Chapter 3. Research Question

In the digital age, public sentiment is increasingly being expressed on online platforms. Websites like YouTube are invaluable to policymakers as they provide a rich data source for measuring public opinion on various topics. This study addresses the question: How do different NLP models work to interpret public sentiment in YouTube comments?

a. Key research questions:

- How effectively can leading NLP models (logistic regression, naive Bayes, SVM, LSTM, and BERT) evaluate and contrast the sentiment of YouTube comments on a demographic issue in South Korea?

b. Sub-question:

- How do traditional machine learning models such as logistic regression and SVM compare to advanced deep learning models such as BERT and Transformers for sentiment analysis tasks?
- What are each model's unique strengths and limitations for discerning public sentiment in text data?
- Which model provides policymakers with the most accurate and actionable insights into the public's view of Korea's demographic challenges?

c. Rationale

Sentiment analysis has long played a pivotal role in determining public perception. As digital data grows exponentially with advances in technology, pinpointing a model that excels at identifying these emotions becomes an even more important variable. The decision to compare these particular models is driven by their current reputation and the different methodologies they implement within NLP. Existing models such as Logistic Regression and SVM are favored for their clarity and simplicity, while innovative models such as BERT and Transformers are redefining the limits of sentiment analysis by understanding the context in ways previously thought out of reach.

The purpose of this study is to provide useful insights for public policy evaluation by identifying the optimal model or combination of models for sentiment analysis on YouTube comments. By studying in depth the efficacy, strengths, and limitations of these models, this study will not stop at revealing public sentiment about Korea's demographic problems. This study will guide future scholars to wisely apply NLP tools to similar studies.

d. Expected Outcomes

The comparative analysis is designed to:

- Highlight models that outperform others in terms of accuracy, processing speed, and knack for detecting subtle emotions.
- Provides a holistic understanding of potential biases or shortcomings inherent in each model.
- Provides policymakers with actionable information about public sentiment and the consequences for demographic strategies.

Chapter 4. Methodology

Recently, as comments on Naver, Korea's leading portal site, were suspended due to political influence, public attention was focused on platforms such as YouTube. Recognized for its extensive user-generated content in the Korean language, YouTube serves as a lens through which to look into public sentiment , providing an essential repository for understanding prevailing opinions. The platform's real-time distribution of various news stories and its interactive nature allows many users to express their feelings. This study is rooted in sentiment analysis of YouTube comments, focusing on Korea's low birth rate policy.

Given YouTube's vast amount of content, it is paramount that researchers adopt a systematic approach to extract relevant data from its wealth. The YouTube Data API plays a key role in this effort. Tailored for efficient data extraction, this API is particularly adept at mining video comments, making it an indispensable tool for sentiment analysis for the Korean public.

The table below is a literature review table of previous studies that performed machine learning and deep learning using NLP dataset. Sentiment analysis using NLP data is being actively implemented, and in line with this trend, this study is expected to contribute to the identification and analysis of public policy public opinion using NLP dataset.

< Table 1 : Literature Survey on Previous Studies on Sentiment Analysis Using NLP Data>

Authors	Used Model	Speicific Features	Data Source	Dataset Size	Accuracy	Key Findings
(Hossain et al., 2021)	Logistic Regression, Naive Bayes, SVM, SGD	Unigram, Bigram, Trigram features	Bengali Book Reviews	2000 reviews	84%	Multinomial Naive Bayes with Unigram outperforms other techniques
(Raza et al., 2019)	Naive-Bayes, SVM, Logistic Regression, Decision Tree, K-Nearest Neighbor, Random Forest	Lemmatization, N-graming, Tokenization, Stop word removal	Scientific Articles	8736 citation sentences	Improved by 9%	Improved accuracy with additional features selection techniques
(Gamal et al., 2018)	Naïve Bayes, SGD, SVM, PA, ME, AdaBoost, MNB, BNB, RR, LR	Unigram	IMDB, Cornell Movies, Amazon, Twitter		87% to 99.96%	Performance of PA with a unigram is the best among other algorithms
(Ali, 2021)	Naïve Bayes, MNB, KNN, Logistic Regression, SVM	Information Gain (IG) as a filtering technique	Arabic Tweets about COVID-19		89.6%	Proposed model performs well in analyzing perception about coronavirus
(Bahri and Suadaa, 2023)	SVM, Complement Naïve Bayes, Logistic Regression, BERT, IndoBERT, mBERT	Aspect-based analysis (attractions, facilities, access, price)	Bromo Tengger Semeru National Park, Indonesia (Google Maps Reviews)		91.48% (IndoBERT), 89.16% (SVM)	IndoBERT achieved the best performance among the models
(Musleh et al., 2023)	SVM, Naïve Bayes, Logistic Regression, KNN, Decision	Arabic comments classification	YouTube Comments	4212 labeled comments	94.62% (NB)	Naïve Bayes achieved the highest accuracy among the models

	Tree, Random Forest					
(Jain et al., 2023)	Logistic Regression, Naïve Bayes, SVM, LSTM, BERT	Bag-Of-Words, TF-IDF	COVID-19 Vaccine Tweets		90.42% (BERT), 88.7989% (SVM)	SVM and BERT achieved the highest accuracies among the models
(Herdiyani and Zailani, 2022)	Random Forest, Naïve Bayes, Lexicon, SVM, NBC, Neighbor Weighted K-NN, Logistic Regression	TF-IDF	Tweets about Indonesia's Capital Relocation	1639 tweets	76%	Random Forest method used to classify tweets into positive, negative, and neutral
(Maree et al., 2023a)	Support Vector Machines, Naive Bayes, Logistic Regression, Random Forest	Lemmatization, knowledge-based n-gram features, WordNet	Movie Reviews	50,000 movie reviews, 10,662 sentences, 300 generic movie reviews	90.43% (SVM)	Coupling lemmatization and knowledge-based n-gram features produced higher accuracy results
(Eom et al., 2022)	SVM, RNNs, LSTMs, BERT, KoBERT	Sentiment analysis regarding vaccination after COVID-19 Omicron variant	South Korean Twitter data		71% (KoBERT)	KoBERT showed the best performance in all predictive performance indicators
(Kastrati et al., 2021)	NLP, Deep Learning	Sentiment analysis in education	Students' feedback in learning platforms	92 studies from 2015-2020		Deep Learning is the most recent trend in sentiment analysis in education
(Ali et al., 2019)	Self-Organizing Map, Principal Component Analysis, Adam Deep Learning	Sentiment classification in social networks	Social network data	Various sizes		Proposed approach is efficient and feasible for Big Data
(Ghorbani	Convolutional	Sentiment			89.02%	ConvLSTMConv

et al., 2020)	Neural Network (CNN), Long Short-Term Memory (LSTM)	analysis in cloud computing				network provides an appropriate solution for sentiment analysis
(Prottasha et al., 2022)	BERT, CNN-BiLSTM	Transfer learning for sentiment analysis	Bangla NLP domain			BERT-based supervised fine-tuning outperforms other embeddings and algorithms

4.1.1. Collecting Public Voice

Detailed methodology for data collection:

- **Setting:**

- Platform registration: To extract data, the project should be registered in the Google Cloud Console, a precursor to accessing the Google API suite.
- API activation: After registration, the YouTube Data API v3 was activated, providing a robust feature set for needs.
- Generate API key: An API key was generated after. This key is pivotal for two reasons. Authenticates data requests and monitors usage to ensure compliance with specified quotas.

