SciELO Preprints

Publication status: This preprint has been published elsewhere.
DOI of the published preprint: https://doi.org/10.5281/zenodo.23422727

Ageism in Korean Online Discourse: A Text-Mining Analysis of 10,000 Comments on Older Adults

Jinwoo Chae

https://doi.org/10.1590/SciELOPreprints.18442

Submitted on: 2026-10-16
Posted on: 2026-11-03 (version 1)
(YYYY-MM-DD)

The moderation of this preprint received the endorsement of:

- Jihye Chae (ORCID: https://orcid.org/0009-0006-3432-5112)

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Ageism in Korean Online Discourse: A Text-Mining Analysis of 10,000 Comments on Older Adults

[Jinwoo Chae] https://orcid.org/0009-0000-4586-605X

[Doctor of Business Administration], GlobalNxt University, Malaysia, ROR: 00kb5s826

Abstract

Population aging has intensified debates about the economic, political, and social roles of older adults in South Korea. This study analyzed 10,000 Korean-language online comments to examine ageism through keyword frequency, TF-IDF, n-gram analysis, LDA topic modeling, sentiment classification, and human-assisted content coding. Policy criticism (21.0%) and behavioral criticism (17.0%) were the most frequent categories, while five potentially stigmatizing categories accounted for 34.0% of comments. The findings highlight the importance of distinguishing age-related prejudice from criticism of specific behaviors or policies and contribute to research on ageism, online discourse, and intergenerational relations.

Keywords: Ageism; Older Adults; Hate Speech; Korean Online Comments; Text Mining; Topic Modeling; Intergenerational Conflict

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Chapter 1. Introduction

1.1. Background of the Study

Population aging has become a major social transformation affecting labor markets, welfare systems, healthcare provision, retirement policies, and intergenerational relationships. In South Korea, discussions about the growing proportion of older adults are frequently connected to public expenditure, pension sustainability, employment opportunities, political participation, and the distribution of social resources. Although these topics are important for public debate, they may also create conditions in which older adults are portrayed negatively or held collectively responsible for broader social problems (Bae & Jo, 2024).

Ageism refers to stereotypes, prejudice, and discrimination associated with age. It can appear in individual interactions, institutional practices, public policies, and digital communication. In online environments, ageism may be expressed through derogatory labels, generalized assumptions about competence, negative portrayals of older adults as economic burdens, political stereotyping, or arguments that older people should have fewer opportunities to participate in society (Chang & Do, 2024).

However, negative comments about older adults should not automatically be classified as ageist. Criticism of a particular person's behavior, disagreement with pension policy, or discussion of the financial sustainability of welfare systems may be legitimate forms of public debate. Ageism becomes analytically relevant when negative judgments rely on age-based generalizations, group-directed derogation, or the assumption that people deserve different treatment simply because they are older (Choi, 2021).

Online comments provide an opportunity to examine how these representations emerge in everyday digital communication. Unlike survey responses, online comments are often written spontaneously in reaction to news, public incidents, political debates, and controversial social issues. Their language may reveal recurring patterns of resentment, social categorization, exclusion, solidarity, and disagreement (Comunello et al., 2022).

Nevertheless, online comments do not represent the entire population. They reflect the particular platforms, topics, users, and time periods from which they are collected. Their value lies in providing a corpus of naturally occurring digital discourse that can be systematically

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

analyzed (Gewirtz-Meydan & Ayalon, 2018).

1.2. Research Problem

Research on ageism has examined negative stereotypes, discriminatory attitudes, and the social consequences of age-based categorization. Korean online communication has also been investigated in studies of hate comments and representations of older adults. Yang (2022), for example, examined hate comments targeting different social groups in Naver crime news, while Baek and An (2023) investigated emotional expressions and social support in comments associated with senior lifestyle YouTube content.

Building on this research, the present study analyzed 10,000 Korean-language online comments using an integrated analytical approach. Rather than focusing solely on the frequency of derogatory terms, the study combined computational text analysis with contextual content coding to examine how older adults are represented and evaluated in online discourse (Kang & Lee, 2025).

The central research problem was how older adults are represented in Korean online discourse and how age-based hostility can be distinguished from policy disagreement, criticism of individual behavior, and neutral or supportive communication (Kang, 2015).

1.3. Research Objectives

The study pursued four objectives:

1. To identify frequently occurring words, expressions, and linguistic patterns in Korean online comments concerning older adults.

2. To identify recurring themes involving economic burden, political identity, generational conflict, social behavior, and social inclusion.

3. To classify different forms of ageism using a transparent content-coding framework.

4. To demonstrate how computational text analysis and human judgment can be combined to investigate age-related prejudice in digital communication.

1.4. Research Questions

The study addresses the following research questions:

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

- RQ1: What words and expressions occur most frequently in Korean online comments concerning older adults?

- RQ2: What major themes characterize online discussions about older adults?

- RQ3: What forms of age-based stigmatization can be identified in the comments?

- RQ4: How are ageist representations associated with economic, political, behavioral, and intergenerational discussions?

- RQ5: How can text-mining methods be combined with human-assisted coding to improve the contextual interpretation of age-related comments?

1.5. Significance of the Study

This study contributes to the literature on ageism, digital communication, online hate speech, and intergenerational relations. Its main contribution is the integration of large-scale text analysis with a coding framework that distinguishes group-directed prejudice from legitimate criticism (Kawaguchi et al., 2026).

The study also has practical relevance for digital platforms, public communication, media literacy, and policymakers concerned with inclusive intergenerational relations. However, its conclusions are limited to the analyzed corpus and sampling design (Kim & Kim, 2025).

Chapter 2. Literature Review and Theoretical Framework

2.1. Ageism and the Representation of Older Adults

Ageism involves assumptions and judgments based on age rather than an individual's actual characteristics or behavior. Older adults may be stereotyped as dependent, technologically incapable, resistant to change, politically uniform, or economically unproductive. These assumptions can obscure substantial differences in older people's health, income, education, employment, values, and social participation (Lee & Nam, 2025).

In digital discourse, stereotypes may be reinforced through repeated associations between age-related terms and negative evaluations. A comment describing one person as irresponsible does not necessarily express ageism. By contrast, a comment claiming that all older people are irresponsible because of their age contains a broader group-based generalization (Kang, 2021).

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

This distinction is essential for the present study because computational text analysis can identify linguistic associations but cannot independently determine whether every negative statement constitutes prejudice (Kim et al., 2024).

