



Construction of Consensus-Labeled Data for the Study of Emotions in Korean Full-Length Novels

WOO-KYU KANG 

DISCUSSION PAPER

 ubiquity press

ABSTRACT

This study proposes a consensus-labeling methodology for constructing emotion data for Korean full-length novels by integrating multiple annotations, model predictions, and researcher verification. Emotion interpretation in literary texts may vary across researchers, while sentence-level emotion prediction models may not fully capture emotions that depend on broader narrative context. First, for 5,282 sentences from *Sohyeonseongrok*, one set of deep learning model predictions and two sets of human annotations were combined. Provisional labels were derived through majority rule and then subjected to researcher verification based on surrounding narrative context. Next, eight emotion prediction models trained under different training-data configurations were applied to 15,426 sentences from *Sossi-Samdaerok*. When five or more models predicted the same emotion, the shared label was treated as a provisional model consensus and subsequently reviewed by the researcher. Model comparisons showed that predictions varied according to training-data configuration, while the model trained on the *Sohyeonseongrok* consensus-labeled data achieved agreement comparable to a model trained on a substantially larger dataset formed by concatenating three annotation datasets. These findings suggest that consensus labeling can provide practical shared reference labels for computational analysis while preserving information about differing annotation perspectives, rather than establishing a single objectively correct emotion label. The resulting datasets are publicly available on Zenodo under a CC-BY 4.0 license for reuse in Digital Humanities and NLP research on Korean classical literature.

CORRESPONDING AUTHOR:

Woo-kyu Kang

Institute of Koreanology,
Kangwon National University,
Chuncheon, South Korea

cuemi97@gmail.com

KEYWORDS:

digital humanities; emotion
annotation; consensus
labelling; Korean full-length
novels; deep learning; human-
in-the-loop

TO CITE THIS ARTICLE:

Kang, W. (2026). Construction
of Consensus-Labeled Data
for the Study of Emotions in
Korean Full-Length Novels.
*Journal of Open Humanities
Data*, 12: 137, pp. 1–14. DOI:
[https://doi.org/10.5334/
johd.626](https://doi.org/10.5334/johd.626)

(1) CONTEXT AND MOTIVATION

(1.1) METHODOLOGICAL LANDSCAPE AND LIMITATIONS OF EMOTION ANALYSIS IN KOREAN FULL-LENGTH NOVELS

Computational emotion analysis can support the investigation of literary genres and narrative structures (Kim et al., 2017). However, applying such approaches to Korean full-length novels requires the construction of emotion data suitable for computational analysis, a process that raises particular methodological challenges because of the scale and narrative complexity of these texts. Korean full-length novels feature extensive, multi-generational narrative structures. While traditional close reading can capture complex psychological shifts and Confucian conflicts, applying this qualitative approach consistently across such extensive texts is difficult in practice. Moreover, individual researchers' emotion annotations may reflect researcher-specific interpretive perspectives, which can contribute to variation across label sets. As recent NLP studies have noted, emotion annotation in textual data may show low inter-annotator agreement because it relies heavily on subjective human interpretation (Öhman, 2020).

Conversely, quantitative deep learning models trained on modern-language datasets (e.g., the Korean Online That-gul Emotions (KOTE) Dataset (Jeon et al., 2022)¹) may have difficulty capturing the archaic syntax and complex discourse contexts characteristic of Korean full-length novels. Historical literary texts can pose challenges for models trained on modern-language data because of linguistic change, archaic vocabulary, and period-specific rhetoric (İlter et al., 2026). In Korean full-length novels, emotion interpretation may additionally depend on broader narrative and ethical contexts, and such information may not always be fully recoverable from an isolated sentence.

(1.2) CONSTRUCTION OF CONSENSUS-LABELED DATA TO OVERCOME DATA SCARCITY

Due to the extensive length of Korean full-length novels and the limited number of researchers engaged in digital humanities research on these texts, only a small number of manually annotated emotion datasets are currently available, including Kang's data,² Kwon's data, and Yoon's data.³ While data augmentation and synthesis can increase the volume of training data (Bayer et al., 2023; Vu et al., 2022), increasing data quantity alone does not necessarily resolve differences in annotation criteria or ensure semantic compatibility with historical literary texts. These limitations are particularly relevant to emotion annotation, where interpretations may vary across annotators and depend on narrative and historical context.

Consensus labeling provides a practical means of integrating multiple annotations in domains where individual judgments may differ. Rather than treating one annotation as inherently authoritative, a consensus procedure identifies shared patterns across multiple sources of annotation evidence. Previous studies have shown that disagreement is a persistent feature of emotion annotation and that consensus-based approaches can be useful for reconciling heterogeneous labels (Öhman, 2020; Shoer & Erzin, 2025). However, conventional consensus-building generally depends on multiple human annotators, which can require substantial time and labor when applied to the full texts of large-scale classical novels. Accordingly, this study proposes a methodology for constructing consensus-labeled data by comparing and reconciling existing human annotations with predictions generated by multiple deep learning models.

This methodology builds on a series of the author's previous studies. The 18-category emotion classification system, developed by reorganizing the 44 KOTE emotion categories to reflect the characteristics of Korean full-length novels, was established in Kang (2024). Based on this classification system, Kang's data was constructed, and the deep learning emotion prediction

¹ Available at <https://github.com/searle-j/KOTE>.

² This is a qualitative emotion identification dataset for the novel Changseongamuirok, annotated by Woo-kyu Kang (Kim & Kang, 2024). Available at OSF: <https://doi.org/10.17605/OSF.IO/XSY8U>.

³ These are qualitative emotion identification datasets for the novel Sohyeonseongrok, annotated independently by Dae-hoon Kwon and Eun-jeong Yoon. Retrieved from <https://osf.io/tcau2>.