- **Video identification:**

- Formulating criteria: Due to the vast repository of Korean language content on YouTube, meticulous criteria were essential to ensure the accuracy of the study. We focused on videos that match Korean research topics with significant user interaction.
- Content source: To secure accurate information, we selected news from famous domestic broadcasters such as JTBC, SBS, and MBC, focusing on policy news related to the low birth rate over the past three years.
- Keyword Implementation: Search parameters were created around the country's approved low birth rate policy. Videos with significant comment interaction were targeted for

analysis according to criteria set by public broadcasters.

- **Extract comments:**

- Tool of choice: PYTHON and COLAB, the primary programming language for these tasks, facilitated annotation extraction. A batch request was made to mine comments using the generated API key and video ID.
- Data retrieval: Each request often gets several comments and a pagination token to ensure comprehensive comment navigation. Expressions, including emoticons, were captured to provide nuanced sentiment analysis. Extraction was not limited to top comments to maintain overall sentiment capture.

- **Data structure and storage:**

- Standardization: Homogenize comments into a unified format before storage.
- Storage Separation: Comments categorized by corresponding video ID were partitioned into separate datasets but maintained a consistent, structured format. This way of working ensures clarity during analysis by ensuring each comment is contextually linked to the original video.

4.1.2. NLP Preprocessing

It is essential to start with clean, structured data for practical sentiment analysis, especially for data sets as varied and complex as YouTube comments on politically charged issues such as South Korea's low birth rate policy. This chapter looks at the preprocessing steps to refine the data set and ensure preparation for subsequent analysis steps.

The raw data set culled directly from YouTube had some problems. First of all, comments sometimes mixed Hangul with Chinese characters or Latin scripts due to the global nature of the platform. In addition, there was a problem that the complexity inherent in Korean, such as regional dialects and deadlock structures, could confuse analysis. In addition to this, comments contained non-essential information such as URLs, user comments, and repetitive phrases that could distort the analysis. Many cases were included, and there were unnecessary elements in constructing and analyzing sentences, such as using emoticons. Given these challenges, pretreatment was not only desirable but essential.

a. Data preprocessing methodology

- Standardize scripts: Before going into deeper language processes, ensuring homogeneous scripts across the data set was crucial. All comments were converted to use Korean as the primary language to standardize scripts and ensure linguistic consistency. In addition, we went through the process of removing emoticons for efficient tokenization.
- Tokenization: The hierarchical morpheme structure of the Korean language required a unique tokenization approach. Utilizing advanced NLP tools tailored for Korean, we subdivided the data set into meaningful words or morphemes, ensuring each unit retains its contextual meaning.
- Eliminate stopwords: We used a list of stopwords curated for the context of the study and the Korean language. We have removed unnecessary expressions and terms that do not contain the primary investigation or meaning used in Korean. By removing these standard terms and particles that lack sentiment analysis value, we ensure that the main content of each comment is the focus of our analysis.
- Headwordization: A circular restoration tool was used to counteract word variations that could introduce redundancy. By standardizing words into root forms, the data set was simplified so that different forms of words were recognized and treated as a single entity.

Performing the preprocessing steps as described above ensures that the data set is representative and clean, laying the foundation for accurate sentiment analysis. By applying these improvements, the subsequent stages of modeling and analysis are established on a solid foundation, ensuring the reliability and robustness of the results.

4.1.2 Process and Scoring

In text analysis, sentiment analysis is essential, especially when measuring public reaction to policies. Sentiment analysis is complex in all languages due to linguistic and cultural nuances, but the complexity of Korean amplifies these challenges. This chapter outlines the approach adopted and highlights the sentiment scoring and data labeling process.

Sentiment analysis in South Korea is unique and complex for several reasons. With its long history, Korean has an agglutinative structure and uses honorifics that can subtly change the sentiment of a sentence. Also, many emotions in the Korean language are deeply rooted in a culture or historical context, and can only be deciphered with an intrinsic understanding of that culture. Also, Korean is characterized by the fact that a sentence can convey both positive and negative meanings depending on the context. Given these difficulties, this study opted for sentiment analysis, which adopts a binary method to classify opinions as positive or negative to ensure clarity and accuracy.

a. Sentiment scoring using 'KNU Korean Emotional Dictionary'

There are various tools and vocabularies for analyzing Korean sentiment, but we chose the 'KNU Korean Sentiment Dictionary,' which comprehensively covers Korean sentiment terms. The sentiment of each comment was scored based on the prevalence and weight of the terms listed in this dictionary.

- Word weight: Words in comments were mapped into a dictionary, assigning each word a predefined sentiment weight.
- Aggregated Sentiment Score: The scores for individual words in a comment are aggregated to provide an overall sentiment score for all comments.

b. Data labeling process for sentiment analysis

- Labeling training data: Comments were initially labeled as 'positive' or 'negative' using the emotion score derived from the 'KNU Korean Emotional Dictionary.' This automated process provided preliminary labeling but required manual intervention to ensure accuracy.
- Manual review: Each comment was carefully reviewed along with an initial sentiment label.
- Fix and Validate: Fixed a miscategorized comment where the automated sentiment did not match the actual sentiment.

c. Creation of labeled data sets:

- After validation, a comprehensive labeled data set was curated. This dataset, full of positive and negative sentiment tags, served as the basis for subsequent predictive modeling.
- Consolidate Labels: Ensure each comment has a single, accurate sentiment label.
- Structuring the dataset: Structure the dataset to be optimally configured for training the machine learning model in subsequent steps.

c. Conclusion

Sentiment analysis, by its very nature, is a very challenging study, but it forms the basis for understanding public sentiment. This study tried to accurately capture and express Koreans' feelings about the low birth rate policy through a meticulous combination of automated sentiment scores and manual verification.

4.1.2 Comparing Models

Uncovering the underpinnings of public sentiment requires a discerning eye and the most accurate and efficient computational models. In the age of digital expression, emotions are no longer confined to surveys or interviews. They resonate in the vast sea of online content. Therefore, this study aims to discover the most powerful model for sentiment analysis, focusing on YouTube comments, which are a treasure trove of opinions related to Korea's low birth rate policy. A careful evaluation was performed on seven models: Naive Bayes, Logistic Regression, SVMs, LSTM, KoBERT, KcBERT, and BERT_Multilingual.

a. Traditional model:

- Naive Bayes: Derived from the fundamental principles of the Bayes theorem, Naive Bayes classifiers have historically been admired for their agility and efficiency, especially on large data sets. Its probabilistic basis provides a clear view of affectability.
- Logistic Regression: With a rich statistical base, it was speculated that this model would provide transparent and insightful probability estimates for different emotion classes.
- SVM: Boasting a unique ability to manage high-dimensional data sets, SVMs have been considered valuable tools for fine-grained analysis of text data.

b. Deep learning model:

- LSTM: A subset of recurrent neural networks, LSTMs have been praised for their proficiency in sequence-based tasks, particularly relevant to sentence-level sentiment analysis.

c. BERT variants:

In this study, we also aimed to compare BERT_Multilingual, designed for application to various languages, and a model specialized for Korean, out of the framework of the existing BERT.

- KoBERT & KcBERT: Given the linguistic complexity inherent in Korean, these specialized models were expected to be more adept at identifying subtle emotional expressions.
- BERT_Multilingual: The ability to identify emotions in multiple languages predicted that this model would be indispensable given YouTube's massive global user base.

