Korean Stereotype Content Model: Translating Stereotypes Across Cultures

Michelle YoungJin Kim, Kristen Marie Johnson
Michigan State University, East Lansing, MI, USA
{kimmic16, kristenj}@msu.edu

Abstract

To address bias in language models, researchers are leveraging established social psychology research on stereotyping. This interdisciplinary approach uses frameworks like the Stereotype Content Model (SCM) to understand how stereotypes about social groups are formed and perpetuated. The SCM posits that stereotypes can be defined based on two dimensions: warmth (intent to harm) and competence (ability to harm). This framework has been applied in NLP for various tasks, including stereotype identification, bias mitigation, and hate speech detection. While the SCM has been extensively studied in English language models and Western cultural contexts, its applicability as a cross-cultural measure of stereotypes remains an open research question. This paper explores the cross-cultural validity of the SCM by developing a Korean Stereotype Content Model (KoSCM). We create a Korean warmth-competence lexicon through machine translation of existing English lexicons, validated by an expert translator, and utilize this lexicon to develop a labeled training dataset of Korean sentences. This work presents the first extension of SCM lexicons to a non-English language (Korean), aiming to broaden understanding of stereotypes and cultural dynamics.

1 Introduction

With the growing emphasis on Responsible and Fair AI, researchers are increasingly addressing the challenge of bias in language models. As this area of study within natural language processing (NLP) is still in its formative stages, scholars are drawing upon the insights of social psychology, a field that has extensively examined bias and stereotypes for many years. By employing the concept of stereotyping, researchers aim to elucidate the underlying mechanisms by which individuals form stereotypes about social groups. A prominent framework in this investigation is the Stereotype Content Model (SCM), which offers critical insights into understanding and addressing stereotypes.

The SCM (Fiske et al., 2002) identifies two key dimensions of stereotypes: warmth and competence. When individuals encounter members of an out-group, SCM suggests they instinctively ask two questions: Do these individuals intend to harm me? And are they capable of causing me harm? The first inquiry assesses warmth (characteristics such as friendliness, good-naturedness, sincerity, and warmth), while the second evaluates competence (traits including capability, skillfulness, confidence, and effectiveness). SCM has been utilized in NLP to develop a computational model for identifying stereotypes (Fraser et al., 2021; Herold et al., 2022; Nicolas and Caliskan, 2024; Schuster et al., 2024; Fraser et al., 2024; Mina et al., 2024), to reduce stereotypical bias in language models (Omrani et al., 2023; Ungless et al., 2022; Gaci et al., 2023), and to enhance hate speech detection (Jin et al., 2024).

There has been substantial research into the application of the SCM in language models, particularly regarding English texts and the stereotypes present in English-speaking cultures. However, the computational analysis of stereotypes in other languages and cultures is underexplored. This raises an important research question: Can the computational approach to the SCM be considered a pancultural measure of stereotypes across diverse societies?

In this paper, we explore the potential of the SCM as a pancultural tool by developing a Korean Stereotype Content Model (KoSCM). We begin by curating a Korean dictionary containing warmth-competence seed words. We translate existing English warmth-competence lexicons into Korean using a machine translation model, subsequently validating this translation with an expert translator. The translated lexicons are then utilized to create the training dataset for the KoSCM.

Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), pages 59–70

This dataset consists of sentences containing the warmth-competence seed words and two labels: warmth and competence directions.

We evaluate KoSCM by applying the model to do a stereotype analysis on social groups. We perform a stereotype analysis on social groups of age, gender, and religion in Korean texts. We observe whether the computational analysis aligns with and validates the social psychology study (Fiske et al., 2002; Cuddy et al., 2009). Further, we investigate the potential of SCM as a computational method for different languages and cultures. Based on the social psychology theory, we test the three hypotheses of the SCM: (1) the two dimensions hypothesis, (2) the ambivalent stereotypes hypothesis, and (3) the social structural correlates hypothesis. To the best of our knowledge, this is the first attempt to expand the SCM lexicons to a different language. Through this study, we aim to provide valuable insights that expand our understanding of stereotypes and cultural dynamics.

Our contributions are as follows:

- We develop a stereotype analysis model in Korean by curating warmth-competence seed words in Korean and generating training data to map texts to warmth-competence dimensions.
- We propose a social psychology-grounded framework for expanding the Stereotype Content Model to other languages and cultures.