‘model v.1.1.1’⁴ trained on this dataset was developed and evaluated in Kang and Kim (2024). The sentence-level agreement patterns among the three *Sohyeonseongrok* label sets used in the present study—predictions from model v.1.1.1, Yoon’s data, and Kwon’s data—as well as the interpretation of these differences in terms of readers’ horizons of expectation, were previously discussed in Kang (2025a). Building on these findings, the present study proceeds from the premise that individual researcher annotations need not be treated as representing a single correct interpretation of emotion and that deep learning model predictions can also serve as one source of annotation evidence.

On this basis, two complementary consensus-labeling procedures were developed. For *Sohyeonseongrok*, the original annotation columns from the previously released label sets were retained, while a new consensus label was derived through majority-rule aggregation and researcher verification. Thus, the present study extends the previous comparison of heterogeneous human and model annotations into a consensus-based data-construction procedure. For *Sossi-Samdaerok*, emotion predictions generated by model v.1.1.1 and genre- and character-related emotional patterns were previously analyzed in Kang (2025b). Building on this prior work, the present study extends the earlier single-model analysis by constructing seven additional emotion prediction models, applying all eight models to *Sossi-Samdaerok*, and deriving consensus labels through a multi-model majority-rule procedure followed by researcher verification. Thus, the primary contribution of the present study lies in constructing and evaluating two complementary consensus-labeling procedures: a three-source consensus combining two human annotations and one AI-model prediction for *Sohyeonseongrok*, and an eight-model majority consensus for *Sossi-Samdaerok*, both followed by researcher verification. This approach can contribute to mitigating the influence of individual annotator subjectivity, supplementing machine predictions with literary context that is difficult for models to capture, and supporting a transparent and reproducible framework for future research in classical literature and digital humanities.

(2) DATASET DESCRIPTION

This section provides specifications for the dataset’s bibliographic information, metadata, and detailed data structure (columns). The constructed consensus-labeled datasets (Kang, 2026) have been released under an open license in the Zenodo public repository, allowing anyone to access and reuse them.

- **Repository location:** <https://doi.org/10.5281/zenodo.23315786>
- **Repository name:** Zenodo
- **Object name:** *Sohyeonseongrok*_consensus-labeled_data.csv (5,282 sentences), *Sossi-Samdaerok*_consensus-labeled_data.csv (15,426 sentences)
- **Format names and versions:** CSV (Comma-Separated Values), Version v3
- **Creation dates:** 2024-03-01 to 2026-08-26
- **Dataset creators:** Woo-kyu Kang (Dataset creator and Author; Institute of Koreanology, Kangwon National University), Eun-jeong Yoon (Annotator/Contributor; Department of Korean Language and Literature, Chung-Ang University, Seoul, South Korea), Dae-hoon Kwon (Annotator/Contributor; Department of Korean Language and Literature, Chung-Ang University, Seoul, South Korea)
- **Language:** Korean (novel text and emotion-label values), English (variable names/column headers, Zenodo metadata, and README description)
- **License:** CC-BY 4.0 (Creative Commons Attribution 4.0 International)
- **Publication date:** 2026-10-13

Both CSV files contain sentence identifiers and text, source-specific human annotations or model predictions, agreement-status information (Label comparison), final consensus labels

⁴ Model v.1.1.1 is based on the Korean comments ELECTRA (KcELECTRA) language model (Lee, 2021), uses the KOTE dataset as the base emotion classification data, and is trained on Kang’s data. Full implementation details and the source code are openly available in the KENT repository (Kim & Kang, 2024; <https://doi.org/10.17605/OSF.IO/XSY8U>). Furthermore, the architecture and codebase of this v.1.1.1 model served as the foundation for constructing the subsequent seven deep learning models used in this study.

(consensus-label), and a consensus_method indicator. The latter records whether each final label was established through majority_consensus, researcher_resolution, or researcher_override; detailed column definitions are provided in the README.

(3) METHOD

The dataset construction methodology of this study focused on building human–AI consensus-labeled data by combining the contextual nuance of qualitative humanities research with the data-processing efficiency of quantitative digital humanities research. To this end, an existing emotion classification system developed to reflect the characteristics of Korean full-length novels was adopted as the labeling standard, and two consensus-labeling methods were applied at successive stages of the research.

(3.1) ADOPTION OF AN 18-CATEGORY EMOTION CLASSIFICATION SYSTEM FOR KOREAN FULL-LENGTH NOVELS

As Öhman (2020) points out, there is no universally accepted taxonomy of human emotions, and different annotation schemes must be proposed depending on specific research objectives and textual modalities. To overcome the categorization errors experienced by previous emotion analysis studies that predominantly adopted modern language-based emotion dictionaries, this study utilized Woo-kyu Kang’s 18-category emotion classification system (Kang, 2024) as the label standard. This system was adopted because it was developed to reflect emotion categories relevant to Korean full-length novels, including Samdaerok-type narratives.

Kang (2024) developed this system by reorganizing KOTE’s 44 emotion categories (Jeon et al., 2024) into 18 categories through quantitative analysis and qualitative review of emotion classifications relevant to Korean full-length novels. The mapping is presented in Table 1.

