Investigation of Topics of Articles from Publishers With Different Political Orientations Through Topic Modeling

Hannah Kim
Wellesley College, United States
hkim22@wellesley.edu

Abstract

In the summer of 2020, medical doctors in South Korea went on a nationwide strike in response to the government's new policies regarding the increase of the admission quotas by four thousand over the next ten years for medical schools. This paper investigated whether there are different trends in the topics of article content and titles by news publishers with different political orientations using topic modeling. After performing topic modeling with LDA models based on TF and TF-IDF vectorizers on article content and titles with five topics, this paper was not able to find any significant differences between articles from publishers with different known political orientations. However, it still remains an important issue to develop ways to expose users of news platforms and readers of articles to objective sources outside of their filter bubble.

1 Introduction

In the summer of 2020, medical doctors in South Korea went on a nationwide strike in response to the government's new policies regarding the increase of the admission quotas by four thousand over the next ten years for medical schools. Many doctors did not think that the government's way of addressing the current problems in Korean health care was appropriate and thus supported the movement. The Korean Medical Association and the Korean Intern Resident Association met with the government multiple times to discuss relevant matters, but they were not able to reach an agreement until early September.

The strike was a significant event that had a notable impact both on the medical and political sectors, especially amidst the spread of COVID-19 this year. Since it was so influential, many patients and even people who were not involved with the medical or political sector had strong opinions about this subject. One of the places where the bias and disputes were most conspicuous was news articles and their comment sections. I became particularly interested in the way news publishers with different known political orientations often appeared to post articles with different focuses and nuances even on the same occurrences.

It is already a somewhat established fact in Korean society and media that certain Korean news publishers have explicit political orientations. Also, Korean media tends to reveal its political orientation through purposefully excluding specific issues [Kim, 2011]. This paper will investigate the question of whether there are different trends in the topics of article content and titles by news publishers with different political orientations using topic modeling, which is also a hypothesis that is expected to be true. Performing topic modeling on titles as well as content seemed like a reasonable analysis to attempt because journalists often tend to put their most explicit opinions that they want to deliver to the readers in the titles of the articles, since titles are the first thing that readers look at and users are most easily exposed to titles more than any other parts of an article.

2 Literature Review

Since the nationwide strike of doctors in the year 2020 is an event that occurred extremely recently, there has been no previous research about this particular case. However, there have been previous works in the analysis of relevant issues.

2.1 The Controversy Between Doctors and the Government in South Korea

The relationship between doctors and the government has not been so smooth during the last few decades; this is not the first time where doctors and the government disagreed on a policy. This is actually the second time that doctors have gone on a strike against a government policy. [Oh, 2003] analyzed how news articles were written around the topic of the doctors' strike in the year 2000. The strike then was due to doctors' opposition to the government's policy of the division of medicine and pharmacy.

To investigate her research question, the author examined various quantitative characteristics of articles from the Korea Daily News, DongA Ilbo, and Hangyoreh. Some features she examined were the number of articles, the average length of the articles, and the number of pages on which relevant articles were featured. The situation seems to have been much more against doctors at the time compared to the recent strike; the author was able to find that there were many more articles that were in favor of the government and also many articles that criticized the doctors. She also found that there were

many articles that emphasized the government's viewpoints more than those of the doctors.

But another aspect of the findings of [Oh, 2003] that was interesting was that most of the changes in the headlines of the articles from that period had been due to the actions of the government. This may be due to the fact that the strength of the doctors and people in general to come together for group action had been limited since there was not as many means of communication across the nation, and the government and thus the contemporary Korean society was much closer to being authoritarian than democratic.

In addition, even in between the year of 2000 and 2020, [Lee et al., 2010] found that the level of trust that Korean physicians had in the government was low. They highlighted the need for the facilitation of open communication between physicians and the government.

