The shaping of the narrative on migration: A corpus assisted quantitative discourse analysis of the impact of the divisive media framing of migrants in Korea

Clara Delort and **Eun-kyoung Jo**
Department of Global Korean Studies, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul 04107, South Korea
clara.delort@gmail.com - jek.cl.nlp@gmail.com

Abstract

This work explores the shaping of public opinion on migration in South Korea by utilizing BERT modeling (Grootendorst 2022) which extends transformer language models to Top2Vec (Angelov 2020) which leverages word semantic embedding to find topic vectors from documents. Data are the public discourse on Twitter and the three biggest local newspapers. The study examines the content of these topics, highlighting key themes and their implications. The findings through BERTopic modeling as a tool of discourse analysis on large data shows that, rather than a simple overall negative media narrative, the news outlets create distinctive concepts of migrants, fragmented into clustered groups, alienated from each other based on their social identities, migration status, and citizenship status. Discriminatory tropes (such as a criminalization frame and a victimization frame) predominant in the Mass Media corpus, are less salient in the New Media corpus and the Public Opinion (Tweets) corpus, where topics of compassion, human rights, union, reports of shared experiences, desire to share culture and communicate, are predominant. With the c-TF IDF formula giving the significance of words per topic, the creation of a divisive concept of refugees is visualized, with the fragmentation of one group (for example, refugees) into vastly distanced topics (either in the victimization frame, with "kid" and "refugee" in one cluster, or the criminalization frame, with "refugee" and "terrorism" in one cluster). This division in the public narrative supports the division in governmental policies. In this case, the Ministry of Justice divides asylum seekers applying for a refugee Visa into "humanitarian" or "economic" refugee categories. Asylum seekers placed in the "economic" refugee category are denied refugee status. The intertopic distance maps illustrate this shaping of divisive semantic meanings.

1 Introduction

Categorizing the recurrent topics in the public migration debate in South Korea allows us to examine the role of media in framing and depicting migrants and to understand the roots of the divisions based on social identities and citizenship status within the working class. This study's aim is to find the role of language in capitalism in shaping societal narratives and influencing perceptions by using a dynamic seeded topic modeling to categorize language data and gain insights in the discourses perpetuating capitalist structures. Scholars developed theories highlighting power dynamics, identity construction, and the importance of understanding global capitalism in the study of media representations of migrants. Stuart Hall (1997) emphasizes the role of media in constructing social hierarchies. Edward Said (1978) highlights how the media perpetuates stereotypes and exoticizes different cultures. Angela McRobbie (2009), explores how media representations contribute to gendered identities and marginalize migrant women. Chandra Talpade Mohanty (2003) examines gendered and racialized stereotypes, including those of migrant women. In the context of South Korea, the three major conservative newspapers, Chosun Ilbo, Joongang Ilbo, and Donga Ilbo, dominate the country's hard news. Smaller newspapers with varying political inclinations are also available as alternatives, but their circulation is lower. Through the discourse analysis of distinct corpora representing the mass media, the new media, and the public opinion, the role of media in the reproduction of class relations is quantitatively studied.

2 Data & Methodology

A corpus of tweets represents the public debate on migration during the 2009-2022 period. The Tweet data of migration-related Korean tweets are collected using a public Twitter scraper, snscrape