The YouTube API served as a conduit for sourcing YouTube comments during data collection. These comments are raw and unfiltered but represent pure public opinion. They were subsequently preprocessed to ensure uniformity and relevance. Tokenization, stemming, and stopword removal were pivotal steps in this preprocessing step. This collection and preprocessing process establishes data collection and methodological accuracy.

This study adopted a systematic approach for each model to ensure accurate analysis.

- Training: A broad subset of YouTube comments previously annotated with sentiment labels were used. The key to this step is to immerse the model into the data set to unravel underlying patterns and complexities.

- Validation: It was essential to ensure that the model was neither overfitting nor underfitting. A separate subset of data is reserved for this purpose, allowing hyperparameter fine-tuning.

- Testing: This step was pivotal in evaluating practical applicability. A series of untouched opinions served as a testing ground for each model's efficiency and accuracy.

In addition, a multifaceted evaluation method has been developed for accurate research performance. Models were evaluated on various aspects such as accuracy, precision, recall, and F1 score. Each metric provided unique insights, collectively painting a comprehensive picture of model effectiveness. It focused on detecting negative sentiment, which is essential for understanding public concerns and concerns.

Every robust methodology has its problems, and this study is no exception.

- Data Imbalance: An imbalance in emotional expression was observed. Some emotions were exaggerated while others were underrepresented. To mitigate this, a properly balanced dataset was adopted and used as labeling data, which became the basis for future emotion prediction of unlabeled data.
- Decoding Complex Language Structures: Comments were sometimes straightforward. There were many examples of satire, metaphor. Existing models in particular suffered from this complexity.
- Dealing with Computational Requirements: Deep learning models, while powerful, were resource intensive. Managing the computational requirements, especially for the BERT variant, required careful planning and infrastructure.

In essence, methodology serves as the basis for all empirical investigations. The models chosen for this exploration represent a harmonious combination of time-tested techniques and newly developed deep-learning approaches. This study aims to provide valuable insights into sentiment analysis by evaluating each model's strengths, limitations, and unique suggestions. Our overarching goal is to provide policymakers and analysts with a robust, stable, and scalable framework to decipher accurately and leverage sentiment expressed on platforms like YouTube.

Chapter 5. Result

5.1 Disclosure of Public Opinion: Results of Data Preprocessing

Sentiment analysis is emerging as a pioneering tool in the modern digital realm. It gives stakeholders, policymakers, and organizations a privileged look into the vast expanse of the public's collective psyche. Capturing genuine emotion on platforms like YouTube that provide unfiltered voices has become both a challenge and a necessity. This effort allowed us to accumulate a data set with opinions on seven policies. Raw, unfiltered comments are grains of gold in the sand. This comprehensive report describes the meticulous preprocessing steps taken to capitalize on these treasures and ultimately sets the stage for in-depth sentiment analysis.

As a melting pot of global narratives, YouTube receives daily comments. A vast commenting environment reflects myriad viewpoints, cultures, and opinions. Seven data sets were carefully extracted from this dynamic digital realm, each reflecting a unique policy. Looking at the overall characteristics related to the entire data set, the data set contains an impressive 45,960 comments. In its initial state, analyses averaged 13.96 words. Repeated terms such as '69 hours', 'more,' and 'possible' were emphasized in the preliminaries. While these phrases reflect popular trends, they can also mask the subtle emotions lurking beneath them and act as potential noise amongst the precious data. The entire data preprocessing process is as follows.

5.1.1 Preprocessing steps and rationale

Tokenization: A fundamental yet pivotal step in text analytics, tokenization is akin to unraveling a fabric by exposing each unique thread or token. These threads often make it possible to identify an underlying emotion, suggesting that sentiment is always in a particular word or phrase and not the sum of opinions. In this study, emotion labeling was prepared based on the KNU emotion dictionary through data organization through tokenization.

Stemming: Stemming acts as a language equalizer. The etymology of words is traced to ensure some level of unity. Considering the complex nature of the Korean language, we avoided this in our analysis. Stemming usually acts as a sieve to filter out linguistic redundancies while preserving essence. In order to apply the KNU Sentimental Dictionary, the basic morphemes in the dictionary and words in the dataset must match. Therefore, this study's words in the emotion dictionary and dataset were homogenized through morpheme analysis.

Eliminate stopwords: Every language has essential words that are structurally important but often dilute the rich meaning of a text. A preliminary scan showed "더" (more) and "진짜" (really) as possible stopwords. Their extraction paved the way for fine-tuned datasets for emotion extraction. Since Korean contains various postpositions such as "의, 을, 를, 한." these search terms are a significant obstacle to NLP data analysis. Therefore, a process was implemented to remove these stop words.

5.1.2. data after preprocessing

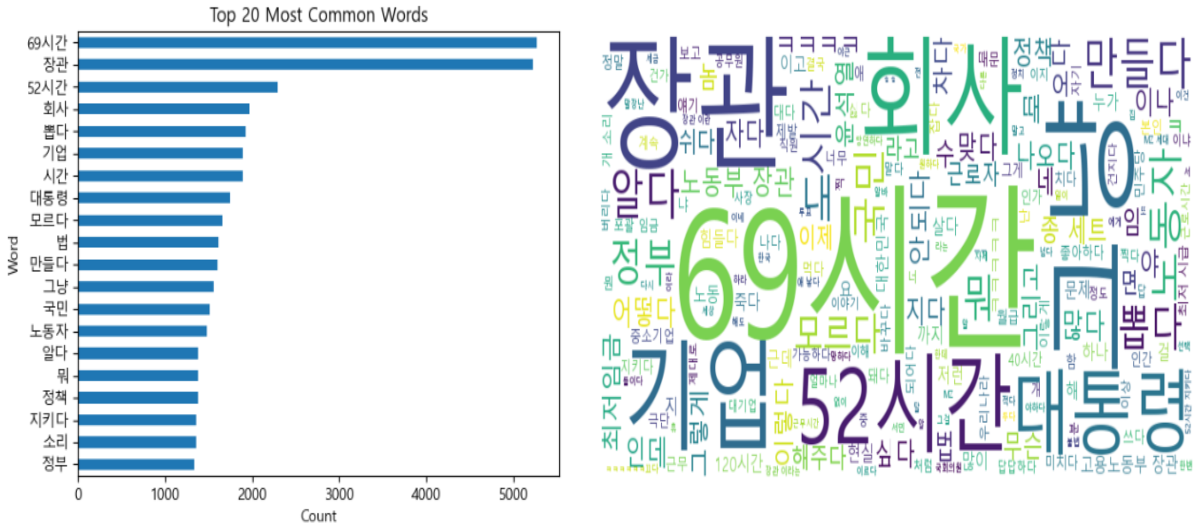
The ramifications of the pretreatment manipulation were evident in the resulting data set. Lighter comments now represent a data set that averages 13.68 words and is ready for scrutiny. Most notably, the linguistic landscape has changed. Words and phrases have gone from generic or contextually ambiguous to emotionally charged. New terms such as 'Walwalmeongmeong', which reminds us of meaningless words and sounds like '다' (all) and '그냥' (just), heralded this change.

A visual representation of data often acts as a bridge, transforming data metrics into understandable insights. In this study, visualization was performed using word clouds and bar graphs. The visualization allowed us to see the underlying themes and interests of the data set. The primary analysis of each of the seven policies was made through visualization.