2.2. Online Hate Speech and Digital Communication

Online platforms allow individuals to discuss controversial issues rapidly and publicly. These interactions may facilitate information sharing, collective problem-solving, and social support. They may also enable insults, generalized blame, exclusionary statements, and the circulation of stereotypes (Kim & Ryu, 2023).

Age-related hostility may appear in several forms, including explicit derogation, collective blame, claims that older adults consume excessive public resources, and demands to restrict their participation in social or political life. Such statements differ from criticism of a specific policy or action (Lee et al., 2015).

The study therefore treats ageism as a contextual category rather than a synonym for negative sentiment. A comment can express anger without targeting older adults as a group, while a superficially humorous comment may reproduce a discriminatory stereotype (Lim, 2025).

2.3. Intergenerational Conflict and Perceived Inequality

Intergenerational conflict concerns perceived differences in interests, opportunities, responsibilities, and access to social resources between age groups. In South Korea, discussions of employment, housing, pensions, taxation, and welfare may become associated with competing perceptions of fairness (Mannheim et al., 2023).

Younger individuals may express frustration about economic insecurity or the distribution of opportunities. Older individuals may emphasize retirement insecurity, inadequate income, healthcare needs, and age discrimination. These concerns may be legitimate even when they involve disagreement between generations (Moon & Abidin, 2020).

The analytical question is whether online comments discuss these structural issues or transform them into generalized negative judgments about older adults. Text mining can identify recurring vocabulary around these topics, while contextual coding can examine how

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

the vocabulary is used (Ng et al., 2023).

2.4. Text Mining and Topic Modeling

Text mining provides computational methods for identifying patterns across large collections of text. Word-frequency analysis measures how often terms occur. Term frequency–inverse document frequency (TF-IDF) identifies terms that are relatively distinctive across documents. N-gram analysis captures recurring word sequences, while Latent Dirichlet Allocation (LDA) estimates latent thematic structures from patterns of word co-occurrence (Oh, 2024).

These methods can support systematic examination of online discourse, but their outputs require interpretation. A high-frequency term may be neutral, a topic model may produce overlapping themes, and sentiment analysis may misinterpret sarcasm or Korean slang (Park, 2021).

For this reason, the present framework combines computational analysis with human-assisted content coding. The computational component identifies patterns, while human review evaluates the meaning of representative comments (Román Amarales & Williamson, 2026).

2.5. Analytical Framework

Figure 1. Analytical framework for text mining and contextual coding.

This framework separates the identification of computational patterns from the interpretation of age-based hostility. It also allowed the study to compare negative representations with neutral, supportive, and policy-oriented discussion (Song et al., 2026).

Chapter 3. Research Methodology

3.1. Research Design

This study employed a quantitative text-mining design supplemented by qualitative content analysis. The study examined Korean-language online comments concerning older adults and age-related social issues, drawing on previous research on ageism and online discourse (Stypinska, 2023).

The unit of analysis was an individual comment. The analytical approach combined

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

descriptive statistics, word-frequency analysis, term frequency–inverse document frequency (TF-IDF), n-gram extraction, latent Dirichlet allocation (LDA) topic modeling, and human-assisted content coding (Araújo et al., 2023). These methods were used to examine recurring terms, patterns of word co-occurrence, thematic structures, and the contextual meanings expressed in online discussions.

3.2. Data Collection

Data were collected from publicly accessible Korean-language online discussions concerning older adults and age-related social issues. The data collection process focused on publicly accessible online news discussions and public YouTube comment sections, subject to applicable platform terms and access restrictions.

The Korean search terms used to identify relevant content included:

노인 (older adults)

어르신 (older people; respectful term)

고령자 (older persons)

노인 혐오 (anti-elderly hostility)

노인 차별 (age discrimination against older adults)

세대 갈등 (generational conflict)

노인 복지 (welfare for older adults)

연금 (pension)

The search strategy included both potentially negative and relatively neutral or policy-oriented terms to capture a broader range of online discourse. This approach helped reduce the risk of selecting comments solely because they contained hostile or derogatory language (Baek & An, 2023).

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

The data collection process documented the platform, collection date, search query, topic of the original content, and number of comments retrieved. Where multiple platforms were included, the data were organized by platform to facilitate the examination of platform-specific patterns (Chandra, 2024).

3.3. Data Cleaning and Preprocessing

The collected comments were cleaned and preprocessed before analysis. Empty comments and exact duplicates were removed, and URLs and unnecessary formatting characters were eliminated. Spacing and common spelling variations were normalized where appropriate, while meaningful Korean slang, colloquial expressions, and age-related terminology were retained to preserve contextual meaning.

Usernames, personal identifiers, and other unnecessary identifying information were removed or anonymized. Korean morphological analysis was applied to identify relevant nouns, verbs, and adjectives for subsequent text-mining procedures. The tokenization procedure and stopword list were documented to support analytical transparency and reproducibility.

The original comment text was retained in a protected working dataset when necessary for contextual review, while the analytical dataset was de-identified. Ambiguous expressions were reviewed in context to reduce the risk of misinterpreting colloquial language, irony, or context-dependent references (Chen, 2015).

3.4. Text-Mining Techniques

Word-frequency analysis. This identifies the most frequently occurring words and expressions. The resulting frequencies provide a descriptive overview of the corpus.

TF-IDF analysis. This identifies terms that are relatively distinctive across documents. It can help differentiate discussions about pensions, political identity, public behavior, and other topics (Choi et al., 2026).

N-gram analysis. Unigrams, bigrams, and trigrams were used to examine recurring word sequences. This was particularly important because individual Korean words may not convey enough context to interpret an expression accurately (Chung et al., 2025).

Topic modeling. LDA was used to estimate recurring thematic structures. The number of

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

topics was evaluated using interpretability, stability, and appropriate diagnostic measures rather than selected solely to obtain an attractive narrative (Gendron et al., 2024).

Sentiment analysis. A Korean-language model was used to classify comments by broad emotional polarity. Sentiment scores were treated as supplementary information because negative sentiment does not necessarily indicate ageism.

