Original scientific article / Izvorni znanstveni članak
Received / Zaprimljen: 21. 10. 2024.
https://doi.org/10.21860/j.16.1.13

Kiseong Lee*

Improving fine-grained emotion classification using LLMs through sequential learning of emotions

SUMMARY

This study proposes an approach to improving emotion classification performance by introducing a Sequential Emotion Learning (SEL) method. Conventional learning methods often struggle with fine-grained emotion categories. To address this, the SEL approach first trains the model on seven basic emotions, which are relatively easier to classify due to their clear distinctions. The model is then fine-tuned using 24 more nuanced emotion labels, enhancing its ability to tackle complex emotion classification tasks. Experimental results suggest that the SEL method performs better than the baseline, achieving higher accuracy from the early stages of training. The SEL model also reaches its peak performance relatively quickly and shows improved classification capabilities on unseen, general sentences, indicating its robustness across different text scenarios. These results suggest that the SEL method can effectively improve emotion classification, particularly in tasks that require distinguishing complex emotions. This sequential learning approach offers a potential advantage over traditional methods and may be applied to other domains that involve intricate classification tasks. Future research can explore the generalizability of this method to other classification problems to further enhance its utility.

Keywords: emotion classification, sequential learning, fine-grained classification, large language model, affective computing.

INTRODUCTION

Recent advancements in artificial intelligence (AI) have significantly enhanced linguistic interactions with humans, particularly due to the emergence of large

* AI Humanities Research Institute, Chung-Ang University, Seoul, Republic of Korea. ORCID: https://orcid.org/0000-0002-0906-9552.

Correspondence Address: Kiseong Lee, AI Humanities Research Institute, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea. Email: goory@cau.ac.kr.

language models (LLMs). LLMs are capable of learning from vast amounts of textual data, enabling them to understand and process complex patterns and language contexts (Mann, 2020). These models have demonstrated remarkable success in natural language processing (NLP) and have produced meaningful results across various domains. Among these, the potential of LLMs in the field of affective computing, especially in emotion analysis, is noteworthy (Kim, 2021). As AI continues to train on increasing volumes of textual data, it is expected to develop the ability to comprehend human emotions and interact based on this understanding.

However, for AI to accurately grasp emotions and appropriately reflect them, it must go beyond merely learning the semantic context of language and develop an understanding of emotional context as well. While emotions are sometimes directly revealed through specific vocabulary, they can also be conveyed indirectly, depending on the context. Metaphorical expressions and indirect methods of conveying emotions, in particular, can undermine the accuracy of emotion analysis (Mohammad, 2013). Therefore, AI models need to comprehend vocabulary related to emotions and learn how these emotions manifest through linguistic context in a complex manner. Additionally, addressing the complexity of emotional range and granularity is crucial for the empirical application of emotion recognition. When emotions are divided into several categories, models may experience performance degradation due to this complexity (Demszky, 2020).

In this study, we propose a sequential learning approach to classify fine-grained emotions. First, a large language model, with its strong ability to understand linguistic context, is trained on seven basic emotions that are relatively easy to distinguish. Then, the model is further trained on 24 fine-grained emotions based on an emotion dictionary. This approach gradually improves the model's ability to understand complex emotions, thereby enhancing its performance in emotion classification.

This research aims to contribute to the advancement of affective computing by presenting a methodology that improves the classification performance of fine-grained emotions. As a result, the potential for AI applications in various emotion-driven services is expected to expand further.

RELATED WORKS

Emotion analysis is a technique for extracting and classifying emotions or opinions from data and is widely used in various fields such as marketing, social media monitoring, and public opinion analysis (Kim, 2019). Emotion analysis and classification have long been critical research topics in natural language processing (NLP). This chapter reviews key studies related to text-based emotion classification.

Emotion Classification and Natural Language Processing

Emotion analysis involves classifying emotions from text. Early research primarily relied on rule-based systems to infer emotions (Pang, 2008). These systems classified text emotions based on a list of emotion-related words, where the presence of a specific emotion word would lead to classification in that emotion category. However, this approach faced challenges in determining which words to prioritize when multiple emotion words appeared together.