5.1.3 Detailed review of pre-processing data for specific policies

Public sentiment on significant topics and policies is illuminated in the preprocessing and post-review. Extracted from extensive YouTube commentary, each dataset provides a rich tapestry of public opinion and attitudes toward the policy. A closer look at these subjects helps reveal more than just emotions. It provides a narrative, an ongoing debate, and a collective pulse that resonates with each policy.

a. Office Hours Policy



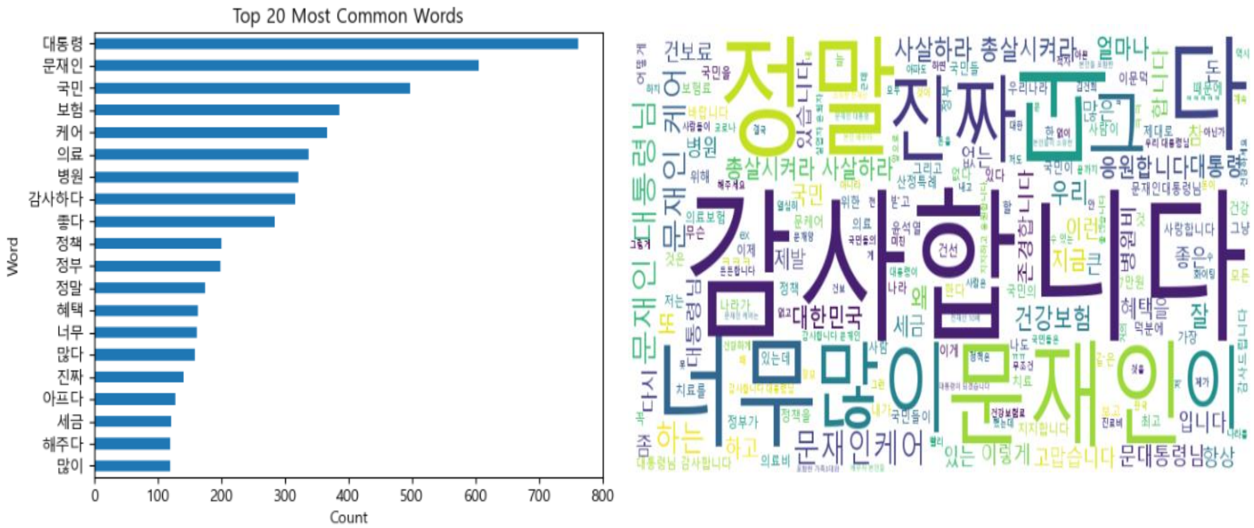
<fig5. Word Cloud & bar chart for Office Hours Policies>

Among the dominant terms in the working hour policy postprocessing dataset, '69 시간' and '52 시간' stand out. South Korea's recent policy of 52-hour work is now considering a 69-hour work week under President Seok-Yeol Yoon, indicating a lively discourse surrounding the details of potentially fixed working hours in the process of shifting the work paradigm.

The frequency of 'ministers' proposing discussions involving ministers or cabinets may indicate government interventions or decisions that have attracted public attention. Interestingly, '기업' (Enterprise) and '회사' (Company) suggest exploring the impact of these policies on businesses and their employees, emphasizing the central role of the corporate sphere in this conversation.

The existence of a '법' (law) means that legal effect or change is expected or actively discussed. These word frequencies suggest that stories of legal, corporate, and policy change intertwine to create a multi-layered discourse.

b. Medical Policy

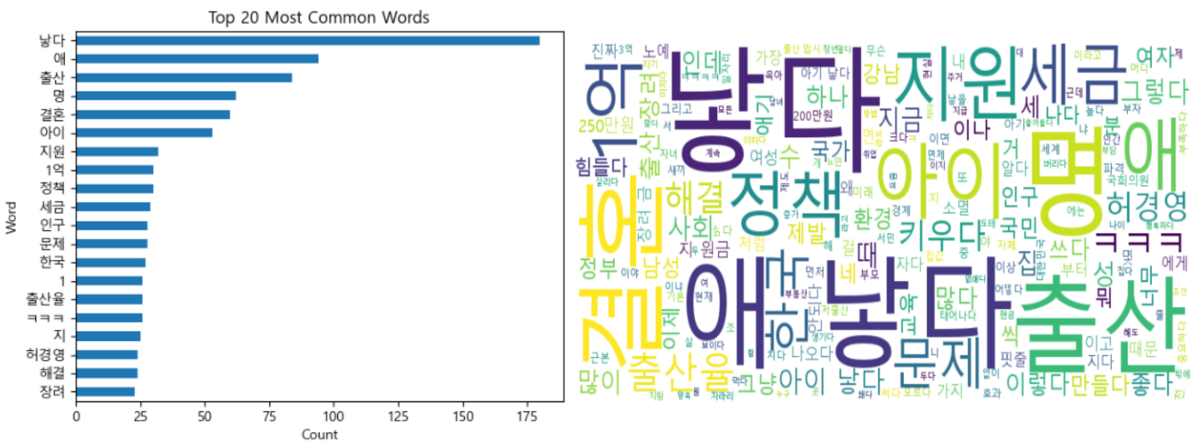


<fig6. Word Cloud & bar chart for Medical Policy>

Health care, universally of significant public interest, gravitated toward a few central themes in the data set. The term 'president' stands out. Political decisions, implications, and potential reforms start at the presidential level. The intersection of '보험' (insurance) and '의료' (medical services) within the discourse highlights the public's interest in the accessibility and quality of health care.

Interestingly, words like '감사합니다' (thank you) and '좋다' (good) in this conversation imply a favorable response to some healthcare reform or initiative, indicating areas where the policy might resonate well with the public.

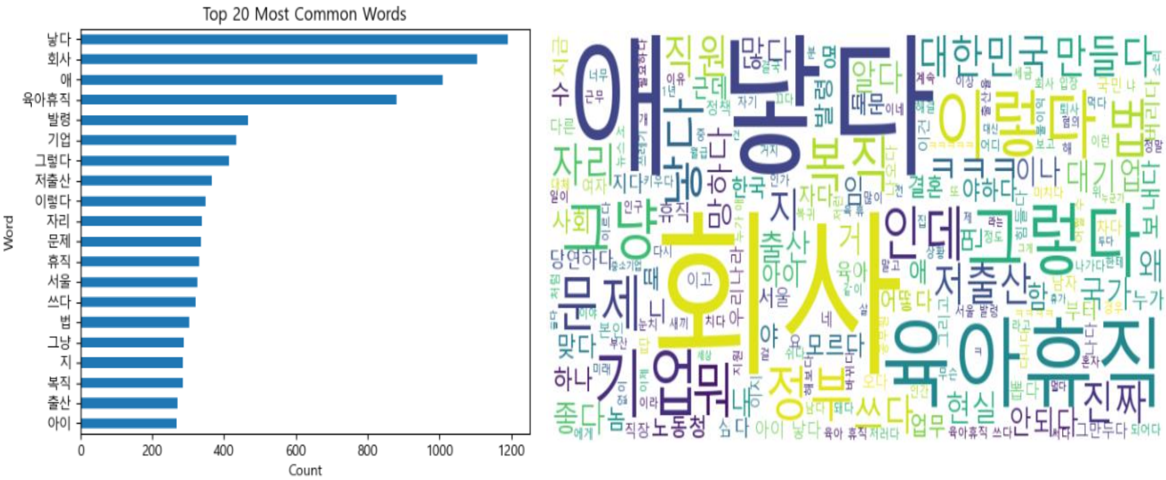
c. Maternity Policy



<fig 7. Word Cloud & bar chart for Maternity Policy >

Terms emerging in infertility policy datasets leverage social culture and policy-driven conversations. '낳다,' '출산,' and '아이' imply a narrative centered on childbirth and parenting. The frequent appearance of '정책' (policy) alongside '100 million' could suggest a financial incentive or support plan to address Korea's well-known low birth rate problem. The term' 결혼' (marriage) adds another layer of pointing to the discussion of the role of marriage in procreation discourse. Are policies framed in such a way as to encourage marriage as a precursor to procreation? Or is there a deeper cultural conversation going on in the commentary?