2 Background and Related Work

In this section, we explore the concept of stereotyping. We begin by examining the definitions and research surrounding stereotypes in social psychology (§2.1). Subsequently, we discuss how NLP researchers have utilized findings from social psychology to detect and evaluate stereotypes within data and models (§2.2).

2.1 Stereotyping in Social Psychology

Stereotyping is a cognitive process in which specific attributes are overly generalized to entire social groups. It is a ubiquitous phenomenon that contributes to the perpetuation of social inequalities. When specific qualities are attributed to entire groups, it reinforces existing power dynamics and legitimizes discriminatory practices.

The perpetuation of stereotypes leads to profound consequences, such as the marginalization of certain groups, increased social inequalities, and significant psychological effects on individuals (Timmer, 2011). Marginalization happens when stereotypes justify the exclusion of specific groups from social, economic, and political opportunities. The increase in social inequalities is further fueled by the distribution of resources in ways that uphold existing power dynamics. Additionally, the internalization of stereotypes can severely affect individuals psychologically, undermining their mental well-being and self-image.

Social stereotypes are complex and multifaceted constructs that influence social perception and interaction. Traditional approaches to understanding stereotypes have relied on simplistic categorizations, such as positive or negative. However, the Stereotype Content Model (SCM) (Fiske et al., 2002; Fiske, 2018) offers a more nuanced framework for understanding social stereotypes. The SCM posits that social perception is guided by two fundamental dimensions: warmth and competence. Warmth refers to the perceived intentions and friendliness of a group, while competence refers to the perceived abilities and effectiveness of a group. These dimensions are orthogonal, allowing for the possibility of positive stereotypes along one dimension and negative stereotypes along the other.

A natural follow-up question for researchers is whether these stereotype studies can be generalized across cultures. Given that stereotypes arise from fundamental human phenomena—namely, the need to distinguish between "friends" and "foes" and the ubiquity of hierarchical status differences and resource competition—it is reasonable to assume that these principles are universally applicable.

To investigate this hypothesis, Cuddy et al. (2009) conducted a cross-cultural study spanning seven European (individualist) and three East Asian (collectivist) nations. The findings suggest that the SCM framework is effective across various cultures, reliably indicating group stereotypes based on structural connections with other groups. Using the SCM, the researchers observed parallels in the basic structures of intergroup relations. Building on this study, we expand the computational social study of SCM from English to Korean, leveraging a computational approach to validate the findings of the social psychology study.

2.2 Stereotype Content Model in NLP

The increasing prevalence of NLP models in various applications has raised concerns about the perpetuation of stereotypical biases in AI systems. Social psychological theories present valuable frameworks for understanding and addressing these biases. Consequently, recent studies have applied established social psychological theories to analyze biases in NLP models. In particular, research has concentrated on stereotype dimensions identified by these theories, notably the SCM.

The SCM has been extensively employed in various NLP applications to identify and mitigate stereotypical biases. For instance, researchers have utilized the SCM to detect stereotype subspaces in word embeddings (Fraser et al., 2021) and debias models by removing stereotype dimensions from the embedding space (Ungless et al., 2022; Omrani et al., 2023). Moreover, the SCM has been applied to assess benchmark datasets for bias (Fraser et al., 2021), examine how NLP models relate SCM dimensions to marginalized groups (Herold et al., 2022; Mina et al., 2024), and develop metrics to investigate biases across demographic and intersectional groups (Cao et al., 2022). Recent studies have further refined the SCM by exploring the construct differentiability of direction and representativeness for warmth and competence dimensions (Nicolas and Caliskan, 2024) and fine-graining stereotype dimensions into six psychologically-motivated categories to study occupation-related stereotypes (Fraser et al., 2024).