2.2 The Analysis of Political Orientations Through Computational Methods

There have been multiple previous work using computational methods to identify the political orientations of various types of texts. [Zhou et al., 2011] uses three semi-supervised learning methods on social news aggregator services with propagation algorithms to successfully classify conservative and liberal articles with a 99.6% accuracy with the best algorithm. [Boudemagh and Moise, 2017] uses GDELT for a case study on "news media coverage of refugees," and [Durant and Smith, 2006] shows that a Naïve Bayes classifier is significantly better at predicting the political category of postings than Support Vector Machines.

Also, [Jiang and Argamon, 2008] attempts to classify a political blog as liberal or conservative by first identifying subjective sentences and then extracting opinion expressions and Bag of Words features to build classifiers. They show that it is possible to reveal opinions that reflect a certain political orientation by "extracting opinion expressions from subjective sentences." Furthermore, [Park et al., 2011] even presents a new approach for the identification, or prediction, of the political orientation of news articles by analyzing commenters' "sentiment patterns towards political news articles."

The work that this project most resembles is [Kang et al., 2011]. Their work uses topic modeling and opinion mining to investigate whether the content of articles published by newspaper publishers with different political orientations show different tendencies. They take three steps to analyze this research question: 1) extract topics from newspaper texts, 2) use network analysis to examine the structure and content of each topic, and 3) use time serial analysis to observe the chronological distribution of the topics in articles by each news publisher. They use the topic modeling algorithm based on Latent Dirichlet Allocation (LDA) and the Pathfinder Network with cosine similarity to conclude that liberal and conservative newspapers tend to select facts in a way that is biased towards their political orientation. This project is heavily motivated by [Kang et al., 2011] in that it will use topic modeling with LDA models and compare topics of articles from publishers with different political orientations to investigate whether they are noticeably different.

3 Data and Methods

News articles about the doctors' nationwide strike from ten major central news publishers were collected from Naver, which is the most commonly used search engine in Korea. The news publishers are Kyunghyang Shinmun, Kukmin Ilbo, DongA Ilbo, Munhwa Ilbo, Seoul Shinmun, Segye Ilbo, Chosun Ilbo, JoongAng Ilbo, Hangyoreh, and Hankuk Ilbo. A total of 1,951 articles was collected from these publishers using Selenium and the HTML parser Beautiful Soup to scrape search results from Naver News. The earliest article had been published on July 23, 2020, and the most recent article had been published on September 21, 2020. Figure 1 shows the number of articles collected from each publisher.

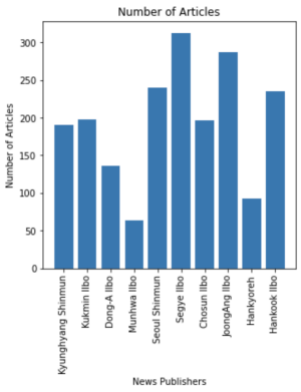

Figure 1. The number of articles collected for each news publisher.

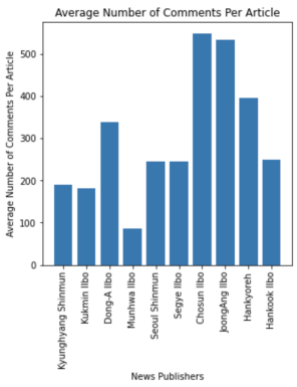

Figure 2. The average number of comments per article for each publisher.

Data collected for each article included the title, datetime at which it was published, number of comments, content, number of each type of reaction, and number of people who recommended the article. There were five types of reactions that a user could express on an article on the Naver News platform: 'good', 'warm', 'sad', 'angry', and 'want'. Apart from that, users could also 'recommend' an article. Figure 2, Figure 3, and Figure 4 respectively show the average number of comments, recommendations, and reactions for each article for each publisher.