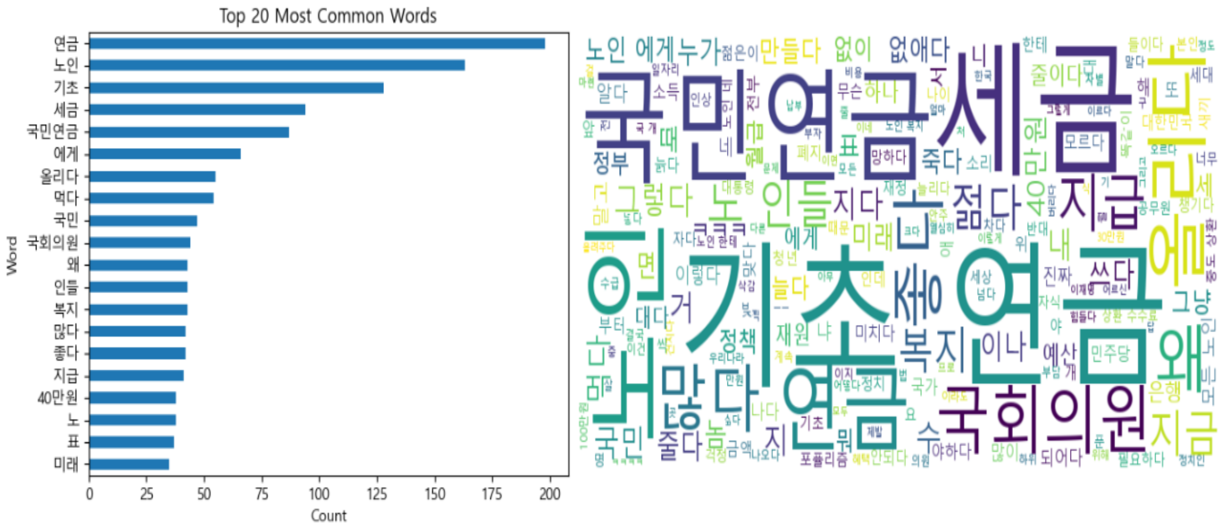
d. Parental leave



<fig 8. Word Cloud & bar chart for Parental leave>

The parental leave system, which is essentially connected to parental leave, directly refers to parental leave by putting terms such as '육아휴직' (Parental leave) and '휴직' (Leave) in the foreground. The prominence of '회사' (Company) indicates that the role of the corporate sector, or perhaps the response to these policies, is a central theme. Reappearing here, '저출산' (low birth) emphasizes the interconnectedness of these policy discussions, weaving narratives that encapsulate personal and professional realms.

e. Senior benefits

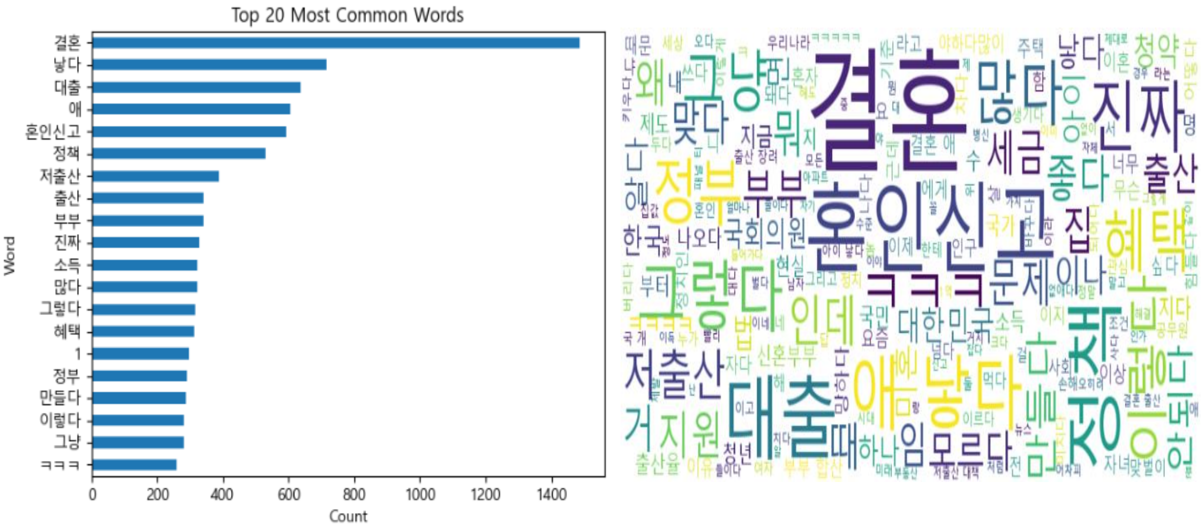


<fig 9. Word Cloud & bar chart for Senior benefits >

In the datasets related to welfare for the elderly, terms such as "연금" (pension) and "노인" (elderly) stand out. There is a clear emphasis on" 국민연금" (National Pension), implying concerns or discussions about the National Pension system.

Phrases such as '기초' may be associated with fundamental support for seniors and describe essential services and benefits. Interestingly, terms like '국회의원' (member of parliament) and "표" (vote) have a political connotation, implying a potential policy change or future decision. Specific references to "400,000 won" may refer to discussions surrounding certain monetary benefits or benefits.