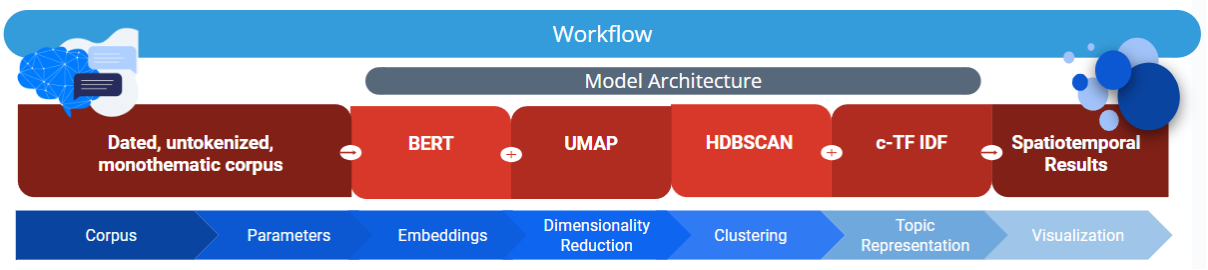

Figure 1: Workflow of the experimental design of Topic Modeling

and tokenized using Mecab, resulting in 3 120 297 Korean tweets that mention irregular immigrants, refugees, illegal immigrants, migrant workers, employment permit system, visa, migrants, immigrants, foreigners, illegal aliens, undocumented migrants, foreign workers. Only words tagged as nouns by Mecab are kept for topic modeling. A second corpus of news articles from the local daily newspapers with the biggest daily circulation, Chosun Ilbo, Joongang Ilbo and Donga Ilbo represents the local mass media. 14 560 articles (Chosun Ilbo, n= 4,678, Joongang Ilbo, n = 6,437, Donga Ilbo, n = 3,445) mentioning refugees, immigrants, immigrants, marriage immigrants, illegal immigrants, migrant women, foreigners, foreign workers, undocumented immigrants, and migrant workers are scraped. The articles were harvested over the 2009-2023 period. A third corpus of descriptions of news articles from Naver represents New Media. The Naver data were accessed using the official Naver News API and used to search for 10,338 articles mentioning refugees, immigrants, immigrants, marriage immigrants, illegal immigrants, migrant women, foreigners, foreign workers, undocumented immigrants, migrant workers. Only the short description of each article and publishing date were obtained, as the official API limits the number of articles scraped by query to 1000 titles and the harvest to the description of the articles rather than the full text. Topic modeling algorithms are used to discover hidden semantic structures, and infer and generate coherent topics by generating contextual word and sentence vector representations. BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018) is based on the encoder component of the Transformer model (Vaswani et al 2017), which reads the text input, and uses it to then generate a language model. In addition, the Class-Based TF-IDF Procedure (Grootendorst 2022) , aggregates all the documents for each topic, to then extract the meaningful words from the entire topic. To distinguish topics from one another based on those cluster words, the class-based TF-IDF (Term Frequency - Inverse Document Frequency) is carried out. This formula is an adaptation of the TF-IDF formula, which measures the importance of a word to a document. To obtain the importance of a word to a topic instead, the c-TF-IDF formula is used. The formula is as follows. class-tfidf(term, class, corpus) = tf(term, class) x idf(term, corpus) There, term is the word for which the weight is calculated, class-tfidf represents the weight of this particular term in a specific class, class is the class or topic to which the term belongs, and corpus is the entire collection of documents being analyzed. The Term Frequency tf(term, class) calculates the frequency of the term within the documents belonging to a specific class. Then, Inverse Document Frequency idf(term, corpus) measures the informativeness of a term by considering its distribution across the entire corpus. By combining both TF and IDF, class-based TF-IDF provides a measure of term importance that considers both the local term frequency within a class and the global term distribution across the corpus. This gives a more accurate and meaningful representation of the importance of terms within specific classes or topics, resulting in an effective topic modeling with BERTopic. Furthermore, to explore the potential hierarchical structure of the topics from the matrix created, hierarchical clustering visualization is performed. The similarity between two c-TF-IDF topics is determined by their distance, where a smaller distance indicates a higher level of similarity. In BERTopic, the merging of topics is achieved through the common linkage method “ward” (Ward J H. , 1963), or “Ward’s minimum variance method”. The formula calculates the increase in variance that would occur if two clusters were combined and compares it to the increase in variance for other potential merges. It selects the pair of clusters with the small-

est increase in variance as the most similar. The tokenizer of the multilingual BERTopic model is changed to the Korean tokenizer Mecab, for a better analysis of the Korean language, and the model is fine tuned with the cleaned, dated, Korean corpus. The number of topics to extract is set to 31. In order to obtain the most coherent topics, a seeded model is performed. Seeded topic modeling is realized by giving the model a list of seed topics with keyword attributes. These guide the topic model to converge towards the topics we want to examine in the documents. However, if those topics do not exist, they will not be modeled. The detailed seed topic list is available alongside the source codes at github.com/clara1del/BERTopic-korean-tweets-newsarticles-migration-discourse. In order to integrate socio-political concepts of class struggle into the language model and combine critical discourse analysis with structural topic modeling, we design a frame of study of the migration topics, which is fed to the model as a seeded topic list. This manual guiding of topics departs from a typical non guided topic modeling, and gives the model a deliberate perspective for a theoretically contextualized text analysis. Using the BERTopic (Grootendorst M., 2022) multilingual model for topic modeling, with the MeCab tokenizer (Kudo, T. ,2005) for the Korean language, and an added step of dynamic topic modeling, the development over time of the semantic meanings of migration related concepts in South Korea is investigated. As a quantitative method of discourse analysis, topic models offer voluminous statistical textual information, which can be used to study the structures of text in their historical and sociopolitical context. Through a study over time and a comparison between the voice of the elites and the voice of the public, we can uncover the relation between media coverage and the assumptions and values towards migrants reflected in the online discourse. Through topic modeling, the shaping of the migration narrative by the mass media, and the root of the hate on migrants is analyzed.