KOTE	KANG
불평/불만(dissatisfaction), 지긋지긋(fed up), 의심/불신(distrust), 한심함(pathetic), 짜증(irritation), 어이없음(preposterous)	불만(dissatisfaction)
감동/감탄(admiration), 고마움(gratitude), 존경(respect)	감동/감사(admiration/gratitude)
환영/호의(welcome), 안심/신뢰(relief)	신뢰/호의(trust/favour)
슬픔(sadness), 힘듦/지침(exhaustion), 서러움(sorrow)	슬픔(sadness)
화남/분노(anger), 역겨움/징그러움(disgust), 증오/혐오(contempt), 경악(shock)	분노/혐오(anger/disgust)
기대감(expectancy), 신기함/관심(interest)	기대(expectation)
아껴주는(care), 호뭇함(귀여움/예쁨)(attracted)	애정(affection)
기쁨(joy), 뿌듯함(pride), 편안/쾌적(comfort), 즐거움/신남(excitement), 행복(happiness)	기쁨(joy)
우쭐댐/무시함(arrogance)	우쭐댐/무시함(arrogance)
안타까움/실망(disappointment)	안타까움/실망(disappointment)
불쌍함/연민(compassion)	불쌍함/연민(compassion)
부끄러움(shame), 절망(despair), 패배/자기혐오 (defeat/self-hatred), 죄책감(guilt)	수치/죄책(shame/guilt)
공포/무서움(fear), 불안/걱정(anxiety)	두려움(fear)
없음(NO EMOTION)	없음(none)
깨달음(realization)	깨달음(realization)
당황/난처(embarrassment), 부담/안_내킴(reluctant)	당황/난처(embarrassment/reluctant)
비장함(resolute)	비장함(resolute)
놀람(surprise)	놀람(surprise)

Table 1 Mapping between the KOTE Dataset Emotion Classification and Woo-kyu Kang’s Emotion Classification (adapted from Kang, 2024).

(3.2) CONSTRUCTION OF CONSENSUS-LABELED DATA FOR SOHYEONSEONGROK

For the first target text, *Sohyeonseongrok* (5,282 sentences), a three-party consensus method involving one deep learning model and two human annotators was adopted.

(3.2.1) Input Label Sets for Consensus Construction

In this study, three existing emotion label sets—prediction data from model v.1.1.1, Yoon’s data, and Kwon’s data—were used as inputs for the consensus-labeling procedure. The emotion labels in these three datasets refer to emotions expressed or implied in each sentence, identifying the emotions of the characters or narrator rather than those experienced by the reader. These three label sets had previously been used and compared in Kang (2025a). In the present study, the three existing label sets were aligned at the sentence level and used as three annotation sources for consensus labeling. Model v.1.1.1 treated each sentence as an independent input unit; preceding and subsequent sentences were not included in the model input. Table 2 presents an example of Yoon’s data, and Table 3 presents an example of prediction data generated by model v.1.1.1.

Table 2 Example of the Pre-existing *Sohyeonseongrok* Emotion Annotation Data (Yoon’s data).

ID	TEXT	EMOTION CLASSIFICATION
1	송나라 태종(太宗) 때 강릉후 겸 구석(九錫) 참지정사(參知政事) 용도각(龍圖閣) 태학사(太學士)인 소공의 이름은 경이고, 자는 자문이였다. During the reign of Emperor Taejong of Song, Lord So, who served as Marquis of Gangneung, Chancellor with Nine Bestowals, and Grand Scholar of the Dragon Pavilion, was named Gyeong, with the courtesy name Jamun.	none
2	그는 날 때부터 산천이 지닌 기운과 해와 달의 정기, 그리고 천지의 조화를 타고났다. From birth, he was endowed with the energy of mountains and rivers, the essence of the sun and moon, and the harmony of heaven and earth.	expectation
3	가슴에는 세상을 다스릴 뜻을 품었고 얼굴에는 어지러운 나라를 평안하게 할 재주가 어려 있었으며, 빼어난 문장과 뛰어난 절개가 당대에 으뜸이었고 수려한 풍채는 천고에 따를 자가 없을 정도였다. In his heart, he harbored the ambition to govern the world; his face reflected the talent to pacify a chaotic nation; his outstanding writing and unyielding integrity were paramount in his era, and his handsome appearance was unmatched throughout history.	expectation

(3.2.2) Majority Rule Consensus

In this study, consensus labels were derived by applying the majority-rule principle at the sentence level to the three existing label sets described in Section 3.2.1. Because emotion categories may overlap and annotator judgments may vary, unanimous agreement can be difficult to achieve in emotion annotation. As discussed by Öhman (2020), majority agreement can provide a practical means of deriving consensus when unanimous agreement is not achieved. Accordingly, the present study adopted majority agreement as the decision rule for consensus construction. Thus, when two or more of the three labels matched, the shared label was treated as the provisional consensus emotion label.

Table 3 Example of the Pre-existing *Sohyeonseongrok* Emotion Prediction Data Generated by Model v.1.1.1.

ID ⁵	ADMIRATION/ GRATITUDE	EXPECTA- TION	JOY	REALIZATION	SURPRISE	EMBARRASSMENT/ RELUCTANT	FEAR	ANGER/ DISGUST	DISSATISFAC- TION
1	.025	.014	.013	.019	.011	.015	.013	.015	.014
2	.672	.029	.028	.201	.012	.011	.017	.023	.018
3	.85	.034	.02	.044	.017	.016	.019	.045	.026
ID	COMPASSION	RESOLUTE	SHAME/ GUILT	SADNESS	TRUST/ FAVOUR	DISAPPOINTMENT	AFFECTION	NONE	ARROGANCE
1	.011	.016	.012	.016	.019	.015	.009	.917	.011
2	.021	.022	.034	.033	.038	.029	.023	.051	.027
3	.018	.022	.038	.031	.033	.021	.021	.099	.024

⁵ The ID refers to the sentence identifier assigned to each sentence in the dataset.

(3.2.3) Researcher Verification

For sentences in which the AI model and the two human annotators assigned three different emotion labels, the researcher reviewed the preceding and subsequent narrative context and the typological characteristics of Korean full-length novels and assigned the final emotion label from the full set of 18 emotion categories adopted in this study. The researcher was not restricted to selecting one of the three labels initially assigned by the AI model and the two human annotators; when the broader narrative context supported a different interpretation, an emotion category not proposed by any of the three annotation sources could be assigned as the final label. As Öhman (2020) points out, emotion annotation becomes particularly challenging when emotional expressions lack explicit surface-level lexical markers and depend heavily on context. Therefore, researcher verification provided an additional step for resolving such context-dependent ambiguities.