As a result, emotion classification expanded to incorporate machine learning techniques. Initially, binary or ternary classification focusing on positive, negative, and neutral emotions was prevalent, with machine learning algorithms such as Naive Bayes and SVM being commonly used (Khairnar, 2013). Since machine learning requires converting text into numerical vectors, common embedding techniques like Bag of Words (Pak, 2010), TF-IDF (Martínez-Cámara, 2014), and Word2Vec (Severyn, 2015) were employed. However, while these statistical embeddings were effective for topic identification, they struggled to capture the full context of sentences, and since emotions often go beyond the literal meaning of words, they faced limitations in emotion analysis.

Recently, advancements in deep learning-based NLP have allowed models like Transformers to understand the semantic context of text (Vaswani, 2017). Large language models (LLMs), pre-trained on vast amounts of textual data, are now capable of grasping subtle emotional nuances and context (Liu, 2019a). LLMs perform exceptionally well in precise emotion classification tasks, learning not only the emotions in text but also the subjects, objects, and causes of these emotions (Yang, 2022).

Transfer learning has proven to be an effective method when applying LLMs to specific problems. Transfer learning accelerates the training process and improves performance by applying models trained in one domain to related tasks (Pan, 2009). This is particularly useful when there is a lack of large datasets or when starting to learn from new datasets. In text classification, transfer learning is often used by fine-tuning pre-trained models, such as LLMs, for specific tasks (Howard, 2018). In this process, a model pre-trained on large general text datasets is applied to new domain data, enabling more accurate classification. This method can deliver better performance while reducing training costs compared to models trained from scratch on specific domain data (Liu, 2019a). Our study explores this potential of LLMs in the context of emotion classification, using a fine-tuning approach on the XLM-RoBERTa model (Conneau, 2019) with a small emotion dataset.

Fine-Grained Emotion Types and Data

Emotion datasets for AI training consist of digital data such as images, text, and audio, labeled to reflect the emotions they contain. For AI to learn from emotion data, a sufficient amount of data is required, and labels must distinguish between at least two types of emotions. The construction of large-scale emotion datasets for AI learning became active in the late 2000s with the popularization of big data. Go and colleagues released a dataset of 1.6 million Twitter comments labeled as positive, neutral, or negative (Go et al., 2009), and Maas and colleagues distinguished between positive and negative reviews on IMDb, a movie review website (Maas et al., 2011). Additionally, SemEva,[1] an international competition for NLP research, has continuously provided datasets, including sentiment analysis datasets for Twitter. However, earlier datasets predominantly collected social media conversation data and focused mainly on polarity (positive versus negative).

In contrast, emotion labels have been expanded to include more specific emotion types. This is referred to as categorical emotion models, which use independent emotion labels representing universal emotions (Kusal, 2021). Prominent models include Plutchik's and Ekman's models. Plutchik proposed eight emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) (Plutchik, 1980), while Ekman defined six emotions (happiness, fear, surprise, sadness, disgust, anger) (Ekman, 1979). Based on these emotions, datasets have been generated from various sources beyond social media, including survey responses (Scherer, 1994), fairy tales (Alm, 2005), blogs (Aman, 2007), news headlines (Strapparava, 2007), election-related tweets (Mohammad, 2015), Facebook posts (Preoţiuc-Pietro, 2016), daily conversations (Li, 2017), Reddit comments (Demszky, 2020), and expressions and responses from authors and readers (Buechel, 2022).

Beyond these universal emotions, research began to expand into more complex and diverse expressions of human emotion. The OCC model proposed twenty-two emotion labels related to events, actions, and objects (Ortony et al., 2022). Yi and colleagues suggested 24 emotions specific to Korean language users, selecting words from an emotion dictionary and publicly labeling them (Yi et al., 2020). However, as emotion labels increase, building datasets becomes more challenging, and data scarcity remains a limitation.