3 Related Work

Pavlova and Berkers (2022) conducted a frame analysis based on topic modeling using LDA clustering, as proposed by Gallagher et al. (2017), to explore the public perception of a divisive concept. They manually defined frames and associated them with top words, which served as the basis for Latent Dirichlet Allocation clustering. This approach facilitated the identification of unique frames for discourse analysis. Building on this methodology, we adopt a similar approach by constructing a theoretical frame, a seed topic list, to extract balanced and insightful topics. In a related study, Nozza, Bianchi, Lauscher, and Hovy (2022) focused on investigating language use towards specific social identities, particularly within the LGBTQIA+ community. They trained a model to complete sentences using LGBTQIA+ related templates and measured harmfulness scores, revealing identity-based attacks. In our work, we use another potential of the BERT model to analyze the language employed in relation to specific social identities, by studying the semantic distance between topics whose subjects are also groups of migrants defined by their social identities.

4 Analysis

Naver's search trends show frequent search terms in Naver's search engine. With 42 millions users, Naver is the most popular search engine in South Korea. Using the keyword research tool Naver Data Lab, the frequency over time of the keywords "migrants" and "refugees" searched in Naver. In figure

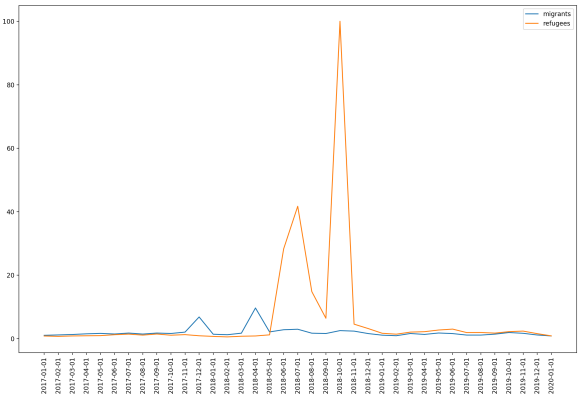

Figure 2: Trends of relative interest in "Refugees" and "Migrants" as frequency of search terms in Naver search engine

2, a peak in relative interest in refugees in 2018 followed the arrival of asylum seekers escaping the Yemeni civil war in Jeju-do, which was heavily covered in the media, portraying the male refugees as dangerous. The public opinion of refugees worsened to the point of the organization of protests to oppose the acceptance of the asylum seekers. In figure 3, the topic modeling of the corpus of news articles harvested from three major newspapers, Chosun Ilbo, Joongang Ilbo and Donga Ilbo, shows

Figure 3: Barcharts of the topics in the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