Researcher verification was also applied to cases in which two of the three labels matched when the majority label appeared inconsistent with the surrounding narrative context. In 52 such cases, the majority-consensus label was replaced with a different emotion category based on contextual review. Through this process, a consensus-based emotion dataset was constructed with additional researcher verification of cases involving annotation disagreement or contextual ambiguity, as shown in Table 4.

Table 4 Example of Sohyeonseongrok Consensus-Labeled Data.

ID	TEXT	V.1.1.1	YOON	KWON	LABEL COMPARISON	CONSENSUS-LABEL	CONSENSUS_METHOD
1	송나라 태종(太宗) 때 강릉후 경 구석(九錫) 참지정사(參知政事) 용도각(龍圖閣) 태학사(太學士)인 소공의 이름은 경이고, 자는 자문이였다. During the reign of Emperor Taejong of Song, Lord So, who served as Marquis of Gangneung, Chancellor with Nine Bestowals, and Grand Scholar of the Dragon Pavilion, was named Gyeong, with the courtesy name Jamun.	none	none	none	C = D = E	none	majority_consensus
2	그는 날 때부터 산천이 지닌 기운과 해와 달의 정기, 그리고 천지의 조화를 타고났다. From birth, he was endowed with the energy of mountains and rivers, the essence of the sun and moon, and the harmony of heaven and earth.	admiration/gratitude	expectation	none	C ≠ D ≠ E	admiration/gratitude	researcher_resolution

(3.3) CONSTRUCTION OF MULTI-MODEL CONSENSUS-LABELED DATA FOR SOSSI-SAMDAEROK

For *Sossi-Samdaerok*, which consists of 15,426 sentences, this study applied a multi-model majority-rule approach to reduce the burden of manual annotation by integrating emotion predictions from multiple models.

(3.3.1) Construction of Seven Additional Models with Diverse Training-Data Configurations

In the present study, model v.1.1.1, developed and evaluated in Kang and Kim (2024) and previously applied to *Sossi-Samdaerok* in Kang (2025b), was incorporated into a multi-model consensus framework. Seven additional deep learning emotion prediction models (v.1.1.2–v.1.1.8) were constructed using different training-data configurations. The three previously released datasets—Kang’s data, Kwon’s data, and Yoon’s data—and the *Sohyeonseongrok* consensus-labeled dataset constructed in this study were used individually or in combination as training data. The eight models were then applied to *Sossi-Samdaerok*, and their emotion predictions were compared to construct multi-model consensus labels.

- ① Model v.1.1.1: Based on ㉠ Woo-kyu Kang’s 18-category emotion classification system, ㉡ KcELECTRA as the base language model (Lee, 2021), ㉢ KOTE as the base emotion classification dataset, and ㉣ Kang’s data as the training dataset. All models used the same codebase and training procedure, with a row-level 90/10 train-validation split (random_state = 200), a random seed of 42, a batch size of 8, a learning rate of 2×10^{-5} ,

a maximum of 15 epochs, and early stopping based on validation loss (patience = 5). The validation data were used for checkpoint selection rather than independent performance evaluation. Each sentence was processed independently without adjacent-sentence context, and combined datasets were concatenated at the row level without reconciling conflicting labels for identical sentences. The effect of such conflicting duplicate instances on model performance was not independently evaluated in the present study. Components ㉔–㉞ were held constant across all model versions, while only ㉔, the training-data configuration, varied. Accordingly, only the version number and training data are presented for models ㉒–㉝ below. Full implementation details are provided in the README and OSF code repository.

- ㉒ v.1.1.2: Model trained on Kwon’s data.
- ㉓ v.1.1.3: Model trained on Yoon’s data.
- ㉔ v.1.1.4: Model trained on Kwon’s data + Yoon’s data.
- ㉕ v.1.1.5: Model trained on Kang’s data + Kwon’s data.
- ㉖ v.1.1.6: Model trained on Kang’s data + Yoon’s data.
- ㉗ v.1.1.7: Model trained on Kang’s data + Kwon’s data + Yoon’s data.
- ㉘ v.1.1.8: Model trained on *Sohyeonseongrok* consensus-labeled data.

(3.3.2) Multi-Model Majority-Rule Consensus

The eight models were applied to *Sossi-Samdaerok*, with each model generating prediction scores for the 18 emotion categories for each sentence. The emotion category with the highest prediction score was selected as each model’s predicted label. If the same emotion label was predicted by five or more of the eight models (majority agreement), the shared label was treated as the provisional model-consensus emotion label for the sentence.

(3.3.3) Researcher Verification Based on Narrative Context

For the 6,791 sentences in which fewer than five of the eight models predicted the same emotion label, the researcher reviewed the preceding and subsequent narrative context and assigned the final emotion label from the full set of 18 emotion categories adopted in this study. The researcher was not restricted to selecting one of the emotion labels predicted by the eight models; when the broader narrative context supported a different interpretation, another category from the 18-category emotion classification system could be assigned as the final label. Researcher verification was also applied to sentences for which five or more models had reached a model consensus when the consensus label appeared inconsistent with the surrounding narrative context. As a result, 297 model-consensus labels were revised based on contextual review. Through this process, the final consensus-labeled dataset incorporated both multi-model majority agreement and researcher verification based on narrative context; an example is shown in Table 5.

Table 5 Example of *Sossi-Samdaerok* Consensus-Labeled Data.