The data for the number of recommendations and reactions per article for each publisher could have been used to analyze the characteristics of the data that have to do with interactions between the articles and the users, but this was not done in this paper partly because the reactions were not thought to be clear indications of the readers of the article. There are seemingly positive and negative reactions, but it is vague whether the reaction means that the article is well-written or badly written, or if it means that users feel good or angry about what happened within the event which is the subject of the article. But what can be seen here is that people tend to engage at similar rates in terms of comments, recommendations, and reactions across all articles for the same publisher. Also, if the number of readers is, in fact,

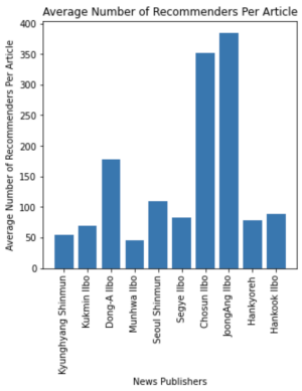

Figure 3. The average number of recommenders per article for each publisher.

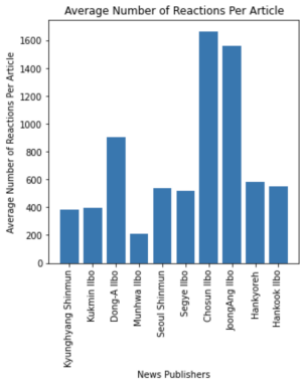

Figure 4. The average number of reactions per article for each publisher.

proportionate to the number of comments, recommendations, or reaction an article gets, this may mean that there are a significantly large amount of readers for Chosun Ilbo and JoongAng Ilbo, followed by DongA Ilbo as well.

Moreover, the datetimes at which all articles were published were analyzed to see the number of articles that were published during each period from July 23 to September 21. Figure 5 shows the distribution of the number of all articles over time, regardless of which publisher the article is from. The peak of the number of articles is from around late August to around early September; this is reasonable because that is when doctors actually went on the strike, and thus everything, including the dispute between the government and doctors and other people's reactions, was at its maximum.

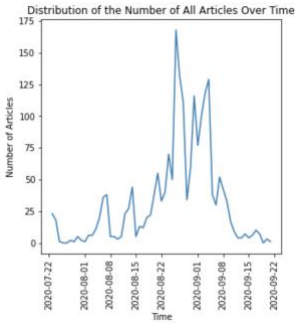

Figure 5. The distribution of the number of all articles over time.

The first variable that was used throughout this paper was the list of publishers for which Naver News articles had been successfully collected, sorted in alphabetical order in Korean. Initially, a list of dictionaries where each dictionary represented an article and its keys represented attributes of the articles such title and content was constructed. Then, for the topic modeling process, two separate lists, one with only the content of the articles and the other with only the titles of the articles, were constructed from the list of dictionaries with all data.

The most significant challenge while preparing the data for topic modeling was the different nature of English and Korean. Unlike in English, the same noun or verb can take many different forms depending on its function in the sentence and tense. It is difficult to obtain satisfying results by simply letting the algorithms consider each split part as one distinct word, because there would be so many forms of words and phrases that should be taken to convey the same meaning. To address this issue, KoNLPy, a natural language processing module for the Korean language, was used. All Korean letters that were not punctuations stayed in the strings; but the strings were morphologically split using the Okt class in KoNLPy and joined again with spaces. This enabled procedures in later steps to focus on a significantly smaller number of different forms of the same words and phrases. Furthermore, in the process of transforming the list of strings into TF vectorizers and TF-IDF vectorizers, many different combinations of the minimum document frequency, maximum document frequency, and n-gram range parameters were experimented with to find the combination that worked best for article content and titles, respectively.

4 Results

Topic modeling, a form of an unsupervised machine learning method, was performed on a total of four Latent Dirichlet Allocation models: the model based on TF of article content, the model based on TF-IDF of article content, the model based on TF of article titles, and the model based on TF-IDF of article titles.

For the vectorizers for the models based on article content, the minimum document frequency was set to 0.05, the maximum document frequency to 0.85, and the n-gram range was set from 2 to 5. The n-gram range was set to start from 2 because it is often the case that multiple nouns and verbs are combined to mean a single entity or phenomenon, and the spacing between the separate elements of such words varies a lot. Using these parameters, 1,276 feature names were extracted from all article content.