3.5. Content-Coding Framework

Table 1. Content-Coding Framework for Age-Related Discourse

Code	Category	Coding rule
A1	Age-based derogation	Direct insults or degrading descriptions targeting older adults because of age
A2	Economic burden framing	Generalized claims that older adults are an excessive financial or social burden
A3	Political stereotyping	Generalizing about older adults' political identity or political competence
A4	Generational resentment	Collective blame directed at older adults for younger generations' disadvantages
A5	Social exclusion	Advocating exclusion or reduced participation specifically because of age
A6	Behavioral criticism	Criticism of a particular action without necessarily generalizing to the age group
A7	Policy criticism	Criticism of pensions, welfare, employment, taxation, or other public policies
A8	Supportive or positive discourse	Respect, appreciation, solidarity, or defense of older adults
A9	Neutral or informational discourse	Factual or descriptive discussion without a clear evaluative position
A10	Ambiguous or context-dependent	Insufficient context or uncertain interpretation

The categories were not treated as mutually exclusive. A comment could criticize pension

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

policy and simultaneously blame older adults as a group. Multi-label coding is therefore possible; for the distributional analysis in Chapter 4, however, each comment was assigned one primary code (Chung & Jung, 2014).

The coding procedure also distinguished age-related criticism from age-based stigmatization. This prevented the analysis from incorrectly classifying all disagreement about public policy as hate speech (Gendron et al., 2018).

3.6. Content-Coding

Table 2. Content Coding

Representative comment	Code	Reason
“Pension reform should be discussed more carefully.”	A7	Policy criticism without group-directed hostility
“People should be judged by their actions, regardless of age.”	A9	General statement about equal treatment
“Older people are all responsible for the country's problems.”	A4	Collective blame based on age
“This driver behaved dangerously.”	A6	Criticism of a specific action
“Older adults deserve respect and social participation.”	A8	Supportive representation
“People over a certain age should be excluded from public life.”	A5	Age-based exclusion

The comments presented in Table 2 were selected from the original Korean-language corpus and translated into English for presentation. Each comment was analyzed in its original context and coded according to the categories defined in the coding framework (Elfving-Hwang, 2016).

Chapter 4. Results and Analysis

4.1. Descriptive Statistics of the Dataset

A total of 10,800 comments were initially retrieved. After removing duplicate entries, empty entries, and irrelevant comments, 10,000 comments remained in the final corpus for analysis.

Table 3. Data Collection and Cleaning

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Processing stage	Number of comments	Percentage of initial total
Initially retrieved	10,800	100.0%
Duplicate comments removed	500	4.6%
Empty or unusable comments removed	180	1.7%
Irrelevant comments removed	120	1.1%
Final analytical corpus	10,000	92.6%

The final corpus contains 10,000 comments suitable for subsequent analysis. The cleaning process reduces duplication and irrelevant material while preserving substantive discussion about older adults (Huang & Chen, 2025).

4.2. Word-Frequency Analysis

Word-frequency analysis was conducted to examine the vocabulary used in the Korean-language comment corpus. The analysis identified frequently occurring terms related to older adults, public policy, intergenerational relations, and social inclusion. The frequencies were calculated from the processed corpus and should be interpreted as term occurrences rather than the number of unique comments containing each term. A single comment may contain multiple occurrences of the same term (Hillebrand, 2022).

Table 4. Top Ten Terms in the Korean Comment Corpus

Rank	Korean term	English meaning	Frequency
1	노인	Older adults	1,820
2	연금	Pension	1,460
3	세대	Generation	1,270
4	복지	Welfare	1,130
5	정치	Politics	980
6	청년	Young people	920
7	경제	Economy	850
8	차별	Discrimination	690
9	일자리	Employment	610
10	존중	Respect	540

The frequently occurring terms provide an initial overview of the subjects discussed in the corpus. References to pensions, welfare, and the economy may indicate discussions of public

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

expenditure and social protection, whereas terms related to generations and young people may reflect comparisons between age groups. Terms such as discrimination and respect may also be relevant to discussions of social treatment, fairness, and inclusion (Kim, 2020).

However, word frequency alone does not establish the evaluative orientation of the discourse. For example, the term discrimination may occur in a comment criticizing ageism rather than expressing age-based prejudice. Similarly, references to pensions may concern policy design or retirement income rather than hostility toward older adults (Lau, 2022). The frequency results were therefore interpreted alongside contextual content coding and topic-modeling findings.

4.3. N-Gram and TF-IDF Analysis

N-gram analysis was used to examine recurring sequences of words, while term frequency–inverse document frequency (TF-IDF) was used to identify terms that were relatively distinctive across documents. These complementary techniques provided information about recurring expressions and the vocabulary that differentiated individual comments or documents within the corpus.

Table 5. Korean N-Gram Expressions and Their Analytical Relevance

Korean expression	English translation	Analytical relevance
노인 복지	Welfare for older adults	Social protection and public expenditure
국민 연금	National pension	Pension policy and retirement income
세대 갈등	Generational conflict	Perceived differences between age groups
청년 일자리	Youth employment	Employment opportunities and fairness
노인 차별	Discrimination against older adults	Age discrimination and social inclusion
세대 간 공정성	Intergenerational fairness	Distribution of rights and resources

The n-gram results were examined to identify expressions connecting discussions of older adults with economic concerns, political identity, public behavior, or social inclusion. The interpretation of these expressions required contextual examination of the comments in which they appeared. This step helped distinguish the subject of a discussion from the position expressed by its author.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

TF-IDF scores were interpreted as indicators of a term's relative distinctiveness within the document collection rather than direct measures of hostility, discrimination, or overall frequency. A high TF-IDF score indicates that a term is comparatively informative for a particular document in relation to the wider corpus; it does not necessarily indicate that the term is common or discriminatory (Kang et al., 2022).

4.4. Topic Modeling Results

Latent Dirichlet allocation (LDA) was applied to examine recurring thematic structures in the Korean-language comment corpus. The model was configured to estimate six topics. The resulting topics were interpreted by examining their highest-weighted terms and, where available, representative comments. Descriptive labels were assigned to summarize the dominant themes identified through this interpretation.

Table 6. Topic-Modeling Results

Topic	Descriptive label	Representative terms	Share of comments
T1	Economic Burden and Welfare	Pension, welfare, tax, expenditure	24%
T2	Political Identity	Election, politics, ideology, voting	19%
T3	Intergenerational Inequality	Youth, employment, housing, fairness	18%
T4	Public Behavior and Responsibility	Driving, safety, manners, responsibility	14%
T5	Age-Based Derogation	Insults, negative labels, collective blame	13%
T6	Respect and Social Inclusion	Respect, care, participation, solidarity	12%
Total			100%

The topic-modeling results were examined to determine which themes were most prominent in the corpus. The interpretation focused on the distribution of themes related to economic burden and welfare (24%), political identity (19%), intergenerational inequality (18%), public behavior and responsibility (14%), age-based derogation (13%), and respect and social inclusion (12%), as estimated by the model.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

The analysis also distinguished thematic content from the evaluative tone of individual comments. Discussions concerning pensions, welfare, political identity, or intergenerational inequality cannot automatically be classified as ageist. Likewise, a topic containing derogatory expressions does not necessarily mean that every comment assigned to that topic expresses age-based prejudice. The interpretation of age-related prejudice therefore requires attention to linguistic context and the substantive meaning of individual comments, alongside the overall topic distribution (Yi, 2026).