distinct noteworthy frames. A first theme (n = 304 articles) shows a focus on Western conservative views on migration, which are reproduced with representative keywords of the topic being: Trump, President, America, Democratic, Party, Republican, Candidate, Election, Biden, House, Congressman, Senate, illegal, immigration,Government, Exile, Minister. The third most predominant topic (topic 3, n = 210 articles) presents a criminalization framework of foreign workers, with the following keywords: Foreigners, Immigration, Policy, Workers, Ministry of Justice, Sojourn, Immigration policies, Employment, Expansion, Manpower, Government, Country, Budget, Population, Visa, Immigration, Employment, Illegal, Libya. The topic describes foreign workers, but not their work conditions. Rather, "Ministry of Justice", "illegal", "Sojourn", "Visa", show a focus on their legal status. This criminalization framework is also found in topic 6 (n = 161 articles) , with the following representative keywords: China, Visa, Jeju, Taiwan, Hong Kong, Lithuania, Italy, Smuggling, Government, Foreigner, illegal stay. The strong association between migrants and crime forms a negative sentiment. This main criminalization framework is present in all topics describing migrant workers. Topic 10 (n = 78) describes migrant workers, and associates them with the "illegal" term. The keywords for topic 10 are: Food, Seasons, Farmers, Vietnam, Labour, Workers, Farming, Corona, Illegal Stay, Grains, Rising, Entry, Potato. The strong association of "illegal" with even the migrants providing the country with provisions of food illustrates how criminalizing migrant workers allows for them to be exploited by the government without public outrage and resistance. Several topics describe refugees with a strong islamophobic association with terrorism. Topic 7 (n = 94 articles), which describes refugees and topic 8 (n = 87 articles), which describes terrorism, are overlapping. The keywords for topic 7 are: Afghanistan, Taliban, Pakistan, Kabul, Refugees, Islam, Humanitarianism, US Army, Reign, Escape, Government, Stay, problem. And for topic 8 are: terror, Islam, France, refugees, Middle East, forces, Muslim, Italy, Paris, Syria, Western Country, Al Qaeda, Bomb, Religion, War. This high coverage of terrorism in the local mass media promotes a fear of terrorism in South Korea. The presence of the terrorism topic (Topic 8) in a corpus of exclusively migrant related articles, and the significance of the word "refugee" in this cluster highlights the islamophobic association with migrants, specifically refugees, and terrorism. In figure 4, the frequency over time of

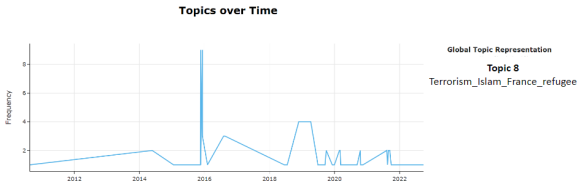

Figure 4: Frequency over time of the Topic 8 from the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

the Topic 8 (Terrorism) in the mass media corpus shows how predominant it is in the media narrative