For the vectorizers for the models based on article titles, the minimum document frequency was set to 0.0015, the maximum document frequency to 0.005, and the n-gram range was again set from 2 to 5. Since the number of words in titles are significantly shorter and there are not so many overlapping words in the first place, the maximum document frequency merely had any effect on the number of feature names being generated when it was higher than 0.005. Using these parameters, 1,829 feature names were extracted from all article titles. All four models were fit on five topics.

The topics generated by all four models were considered in order to find the most reasonable set of topics. The topic models were visualized using pyLDAvis. Also, the proportions of each topic for all article content and titles for each publisher were examined to answer the research questions for this paper.

4.1 Topic Models Based on Content

Figure 6 and Figure 7, respectively, are topic models that were formed from TF and TF-IDF of all article content.

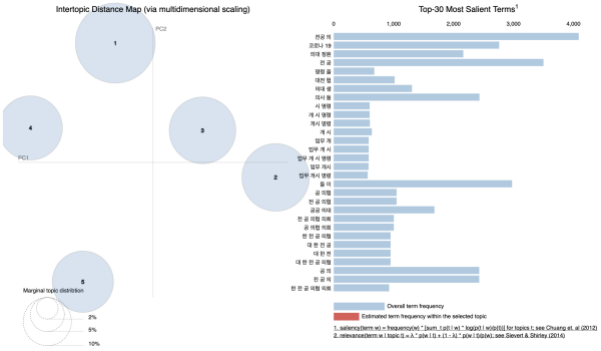

Figure 6. Topic model based on TF of article content.

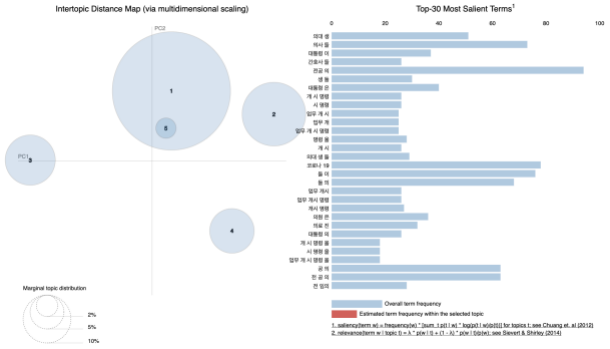

Figure 7. Topic model based on TF-IDF of article content.

The topic model formed on just the term frequency of the article content was more clearly divided in terms of its topic. Table 1 is the list of extracted words and names inferred from the words for each topic for the TF model.

Topic	Name	Words
1	Doctors	Korean Intern Resident Association, Park Ji-hyun, Choi Dae-zip, agreement, back to the origin, reexamine
2	COVID-19	social distancing, People Power Party, speaking out, confirmed cases, re-spread of COVID-19, Central Disaster and Safety Countermeasures Headquarters, National Assembly, politicians
3	The controversial policy that started the strike	number of doctors in local area, Organization for Economic Cooperation and Development, medical school admission quotas, 400 people, for 10 years, doctors of oriental medicine, government, policy, provision of medical herbs
4	The order for commencement of duty	order for commencement of duty, residents and fellows, Korean Medical Student Association
5	The president's negative framing	doctors on strike, undertake the burden of, nurses, hospital staff, the Catholic University of Korea Seoul St. Mary's Hospital

Table 1. Extracted words and inferred topics for the TF model based on article content.