In recent years, researchers in NLP have expanded the study of bias and fairness to include non-English languages such as Arabic, Bengali, Chinese, Dutch, French, German, Hindi, Japanese, Korean, Spanish, and Telugu (Zhou et al., 2019; Chávez Mulsa and Spanakis, 2020; Kurpicz-Briki, 2020; Lauscher et al., 2020; Liang et al., 2020; Moon et al., 2020; Pujari et al., 2020; Takeshita et al., 2020; Zhao et al., 2020; Malik et al., 2021; Jeong et al., 2022), mirroring developments in social psychology. Bhutani et al. (2024) have expanded the number of languages by releasing a multilingual stereotype dataset that includes 20 languages across 23 regions. Acknowledging that biases are influenced by societal constructs, socio-cultural structures, and historical contexts, researchers are also seeking to adopt a more holistic approach to NLP fairness by taking the geo-cultural context into consideration (Sambasivan et al., 2021; Bhatt et al., 2022). The SeeGULL dataset (Bhutani et al., 2024) includes Korean but differs from our work in that it consists of pairs of associations between an identity term and an attribute generated by a language model. In contrast, our dataset and method are based on stereotyping theory from social psychology, utilizing seed words to identify stereotypes. This approach allows for broader applicability to various identity terms and social groups.

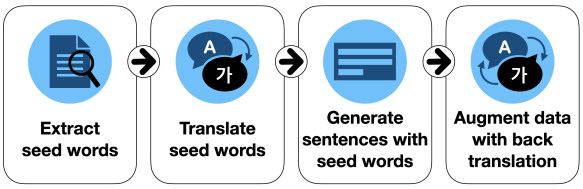

Figure 1: **Stereotype Translation Framework.** This figure illustrates the four steps for generating the data for KoSCM.

3 Translating Stereotype

This section presents a framework for expanding the SCM to a different language. As shown in Figure 1, we adopt the four steps to translate English SCM to Korean and create the dataset for KoSCM [1].

Step 1. Extract seed words The first step is to extract seed words for the stereotype content dictionary (Nicolas et al., 2019). The stereotype content dictionary is a collection of theory-driven seed words used to measure sociability, morality/trustworthiness, ability, assertiveness/dominance, status, political beliefs, and religious beliefs in relation to social groups. The list contains 341 words with their respective theoretical direction—either high or low—on their relevant dimension.

From the list, we select seed words that reflect warmth and competence dimensions. Specifically, words representing sociability and morality measures are classified as warmth seed words, and those related to ability and agency are categorized as competence seed words. There are a total of 157 seed words associated with the warmth dimension and 128 for the competence dimension. Each seed word is labeled with a direction within its respective dimension. For example, the word "warm" is a high-direction seed word in the warmth dimension, whereas "cold" represents a low-direction

[1]The dataset is available in github.com/MSU-NLP-CSS/KoSCM.

Dim	Dir	#	Example
W	high	75	친절한friendly, 호감이 가는likable
	low	82	불친절한unfriendly, 냉담한cold
C	high	68	유능한competent, 영리한clever
	low	60	무능한incompetent, 멍청한stupid

Table 1: **Statistics of Korean Seed Words.** The table shows statistics of translated seed words for KoSCM. The first column denotes dimensions: warmth and competence. The second column indicates a direction in each dimension. The next column lists the number of data points, while the final column provides examples of seed words in Korean.

seed word within the same dimension. Similarly, the word "competent" is an example of a high-direction seed word in the competence dimension, while "incompetent" is classified as having low direction in that dimension.

Step 2. Translate seed words Next, the extracted seed words are translated into Korean. The first step of translation is to adopt a machine translation model. We choose Naver Papago [2], one of the most popular Korean-English AI translators in Korea, to translate English seed words to Korean. Afterward, we validate the translation with an expert translator. The translator is asked to validate the translation by answering the following questions: (1) Is the translation grammatically correct (e.g., a noun is translated as a noun)? (2) Is a word translated into a distinct word (i.e., no recurrence in the translated list)? Through validation, we verify 285 Korean seed words labeled with stereotype dimension and direction in their corresponding dimension. See Table 1 for statistics and examples of seed words.

Step 3. Generate sentences with seed words With the translated stereotype seed words, we generate sentences based on a template. Similar to May et al. (2019), sentences are generated by inserting individual seed words from the list of Korean stereotype words into simple templates such as "그 사람은 <seed word> 사람이다" (That person is a[n] <seed word> person). The templates are selected according to the part-of-speech (POS) tagging of the seed words. Further, The template words are chosen carefully to prevent the generated sentences from referencing specific social groups. For example, the pronouns "he" and "she" indicate a person's gender. We intentionally refrain from using these pronouns as subjects because we aim to create a dataset centered on understanding the dimensions of warmth and competence. For more details, see Appendix A.