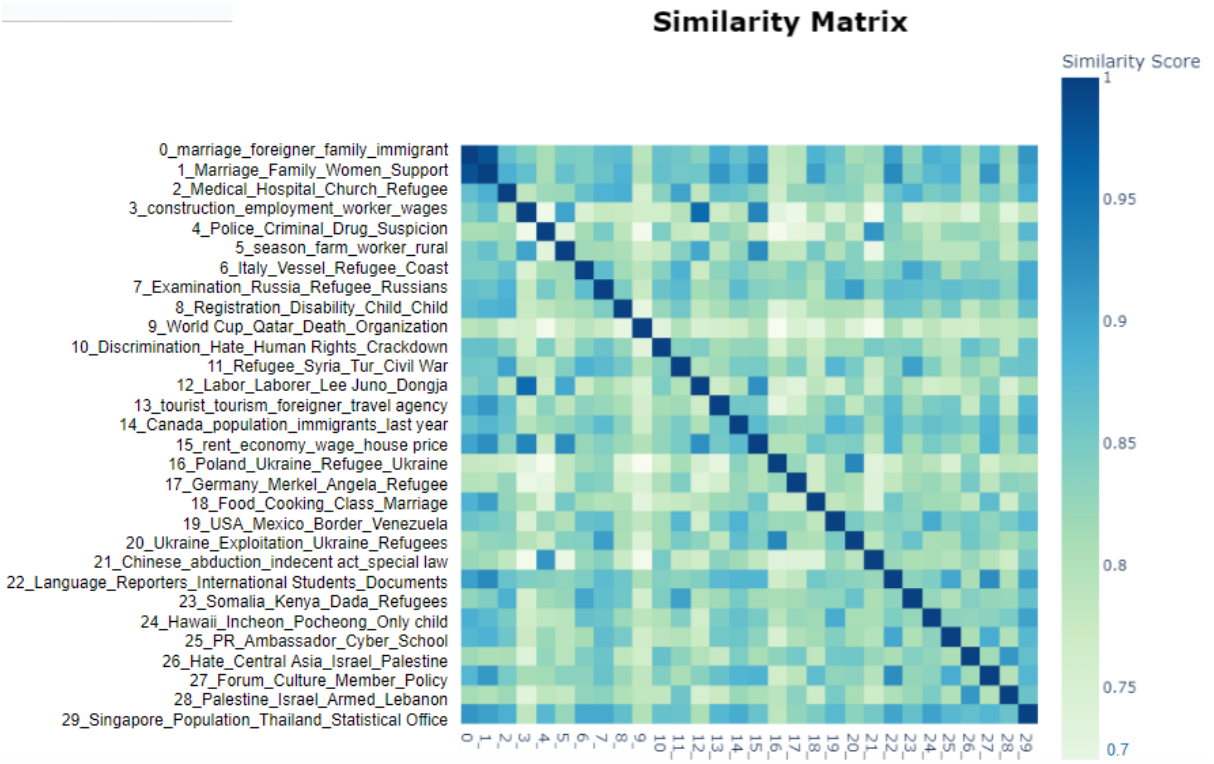

Figure 5: Similarity Matrix of the topics in the articles harvested from Naver News

on migration. Another important framework is the victimization framework, painting women migrants as victims. Topic 21 (n = 32 articles) describes migrant women with the following keywords: Women , Prostitution, Business owner, Police, Violence, suicide, victim, assault, male, report, Husband, Crime, Incident, Business, Sexual assault, Damage, punishment. Women migrants are both painted as victims of "violence", and as criminals, with the criminalization of sex work, with "prostitution" and "police". This victimization narrative puts women as victims of individuals, ("husband", "male"), rather than systemic exploitation. Combined with the criminalization narrative, women migrants are distanced from claims to citizenship. A prejudiced association with drugs is also found in top 29 (m = 12 articles), grouping migrants with the following keywords: Drugs, Thailand, Possession, cultivation, firearms, production, Southeast Asia, crime, Myanmar, Suspicion, Criminal, Regulation. The Mass Media narrative shows three primordial characteristics. First, migrants are separated into specific, and distanced groups, based on their social identities, such as gender. Then, a criminalization framework is applied, in particular to foreign workers and, or, a victimization framework, in particular to marriage migrants. Finally, an accrued coverage of Western conservative migration policies, namely USA and Germany's policies, passes on Western conservative views on immigration. In figure 5, from the topic modeling of the New Media corpus, the criminalization of migrants through the keyword "illegal" shows a strong association of specific subgroups of migrants with illegal status. In topic 4 (n = 302 descriptions of articles), violent police intervention is justified with the following keywords: Police, Crime, Drugs, Suspicion, illegal, assault, nationality, police station, stay, Thailand, violation, police agency, arrest, police officer, foreigner, Male, Jeju . Specifically, male migrants are covered as illegal. In contrast, women migrants are associated with "support" , in topic 1 (n = 656) , with the following keywords: Marriage, Family, Women, Support, Center, Education. This shows how both the criminalization frame and victimization frame restricts the rights to citizenship for both groups of migrants. Less salient topics however, do offer a coverage focusing on social justice and human rights. Topic 7 (n = 129) shows a high coverage of the situation of refugees waiting at the Incheon airport before being allowed to apply for the refugee status (keywords = Examination Russia Refugee Russians Conscript Ministry of Justice Incheon Recognition Litigation Airport Referral