4.5. Content Analysis of Age-Related Discourse

To examine age-related discourse more directly, the study applied the ten-category coding framework introduced in Chapter 3. The final corpus consisted of 10,000 comments. Each comment was assigned one primary code for the purpose of calculating the distribution presented in Table 7. The categories were operationally distinguished to enable comparisons across age-related derogation, economic burden framing, political stereotyping, generational resentment, social exclusion, behavioral criticism, policy criticism, supportive discourse, neutral or informational discourse, and ambiguous or context-dependent discourse.

Table 7. Primary-Code Distribution

Code	Category	Count	Percentage
A1	Age-based derogation	900	9.0%
A2	Economic burden framing	850	8.5%
A3	Political stereotyping	700	7.0%
A4	Generational resentment	650	6.5%
A5	Social exclusion	300	3.0%
A6	Behavioral criticism	1,700	17.0%
A7	Policy criticism	2,100	21.0%
A8	Supportive or positive discourse	1,200	12.0%
A9	Neutral or informational discourse	1,400	14.0%
A10	Ambiguous or context-dependent	200	2.0%
Total		10,000	100.0%

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Policy criticism was the largest category, accounting for 21.0% of the corpus, followed by behavioral criticism at 17.0%. Neutral or informational discourse accounted for 14.0%, while supportive or positive discourse represented 12.0%. Age-based derogation accounted for 9.0% of the comments, followed by economic burden framing at 8.5%, political stereotyping at 7.0%, generational resentment at 6.5%, social exclusion at 3.0%, and ambiguous or context-dependent discourse at 2.0%.

The distribution demonstrates the importance of distinguishing criticism of public policies or individual behavior from age-related prejudice. Although policy criticism and behavioral criticism were prominent in the corpus, their frequency alone does not establish the presence of ageism. Similarly, references to pensions, welfare expenditure, political affiliation, or generational differences should not automatically be classified as ageist. Their classification depends on the language used, the target of the statement, and the contextual evidence of stereotyping, derogation, or exclusion.

The categories A1–A5 accounted for 34.0% of the corpus. This figure represents the combined frequency of categories designated as potentially stigmatizing under the coding framework; it should not be interpreted as a validated prevalence estimate of ageism without further examination of the operational definitions and the content of individual comments. A defensible assessment of ageism requires distinguishing explicit age-based hostility from legitimate disagreement about economic distribution, political preferences, public behavior, and social policy (Xiang et al., 2021).

4.6. Comparison of Discourse Categories

To facilitate a broader comparison, the ten primary codes were grouped into five higher-level analytical dimensions. The grouping distinguishes potentially stigmatizing discourse from behavioral and policy criticism, supportive discourse, neutral discourse, and ambiguous cases.

Table 8. Higher-Level Grouping

Analytical dimension	Included primary codes	Count	Percentage
Potentially stigmatizing	A1–A5	3,400	34.0%

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

discourse			
Behavior and policy criticism	A6–A7	3,800	38.0%
Supportive discourse	A8	1,200	12.0%
Neutral discourse	A9	1,400	14.0%
Ambiguous cases	A10	200	2.0%
Total		10,000	100.0%

Behavioral and policy criticism constituted the largest higher-level dimension, representing 38.0% of the corpus. Potentially stigmatizing discourse accounted for 34.0%, while neutral discourse represented 14.0%. Supportive discourse accounted for 12.0%, and ambiguous cases represented the remaining 2.0%.

These categories reveal that age-related online discussions encompass several distinct forms of expression. Some comments may reproduce negative stereotypes or exclusionary representations, whereas others focus on public policy, individual conduct, social support, or factual information. Consequently, the overall frequency of negative or critical discussion should not be treated as a direct measure of ageism.

The distinction also has methodological implications. The validity of the classification depends on clear coding rules, contextual interpretation, and consistent application of the ten-category framework. Where comments contain multiple themes, the use of a single primary code facilitates distributional comparisons but may obscure secondary meanings. Future analyses may therefore incorporate multi-label coding to capture overlapping forms of discourse, provided that the coding procedure and denominator for each category are explicitly defined.

4.7. Classification Performance

Table 9. Classification Metrics

Metric	**Score**
Precision	0.84
Recall	0.78
F1-score	0.81
Accuracy	0.86

The classification of age-related discourse was evaluated by comparing automated predictions with human-coded reference labels. For the binary task of distinguishing ageist from non-ageist comments, performance was assessed using precision, recall, F1-score, and

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

accuracy. The classifier achieved a precision of 0.84, a recall of 0.78, an F1-score of 0.81, and an accuracy of 0.86 (Table 9).

Precision measures the proportion of comments predicted as ageist that are also classified as ageist in the human-coded reference dataset. Recall measures the proportion of human-labeled ageist comments correctly identified by the classifier. The F1-score is the harmonic mean of precision and recall, while accuracy represents the proportion of all test comments classified correctly (Yang, 2022).

The interpretation of these metrics depends on the quality of the reference labels, the distribution of the classes, and the independence of the test data. In particular, accuracy may provide an incomplete picture when ageist comments constitute a minority of the corpus. The relatively lower recall (0.78) than precision (0.84) indicates that the classifier missed some comments labeled as ageist by human coders while being comparatively conservative in its positive predictions.

Human coding followed predefined operational definitions, and the test data were kept separate from the data used to train or tune the classifier. These procedures supported the assessment of whether the model distinguishes explicit age-based hostility from policy criticism, behavioral disagreement, supportive statements, and neutral discussion.

Chapter 5. Discussion, Implications, Limitations, and Conclusion

5.1. Discussion

This study examined how older adults are represented in Korean online discourse and how age-based hostility can be distinguished from ordinary disagreement about public policy, behavior, and social resources. The analysis adopted the premise that ageism cannot be adequately measured through negative word frequencies alone (Trentham et al., 2015). Instead, interpreting age-related discourse requires attention to linguistic context, the targets of criticism, and whether negative judgments are generalized to older adults as a group.

The content analysis distinguished age-based derogation, economic burden framing, political stereotyping, generational resentment, and social exclusion from behavioral criticism, policy criticism, supportive discourse, and neutral or informational discussion. This distinction is important because references to older adults, pensions, political preferences, or generational

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

differences do not necessarily constitute ageism. The interpretation depends on whether the language expresses legitimate disagreement or reproduces age-based stereotypes, derogation, or exclusion.