Figure 8 is a set of barplots that show the proportions of content topics for each publisher. All of the topics for all publishers were spread out relatively evenly, compared to

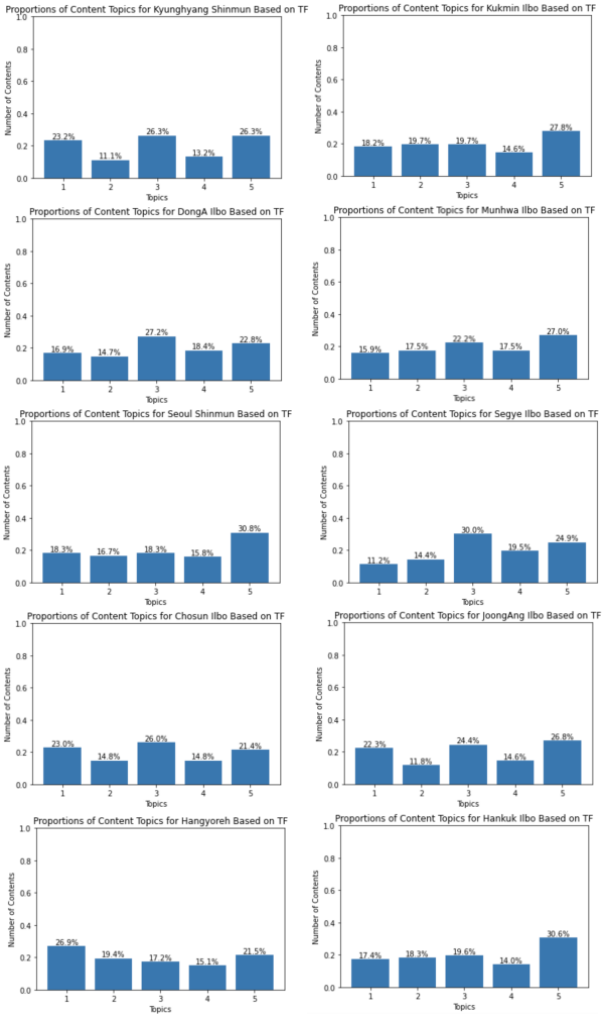

Figure 8. Proportions of content topics for each publisher, based on TF.

what had been expected; there were no stark differences between any of the distributions of topics.

However, there were publishers that had a similar pattern in their topic proportions: 1) Kyunghyang Shinmun, Chosun Ilbo, and JoongAng Ilbo, 2) Kukmin Ilbo, Seoul Shinmun, and Hankuk Ilbo, and 3) DongA Ilbo and Segye Ilbo. The first group has their peaks at Topics 1, 3, and 5, and even the proportions of the articles about each topic are similar among these three publishers. Chosun Ilbo and JoongAng Ilbo are well-known for having a politically conservative stance, while Kyunghyang Shinmun is well-known as one of the liberal publishers. It is difficult to predict what this trend among the three publishers means due to their different political orientations. Also, it is surprising that Kyunghyang Shinmun, a publisher with the same stance as the ruling party, has a lot of articles about Topic 5, which is one of the most negatively viewed actions of the president regarding the strike. But perhaps it is also possible that Kyunghyang Shinmun supports the president's view; there have been some articles about controversial statements by the president in

support of his statements, as well as those written to criticize his statements.

The second group, Kukmin Ilbo, Seoul Shinmun, and JoongAng Ilbo, has noticeable peaks at Topic 5, all of them around 30% of all articles published by each publisher. One common characteristic they have is that all of them tend to have only a weak or neutral political stance. But again, for similar reasons, it is difficult to interpret what each publisher was trying to imply with Topic 5. At this point, it may just be the case that Topic 5 was a popular one among all topics because the president's statement explicitly criticizing doctors was a controversial issue that would make a good headline.

Both publishers in the third group, DongA Ilbo and Segye Ilbo, have a peak at Topic 3 and similar proportions of all of Topics 3, 4, and 5. Both of them happen to be conservative publishers, although DongA Ilbo is much more so than Segye Ilbo. Although it could have been possible to use Topic 3 in both positive and negative ways, just like Topic 5, in this case it may be possible that DongA Ilbo and Segye Ilbo took more articles to explain the controversy between the government and doctors, perhaps with more emphasis on the doctors' side. From what was observed outside of this study, even when their publishers had a relatively clear political orientation, all news articles tended to explain the government's new policy and their rationale for presenting such a policy; however, the difference was that articles with strongly leaning opinions towards the conservative party would take one step further and explain why most doctors believed that the government's approach of solving the problem was wrong, while those with opposite views would just end the article after reasoning that the government had their reasons for pushing the policy. It could be the case that DongA Ilbo and Segye Ilbo explained the policy in many articles to make further points about the policy.