Step 4. Augment data with back-translation To tackle the limitation of available Korean seed words and address challenges associated with low-resource scenarios, we utilize data augmentation. Sentences generated in Step 3 are augmented using back-translation (Sennrich et al., 2016; Domhan and Hieber, 2017; Belinkov and Bisk, 2018). Back-translation generates paraphrases by leveraging translation models. Initially, a text is translated into another language (forward translation) and then translated back into the original language. This process creates paraphrased sentences, introducing greater variety by allowing for diverse choices in terminology and sentence structure. While the content remains intact, stylistic features that reflect the author's specific traits may be adjusted or omitted during translation.

For our dataset, we first translate the Korean sentences from Step 3 into English and then translate them back into Korean. We use the No Language Left Behind model (Team et al., 2022), a multilingual model that supports translation for 202 languages, for the back-translation step. This model is selected for two key reasons. Firstly, it was designed to assist with low-resource language translations. Secondly, it supports both Korean and English languages. As a result of the back-translation, we obtain a dataset containing 3,420 sentences.

4 Korean Stereotype Content Model

In this section, we detail how the KoSCM dataset, collected through the four steps of the stereotype translation framework, is utilized to build the SCM model. By fine-tuning a model with the dataset, we build KoSCM, which predicts the warmth and competence scores of given Korean sentences.

4.1 Method

We suggest a systematic method to develop a SCM model specific to the language model employed. We introduce two SCM classifier frameworks: the first is designed for embedding models like BERT (Devlin et al., 2019), which excel in processing context-rich information, while the second framework targets large language models (LLMs), leveraging their expansive capabilities in understanding and generating human-like text.

[2] https://papago.naver.com/

POS	Template	English Translation
NOUN	[SUBJECT]은/는 <seed word>이/가 있다.	[SUBJECT] has <seed word>.
ADJECTIVE	[SUBJECT]은/는 <seed word> 사람이다.	[SUBJECT] is a[n] <seed word> person.

Table 2: **Templates for Sentence Generation.** The table shows two different sentence templates based on the POS tagging of a seed word. English versions of Korean templates are provided for reference.

The first framework utilizes an embedding model as its base, adding two classifiers on top. Each classifier predicts the directions of a given text in the warmth and competence dimensions, respectively. Namely, the two classifiers perform multi-class classification, identifying one of three potential directions: high, low, or none. Formally, we use two classifiers, f_w and f_c, to predict warmth and competence directions, respectively. These prediction tasks are formulated as multi-class classification problems with cross-entropy losses, $\mathcal{L}_w$ and $\mathcal{L}_c$; $\mathcal{L}_w = -\sum_{t \in D} W(t) \cdot \log(f_w(t))$ and $\mathcal{L}_c = -\sum_{t \in D} C(t) \cdot \log(f_c(t))$, where t is a text in the dataset D, and $W(t)$ and $C(t)$ are warmth and competence directions of the text t. The final loss of the model is the sum of the prediction losses: $\mathcal{L} = \alpha\mathcal{L}_w + \beta\mathcal{L}_c$. where α and β are hyperparameters.

As for the second framework, we implement in-context learning with LLMs such as Llama. (Touvron et al., 2023). A small number of samples selected from the KoSCM dataset is provided to an LLM in the prompt. We select four samples for our experiment. In-context learning performance is sensitive to factors such as the selection and order of demonstration examples (Dong et al., 2024). To address this, we test the model using two approaches: first, by utilizing carefully selected samples based on a distance metric, and second, by randomly selecting samples from the KoSCM dataset to eliminate selection bias. The prompt utilized for the experiment is displayed in Table 3.