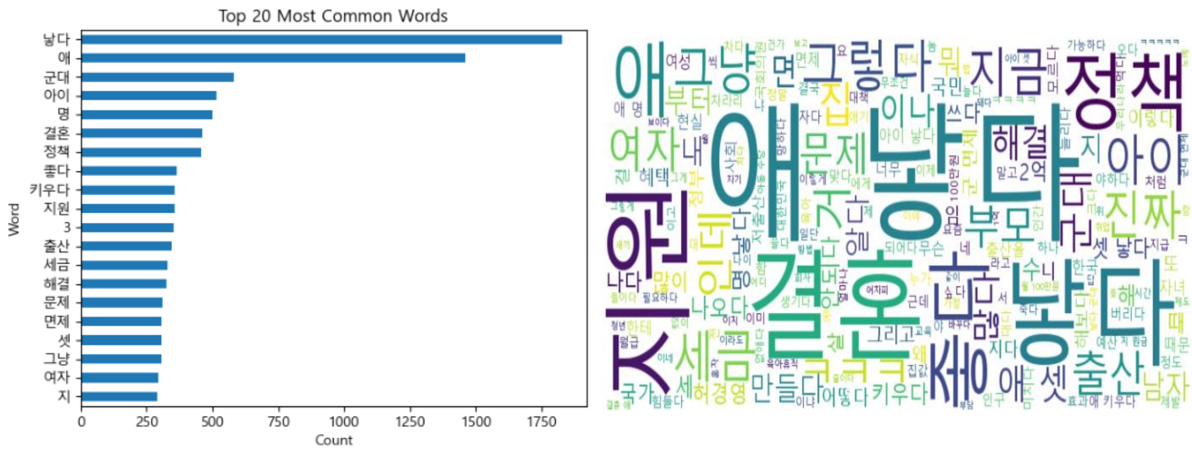
f. Loans for new couples



<fig 10. Word Cloud & bar chart for Loans for Newlyweds >

In the field of support for newlyweds, of course, '결혼' (marriage) stands out, followed by '대출' (loan) and 'marriage registration.' These terms emphasize the government's efforts to support young couples in the housing sector, as '집' (house) implies. Interestingly, '저출산' (low birth) reappears here, showing that marriage, housing, and fertility policies are intertwined. Terms like '소득' (income) and '혜택' (benefits) suggest an evaluation conversation about the financial thresholds or incentives associated with these loans. Recurring '문제' (problems) suggest potential issues or controversies related to this policy.

g. Child Care Assistance



<fig 11. Word Cloud & bar chart for Child Care Assistance >

The discourse surrounding childcare support is vividly expressed in terms such as '군대' (Army) and '키우다' (raise). It suggests stories about raising children and social expectations or roles related to children. '지원' (support) is evidence of the support provided to the family or the discussion surrounding support.

Also, terms like '면제' (exemption) can refer to certain privileges or exemptions available to families with multiple children or those who meet specific criteria. The frequency of terms such as '남자' (male) and '여자' (female) indicate gender discussions about the roles, responsibilities, or issues each gender faces in the context of raising children.

5.2 Public Sentiment Revealed: Sentiment Analysis Results

5.2.1 Sentiment Distribution: Deep

In the era of digital proliferation, public opinion is not confined to cafes and parlors. They permeate the vast expanse of the internet, empathizing with tweets, blogs, reviews, and comments. Like a digital footprint, these emotions provide a rich tapestry of human thoughts and feelings. Using sentiment analysis, we explore the core of these expressions and decipher the comments that make up the global or national psyche on various topics—the following results from classifying the collected comments into positive or negative comments using the KcBERT model.

a. Office Hours Policy

The positive response to the Office Hours Policy remained at only 5.8%, indicating low satisfaction with working hours. Responding to this response can be attributed to the industry embracing a "work from home" culture, allowing employees to schedule and provide a work-life balance. In addition, a plan to reduce the current 52 hours of work can be discussed. The national opposition and response to the Yoon Seok-yeol government's 69-hour working hour policy reflect public opinion on the working environment. The negative sentiment towards the Office Hours Policy was 94.2%, which is interpreted as a sign of great dissatisfaction. These negative emotions stem from the current prevalence of chronic fatigue, reduced personal time, and encroachment on family time. This sentiment warns the industry to rethink operating hours and adopt a more holistic approach to employee well-being.

b. Medical Policy

In the healthcare sector, positive emotions are confirmed at 81%. The overwhelming majority of satisfaction with health care reflects the recent developments that our society has witnessed, and in particular, Moon Jae-in Care, which was introduced during the Moon Jae-in administration, has had a significant impact on this. A representative positive foundation could be contributing factors: improved medical research, AI integration in diagnostics, more comprehensive medical coverage, and more responsive emergency services. Negative emotions are a small but not negligible part, at around 19%. This negative sentiment can strongly affect areas with poor healthcare infrastructure, exorbitant medical bills, malpractice stories, and bureaucratic red tape delaying essential services.

c. Maternity and child support:

The positive response to childbirth and child support is low at 1.9%.
On the other hand, 98.1% of negative responses are analyzed as areas requiring immediate attention. Comprehensive maternal care, postnatal counseling, parenting support, and, where possible, educational support can help with this policy. In addition, high medical costs associated with childbirth, insufficient maternity leave, lack of childcare facilities, and possibly social pressures are pressing issues.

d. Parental leave:

The positive response to parental leave was 3.2%, showing little satisfaction. In order to improve this, it is necessary to create an environment in which parental leave can be implemented, such as an organizational culture that provides ample childcare benefits such as extended leave, childcare facilities, and flexible working hours. It is analyzed that the concern is very high.

In addition to the length of leave, factors leading to these outcomes may include underlying fears such as stigma over extended leave, concerns about job security, or the financial impact of unpaid leave. It is expected that these areas can be improved through the environment and national consensus in which the state can lead parental leave and through political, economic, social, and cultural support.

e. Senior Support:

10.7% of positive comments about senior support suggest that some districts or communities may be doing it right. These feelings can arise from areas with excellent geriatric care, communities that promote intergenerational interaction, or policies that provide financial and medical support to the elderly. However, with 89.3% of negative responses, supporting the elderly seems to be an urgent issue. When caring for the elderly, inadequate nursing homes, a lack of skilled aged care facilities, social apathy, and financial insecurity face numerous challenges for older people. Governments and society need to overcome these difficulties and approach the problem from a broader low birthrate system perspective.

f. Loans for Newlyweds:

Responses to loans for newlyweds are faintly positive at 1.4%. These may be the beneficiaries of a simple loan approval process, low-interest rates, or the counseling sessions some banks offer to new couples. However, negative emotions reached 98.6%, so improvement is needed. Strict documentation requirements, high collateral requirements, fluctuating interest rates, and a lack of awareness of available schemes can be several problems, and a response to address them is urgent.

5.2.2 Qualitative Insights: The Soul Beyond Numbers

All abilities of sentiment analysis on Korea's low birth rate policy based on the NLP dataset are accompanied by a sea of emotions. The anguish of a mother who cannot spend quality time with her child due to hard labor, the relief of a patient receiving treatment at a hospital, and the quiet despair of an abandoned old man, these are the stories created. Looking at the results of this sentiment analysis, the current low birth rate phenomenon in Korea is not a problem in one field, but all parts of society are closely related, and it should be seen as a result of organic relationships.

This characteristic of the sensibility of the low fertility policy further emphasizes the power of information. Knowledge is power, but lack of knowledge can lead to negative emotions. Many people need to be made aware of policies or support systems, which can lead to misinformation and frustration. When conducting sentiment analysis, it is also confirmed that there needs to be more information on the current policy or a negative perception due to a lack of accuracy. A robust information dissemination mechanism accessible to all levels of society is important. The government's intervention and efforts are essential to provide a channel for disseminating such information.