4.2 Topic Models Based on Titles

Figure 9 and Figure 10, respectively, are topic models that were formed from TF and TF-IDF of all article titles. The topic model formed on the TF-IDF of the article titles was slightly more clearly divided in terms of its topic. Table 2 is the list of extracted words and names inferred from the words for each topic for the TF model.

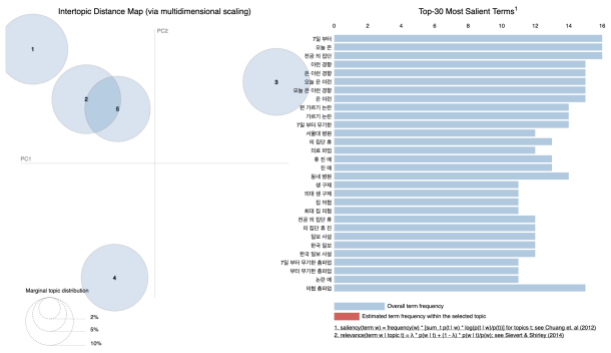

Figure 9. Topic model based on TF of article titles.

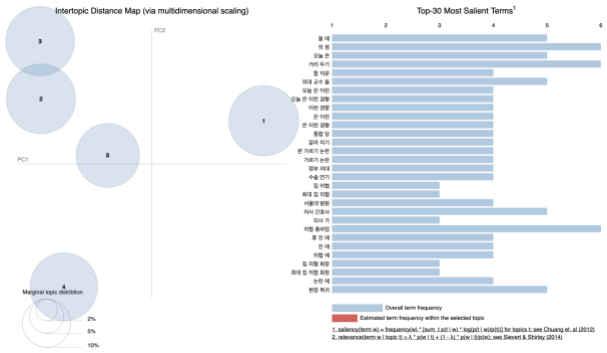

Figure 10. Topic model based on TF-IDF of article titles.

Topic	Name	Words
1	The effect of the strike	operations postponed, medical sector closed business, residents, strike, president, opposition to increasing admission quotas for medical school, surrender, health care in non-capital regions, agreement, support strike
2	Doctors' response to the strike	medical school professors, forewarn, strike, residents closed appointments, Korea Medical Association (KMA) medical school quotas, emergency room, top student, medical school students national examination, increase quotas, professors as well
3	The disagreement between doctors and the government amidst the strike	polarization, dispute, KMA, fake news, prosecute residents, social distancing level, start strike, return, re-spread of COVID-19, indefinite from September 7th, patients, expansion of medical school, Ahn Cheol-soo, president, strict response, confrontation between great powers, operation
4	The government's polarization attempt	polarization, government, COVID-19, found public medical schools, Korean Medical Student Association, angry, from the 21st, the 26th and 28th, on the side of nurses, doctors, participate in strike, medical sector, refuse to take examination, strict response
5	Medical licensing examination	agreement, government, medical school, Choi Dae-zip, president of KMA, back to the origin, reexamine, doctors, order of the mayor, Seoul National University Medical School, national examination, postponed for a week, strike, residents' opposition

Table 2. Extracted words and inferred topics for the TF-IDF model based on article titles.

There are two groups of publishers that have extremely similar trends in their proportions of title topics: 1) Munhwa Ilbo and Hangyoreh and 2) Kyunghyang Shinmun and Kukmin Ilbo. The first group has the most articles about Topic 5 and the second most articles about Topic 3; the second group has the most articles about Topic 5 and the second most articles about Topic 2. Even all of the topic proportions among each group are similar. However, both groups are mixtures of publishers that are known to be conservative and liberal, and all of Topics 2, 3, and 5 are close to neutral. Thus, it is difficult to conclude anything from the proportion topics of titles in this case regarding different political viewpoints of the news publishers of concern.