4.2 Experimental Setup

We evaluate the proposed methods on the following models:

- Multilingual BERT (mBERT): A masked language model pre-trained on 104 languages with 110M parameters. We use `BERT-Base, Multilingual Cased`, which is recommended for languages with non-Latin alphabets.
- Multilingual Sentence Transformer (mST): A modification of the Sentence Transformer (Reimers and Gurevych, 2019) aimed at adapting it for a new language using multilingual knowledge distillation. The pre-trained model `paraphrase-multilingual-MiniLM-L12-v2` is used for the experiment.
- Korean Llama (kLlama): Llama 3.2 fine-tuned with Korean texts using instruction tuning. The model `Bllossom/llama-3.2-Korean-Bllossom-3B` (Choi et al., 2024) uploaded to the HuggingFace models is utilized.
  - $\text{kLlama}_{\text{random}}$: kLlama evaluated using randomly selected samples. In all few-shot learning experiments with kLlama, we utilize four samples.

All models are implemented with PyTorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2020). For training, we use the Adam optimizer (Kingma and Ba, 2015) and set the learning rate from $\{1e-05, 2e-05\}$, an epoch as 10, a dropout rate of 0.3, and hyperparameters α and β as 1 . All experiments are conducted on a Nvidia Quatro RTX 5000, 16 GB memory GPU in a machine with Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz.

4.3 Evaluation

Using our proposed method, we evaluate how effectively models trained on the KoSCM dataset predict stereotypes. To assess the effectiveness of these models, we measure the accuracy of warmth and competence prediction on the test data. The results are presented in Table 4, which illustrates both the average and standard deviation of the prediction accuracies.

In our analysis, we observe a significant gap between the first and second frameworks. The first framework, which utilizes an embedding model, demonstrates competitive performance with high prediction accuracies for both warmth and competence. Notably, mBERT is the best-performing

Prompt
{"role": "system", "content": "당신은 편견 판별 서비스를 제공합니다. 주어진 글의 온기와 능력 점수를 -1, 0, 1 중에 하나로 매깁니다. You provide a stereotype detection service. Score the warmth and competence of a given sentence in one of the three scores: -1, 0, or 1."}, {"role": "user", "content":<sentence>}, {"role": "assistant", "content": "온기 <warmth>, 능력 <competence>"} ⟩× num. samples

Table 3: **Prompt for Few-Shot Learning.** The table above shows the prompt used for few-shot learning with LLMs. After the system prompt, a sentence is provided as the user prompt, and the assistant predicts warmth and competence directions. The format of the last two prompts can be duplicated based on the chosen number of samples.

Model	Warmth	Competence
mBERT	0.9230 (0.006)	0.9376 (0.005)
mST	0.9172 (0.010)	0.9240 (0.006)
kLlama	0.5376 (0.012)	0.5889 (0.002)
$\text{kLlama}_{\text{random}}$	0.5002 (0.003)	0.5031 (0.005)

Table 4: **Evaluation of KoSCM.** The evaluated performance of the three selected models is displayed. The average accuracy of warmth and competence predictions is presented. The standard deviation is indicated within the parentheses.

model, achieving accuracies of 0.9230 for warmth and 0.9376 for competence prediction. In contrast, the second framework designed for LLMs exhibits much lower performance, with accuracy scores of around 0.5 across all cases. The performance is particularly poor when using prompts with randomly chosen samples for each prediction. Although carefully curating the samples does enhance the performance slightly, the accuracies still remain modest at 0.5376 and 0.5889 for warmth and competence prediction, respectively. We surmise that the performance may have been affected by the limited data distribution of kLlama, as research shows that the diversity of pretraining corpora significantly impacts in-context learning performance (Shin et al., 2022; Raventós et al., 2023).

To evaluate the generalization capacity of the KoSCM, we conduct additional tests to determine whether the computational analysis aligns with and supports the results obtained from the SCM survey conducted in South Korea (Cuddy et al., 2009). We leverage the best-performing model, mBERT, from the evaluation to measure the stereotype directions of various social groups. For this analysis, we utilize the Korean Offensive Language Dataset (KOLD) dataset (Jeong et al., 2022). The dataset consists of comments collected from news articles and videos, with labels indicating group information among the 21 target group labels tailored to Korean culture. We use this group information for analysis. From the existing group labels, we select 19 groups that intersect with the 23 social groups in the survey.