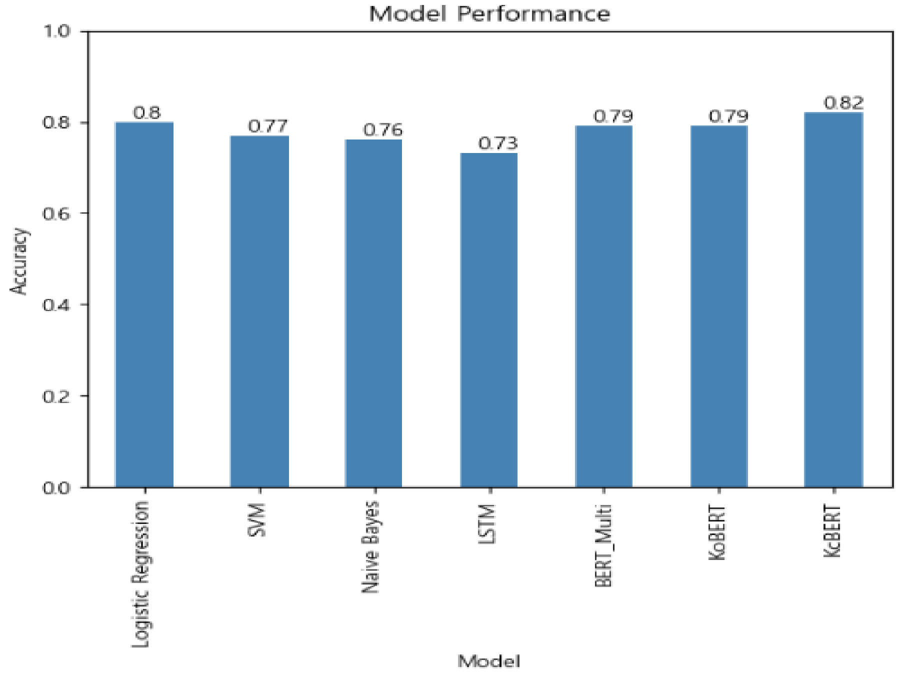
Culture is an important variable that influences people's perceptions and emotions. Culture, often the invisible hand, shapes perceptions. In a society that regards overwork as a mark of honor, negative feelings about working long hours are a clear call for cultural reflection. Likewise, in societies with strong family values, we can witness fervent cries for better parents and support for the elderly. Such cultural remarks of the people appearing in YouTube comments will be an essential criterion in determining the government's direction and focus on policy, and various channels must be prepared to confirm the public's perception through these cultural remarks.

As we weave this complex web of emotions, we find opinions, stories, hopes, dreams, and sometimes silent cries for change. When flowing in the right direction, these insights can be catalysts for transformative social change.

5.3 Performance Comparison of NLP Models for Sentiment Analysis of Korea's Low Fertility Policy: A Comprehensive Review

5.3.1 Overview of comparison training models

As we journey into the sophisticated realm of natural language processing (NLP), the myriad techniques available to decode emotions are a testament to the rapid progress in this field. Our primary goal in this study is to juxtapose deep learning paradigms with traditional machine learning methods while focusing on sentiment analysis of Korea's low birth rate policy. A selected dataset from YouTube comments provides the spontaneity and candor typical of public discourse.



<Fig 12. Comparison of Training Models>

An overview of model accuracy is introduced before looking at each model in detail. A comparison of model accuracies reveals varying performance levels across the seven models tested for sentiment analysis. Accuracy ranged from about 73% to 82%, highlighting the difference in accurately classifying emotions.

KcBERT leads the way with the highest accuracy of 82% and exhibits an excellent ability to capture complex text patterns, especially in language-specific contexts such as Korean.
BERT_Multi and KoBERT follow closely with 79% accuracy each, demonstrating the strength of transformer-based models in handling complex language tasks.
Naive Bayes was surprised with an impressive 75.9% accuracy. This is commendable for existing algorithms and indicates its adaptability to text classification tasks.
Logistic regression provides a respectable accuracy of 79.9%, highlighting its usefulness as a robust baseline model for binary classification tasks.
SVM has an accuracy of 76.8%, which is a testament to its ability to find optimal boundaries in high-dimensional data, even if slightly better than other models.
Despite its sequence modeling ability, LSTM gives an accuracy of 73.0%, which may suggest the need for model optimization or problems inherent to the data set.

Visually, the bar chart effectively shows the marginal difference in performance, with the highest bar for KcBERT and the bars for BERT_Multi, KoBERT, and Naive Bayes highlighting competing accuracies.

a. The Traditional Algorithmic Model: Revisiting the Basic Skills

- Logistic regression:

<Table 3 : Classification Report – Logistics Regression>

	precision	recall	F1-score	Support
0	0.83	0.55	0.66	375
1	0.79	0.94	0.86	670
Accuracy			0.80	
Macro avg	0.81	0.74	0.76	1045
Weight avg	0.80	0.80	0.79	1045

The logistic regression model's achievement of 80% accuracy is a testament to the resilience and reliability of this veteran model. The balance of precision and recall inferred from the f1 score indicates balanced classification power. Although logistic regression may seem rudimentary to some, its performance highlights its ability to be a reliable criterion. However, we must recognize the potential downside of not being able to capture the more complex linguistic nuances embedded in YouTube comments. Laudable accuracy on negative sentiment means potentially harmful content can be quickly flagged on large data sets where rapid evaluation is essential. However, the model's decline in recall of negative emotions is a caveat. This means it can underperform in applications such as content alerts, where identifying all negative emotions is critical. The model's efficiency and computational lightness make it an invaluable tool for fast sentiment decisions on massive data sets.

- Support Vector Machine (SVM):

<Table 4 : Classification Report – Support Vector Machine>

	precision	recall	F1-score	Support

0	0.88	0.41	0.56	375
1	0.75	0.97	0.84	670
Accuracy			0.77	
Macro avg	0.81	0.69	0.70	1045
Weight avg	0.79	0.77	0.74	1045

SVM shows a remarkable strength in discriminating positive sentiment with an accuracy of 77%, especially the f1 score. What is interesting about SVMs is their inherent flexibility. The choice of kernel and hyperparameters can significantly impact performance, mainly when applied to multifaceted topics such as low birthrate policies. The guiding principle of maximizing margins between classes is adept at separating complex data clusters, especially when sentiment is polarizing. However, recall of negative emotions is a concern. Missing complaint detection in applications such as automated customer support can lead to significant customer complaints, indicating that relying solely on SVMs can be a gamble.

- Naive Bayes:

<Table 5 : Classification Report – Naïve Bayes>

	precision	recall	F1-score	Support
0	0.65	0.73	0.69	375
1	0.84	0.78	0.81	670
Accuracy			0.77	
Macro avg	0.74	0.75	0.75	1045
Weight avg	0.77	0.76	0.76	1045

With a respectable accuracy of 76%, Naive Bayes reminds us of the beauty of simplicity. Performance is good despite the default hypothesis of functional independence. However, there is a fundamental question. Can these assumptions be applied consistently for interconnected and context-rich YouTube comments? If emotion is deeply embedded within the linguistic nuances of a text, such as poetry or complex prose, Naive Bayes may not be the best choice.

b. Deep learning models: venture into advanced paradigms

- LSTM (Long Short Term Memory):

<Table 6 : Classification Report – LSTM>

	precision	recall	F1-score	Support
0	0.63	0.63	0.63	375
1	0.79	0.79	0.79	670
Accuracy			0.73	
Macro avg	0.71	0.71	0.71	1045
Weight avg	0.74	0.73	0.74	1045

The LSTM architecture showed the ability to learn over multiple iterations, reaching a final test accuracy of 73%. This indicates the ability to learn complex patterns and dependencies within the text when juxtaposed with an initial accuracy of 65.33%. LSTMs with memory cells are designed to process sequences in nature, making them suitable for annotation analysis. However, a closer look at the validation loss and accuracy metrics over different epochs reveals potential overfitting. The model may be too tuned to the training data, reducing its capacity to generalize to unseen data. Considering that this model is a deep learning model and is suitable for a much more complex and large data set than conventional machine learning models, the potential overfitting shows limitations in choosing an LSTM model.

c. BERT variant

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a Google invention tailored to understand the context of words in a sentence. Multilingual variants provide a uniform solution for multiple languages, but for specific languages such as Korean, custom models such as KoBERT and KcBERT have been trained. This study used a BERT multilingual variant model and three BERT variant models, KoBERT and KcBERT. These models' training dynamics and performance metrics provide insight into their efficiency and nuance. BERT_Multilingual Training Dynamics, KoBERT and KcBERT, say a lot about the consistency of BERT-based architectures with over 79% parallel performance. Details for each model are as follows.