5 Discussion

Overall, the topic modeling process for both article content and titles worked well. Both models were able to each extract five distinct topics among the input data they were given. However, the topics that resulted from the models and the proportions of topics that news publishers used did not align with my initial hypothesis that there would be clear differences between articles published by publishers with different political stances. Although there were some groups

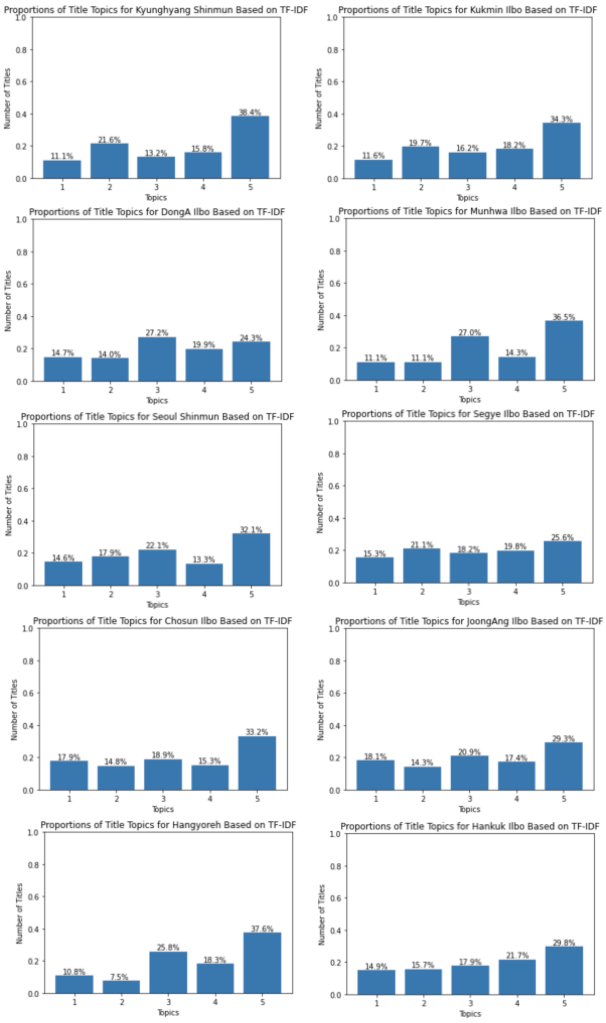

Figure 11. Proportions of title topics for each publisher, based on TF-IDF.

with similar trends, in the proportions of extracted topics, in most cases, they did not share the same view.

Furthermore, the topics themselves did not contain a strong sentiment about any one group in most cases, unlike what was expected; and even if it was possible that they did, like Topic 5 in the topic model based on the term frequencies of article content, the topic appeared with similar proportions in all news publishers.

From this particular work, it may seem as though topics and facts are not being selectively chosen by publishers for the expression of their political orientations. After all, all publishers are writing articles with similar topic proportions, and that may be taken to imply that all publishers prioritize the delivery of objective facts. However, the case that should be considered more seriously is one in which there is, in fact, political bias present in the articles, but it is just that the methodology of this work is not able to catch the biases present. Based on the review of previous literature and work done on this particular subject, I believe that this is the more probable situation.

Therefore, ways for platforms that provide news articles to its users, such as news sites like Naver News or various social media, to detect political bias in news articles and provide them in a way that enables their users to get exposed to diverse political viewpoints should be considered. This is an important issue for two reasons. First of all, news articles are meant to be sources from which objective facts about the world around us can be obtained. Arguing in support of or in opposition to a reasonable or an incomprehensible policy may be inclining but is definitely something that a news publisher should not be doing. It should be up to the readers to absorb objective facts about a happening and develop their own solid reasons for believing one thing or another. Second, the fact that certain publishers can have fixed political views that affect all articles that they publish is dangerous to its users because it can trigger the formation of filter bubbles. Not only will search engines and news platforms start analyzing the users' political stance to only show them articles that relate to their viewpoints, but the users being aware of their viewpoints and not caring enough to consider trying to read articles written from different stances will result in their being trapped in their filter bubble at their own will. This is not a matter that can be taken lightly.