We assess the warmth and competence directions of texts that comment on a target group and calculate the average warmth and competence directions. Then, the groups are clustered using hierarchical cluster analysis, following the method of Cuddy et al. (2009). The results are illustrated in the SCM dimension in Figure 2. In general, we observe a significant overlap between our results and the survey findings. For instance, social groups such as "women," "blue-collar," and "Protestants" fall into the low-competence/high-warmth cluster, while groups like the "poor" and "unemployed" are categorized as low-competence/low-warmth. However, there are also outliers. For example, the group "public functionaries" is positioned in the high-competence/high-warmth cluster in our figure, but it falls within the low-competence/low-warmth cluster in the survey plot. This discrepancy may come from the lack of data since outliers like "public functionaries" have insufficient data, with only nine text samples contributing to their classification.

5 SCM as a Pancultural Tool

In this section, we explore the applicability of the proposed computational method of the SCM for analyzing stereotypes across various languages and cultures. Based on the survey in Cuddy et al. (2009), we examine three key hypotheses of SCM: (1) the two dimensions hypothesis, (2) the ambivalent stereotypes hypothesis, and (3) the social structural correlates hypothesis.

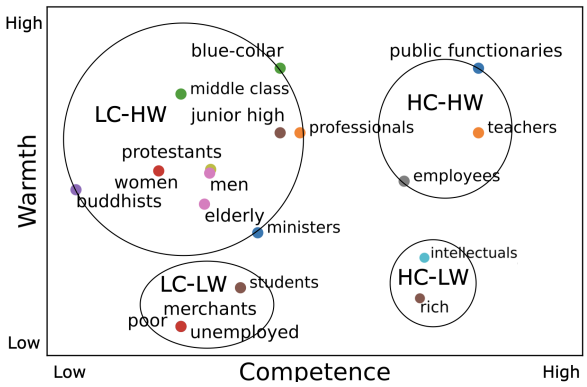

Figure 2: **Stereotypes of Groups Projected to the SCM dimension.** Social groups are mapped onto the SCM dimension according to their predicted warmth and competence by KoSCM.

Two Dimensions Hypothesis The first hypothesis posits that (1) within each sample, groups will be positioned along the dimensions of warmth and competence and that (2) based on their warmth and competence scores, groups will form multiple clusters, including some at both the high and low ends of each dimension. As shown in Figure 2, our results support this hypothesis, as groups are mapped along the warmth and competence dimensions. The figure reveals a structure that aligns with the SCM survey. Notably, we identify four distinct clusters that reflect both high and low scores on each dimension. Consistent with the survey findings, the largest cluster is the low-competence/high-warmth group, which encloses the majority of the sampled groups. Yet we observe that the high-competence/high-warmth cluster in the survey has a lower average warmth score compared to our findings. As discussed in Section 4, this dissimilarity may be attributed to outliers, such as the "public functionaries" category, which suffered from insufficient data.

Ambivalent Stereotypes Hypothesis This hypothesis proposes that (1) within any given sample, there will be significant variations in perceptions of warmth and competence across different social groups and that (2) it predicts that cluster analyses will reveal at least one high-competence/low-warmth cluster and one low-competence/high-warmth cluster. This indicates that numerous groups are characterized as being adept in one area—either warmth or competence—while being perceived as lacking in the other.

Figure 2 shows four distinct clusters at each end, which supports the hypothesis that the four clusters of stereotype content, defined within the warmth-competence space, have universal characteristics. We observe that the groups "women" and "elderly" fall within the low-competence/high-warmth group. This supports the theory that groups seen as gentle but useless—often associated with a "pitying" prejudice—frequently include traditional women and older people. These groups are often viewed as having strong communal traits but lacking agentic qualities, representing a significant stereotype identified in the existing literature. (Jackman, 1994; Glick and Fiske, 2001b,a). In contrast, another significant stereotyped group includes those seen as skilled yet dishonest. Our analysis emphasizes individuals labeled as "intellectuals" and "rich" in this group. It shows that "envious" prejudice frequently targets those considered alarmingly skilled yet untrustworthy (Glick and Fiske, 2001b,a; Fiske et al., 2002; Glick, 2002). This dynamic highlights the complex relationship between admiration and disdain influencing societal perceptions.

Social Structural Correlates Hypothesis The social structural correlates hypothesis suggests that (1) within each sample, perceived status is expected to positively correlate with competence and that (2) perceived competition is anticipated to negatively correlate with warmth. In the survey, participants are asked to evaluate the perceived status and competition of various social groups. As we cannot access the information of commentators in the KOLD dataset, we focus on validating the first part of the hypothesis by examining the relationship between perceived status and competence ratings.