ID	TEXT	V.1.1.1	V.1.1.2	V.1.1.3	V.1.1.4	V.1.1.5	V.1.1.6	V.1.1.7	V.1.1.8	LABEL COMPARISON	CONSENSUS- LABEL	CONSENSUS_ METHOD
1	대송 시절에 승상 소현성의 이름은 경이고 자(字) 는 자운이었다. During the Great Song Dynasty, Chancellor So Hyeon-seong’s given name was Gyeong and his courtesy name was Jamun.	none	none	none	none	none	none	none	none	v.1.1.1 = v.1.1.2 = v.1.1.3 = v.1.1.4 = v.1.1.8 = v.1.1.5 = v.1.1.6 = v.1.1.7	none	majority_ consensus

(4) RESULTS AND DISCUSSION

This section examines model-agreement patterns and the roles of training-data configuration, majority-rule consensus, and researcher verification in literary emotion-data construction.

(4.1) INFLUENCE OF ANNOTATOR-SPECIFIC PERSPECTIVES ON MODEL PREDICTIONS

The multi-model comparison suggests that deep learning models may reflect annotator-specific perspectives present in their training data. As a starting point for examining this pattern, it is useful to consider the differences previously observed among the three *Sohyeonseongrok* label sets used in the present study.

According to Kang (2025a), among the three *Sohyeonseongrok* label sets—predictions from model v.1.1.1, Yoon’s data, and Kwon’s data—all three labels matched in 22.30% of the sentences, whereas all three labels differed in 36.88%. The pairwise agreement rates were 33.96% between model v.1.1.1 and Yoon’s data, 32.64% between model v.1.1.1 and Kwon’s data, and 41.12% between Yoon’s and Kwon’s data. Kang (2025a) interpreted these differences as reflecting distinct horizons of expectation represented in the three label sets.

To examine the relationship among the three label sets beyond simple agreement rates, the present study calculated Fleiss’ Kappa for the three emotion label sources—predictions from model v.1.1.1, Yoon’s data, and Kwon’s data. The resulting inter-source agreement coefficient was 0.2978, indicating limited agreement among the three label sets. As Öhman (2020) notes, agreement in emotion annotation can be constrained by the overlap among emotion categories and variation in annotation judgments. The observed level of agreement provides quantitative evidence of the potential influence of annotator- or model-specific perspectives on emotion labeling.

The eight models trained on Kang’s, Kwon’s, and Yoon’s data, their combinations, or the consensus-labeled data derived from these annotation sources showed variation in their emotion prediction agreement rates. This suggests that annotator-specific perspectives reflected in the training data may influence model predictions.

To quantify the overall level of agreement among the predictions of the eight models, Fleiss’ Kappa was calculated. The resulting coefficient was 0.4390, indicating a moderate level of agreement among the models. When considered together with the variation in pairwise agreement rates shown in Table 6, this result indicates that model predictions varied according to training-data configuration and is also consistent with the possibility that annotator-specific perspectives reflected in the training data may influence emotion predictions.

To examine the relationship between training-data configuration and model agreement more systematically, the mean agreement rate of each model with the other seven models was calculated. Among the three models trained on single annotation datasets, the mean agreement rates were 44.56% for model v.1.1.1, 40.98% for model v.1.1.2, and 48.57% for model v.1.1.3, with an overall mean of 44.70%. By comparison, models v.1.1.4, v.1.1.5, and v.1.1.6, each trained on a combination of two annotation datasets, showed mean agreement rates of 51.98%, 49.18%, and 49.54%, respectively, with an overall mean of 50.23%. Model v.1.1.7, trained on the combination of all three annotation datasets, showed the highest mean agreement rate among the eight models at 53.37%. These comparisons suggest two main implications.

MODEL	V.1.1.1	V.1.1.2	V.1.1.3	V.1.1.4	V.1.1.5	V.1.1.6	V.1.1.7	V.1.1.8
v.1.1.1	100	31.97	43.8	45.42	44.14	46.91	51.34	48.31
v.1.1.2	31.97	100	35.79	45.49	54.72	33.18	46.44	39.26
v.1.1.3	43.8	35.79	100	52.09	40.47	56.68	50.19	60.97
v.1.1.4	45.42	45.49	52.09	100	52.41	52.78	60.16	55.52
v.1.1.5	44.14	54.72	40.47	52.41	100	46.81	59.39	46.34
v.1.1.6	46.91	33.18	56.68	52.78	46.81	100	53.87	56.58
v.1.1.7	51.34	46.44	50.19	60.16	59.39	53.87	100	52.22
v.1.1.8	48.31	39.26	60.97	55.52	46.34	56.58	52.22	100

Table 6 Emotion Prediction Agreement Rates Among Models.

First, models trained on more diverse annotation sources tended to show higher mean agreement with the other models. However, this tendency may be associated not only with greater diversity of annotation sources but also with increased training-data volume. Because the number of annotation sources and the amount of training data increased simultaneously in the present model configurations, the individual effects of these two factors cannot be separated on the basis of the present comparison alone.

Second, model v.1.1.8, trained on the *Sohyeonseongrok* consensus-labeled data, showed a level of agreement comparable to that of models trained on larger combinations of annotation datasets. This does not mean that consensus-labeled data consistently outperform combined training data; rather, it suggests that training data refined through a consensus process may provide a viable alternative to simply increasing the volume of training data.

Taken together, these results are consistent with the view that emotion labels are not fixed correct answers but may reflect annotator-specific interpretive perspectives. Deep learning models trained on such labels may likewise reproduce aspects of those perspectives. Accordingly, comparing predictions across multiple models can help reveal not only differences in emotion recognition across annotation sources but also areas of shared recognition among them.