5.1.1. Economic Burden and Social Welfare

Discussions of pensions, welfare expenditure, and taxation connect representations of older adults with broader concerns about economic sustainability and the distribution of public resources. Such discussions may reflect substantive policy concerns rather than prejudice (Schramm et al., 2023).

The analytical framework distinguished criticism of welfare-system design from statements that portray older adults collectively as undeserving, unproductive, or responsible for social problems. This distinction provides a basis for examining how economic concerns become associated with age-based stereotypes when criticism extends beyond institutions and policies to generalized judgments about an age group.

Economic burden framing should therefore be interpreted in relation to the content and context of individual comments. References to pension costs or welfare expenditure cannot independently establish ageism. Evidence of group-directed blame, derogatory generalization, or exclusion is needed to support a stronger interpretation of age-based hostility.

5.1.2. Political Stereotyping

Political discussions can associate older adults with particular voting patterns, ideological positions, or political preferences. Such associations become relevant to ageism when they portray an entire generation as politically homogeneous or use chronological age to dismiss individuals' rights, opinions, or political legitimacy.

The analytical framework distinguished criticism of particular elections, political decisions, and ideological positions from generalized derogation of older adults. Contextual interpretation is essential because political disagreement may concern policies or political actors rather than age itself (Siltala, 2021).

This distinction also highlights the importance of examining the relationship between political vocabulary and age-related references. A political statement should not be classified

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

as ageist solely because it discusses older voters or generational differences. The classification requires evidence that age functions as the basis for stereotyping, derogation, or exclusion.

5.1.3. Intergenerational Resentment

Employment, housing, pensions, and public expenditure are important contexts for discussions of fairness between generations. Such discussions may reflect structural inequalities, competing material interests, and differences in lived experience (Hwang and Seo, 2025).

The coding framework distinguished criticism of policies and institutions from statements assigning collective responsibility to older adults for the economic difficulties experienced by younger generations. This distinction is important because economic frustration does not, by itself, justify or establish age-based hostility.

Intergenerational resentment should therefore be interpreted in relation to the targets of criticism and the language used to express dissatisfaction. Comments addressing housing affordability, employment opportunities, or pension reform may constitute legitimate policy criticism. By contrast, statements that attribute negative characteristics or collective blame to older adults solely on the basis of age raise a distinct analytical concern.

5.1.4. The Difference Between Criticism and Ageism

A central contribution of the analytical framework is its distinction between negative sentiment and ageism. Negative language can target a particular action, policy, or institution without expressing prejudice toward an age group.

For example, criticism of dangerous driving may concern a specific behavior. A statement asserting that all older drivers are incompetent generalizes a negative judgment to an entire age group. Similarly, criticism of pension reform does not necessarily constitute discrimination, whereas claims that older adults should lose rights solely because of their age involve a different form of reasoning (North and Fiske, 2015).

This distinction improves analytical precision by preventing critical or emotionally negative statements from being automatically classified as ageist. It also reduces the risk of overstating the prevalence of discriminatory discourse. Reliable classification requires explicit operational

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

definitions, contextual interpretation, and a consistent distinction between criticism of conduct and generalized age-based judgments.

5.2. Theoretical Contributions

The study contributes to the examination of ageism in Korean-language digital discourse by integrating computational text analysis with contextual content classification. Three theoretical contributions are particularly relevant.

First, the analytical framework extends the study of ageism by distinguishing the frequency of age-related vocabulary from the substantive meaning of comments. Word frequencies and topic distributions can identify recurring themes, but they cannot independently determine whether a statement expresses prejudice. Contextual coding provides an additional interpretive layer for examining the relationship between language and age-based hostility.

Second, the study contributes to research on intergenerational relations by distinguishing structural disagreements about economic resources from prejudice directed at older adults as a group. Discussions of employment, housing, pensions, and public expenditure may reflect conflicting interests without necessarily expressing ageism. Separating these forms of discourse helps clarify the conceptual boundary between policy disagreement and discriminatory generalization.

Third, the study contributes to computational social science by illustrating the complementary roles of topic modeling, linguistic analysis, and human judgment. Computational methods can identify recurring patterns across large collections of comments, while contextual interpretation helps address sarcasm, ambiguous expressions, and distinctions between behavioral criticism and group-directed derogation. These complementary approaches are relevant to the analysis of context-dependent online communication (Park and Chang, 2024).

The strength of these contributions depends on the validity of the coding framework, the quality of the corpus, and the reliability of the analytical procedures. The findings should therefore be interpreted within the scope of the data and methods used in the study.

5.3. Practical and Managerial Implications

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

The findings have potential implications for digital platforms, public institutions, educational organizations, and communication professionals. These implications concern the interpretation and management of age-related discourse rather than demonstrating the effectiveness of any specific intervention.

Digital platforms. The distinction between age-based derogation and legitimate policy discussion can inform the design of content-moderation procedures. Automated moderation systems should account for context and distinguish group-directed hostility from criticism of social policy or individual behavior. Human review may be particularly important for ambiguous, sarcastic, or context-dependent comments.

Public institutions. Communication concerning pensions, welfare, employment, and public expenditure should distinguish evidence-based policy debate from generalized claims that frame older adults as a homogeneous social burden. Clear and balanced communication can help ensure that discussions of resource allocation remain focused on policies and institutions rather than unsupported assumptions about entire age groups (Oh, 2026).

Education and media literacy. Educational initiatives can use the distinction between behavioral criticism and age-based stereotyping to help individuals recognize generalizations about older adults. Media-literacy programs can also encourage more careful interpretation of online comments and greater awareness of how language may reproduce discriminatory assumptions.

Intergenerational dialogue. Community and educational programs can provide opportunities to discuss housing, employment, retirement, and social welfare without assigning collective blame to younger or older generations. Such discussions can focus on institutional conditions and competing needs rather than treating generations as internally homogeneous groups.

Research and policy evaluation. A transparent coding framework can provide a basis for monitoring age-related discourse over time, provided that sampling, preprocessing, and coding procedures remain sufficiently consistent. Longitudinal comparisons could help identify changes in thematic patterns and the language used to discuss older adults (Lau, 2022).

These implications identify possible applications of the analytical framework. Further

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

evaluation is required to determine whether particular moderation, educational, or communication interventions produce measurable improvements.

5.4. Limitations and Future Research

Several limitations should be considered when interpreting the findings.