In our analysis, we utilize average wage statistics as a measure of perceived status, recognizing that socioeconomic status is a complex, multidimensional construct influenced by various factors, with income being a key component (Havranek et al., 2015). Individuals with lower incomes often face a lack of economic resources, which leads to social disadvantages such as limited access to quality education, poor working conditions, housing insecurity, and living in unsafe neighborhoods. These factors collectively contribute to a lower perceived status within society. Thus, we use income information as a symbolic indicator of perceived status, emphasizing its significant effect on individuals' overall social standing.

The Korean Ministry of Employment and Labor publishes the Current Status of Wage Distribution

	status-competence corr.
koSCM	0.71
South Korea	0.64
Universal Average	0.79

Table 5: **The correlation between perceived status and competence.** The table displays the correlation coefficient between perceived status and competence

by Business Characteristics every year [3]. We reference the 2024 report to extract the average income across different social groups. This report offers average wage data categorized by labor industry, gender, and years of experience. Due to the ambiguity in categorizing jobs within non-occupational social groups like "intellectuals" and "rich," we exclude these groups from this analysis. The report includes gender data for all jobs, so the average income for each gender is computed to represent the perceived status of the groups "women" and "men."

Next, we calculate the correlation coefficient between the average wage and competence for the social groups. The correlation coefficient is computed as $\text{cov}(\text{wage}, \text{competence})/(\sigma_{\text{wage}} \cdot \sigma_{\text{competence}})$. As shown in Table 5, the calculated correlation value is 0.71, a positive correlation that supports the hypothesis. In the survey, South Korea has a correlation of 0.64, and the average of all 13 surveys shows a correlation of 0.79.

6 Conclusion

In this paper, we propose the Korean Stereotype Content Model (KoSCM), a theory-grounded stereotype model that adapts the existing SCM for the Korean language and culture. We develop a Korean warmth-competence lexicon by translating existing English lexicons and curating a Korean dictionary of seed words. This translated lexicon is used to train the KoSCM, a classification model for predicting directions in warmth and competence dimensions. Then, we utilize KoSCM to analyze stereotypes of age, gender, and religious groups in Korean texts, comparing the results to the social psychology survey. To test whether the computational approach of SCM can be applied cross-culturally, we examine three core hypotheses of the SCM: the two-dimensional structure of stereotypes, the presence of ambivalent stereotypes, and the relationship between stereotypes and social structure.

This study marks the first attempt to adapt the SCM to the Korean language, aiming to enhance the understanding of stereotypes across languages. In the future, we plan to expand our research by incorporating additional languages and utilizing the warmth-competence framework to develop an algorithm that can guide and transform stereotypes present in sentences.

Limitations

We recognize several limitations that may impact the validity of our findings. Despite our efforts to minimize authorial bias, there remains a possibility for such bias to influence both the experimental design and analysis. For example, the process of clustering social groups is inherently affected by the selection of hyperparameters, which can significantly alter the resulting clusters. Additionally, our decisions in curating prompts for sampling from the dataset and crafting the prompt texts introduce further elements of bias. Hence, these decisions may result in selection bias, which could ultimately impact the conclusions drawn from our study.

Furthermore, our data and experiments are limited by scale constraints. Unlike the abundance of resources available for English models and datasets, there is a significant lack of open-source Korean datasets and models, which has limited our efforts. This insufficient data may suggest that the models utilized in this research are not performing at the same level as their English counterparts. For instance, while conducting back-translation in the data curation process, we observed significant noise in the generated data, which might indicate the difficulties posed by limited resources.

Ethical Considerations

We curate and publish the KoSCM dataset, which is used for training and evaluating KoSCM. This dataset is based on a specific social psychology theory known as the SCM, meaning our research investigates stereotypes within this particular framework. As a result, our dataset and analysis do not encompass the complete range of perspectives on stereotypes. Therefore, we advise researchers utilizing the KoSCM dataset and the proposed translation framework to be mindful of these limitations and encourage them to explore additional methodologies to gain a more comprehensive understanding of stereotypes.