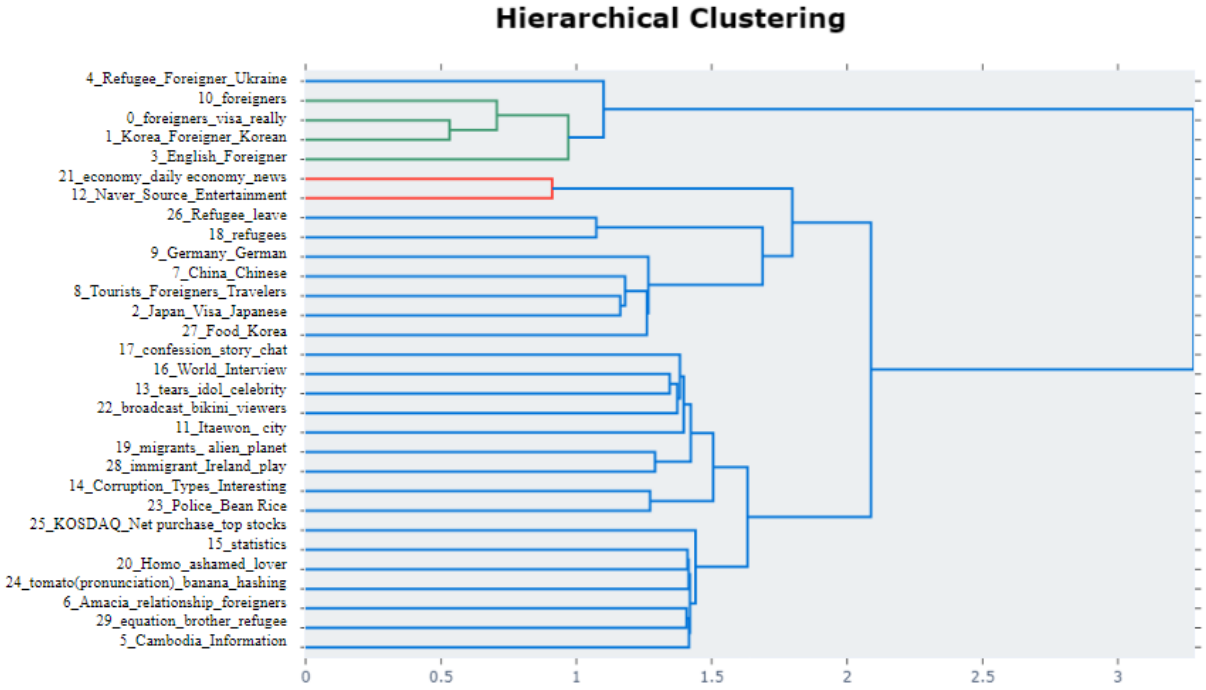

Figure 6: Hierarchical clustering of the topics in the tweets harvested in 2022

Court Forced Decision War Victory Korea Opponent cancel reject) . Topic 3 (n = 392) shows a coverage of migrant workers in exploitative work conditions (keywords = Construction Employment Workers Wages Foreigners Employment site accident survivors work foreign children Juno Lee Worker Constitution Hanam Factory Manufacturing Late Payment Provision). Topic 5 (n = 261) also mentions the fatal consequences of the exploitation of migrant workers (keywords = Season Farm Worker Rural Pig Foreigner Professor Farmhouse Agriculture Organic Cadaver Batch worker remark employment farmer entry manpower shortage when work). Topic 10 (n = 113) focusing on the repressive refugee application process (keywords = discrimination hate human rights regulation registration minorities society residents government halfhuman race immigrants Refugees Illegal Equality Deportation Groups Women Respect Suggestion), and topic 12 (n = 75) even shows compassion and union, not pity, with the immigrants undergoing this administrative process (keywords = labor worker Juno Lee dongja employment illegal problem field union discrimination human rights violence environment workplace regulation condition wage relocation registration construction). In figure 6, the topic modeling of the corpus of Tweets harvested in 2022 generated several remarkable topics. The first topic (n = 7641 tweets) in the public debate on migration focuses on South Korean locals migrating to Japan (keywords = Japan Visa Japanese Tourist Visa Travel Immigration). Locals are describing their own experiences as immigrants, troubles with visa processing, administration, integration in the country. This reveals a common experience as migrants between locals and immigrants. This is a primordial source of understanding. The second topic (n= 8069) shows a desire for communication with foreigners, as class friends. (keywords = English School Speak I Today Foreigner Class Friend) . The third topic (n = 7622), shows compassion with migrants in vulnerable situations (keywords = Refugees Foreigners Ukraine Women Marriage) . However, the topic 26 (n = 615), with victimization keywords (keywords = Refugees Syria Ukraine United Nations Children UNICEF) , presenting a focus on children, shows how this compassion is not turned into political activism, but distracted towards pity, charity, and an individual responsibility to donate to NGOs. In figure 7, the intertopic distance map from the topic model of Mass Media articles show clusters of topics distinctively distanced from each other. On the right, migrants in charge of education are vastly separated from groups of migrants on the top rights, associated with drugs. On the contrary, topics of refugees and terrorism are overlapping. The intertopic map shows the associations between refugees