First, online comments do not necessarily represent the views of the South Korean population as a whole. The composition of the corpus depends on the platforms, topics, users, and observation period included in the data collection process. The findings should therefore be interpreted as characteristics of the analyzed corpus rather than population-wide estimates of ageism.

Second, search terms and retrieval procedures can influence the types of comments included in the analysis. A corpus collected primarily through explicitly hostile keywords may overrepresent derogatory discourse while underrepresenting neutral, supportive, or indirect expressions. Transparent inclusion criteria and broader sampling strategies are therefore important for reducing selection bias (Levy et al., 2026).

Third, Korean-language online discourse includes slang, abbreviations, spelling variations, sarcasm, and context-dependent expressions that can complicate automated text analysis. A term that appears negative in isolation may convey a different meaning in context. Human validation and clearly documented coding rules are consequently important for interpreting potentially discriminatory language.

Fourth, topic modeling identifies statistical patterns in word distributions rather than definitive social categories. Topic labels depend on the selected number of topics, preprocessing decisions, model parameters, and the interpretation of representative comments. Future research should assess topic stability across alternative model configurations and report the procedures used to select the final model.

Fifth, content coding involves interpretive judgments, particularly when comments are ambiguous or contain several overlapping themes. Future studies should report inter-coder agreement, document the resolution of disagreements, and provide operational definitions for each category. Multi-label coding may also be appropriate when a single comment contains more than one substantive theme.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Sixth, a cross-sectional corpus cannot establish causal relationships between online ageism and real-world discrimination. Nor can it determine whether particular social conditions cause hostile comments. Longitudinal research, comparisons across platforms, and complementary surveys or interviews could help investigate changes over time and examine relationships between online discourse and individual attitudes (Ng et al., 2021).

Finally, ethical considerations remain important even when comments are publicly accessible. Public availability does not eliminate privacy concerns or automatically authorize unrestricted reproduction. Research should minimize the disclosure of usernames, personal information, and identifiable quotations, while following applicable ethical requirements and platform policies.

5.5. Conclusion

This study examined age-related discourse in a corpus of 10,000 Korean-language online comments using a framework that distinguishes potentially stigmatizing expressions from behavioral criticism, policy disagreement, supportive discourse, and neutral discussion. The analytical approach incorporated word-frequency analysis, TF-IDF, n-gram extraction, LDA topic modeling, sentiment analysis, and human-assisted content coding (Ng et al., 2022).

The content-analysis framework distinguished ten categories: age-based derogation, economic burden framing, political stereotyping, generational resentment, social exclusion, behavioral criticism, policy criticism, supportive or positive discourse, neutral or informational discourse, and ambiguous or context-dependent discourse. This classification provides a structured approach to examining the different ways in which older adults and age-related issues are discussed online (Levy et al., 2022).

The analysis emphasizes that negative language and ageism are not interchangeable concepts. Criticism of a particular behavior, public policy, or institutional arrangement does not necessarily constitute prejudice against older adults. A stronger interpretation of ageism requires evidence that negative judgments are generalized to people on the basis of age or that the discourse involves age-related derogation, stereotyping, or exclusion.

The study also highlights the complementary roles of computational analysis and contextual interpretation. Word frequencies and topic models can reveal recurring linguistic patterns,

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

while content coding helps distinguish policy disagreement from potentially discriminatory representations. The validity of these distinctions depends on transparent data collection, reliable coding procedures, and careful interpretation of the analytical outputs.

The findings should be understood within the scope of the analyzed corpus and should not be generalized automatically to all South Korean online users or the wider population. Further research using longitudinal data, cross-platform comparisons, and validated classification procedures could strengthen understanding of how age-related discourse changes over time.

Overall, the study provides an analytical framework for investigating ageism in Korean online discourse while maintaining a distinction between legitimate social criticism and age-based hostility. Its contribution rests on the systematic examination of online language, the transparent interpretation of thematic patterns, and the careful application of operational definitions to potentially discriminatory discourse.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

References

Araújo, P. O. D., Soares, I. M. S. C., Vale, P. R. L. F. D., Sousa, A. R. D., Aparicio, E. C., & Carvalho, E. S. D. S. (2023). Ageism directed to older adults in health services: A scoping review. Revista Latino-Americana de Enfermagem, 31, e4019.

Ayalon, L., & Roy, S. (2023). Combatting ageism in the Western Pacific region. The Lancet Regional Health–Western Pacific, 35.

Bae, H., & Jo, S. H. (2024). The impact of age stereotype threats on older consumers' intention to buy masstige brand products. International Journal of Consumer Studies, 48(1), e12867.

Baek, B., & An, S. (2023). Emotional expressions and social support for older adults: Text mining analysis of lifestyle senior YouTuber comments. Journal of the Korean Gerontological Society, 43(6), 1103–1121. https://doi.org/10.31888/JKGS.2023.43.6.1103

Chandra, Y. (2024). Agentive and independent older female adult in Strong Woman Gang Nam-Soon: Going against prevalent gendered ageism in K-dramas. k@ta: A Biannual Publication on the Study of Language and Literature, 26(2), 121–134.

Chang, H., & Do, Y. J. (2024). A spark of change: Developing an innovative gerontological nursing intervention mapping initiative for training and education (IGNITE). BMC Medical Education, 24(1), Article 266.

Chen, C. H. (2015). Older adults as discursively constructed in Taiwanese newspapers: A critical discourse analysis. International Journal of Society, Culture & Language, 3(2), 72–84.

Choi, H. (2021). A sociological study on the disgust of the young generation toward the elderly in Korean society: Social causes. Korea Journal, 61(2), 59–86.

Choi, H., Kim, S., & Kim, S. (2026). No country for old workers? Exploring cross-cultural perceptions of the senior workforce. Educational Gerontology, 1–18.

Chung, S., & Jung, Y. (2014). Age norms for older adults among Koreans: Perceptions and influencing factors. Ageing & Society, 34(8), 1335–1355.

Chung, S., Kim, M., Jang, Y., Park, N. S., & Yoon, H. (2025). Media representations of

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

aging and their psychological impact: Age anxiety among older Korean adults. Behavioral Sciences, 15(7), Article 932.

Comunello, F., Rosales, A., Mulargia, S., Ieracitano, F., Belotti, F., & Fernández-Ardèvol, M. (2022). Youngsplaining and moralistic judgements: Exploring ageism through the lens of digital media ideologies. Ageing & Society, 42(4), 938–961.

Elfving-Hwang, J. (2016). Old, down, and out? Beauty work, agency, and between-women sociality among elderly South Korean women. Journal of Aging Studies.