(4.2) VALUE OF CONSENSUS-LABELED DATA AS AN EFFICIENT TRAINING RESOURCE

Section 4.1 examined how annotator-specific perspectives reflected in the training data may influence model predictions and how model predictions may vary according to the composition of the training data. In general, models trained on multiple datasets showed higher agreement rates with other models than those trained on a single dataset, with model v.1.1.7, trained on the combined data of Kang, Kwon, and Yoon, showing the highest agreement rate. However, model v.1.1.8, trained on the *Sohyeonseongrok* consensus-labeled data, also showed an agreement rate with the other models comparable to that of model v.1.1.7. This pattern suggests the potential efficiency of consensus-labeled data as a training resource. To examine this possibility more closely, this section focuses on a comparison between models v.1.1.7 and v.1.1.8 and considers the value of consensus-labeled data as an efficient training resource.

Model v.1.1.7 was trained on 14,872 sentences obtained by combining Kang's, Kwon's, and Yoon's annotation datasets, whereas model v.1.1.8 was trained on 5,282 *Sohyeonseongrok* consensus-labeled sentences constructed by reconciling differences among the three annotation sources through majority-rule consensus and researcher verification. Table 7 shows the extent to which each model's emotion predictions agree with the provisional model-consensus labels derived through majority voting across the eight models prior to researcher verification. Accordingly, the agreement rates in Table 7 should be understood as measures of consistency with the multi-model consensus rather than as independent estimates of predictive accuracy. As shown in Table 7, model v.1.1.7 showed the highest agreement with the provisional multi-model consensus labels at 91.11%, while model v.1.1.8 showed a comparable agreement rate of 89.28%. These results do not indicate that consensus-labeled data outperform combined training data. Rather, it is noteworthy that model v.1.1.8, trained on approximately one-third of the training data used for model v.1.1.7, achieved a comparable level of agreement. This pattern suggests that, in addition to the quantitative expansion of training data, the way

MODEL	TRAINING DATA VOLUME	MATCHES	VALID SENTENCES (MATCHED BY FIVE OR MORE)	AGREEMENT RATE (%)
v.1.1.1	4,308	6,705	8,635	77.65%
v.1.1.2	5,282	5,683	8,635	65.81%
v.1.1.3	5,282	7,298	8,635	84.52%
v.1.1.4	10,564	7,740	8,635	89.64%
v.1.1.5	9,590	7,019	8,635	81.29%
v.1.1.6	9,590	7,324	8,635	84.82%
v.1.1.7	14,872	7,867	8,635	91.11%
v.1.1.8	5,282	7,709	8,635	89.28%

Table 7 Number and Rate of Agreements between Individual Model Predictions and Provisional Multi-Model Consensus Labels.

in which differences among annotation sources are reconciled may also be associated with consistency in model predictions.

A similar pattern was observed when the predictions of each model were compared with the final consensus labels after researcher verification. Table 8 evaluates the agreement between each model's predictions and the final consensus labels across all 15,426 sentences in the *Sossi-Samdaerok* dataset.

As shown in Table 8, model v.1.1.7 showed an agreement rate of 61.52% with the final *Sossi-Samdaerok* consensus labels, whereas model v.1.1.8 showed a slightly higher agreement rate of 61.75%, a difference of only 0.23 percentage points. This pattern differs from the comparison with the provisional model-consensus labels before researcher verification, where model v.1.1.7 showed the higher agreement rate. However, this very small difference should not be interpreted as evidence that model v.1.1.8 is superior. Rather, the more important point is that model v.1.1.8, trained on 5,282 *Sohyeonseongrok* consensus-labeled sentences, achieved a level of agreement with the final consensus labels that was virtually identical to that of model v.1.1.7, which was trained on 14,872 sentences obtained by combining three annotation datasets. In the context of Korean full-length novel research, where annotated emotion data remain limited and the construction of large-scale new qualitative annotation datasets requires substantial time and resources, this finding suggests that reconciling differences among existing annotation sources through consensus and researcher verification may provide a practical alternative to relying solely on continued quantitative expansion of annotated training data.

MODEL	TRAINING DATA COMPOSITION	TRAINING DATA VOLUME	MATCHED COUNT	AGREEMENT RATE (%)
v.1.1.1	Kang's data	4,308 sentences	8,327	53.98%
v.1.1.2	Kwon's data	5,282 sentences	6,682	43.32%
v.1.1.3	Yoon's data	5,282 sentences	8,670	56.20%
v.1.1.4	Kwon's & Yoon's data	10,564 sentences	9,126	59.16%
v.1.1.5	Kang's & Kwon's data	9,590 sentences	8,359	54.19%
v.1.1.6	Kang's & Yoon's data	9,590 sentences	8,962	58.10%
v.1.1.7	Kang's & Kwon's & Yoon's data	14,872 sentences	9,490	61.52%
v.1.1.8	<i>Sohyeonseongrok</i> consensus-labeled data	5,282 sentences	9,525	61.75%

Table 8 Agreement Rate between *Sossi-Samdaerok* Consensus-Labeled Data and Eight Models' Emotion Predictions.

(4.3) CONTEXT-DEPENDENT EMOTION INTERPRETATION AND THE ROLE OF RESEARCHER VERIFICATION

The preceding section showed that consensus-labeled data may serve as an efficient training resource by achieving a comparable level of consistency across model predictions even with a relatively smaller volume of training data. However, consensus across multiple models did not always correspond to interpretations based on the surrounding narrative context. During the construction of the *Sossi-Samdaerok* consensus-labeled dataset, provisional model consensus was reached for 8,635 sentences in which five or more of the eight models predicted the same emotion. After reviewing the preceding and subsequent narrative context, the researcher revised the consensus labels for 297 of these sentences (approximately 3.44%). This indicates that majority-rule consensus among models can be used to derive emotion labels efficiently, while in some cases the resulting labels may not sufficiently reflect the broader narrative context, making researcher verification a useful complementary step. Because the models received one sentence at a time without adjacent-sentence context, these discrepancies should also be interpreted in relation to the sentence-level input design rather than as an inherent limitation of transformer-based models.