But under the realistic assumption that the characters of news publishers cannot be changed overnight and that this is not on the list of priorities for any publisher, the first step in systematically targeting this problem would be to devise an algorithm or system that is able to judge the political opinion of news articles and to what extent the opinions are strongly biased accurately and efficiently. It would be even more problematic if this system does not work properly, because then articles would be filtered or highly exposed to users for no reason at all. Another way to approach this problem would be to develop a system that extracts only facts from news articles or similar factual sources and provides them to users in a coherent way that doesn't limit the users' view with a frame like a news article might, but is logical and smooth enough for the users to be able to read over the series of facts and understand what is happening. It should not even be allowed for such systems to skip over certain facts; all possibly identifiable facts from a source of news should be extracted and provided to the user to prevent selective choice of facts, which can lead to the implicit expression of support or opposition for a certain political matter.

6 Conclusion

In conclusion, what was found through topic modeling and observing the proportions of topics of article content and titles of each publisher was that it is a more sophisticated problem than expected to detect political bias in news articles. This paper was not able to find particularly noticeable patterns or similarities among topic proportions of articles written by publishers with similar political stances. Nonetheless, this paper was able to provide insights into the general flow and topics constituting the significant issue of the nationwide strike of doctors in South Korea, which exerted influence over the medical and political sectors in the summer of 2020 through topic modeling.

One further research question worth exploring would be whether it would be possible to come to a different conclusion if the amount of data is increased, by expanding the dataset with articles by other publishers as well or collecting more articles over a wider time range. Moreover, it would be interesting to use other machine learning methods such as sentiment analysis or clustering on a similar dataset to examine how the methodology can change our insights about the way in which political bias is expressed in Korean news articles.

References

[Boudemagh and Moise, 2017] Emina Boudemagh and Izabela Moise. *News Media Coverage of Refugees in 2016: A GDELT Case Study*. AAAI Technical Report WS-17-17: News and Public Opinion, 2017.

[Durant and Smith, 2006] Kathleen Durant and Michael Smith. Mining Sentiment Classification from Political Web Logs. *WEBKDD'06*, August 20, 2006.

[Jiang and Argamon, 2008] Young-wook Kim. Political Leaning Categorization by Exploring Subjectivities in Political Blogs. Conference: Proceedings of The 2008 International Conference on Data Mining, *DMIN*, July 14-17, 2008.

[Kang et al., 2011] Beomil Kang, Min Song, and Whasun Jho. A Study on Opinion Mining of Newspaper Texts based on Topic Modeling. *Journal of the Korean Society for Library and Information Science*, 47(4):315-334, November, 2011.

[Kim, 2011] Young-wook Kim. The Political Orientation of Korean Media and Crisis in Social Communication. *The Korean Society for Journalism & Communication Studies*, 107-136, 2011.

[Lee et al., 2010] Sun-Hee Lee, Gun-Mo Yang, Ju-Hyun Seo, and Ju-Hye Kim. A Study of Factors Related to Korean Physicians' Trust in the Government: On the Target for Board Members of Physicians' Associations. *Journal of Preventive Medicine & Public Health*, 43(5):411-422, 2010.

[Oh, 2003] Hye-young Oh. *Analysis of Newspapers of Government, Korea Medical Association, Civic Groups in Conflicting Organizations from June 2000 to September 2000*. The Graduate School of Ewha Womans University, 2003.

[Park et al., 2011] Souneil Park, Minnsam Ko, Jungwoo Kim, Yinng Liu, and Junehwa Song. The Politics of Comments: Predictinng Political Orientationn of News Stories with Commenters' Sentiment Patterns. *CSCW*, March 19-23, 2011.

[Zhou et al., 2011] Daniel Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the Political Leaning of News Articles and Users from User Votes. *ICWSM*, 2011.