- BERT_Multilingual

<Table 7 : Classification Report – BERT_Multilingual >

	precision	recall	F1-score	Support
0	0.69	0.72	0.71	375
1	0.84	0.82	0.83	670
Accuracy			0.79	
Macro avg	0.77	0.77	0.77	1045
Weight avg	0.79	0.79	0.79	1045

The BERT_Multilingual Training Dynamics model performance showed significant loss reduction during early epochs. It started at a loss of 0.6867 in the first epoch and experienced a sharp decline, with a loss of 0.1862 at the start of the second epoch. This behavior demonstrates the model's proficiency in quickly adapting to the complexity of the data set. However, the loss values oscillated as training continued, suggesting continuous model recalibration to fine-tune the emotion classification function. The eighth epoch witnessed a meager loss value of 0.0017, a testament to the model's profound understanding of the data set. However, subsequent epochs, especially the ninth, show an increase in loss, which may indicate overfitting. Early dismissal was introduced in response, and the final accuracy was 79%.

- KoBERT

<Table 8 : Classification Report – KoBERT >

	precision	recall	F1-score	Support
0	0.71	0.71	0.71	375
1	0.84	0.83	0.84	670
Accuracy			0.79	
Macro avg	0.77	0.77	0.77	1045
Weight avg	0.79	0.79	0.79	1045

Designed explicitly for Korean text, KoBERT consistently performed well throughout the learning process. In the early stages of training, the ability of the model to quickly adapt to the structure and nuances of the data set was evident. This may be due to the Korean-optimized architecture, which can help capture complex language patterns more effectively than typical BERT models. A significant decrease in loss values was observed in early epochs. This fast adaptability is a model that can benefit from prior knowledge and context by fine-tuning pre-trained weights to the Korean corpus. The stability of loss reduction, especially when compared to BERT_Multilingual, suggests a more focused and streamlined training trajectory. However, as with most deep-learning models, later epochs began to show minimal signs of oscillation in the loss curve. This behavior is not uncommon, especially when the model reaches an optimal state, and can sometimes be exceeded in gradient updates. A slight increase in the loss by the ninth epoch could be an early indicator of overfitting. This is where the model starts memorizing the training data rather than generalizing it. Implementing strategies such as early dismissal can be an effective countermeasure against such pitfalls. An accuracy of 0.79% is

It is not different from BERT_Multilingual Training Dynamics, but given the characteristics of a Korean-specific model, it is evidence of solid performance and the effectiveness of the applied measures.

- KcBERT

<Table 9 : Classification Report – Naïve Bayes>

	precision	recall	F1-score	Support
0	0.89	0.66	0.76	100
1	0.78	0.94	0.86	132
Accuracy			0.82	232
Macro avg	0.84	0.80	0.81	232
Weight avg	0.83	0.82	0.81	232

The training dynamics of KcBERT presented an exciting trajectory. The model starts with a loss of 0.3939 on the first batch of the first epoch and rises slightly to 0.4384 on the 200th batch. This initial performance suggests that the model understands the structure and complexity of the data set. In the second epoch, the model started with a loss close to the second half of the first epoch (0.4266) but managed a significant decrease, reducing the loss to 0.1895 by the end. The third epoch showed a much more pronounced improvement, with the loss dropping to 0.0937 but increasing to 0.2674. The trend of significant loss reduction continued, reaching the lowest values at the fifth and sixth epochs, where the loss decreased to values such as 0.0056 and 0.0013, respectively. However, spikes such as 0.3875 in the sixth epoch are placed between these troughs, demonstrating the oscillatory behavior of the loss value during training. Since the seventh epoch, the model has maintained low loss figures with intermittent slight spikes. At the 9th and 10th epochs, the loss values consistently remained below 0.004, indicating that the model achieved an optimal understanding of the data distribution.

Regarding classification performance, KcBERT provided an accuracy of 0.82. Precision, recall, and f1-score show that the model does an excellent job at detecting 'negative' sentiment (f1-score of 0.86), while it lags slightly behind at identifying 'positive' sentiment, with an f1-score of 0.76. Nevertheless, the accuracy of the 'positive' class is surprisingly high at 0.89, but the recall of 0.66 suggests that there are certain positive instances that the model missed.
In summary, KcBERT's training showed fast adaptability, scattered loss variability, and commendable performance in the second half. The classifier's high precision and slightly lower recall for the 'positive' class suggests that future tuning can focus on increasing the model's sensitivity to positive sentiment without compromising specificity.

5.4 Final Conclusion:

Selecting the ideal model for sentiment analysis requires an evaluation that combines training dynamics, adaptability, performance metrics, and specific use-case relevance.

a. Training mechanics:

- BERT_Multilingual : Experienced fluctuations in loss values, which may suggest potential overfitting in later epochs.
- KoBERT: More reliable than BERT_Multilingual, but hints at potential overfitting in later steps.
- KcBERT: shows rapid adaptability with significant loss reduction, especially in late epochs, indicating learning ability.
- LSTM: In general, LSTMs tend to learn sequential patterns well, but can be computationally intensive and struggle with long sequences unless combined with other

mechanisms.

- SVM, Logistic Regression, and Naive Bayes: As traditional algorithms, these models may not capture complex patterns in text data as effectively as deep learning models, but they can serve as solid baselines.

b. Performance metrics:

- BERT_Multilingual and KoBERT: Both rendered an accuracy of 0.79%. However, given the nature of deep learning, the computational cost may not justify the performance.
- KcBERT: achieved the highest accuracy of 0.82%. Balanced performance across classes (precision, recall and F1 score) stands out.
- LSTM: LSTMs can be robust, but without details they can generally perform well on sequence data, but can outperform transformer-based models such as BERT derivatives.
- SVM and Logistic Regression: Deep learning models may have provided significant accuracy, but they have demonstrated the ability to outperform these models on large and complex data sets.
- Naive Bayes: Generally good for text classification, but may make limited assumptions about feature independence.

c. Model specificity:

- KcBERT: Specialized KcBERT is more likely to capture nuances specific to a dataset than more generalized models such as BERT_Multilingual.
- LSTM: Provides sequential modeling, but lacks the depth and complexity of transformer architecture.
- SVM, Logistic Regression, and Naive Bayes: May not capture the contextual relevance of text data because they are non-sequential.

d. Computational considerations:

- Deep learning models (BERT variants, LSTMs): require significant computing resources, especially during training. However, the ability to process large data sets and capture complex patterns can justify the computational cost.

- Traditional models (SVM, Logistic Regression, Naive Bayes): Not computationally intensive and fast to learn. Excellent for setting baselines or when computational resources are limited.

youWhile traditional models such as LSTM and SVM, Logistic Regression and Naive Bayes have their advantages, KcBERT's consistent performance combined with its specialization provides the best combination of accuracy and model complexity for that dataset. This comprehensive analysis supports the decision to select KcBERT for the study.