Table 9 presents examples in which provisional model-consensus labels were revised through researcher verification. For ID 369, seven of the eight models predicted 'expectation.' However, when the utterance is examined in its surrounding narrative context, it can be interpreted not simply as a question seeking information about Mr. Bang, but as an expression of Sohyeonseong's hostility and dissatisfaction toward him for having put Lady Wi in danger. Accordingly, the

ID	TEXT	V.1.1.1	V.1.1.2	V.1.1.3	V.1.1.4	V.1.1.5	V.1.1.6	V.1.1.7	V.1.1.8	CONSENSUS-LABEL
369	방생은 어떤 놈인가? What kind of scoundrel is Mister Bang?	none	expectation	expectation	expectation	expectation	expectation	expectation	expectation	dissatisfaction
854	“네가 옛날 악래(惡來)보다 더하구나. You are stronger than Akrae (惡來) of old.	anger/disgust	admiration/gratitude	anger/disgust	anger/disgust	anger/disgust	anger/disgust	admiration/gratitude	anger/disgust	admiration/gratitude
1760	위씨 · 강씨 · 형씨 세 부인이 각각 그 지아비의 풍모를 보니 그들이 공주의 눈에 들 것이 뻔하였다. When the three wives, Wi, Kang, and Hyeong, each saw the appearance of their husbands, it was obvious that they would catch the princess’s eye.	dissatisfaction	arrogance	expectation	joy	expectation	expectation	expectation	expectation	fear

researcher assigned ‘dissatisfaction’ as the final emotion label. This case suggests that surface-level formal features of a sentence, such as an interrogative structure, do not necessarily correspond to the emotion conveyed in its narrative context.

Next, for ID 854, six models identified the emotion as ‘anger/disgust.’ Akrae (惡來)⁶ has traditionally been represented as a negatively evaluated historical figure while also serving as a conventional figure associated with extraordinary physical strength. This historical and cultural ambiguity may have influenced the models’ emotion predictions. However, when the utterance is read together with the immediately following sentence, ID 855, “Thereafter, [he] cherished Un-seong very affectionately,” the comparison with Akrae can be understood not as a criticism of Un-seong but as a positive evaluation of his strength. Accordingly, the researcher assigned ‘admiration/gratitude’ as the final emotion label. This case suggests that the speaker’s evaluative stance may not always be adequately understood from the lexical or surface meaning of an individual sentence alone and that adjacent narrative context can provide an important basis for emotion interpretation.

Finally, for ID 1760, five models identified the emotion as ‘expectation.’ In the immediately following sentence, ID 1761, however, the three wives are described as becoming deeply worried and anxious. When this narrative sequence is taken into account, ID 1760 can be interpreted not simply as expressing a positive expectation about the future, but as describing the wives’ recognition of a possible threat to their marital relationships arising from the princess’s interest in their husbands. Accordingly, the researcher assigned ‘fear’ as the final emotion label. This case demonstrates that the emotion conveyed by a particular sentence may become clearer when interpreted in relation to subsequent narrative developments rather than solely on the basis of expressions contained within the sentence itself.

Although these three cases are not intended to represent all 297 instances revised by the researcher, they show that information extending beyond an individual sentence—such as sentence form, the speaker’s evaluative stance revealed in adjacent sentences, and emotions clarified through subsequent narrative development—can influence emotion labeling in literary texts. Emotion interpretation in literary texts may therefore depend not only on expressions within an individual sentence but also on narrative progression, implicit evaluation, historical and cultural meanings, and relations with the surrounding context. The existence of 297 revised cases indicates that researcher verification functioned as a substantive complementary stage to multi-model consensus, while the examples above illustrate more specifically how some of these revisions involved context-dependent emotion interpretation.

Table 9 Examples of Researcher Verification and Modification of Deep Learning Model Consensus Results.

6 Akrae (惡來) has ambivalent cultural associations. The Shiji portrays him negatively as a retainer of King Zhou but also records his exceptional strength, while Romance of the Three Kingdoms invokes him as a model of extraordinary physical power in Cao Cao (曹操)’s praise of Dian Wei (典韋).

In this respect, the consensus-labeling methodology proposed in this study can be understood as a complementary integration of AI-based large-scale text processing and the contextual review of humanities researchers. It offers a practical procedure for considering both computational consistency and contextual interpretation in the construction of literary emotion data.

(5) IMPLICATIONS/APPLICATIONS

The consensus-labeled datasets constructed for *Sohyeonseongrok* and *Sossi-Samdaerok*, together with the methodology used to construct them, have several methodological implications for digital emotion research on Korean full-length novels.

First, rather than treating the emotion judgment of an individual researcher or a single model as a single correct answer, this study proposes a procedure for constructing consensus labels by comparing multiple human annotations and model predictions and integrating them through majority-rule consensus and researcher verification. This approach is meaningful in that it makes it possible to construct shared labels for computational analysis while preserving differences among annotation perspectives.

Second, consensus-labeled data may serve as an efficient training resource in the research environment of Korean full-length novels, where annotated emotion data available for model training remain limited. Model v.1.1.8, trained on 5,282 *Sohyeonseongrok* consensus-labeled sentences, showed a level of agreement comparable to that of model v.1.1.7, which was trained on 14,872 sentences combining three annotation datasets. This does not mean that consensus-labeled data are inherently superior to simply combining larger amounts of training data. Rather, it suggests that reconciling differences among existing annotation sources through consensus and verification may provide a practical way to construct useful training resources in data-limited research settings. In addition, the 297 cases revised by the researcher in *Sossi-Samdaerok* show that multi-model consensus can function as an efficient first-stage labeling procedure, while researcher interpretation can complement it in some context-dependent cases.