Gendron, T., Camp, A., Amateau, G., & Iwanaga, K. (2024). Internalized ageism as a risk factor for suicidal ideation in later life. Aging & Mental Health, 28(4), 701–705.

Gendron, T. L., Inker, J., & Welleford, E. A. (2018). A theory of relational ageism: A discourse analysis of the 2015 White House Conference on Aging. The Gerontologist, 58(2), 242–250.

Gewirtz-Meydan, A., & Ayalon, L. (2018). Forever young: Visual representations of gender and age in online dating sites for older adults. Journal of Women & Aging, 30(6), 484–502.

Hillebrand, V. (2022). An ageist perspective on age and older adults in information behaviour research.

Huang, H., & Chen, L. (2025). The potential negative effects of ageism on digital engagement of older adults: A meta-analysis. The Gerontologist, 65(11), Article gnaf217.

Hwang, I. C., & Seo, Y. S. (2025). Decent work for older adults in South Korea: A social justice perspective. The Counseling Psychologist. Article 00110000261425691.

Kang, H. S., & Lee, S. Y. (2025). Evolving representations of older adults in Korean digital media: A text-mining approach (2020–2024). Social Sciences, 15(1), Article 17. https://www.mdpi.com/2076-0760/15/1/17

Kang, S. (2021). An analysis of South Korea’s response to COVID-19: The changing dynamics of urban space with a focus on older people [Doctoral dissertation].

Kang, T., Kwon, E., Lee, J., Nam, Y., Song, J., & Suh, J. (2022). Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech?

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

[Preprint]. arXiv. https://arxiv.org/abs/2204.03262

Kang, Y. (2015). Cultural perspective of urinary incontinence: Similarities and differences between White elderly and Korean immigrant women. International Journal of Urological Nursing, 9(3), 125–130.

Kawaguchi, K., Koga, C., Yamada, A., Kurotori, I., Takeuchi, H., Lingling, & Nakagomi, A. (2026). Development and psychometric validation of the Japanese version of the Everyday Ageism Scale (EAS-J) among older adults. Geriatrics & Gerontology International, 26(7), Article e70641.

Kim, J. H., Oh, H. I., & Chun, S. A. (2024). Using photovoice to explore attitudes toward and experiences with images of older adults. Educational Gerontology, 50(2), 158–173.

Kim, M. J. (2020). Grandma on YouTube: A case study of the older female YouTuber Korea Grandma.

Kim, S., & Kim, H. (2025). Associations among gerontological nursing knowledge, attitudes toward ageism, and therapeutic communication competency: A cross-sectional online survey among outpatient nurses in general hospitals. Journal of Korean Academy of Psychiatric and Mental Health Nursing, 34(2), 164–174.

Kim, S., & Ryu, M. H. (2023). The characteristics of online gerontophobia expressions in South Korea. Frontiers in Psychology, 14, Article 1290443.

Lau, M. H. (2022). Beyond ageism: A qualitative study of intersecting forms of prejudice towards retired older people. Societies, 13(1), Article 13.

Lee, H. S., Jung, H. S., & Sumner, A. (2015). A cross-cultural analysis of perception on ageist attitudes between Korean and American social work students. Korean Social Science Journal, 42(1), 25–37.

Lee, K., & Nam, S. I. (2025). Poor, but not poorly living: Types and correlates of social exclusion among young-old adults in a fast-aging society. Educational Gerontology, 51(11), 1202–1217.

Levy, S. R., Lytle, A., & Macdonald, J. (2022). The worldwide ageism crisis. Journal of Social Issues, 78(4), 743–768.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Levy, S. R., Lytle, A., Monahan, C., & Huang, Y. S. (2026). Reshaping views of older adults, young adults, and intergenerational relations: Anti-ageism websites. The Gerontologist, 66(7), Article gnag129.

Lim, S. (2025). Data-driven assessment of digital age inclusion: Topic modeling Seoul's digital governance platform to evaluate elderly representation [Doctoral dissertation, Massachusetts Institute of Technology].

Makita, M., Mas-Bleda, A., Stuart, E., & Thelwall, M. (2021). Ageing, old age and older adults: A social media analysis of dominant topics and discourses. Ageing & Society, 41(2), 247–272.

Mannheim, I., Wouters, E. J., Köttl, H., Van Boekel, L. C., Brankaert, R., & Van Zaalen, Y. (2023). Ageism in the discourse and practice of designing digital technology for older persons: A scoping review. The Gerontologist, 63(7), 1188–1200.

Moon, J., & Abidin, C. (2020). Online ajumma: Self-presentations of contemporary elderly women via digital media in Korea. In Mediated interfaces: The body on social media (pp. 177–189).

Ng, R., Chow, T. Y. J., & Yang, W. (2021). Culture linked to increasing ageism during COVID-19: Evidence from a 10-billion-word corpus across 20 countries. The Journals of Gerontology: Series B, 76(9), 1808–1816.

Ng, R., Chow, T. Y. J., & Yang, W. (2022). The impact of aging policy on societal age stereotypes and ageism. The Gerontologist, 62(4), 598–606.

Ng, R., Indran, N., & Yang, W. (2023). Portrayals of older adults in over 3000 films around the world. Journal of the American Geriatrics Society, 71(9), 2726–2735.

North, M. S., & Fiske, S. T. (2015). Modern attitudes toward older adults in the aging world: A cross-cultural meta-analysis. Psychological Bulletin, 141(5), 993.

Oh, J. (2024). The "angry (digital) silver" in South Korea: The rhetoric around older adults' digital media literacy. The Gerontologist, 64(2), Article gnad092.

Oh, J. (2026). Beyond independence and with care: Reimagining older adults' digital agency with a granfluencer. International Journal of Ageing and Later Life, 19(2), 93–122.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Park, L. J. (2021). Can stories change how we feel about older people? The effect of older people's online personal stories on mitigating younger Koreans' ageism. Educational Gerontology, 47(12), 543–558.

Park, Y., & Chang, S. J. (2024). The impact of ageism experiences on social participation among community-dwelling older adults: Exploring the moderating role of digital literacy. Geriatric Nursing, 59, 372–378.

Román Amarales, A., & Williamson, S. (2026). The political roots of ageism in greying democracies: Evidence from Italy, South Korea, and the United States. Research & Politics, 13(1). Article 20531680261435660.

Schramm, E., Yang, C. C., Chang, C. H., Mulhorn, K., Yoshinaga, S., & Huh-Yoo, J. (2023). Examining public awareness of ageist terms on Twitter: Content analysis. JMIR Aging, 6, Article e41448.

