MBC, with differences of up to about 2000 entries. The total number of entries is 25646. When observing the word cloud of the crawled data, it appears that there are many promotional and provocative comments.We can observe a high frequency of irrelevant words such as "1", "a", "URL", "href", etc. Comparing the word clouds before and after removing stopwords, we have the following observations. When stopwords are removed, words related to "jeonse," "scam," and "money" become more prominent, highlighting terms associated with jeonse scams.

Figure 2: Wordcloud of Original Comments

Figure 3: Wordcloud after Removing Stopwords

4.2 Latent Dirichlet Allocation

The second topic can be named 'Jeonse Scams and Institutional Issues'. Key words include 'law', 'country', 'victim', 'government', 'system', and 'fraudster', suggesting discussions not only about conflicts between victims and fraudsters, but also about the responsibility and role of the government and institutions. There is a need for strengthening laws and systems related to the South Korean jeonse system, and support and protection measures need to be provided to victims. Investigation into the causes and backgrounds of jeonse scams is also necessary.
The third topic can be named 'YouTube Advertising Comments'. These are comments unrelated to jeonse scams but are commonly found on YouTube, promoting specific products or channels.
The fourth topic can be named 'Criminal Organizations and Jeonse Scams'. Key words like 'prosecution', 'country', 'apartment', 'price', and 'organization' suggest involvement of criminal organizations in jeonse scams, demanding thorough investigation.
The fifth topic can be named 'Jeonse Scams and Consumer Protection Issues'. Discussions are actively held regarding issues related to jeonse security deposits and insurance, indicating interest in consumer protection.
The sixth topic can be named 'Jeonse Scams and Measures for Victims'. Discussions on coping strategies for jeonse scam victims show negative reactions from some people who believe it may benefit them.
The final topic can be named 'South Korea's Real Estate Transaction-Related Accident and Management Issues'. Interest is shown in issues related to real estate transactions and management, along with discussions on dealing with accidents and managing them in the real estate market. The first topic can be named 'South Korea's Jeonse Scams and Real Estate Market Issues'. Key words include 'jeonse', 'real estate', 'fraud', 'certified real estate agent', 'tenant', and 'problem', indicating the occurrence of jeonse scams in the South Korean real estate market, leading to issues between tenants and landlords. Additionally, words like 'fraudster', 'money', 'tax', 'monthly rent', 'property price', and 'loan' suggest high interest in macroscopic issues in the real estate market.
The second topic can be named 'Jeonse Scams and Institutional Issues'. Key words include 'law', 'country', 'victim', 'government', 'system', and 'fraudster', suggesting discussions not only about conflicts between victims and fraudsters, but also about the responsibility and role of the government and institutions. There is a need for strengthening laws and systems related to the South Korean jeonse system, and support and protection measures need to be provided to victims. Investigation into the causes and backgrounds of jeonse scams is also necessary.
The third topic can be named 'YouTube Advertising Comments'. These are comments unrelated to jeonse scams but are commonly found on YouTube, promoting specific products or channels.
The fourth topic can be named 'Criminal Organizations and Jeonse Scams'. Key words like 'prosecution', 'country', 'apartment', 'price', and 'organization' suggest involvement of criminal organizations in jeonse scams, demanding thorough investigation.
The fifth topic can be named 'Jeonse Scams and Consumer Protection Issues'. Discussions are actively held regarding issues related to jeonse security deposits and insurance, indicating interest in consumer protection.
The sixth topic can be named 'Jeonse Scams and Measures for Victims'. Discussions on coping strategies for jeonse scam victims show negative reactions from some people who believe it may benefit them.
The final topic can be named 'South Korea's Real Estate Transaction-Related Accident and Management Issues'. Interest is shown in issues related to real estate transactions and management, along with discussions on dealing with accidents and managing them in the real estate market.

Topics	1	2	3	4	5	6	7
1st	Jeonse	법	바람	검찰	세상	집	개꿀
2nd	부동산	나라	수사	717	이해	빌라	바지
3rd	사기	피해자	석촌호수	국가	의심	사람	모르겠
4th	사기꾼	사람	해주세요	집안	나	구멍	어그
5th	돈	인간	한국	놈	구조	나왔네요	경찰
6th	공인중개사	정부	20년	아파트	아쉽	바지사장	지리
7th	빌라왕	제도	얼마	가격	업체	집안	사고
8th	세입자	집	30만원	살	가격	욕심	이름
9th	문제	사기꾼	이번제품	조직	겉	돈	관리
10th	얼굴	구매	넘비	사람	반납	1인	일반
11th	이유	대한민국	미사	상식	제거	다들	처음
12th	세금	사용	추워여	사업	판매	혜택	돈
13th	월세	전세금	꽁으	50만원	저렴	한국인	집
14th	사람	책임	독특하	문구	뻥튀기	대처	같이
15th	처벌	전세사기	독특하네	아래	보증보험	디	주기
16th	집주인	조사	될거라고는	마감	소비자	너무	내놓으라고
17th	법	강력	인트로	와우	부류	최초	칭찬
18th	집값	피해	현대	콜라	낚시	장만	대충
19th	대출	배후	자동차	도착	뿌	인가요	언케연상
20th	집	원인	스피커인줄	콜라보레이션이군요	콜센터	키	온나

Table 2: Visualize Top 20 Words of Each Topics of Best LDA Model

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, *3(Jan):993–1022.*

Insight, G. M. (2024). Youtube statistics 2024: Demographics, users by country more.

Lee, K., Kim, T.-J., Cefa Sari, B., and Bozkurt, A. (2023). Shifting conversations on online distance education in south korean society during the covid-19 pandemic: A topic modeling analysis of news articles. International Review of Research in Open and Distributed Learning, *24(3):125–144.*

Oh, Y. W. and Kim, J. (2023). Insights into korean public perspectives on urology: Online news data analytics through latent dirichlet allocation topic modeling. International Neurourology Journal, *27(Suppl 2):S91.*

Sun, S., Kim, J.-H., Jung, H.-S., Kim, M., Zhao, X., and Kamphuis, P. (2023). Exploring hype in metaverse: Topic modeling analysis of korean twitter user data. Systems, *11(3):164.*

권경선 (2023). 대규모 전세사기 (빌라왕) 에 대한 공법적 규제방안. 부패방지법연구, *6(2):41–70.*

권경인 and 신성미 (2023). 토픽모델링을 활용한 '자해'관련 유튜브 댓글의 내용분석. 학습자중심교과교육연구, *23(16):449–466.*

김가은 (2023). 스트릿 댄스 서바이벌 프로그램< 스트릿 우먼 파이터> 와< 스트릿 맨 파이터> 의 유튜브 댓글 비교분석. 한국무용학회지, *23(3):51–63.*

김성용 and 신광문 (2023). 전세사기 예방을 위한 공인중개사의 역할과 개선방안. 부동산경영, *28:121–139.*

최윤정 (2023). 유튜브 영어 학습 콘텐츠의 진행자 요인에 따른 이용자 인식: 토픽모델링을 이용한 댓글 분석을 중심으로. 영어평가, *18(1):141–161.*

최재서, 정유미, and 김정환 (2023). 대학 기본역량 진단에 관한 유튜브 뉴스 및 댓글 연구-내용분석과 토픽모델링의 혼합방법론을 중심으로. 지역과 커뮤니케이션, *27(4):123–169.*

한국언론진흥재단 (2023). 주요 결과 및 시사점. In 2023 언론수용자 조사, *pages 18–39. 한국언론진흥재단.*