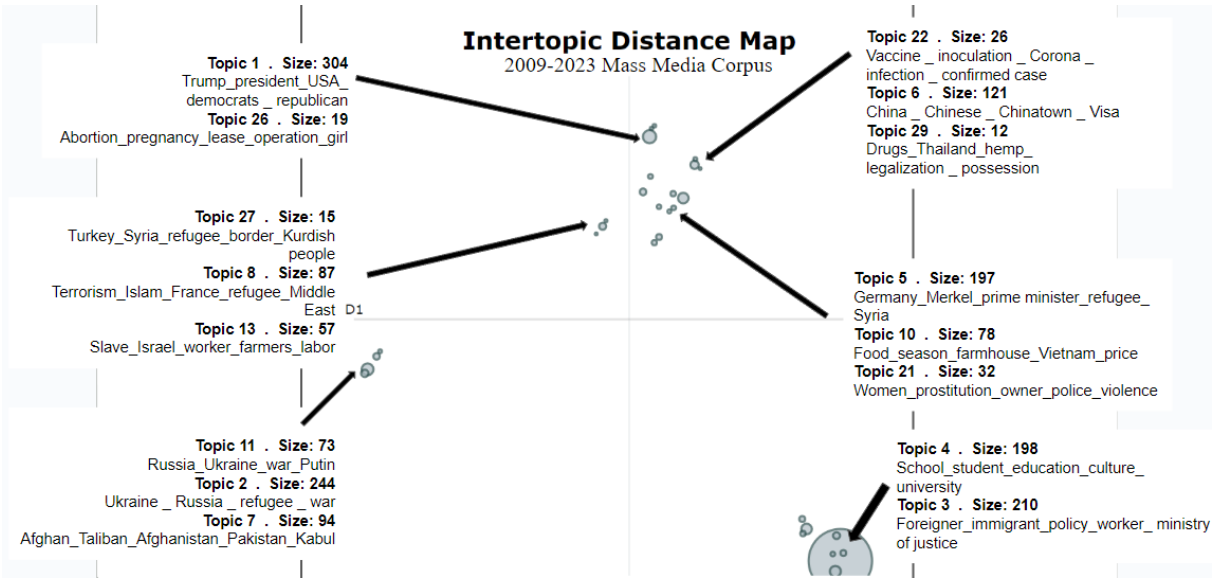

Figure 7: Intertopic Distance Map Topic in the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

and Islamophobic tropes, and the fragmentation of the groups of migrants in the discourse. In figure 8,

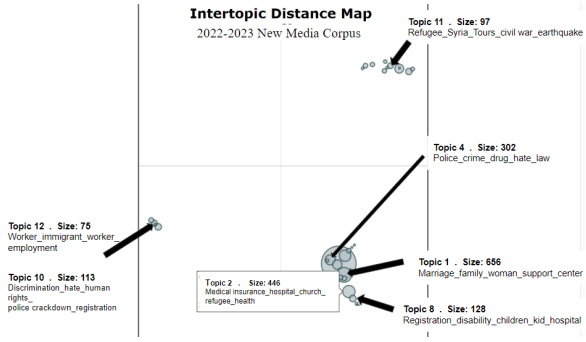

Figure 8: Intertopic distance map of the topics in the articles harvested from Naver News

the intertopic distance map from the topic model of New Media articles shows a strong separation between refugees (on the top right of the map) and migrant workers with the description of their exploitation (bottom left of the map) . The victimization frame (with "women" and "refugees") and the criminalization frame (with "police" and "drugs") are close. In figure 9, the intertopic distance map from the topic model of 2022 Tweets show clusters with overlapping topics on the right. Twitter users talk about their shared experiences (with visa, and as learners of English, Korean, Japanese). It is a source of union through shared experiences in the same country. On the left, separated topics are distanced based on social identities, such as nationality and sexual orientation.