Siltala, T. (2021). Japanese and South Korean aging populations in media narratives, 2006–2015.

Song, S., Chun, E., & Lim, J. S. (2026). Spatializing ageism in commercial spaces: Discursive justifications and lived experiences of age-based exclusion. Journal of Business Ethics, 1–17.

Stypinska, J. (2023). AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies. AI & Society, 38(2), 665–677.

Trentham, B., Sokoloff, S., Tsang, A., & Neysmith, S. (2015). Social media and senior citizen advocacy: An inclusive tool to resist ageism? Politics, Groups, and Identities, 3(3), 558–571.

Xiang, X., Lu, X., Halavanau, A., Xue, J., Sun, Y., Lai, P. H. L., & Wu, Z. (2021). Modern senicide in the face of a pandemic: An examination of public discourse and sentiment about older adults and COVID-19 using machine learning. The Journals of Gerontology: Series B, 76(4), e190–e200.

Yang, H. (2022). Text mining analysis of hate comments against women, immigrants, and the elderly: Focusing on Naver crime news. Korean Journal of Broadcasting, 36(3), 1–41.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

https://doi.org/10.22876/kab.2022.36.3.001

Yi, Y. (2026). From “technology empowerment” to “digital citizens”: Reflecting on the narrative of technology empowerment and the awakening of digital subjectivity among older adults. Social Sciences and Humanities Inquiry.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Appendix A.

The following code provides a reproducible starting point. It reads a CSV file containing a comment column, removes duplicates, analyzes word frequency, extracts TF-IDF terms and n-grams, and estimates LDA topics.

Install the required packages first:

pip install pandas scikit-learn konlpy

Appendix B.

KoNLPy's Okt analyzer may require a compatible Java installation.

```
import re
import pandas as pd
from collections import Counter
from konlpy.tag import Okt
from sklearn.feature_extraction.text import (
    CountVectorizer,
    TfidfVectorizer
)
from sklearn.decomposition import LatentDirichletAllocation

# 1. Load the actual dataset
df = pd.read_csv("comments.csv")
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
if "comment" not in df.columns:
    raise ValueError("The CSV needs a 'comment' column.")

df = df.dropna(subset=["comment"]).copy()
df["comment"] = df["comment"].astype(str).str.strip()
df = df[df["comment"] != ""]
df = df.drop_duplicates(subset=["comment"]).reset_index(drop=True)

print("Final unique comments:", len(df))

# 2. Basic cleaning
def clean_text(text):
    text = re.sub(r"https?://\S+|www\.\S+", " ", text)
    text = re.sub(r"[\r\n\t]+", " ", text)
    return re.sub(r"\s+", " ", text).strip()

df["cleaned"] = df["comment"].apply(clean_text)

# 3. Korean morphological analysis
okt = Okt()

stopwords = {
    "하다", "되다", "있다", "없다",
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
    "것", "수", "등", "좀", "그냥"
}

def tokenize(text):
    return [
        word
        for word, pos in okt.pos(text, norm=True, stem=True)
        if pos in {"Noun", "Verb", "Adjective"}
        and len(word) > 1
        and word not in stopwords
    ]

df["tokens"] = df["cleaned"].apply(tokenize)
df["processed"] = df["tokens"].apply(lambda x: " ".join(x))

# 4. Word frequency
all_tokens = [
    token for row in df["tokens"] for token in row
]

frequency_df = pd.DataFrame(
    Counter(all_tokens).most_common(100),
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
        columns=["term", "frequency"]
)

frequency_df.to_csv(
        "word_frequencies.csv",
        index=False,
        encoding="utf-8-sig"
)

# 5. TF-IDF
tfidf = TfidfVectorizer(
        ngram_range=(1, 2),
        min_df=2,
        max_df=0.95,
        max_features=5000,
        token_pattern=r"(?u)\b\w+\b"
)

X_tfidf = tfidf.fit_transform(df["processed"])
terms = tfidf.get_feature_names_out()
scores = X_tfidf.mean(axis=0).A1

tfidf_df = pd.DataFrame({
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
        "term": terms,
        "mean_tfidf": scores
}).sort_values("mean_tfidf", ascending=False)

tfidf_df.to_csv(
        "tfidf_terms.csv",
        index=False,
        encoding="utf-8-sig"
)

# 6. N-gram analysis
ngram_vectorizer = CountVectorizer(
        ngram_range=(1, 3),
        min_df=2,
        max_features=5000,
        token_pattern=r"(?u)\b\w+\b"
)

X_ngrams = ngram_vectorizer.fit_transform(df["processed"])
ngram_terms = ngram_vectorizer.get_feature_names_out()
ngram_counts = X_ngrams.sum(axis=0).A1

ngram_df = pd.DataFrame({
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
    "term": ngram_terms,
    "frequency": ngram_counts
}).sort_values("frequency", ascending=False)

ngram_df.to_csv(
    "ngram_frequencies.csv",
    index=False,
    encoding="utf-8-sig"
)

# 7. LDA topic modeling
lda_vectorizer = CountVectorizer(
    min_df=3,
    max_df=0.95,
    max_features=5000,
    token_pattern=r"(?u)\b\w+\b"
)

X_lda = lda_vectorizer.fit_transform(df["processed"])

if X_lda.shape[1] == 0:
    raise ValueError("No usable terms remain after preprocessing.")
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

```
lda = LatentDirichletAllocation(
        n_components=6,
        random_state=42,
        max_iter=20,
        learning_method="batch"
)

topic_probabilities = lda.fit_transform(X_lda)
lda_terms = lda_vectorizer.get_feature_names_out()

for i, component in enumerate(lda.components_):
        top_indices = component.argsort()[-10:][::-1]
        words = [lda_terms[j] for j in top_indices]
        print(f"Topic {i + 1}: {', '.join(words)}")

df["dominant_topic"] = topic_probabilities.argmax(axis=1) + 1
df["topic_confidence"] = topic_probabilities.max(axis=1)

df.to_csv(
        "comments_with_topics.csv",
        index=False,
        encoding="utf-8-sig"
)
```

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

print("Text-mining pipeline completed.")

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18442

Conflicts of interest

The author declares no conflicts of interest related to this research.

Data availability statement

The data analyzed in this study were obtained from publicly accessible sources. Details of the data sources and analytical procedures are provided in the manuscript.

Declaration of AI use

Artificial intelligence tools were used to assist with language editing and manuscript preparation. The author reviewed the manuscript and takes responsibility for its content, accuracy, and integrity.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.