[3] Ministry of Employment and Labor website

We strongly recommend against using this research for harmful purposes, including the promotion and dissemination of stereotypical biases.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and Kyung-Tae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12514–12526, Torino, Italia. ELRA and ICCL.

Amy J. C. Cuddy, Susan T. Fiske, Virginia S. Y. Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, Tin Tin Htun, Hyun-Jeong Kim, Greg Maio, Judi Perry, Kristina Petkova, Valery Todorov, Rosa Rodríguez-Bailón, Elena Morales, Miguel Moya, Marisol Palacios, Vanessa Smith, Rolando Perez, Jorge Vala, and Rene Ziegler. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

S. T. Fiske, A. J. C. Cuddy, P. Glick, and J. Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902.

Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73. PMID: 29755213.

Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How does stereotype content differ across data sources? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2023. Societal versus encoded stereotypes in text encoders. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 46–53.

Peter Glick. 2002. Sacrificial lambs dressed in wolves' clothing: Envious prejudice, ideology, and the scapegoating of jews.

Peter Glick and Susan T Fiske. 2001a. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Peter Glick and Susan T. Fiske. 2001b. Ambivalent sexism. volume 33 of *Advances in Experimental Social Psychology*, pages 115–188. Academic Press.

Edward P. Havranek, Mahasin S. Mujahid, Donald A. Barr, Irene V. Blair, Meryl S. Cohen, Salvador Cruz-Flores, George Davey-Smith, Cheryl R. Dennison-Himmelfarb, Michael S. Lauer, Debra W. Lockwood, Milagros Rosal, and Clyde W. Yancy. 2015. Social determinants of risk and outcomes for cardiovascular disease. *Circulation*, 132(9):873–898.

Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 58–65, Dublin, Ireland. Association for Computational Linguistics.

Mary R Jackman. 1994. *The velvet glove: Paternalism and conflict in gender, class, and race relations*. Univ of California Press.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiping Jin, Leo Wanner, and Aneesh Moideen Koya. 2024. Disentangling hate across target identities. *Preprint*, arXiv:2410.10332.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. *Arbor-ciencia Pensamiento Y Cultura*.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. Socially aware bias measurements for hindi language representations. *CoRR*, abs/2110.07871.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario Mina, Júlia Falcão, and Aitor Gonzalez-Agirre. 2024. Exploring the relationship between intrinsic stigma in masked language models and training data using the stereotype content model. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 54–67, Torino, Italia. ELRA and ICCL.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Gandalf Nicolas, Xuechunzi Bai, and Susan Fiske. 2019. Automated dictionary creation for analyzing text: An illustration from stereotype content.

Gandalf Nicolas and Aylin Caliskan. 2024. Directionality and representativeness are differentiable components of stereotypes in large language models. *PNAS Nexus*, 3(11):pgae493.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie

Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2020. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 450–456, New York, NY, USA. Association for Computing Machinery.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Carolin M. Schuster, Maria-Alexandra Dinisor, Shashwat Ghatiwala, and Georg Groh. 2024. Profiling bias in llms: Stereotype dimensions in contextual word embeddings. *Preprint*, arXiv:2411.16527.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Barcelona, Spain (Online). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Alexandra Timmer. 2011. Toward an Anti-Stereotyping Approach for the European Court of Human Rights. *Human Rights Law Review*, 11(4):707–738.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

A Templates for Sentence Generation

In this section, we describe the details of the templates used for generating sentences in Section 3. The templates are curated based on the part-of-speech (POS) tagging of the seed words. The curated seed words contain noun and adjective tags. Based on those tags, we utilize the two templates in Table 2. The subject words for the templates are chosen carefully to ensure that the generated sentences do not contain information about specific social groups. For instance, the pronouns "he" and "she" indicate a person's gender. We chose to avoid using these pronouns as subjects because our objective is to develop a dataset focused on learning the dimensions of warmth and competence. The subject words used for the templates are: ["나 I", "너 You", "우리 We", "그 사람 A person", "저 사람 That person", "이 사람 This person"]. With the curated templates, a total of 1,710 sentences are generated. Here are sample sentences generated using the templates: "나는 능력이 있다. I have competence.", "그 사람은 친절한 사람이다. A person is a friendly person.".