Third, the released datasets retain source-specific annotations or model predictions, agreement information, and a consensus_method indicator, enabling reuse for domain-adaptation benchmarks, character-level emotional-trajectory analysis, and studies of annotation divergence. Because final verification was conducted by a single researcher, the consensus labels should be treated as operational reference labels that may still reflect researcher-specific interpretation. As the data derive from long, multi-generational historical narratives with archaic language and context-dependent emotion expression, cross-genre and cross-cultural reuse should account for these characteristics. Future research should examine the applicability of this methodology to other literary works and, where feasible, also explore consensus-labeling procedures involving multiple researchers in the final verification stage.

In conclusion, the proposed methodology constructs computationally usable consensus labels while preserving interpretive plurality and combining AI-based processing with contextual researcher review.

SUPPLEMENTARY FILES

The consensus-labeled dataset and README document are available in the public Zenodo repository (DOI: <https://doi.org/10.5281/zenodo.23315786>). Additionally, the auxiliary dataset and deep learning model building code referenced in this study are accessible through the OSF repository (Kang, 2025c; Kim & Kang, 2024).

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to the Institute of Koreanology at Kangwon National University for providing the research environment and administrative support necessary for this project. The author also extends sincere thanks to Eun-jeong Yoon and Dae-hoon Kwon for their valuable contributions to the qualitative annotation of emotional expressions in *Sohyeonseongrok*. Additionally, the author thanks colleagues who provided constructive feedback during the preliminary discussions of this dataset.

FUNDING STATEMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5B5A16075206).

Kang
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.626

13

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Woo-kyu Kang: Conceptualization; Methodology; Data curation; Formal analysis; Investigation; Writing – original draft; Writing – review & editing.

AUTHOR AFFILIATIONS

Woo-kyu Kang  orcid.org/0009-0005-2535-6083
Institute of Koreanology, Kangwon National University, Chuncheon, South Korea

REFERENCES

- Bayer, M., Kaufhold, M. A., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2023). Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics*, 14, 135–150. <https://doi.org/10.1007/s13042-022-01553-3>
- İlter, M., Özcan Gönülal, Y., Erşahin, B., Evecen, D., Karabulut, S., Tekir, S., Onuç, E., & Berci, İ. (2026). Cross-individual sentiment analysis for historical text processing: detecting author-based relationships in Ottoman-Turkish memoirs. *Digital Scholarship in the Humanities*, 41(3), 1383–1400. <https://doi.org/10.1093/dlc/fqag063>
- Jeon, D., Lee, J., & Kim, C. (2022). KOTE (Korean Online That-gul Emotions) dataset [Data set]. GitHub. Retrieved from <https://github.com/searle-j/KOTE> (last accessed: July 21, 2026).
- Jeon, D., Lee, J., & Kim, C. (2024). User guide for KOTE: Korean Online That-gul Emotions Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 17254–17270). ELRA and ICCL. <https://doi.org/10.63317/2snmd5hc4bx>
- Kang, W. (2024). Consideration of realistic ways to model sentiment data for Korean full-length novels: Centered on establishing an emotion classification system. *Journal of Korean Literary History*, 84, 397–432. <https://doi.org/10.23296/minmun.2024.84.397>
- Kang, W. (2025a). A study on the meaning of digital emotion analysis from a reception aesthetics perspective: Focusing on Sohyeonseongrok. *Multi-Cultural Contents Studies*, 53, 55–91. <https://doi.org/10.15400/mccs.2025.09.53.02>
- Kang, W. (2025b). A comparative study of emotions in the Sohyeonseongrok and Yuhyogongseonhaengnok series. *Korean Language and Literature in International Context*, 107, 9–42. <https://doi.org/10.31147/IALL.107.01>
- Kang, W. (2025c). A study on the meaning of digital emotion analysis from a reception aesthetics perspective [Data set]. OSF. Retrieved from <https://osf.io/tcau2> (last accessed: July 21, 2026).
- Kang, W. (2026). *Consensus-labeled Emotion Dataset for Korean Full-Length Novels* [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.23315786>
- Kang, W., & Kim, B. (2024). Qualitative and quantitative research on emotions in Korean full-length novels: Focusing on Changseongamuirok and Sohyeonseongrok. *The Research of Old Korean Novel*, 58, 223–258. <https://doi.org/10.23836/kornov.2024.58.223>
- Kim, B., & Kang, W. (2024). KENT (Kang woo kyu Emotion for Novel Text) [Data set]. OSF. <https://doi.org/10.17605/OSF.IO/XSY8U>
- Kim, E., Padó, S., & Klinger, R. (2017). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 17–26). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-2203>
- Lee, J. (2021). KcELECTRA: Korean comments ELECTRA [Computer software]. GitHub. Retrieved from <https://github.com/Beomi/KcELECTRA> (last accessed: July 21, 2026).
- Öhman, E. (2020). Emotion annotation: Rethinking emotion categorization. *CEUR Workshop Proceedings*, 2865, 134–144. Retrieved from <https://ceur-ws.org/Vol-2865/short5.pdf>

- Shoer, I., & Erzin, E. (2025). *Learning annotation consensus for continuous emotion recognition* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2505.21196>
- Vu, D. T., Yu, G., Lee, C., & Kim, J. (2022). Text data augmentation for the Korean language. *Applied Sciences*, 12(7), 3425. <https://doi.org/10.3390/app12073425>

Kang
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.626

14

TO CITE THIS ARTICLE:

Kang, W. (2026). Construction of Consensus-Labeled Data for the Study of Emotions in Korean Full-Length Novels. *Journal of Open Humanities Data*, 12: 137, pp. 1–14. DOI: <https://doi.org/10.5334/johd.626>

Submitted: 27 July 2026

Accepted: 10 September 2026

Published: 16 October 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.