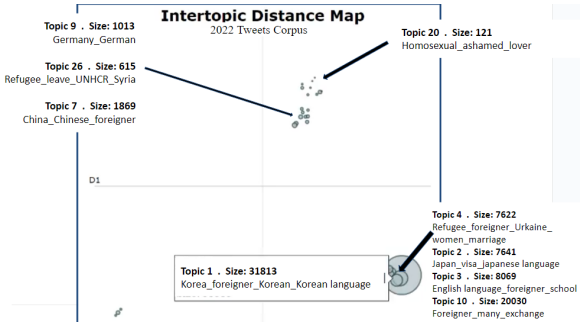

Figure 9: Intertopic distance map of the topics in the tweets harvested in 2022

5 Discussion

By framing migrants as criminals or threats to social order through the criminalization framework, the media perpetuates a narrative that justifies oppressive immigration policies and reinforces divisions within the working class. With the charitable framework, the media frames refugees and women migrants as passive victims, reducing them to non-political recipients of aid. Migrant women's victimization in the media undermines systemic oppression: their experiences are reduced to instances of personnel, individual suffering, diverting attention from the systemic factors that contribute to their exploitation. Similarly, mass media's appeal for charity and individual donations to aid refugees abroad, while neglecting to address the issue of visa recognition, individualizes and depoliticizes the refugee crisis, shifting the responsibility to individual acts of compassion. The mass media's

categorization of migrants into separate groups, dividing them into simplistic and stereotypical roles such as women as victims, or men as violent criminals, perpetuates a distorted narrative. By focusing on certain subgroups of migrants, the media obscures the systemic causes of migration, such as economic exploitation, political instability, and imperialist policies. This selective portrayal creates a false dichotomy of "good" versus "bad" migrants, perpetuating divisions among the working class. The study finds that the public shares experiences with immigrants, specifically struggles with visa regulations and language learning. It does not passively accept the divisive portrayal of foreigners by the mass media, and seek alternative narrative in new media, which covers the experience of immigrants with a human rights framework. To encourage this potential for union, it is necessary to challenge the categorizations of migrants shaping the narrative in the mass media.

6 Limitations

The mono-thematic corpora were centered around the migration theme, overlapping topics remained. While the seed topic list improved the definition of topics, the majority of the data was still categorized in the topic -1, 0 and 1. Modifying the parameters of the model, particularly of the UMAP dimensionality reduction model, slightly improved this issue. The predominance of topic -1 is an important limitation in this experience, as the top three words clustered in topic -1 included “women”, “marriage” in the Mass Media corpus, and “married”, “female” in the New Media corpus. Efficiently decreasing the size of topic -1 may provide information on the shaping of the narrative on gender and migration.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A8065237)

References

Ana M. Aranda, Kathrin Sele (2021). From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling. European Management Review Agarwal, N. Dokoohaki, and S. Tokdemir (2019). Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining (209–230). Berlin: Springer Baudrillard, J. (1970). The Consumer Society. Paris: Gallimard Cesare, N., Lee, H., McCormick, T., Spiro, E., Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research (55(5), 1979–1999). Demography Chandra Talpade Mohanty (2003). Feminism Without Borders: Decolonizing Theory, Practicing Solidarity. Duke University Press Curran, J., Couldry, N. (2003). Contesting Media Power: Alternative Media in a Networked World. New york: Rowman and Littlefield Publishers Inc. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. CoRR Dimo Angelov (2020). Top2Vec: Distributed Representations of Topics Edward Said (1978). Orientalism. Pantheon Books Grootendorst, Maarten(2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure Jaradat S., Matskin M. (2019). On dynamic topic models for mining social media. Lecture Notes in Social Networks JustAnotherArchivist (2018). snscrape: A social networking service scraper in Python Kudo, T. (2005). MeCab : Yet Another Part-of-Speech and Morphological Analyzer Latour, Bruno (2007). Beware, your imagination leaves digital traces. Times Higher Literary Supplement 6th Lyotard, Jean-François (1991). Phenomenology. Presses Universitaires de France McRobbie, Angela (2009). The Aftermath of Feminism: Gender, Culture, and Social Change. Sage Publications Ltd. Miliband, Ralph (1973). The state in capitalist society. London : Quartet Books Nozza et al., LTEDI (2022). Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals Pavlova, A., Berkers, P. (2022). Mental Health as defined by Twitter: Frames, emotions, stigma (37(5), 637-647). Health Communication Quoc V. Le, Tomas Mikolov (2014). Distributed Representations of Sentences and Documents. https://doi.org/10.48550/arXiv.1405.4053 Stuart Hall (1997). Representation: Cultural Representations and Signifying Practices. Sage Publications, Inc; Open University Press Törnberg, A., Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum (27(4), 401–422). Discourse and Society Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Aidan, Kaiser, Polosukhin (2017). Attention Is All You Need. CoRR