Emotion classification generally does not achieve as high performance as general text classification, and despite various efforts to improve it, dramatic performance improvements are rarely observed. For instance, Google's GoEmotions dataset contains 58,009 Reddit comments labeled with 27 fine-grained emotions and

[1] https://semeval.github.io/

neutral emotion. When classified using the BERT model, the F1-score was only 0.46, indicating the difficulty of identifying fine-grained emotions accurately (Demszky, 2020). Singh and colleagues increased the classification performance on GoEmotions from an F1-score of 0.46 to 0.51 by teaching the model the dictionary definitions of emotions (Singh et al., 2021).

Fine-grained emotion classification aligns with the challenges inherent in multi-class text classification. In multi-class text classification, accuracy consistently decreases as the number of classes increases (Liu, 2019b). This trend varies depending on the type of text. For example, legal document classification has reported an F1-score of 0.72 for six classes (Luz de Araujo, 2022), while technical document classification achieved an F1-score of 0.53 for 33 classes (Hwang, 2020). Emotion classification, in particular, goes beyond understanding text to capturing subtle intrinsic tendencies that can vary significantly, making it inherently challenging.

In such a context, recent studies have continued to explore the performance of emotion classification models across diverse datasets and approaches. For instance, a study introduced a hybrid approach combining human-engineered features and deep learning-based representations, achieving a Jaccard accuracy of 68.40% on the SemEval-2018 dataset and 53.45% on the GoEmotions dataset (Ahanin, 2023). Similarly, the Transformer Transfer Learning (TTL) model demonstrated strong performance in emotion detection, achieving an average F1-score of 0.84 across test datasets, with 0.87 on annotator-rated emotions and 0.79 on self-reported emotions, showcasing its effectiveness in emotion classification (Lee, 2023).–Additionally, a study leveraging emotion-specific features to enhance transformer-based models achieved an accuracy of 61.9% and a macro F1-score of 0.52 on the WASSA 2022 emotion classification shared task (Desai, 2022).

These studies collectively demonstrate the ongoing efforts to address the inherent challenges of emotion classification, particularly in fine-grained and multi-label scenarios, and further highlight the importance of combining model architecture innovation with dataset-specific adaptations to achieve better outcomes.

This study challenges the task of fine-grained emotion classification using Korean text. We aimed to improve performance by applying transfer learning to LLMs, first training on seven basic emotions and then performing fine-grained classification of 24 specific emotions. This approach simultaneously considers both generalization and granularity in emotion classification, differentiating it from previous studies.

RESEARCH METHOD

In this study, we propose a learning method to improve classification performance for fine-grained emotion labels, categorized into 24 distinct emotions.

Emotion Dataset

To recognize fine-grained emotions, it is necessary to have a dataset for training. Most emotion datasets are based on Ekman's or Plutchik's emotion labels, consisting of 6 or 8 categories (Ekman, 1979; Plutchik, 1980). However, large amounts of text labeled with detailed emotions are rare. For this study, we utilized the *Multi-Label Korean Emotion Word Dictionary with 24 Emotions*[2] to ensure high-quality text suitable for training. This dataset consists of 19,617 Korean words, each labeled with one of 24 emotion categories. Since the words alone do not provide context, they were expanded into sentence form. Using ChatGPT, sentences were generated that express emotions using the specific words from the dictionary, resulting in approximately 480,000 sentences across the 24 emotions.

The emotion labels are: *disgust, anger, passion, distress, anxiety, sadness, happiness, peace, surprise, jeong (affection), fun, love, achievement, moved, depression, fear, shame, tedium, regret, excitement, compassion, jealousy, guilt,* and *neutral*. For each emotion, 20,000 sentences were generated to ensure a balanced dataset. For model training, 80% of the dataset was used as the training set, and the remaining 20% was used as the validation set. Additionally, two external test sets were created to more precisely evaluate the model's performance.

- Test Set-1: This test set consisted of sentences presented as "examples" in the dictionary. It contained 9,514 sentences collected from general dictionaries, which were direct uses of emotion words from the *Multi-Label Korean Emotion Word Dictionary with 24 Emotions*. As these were example sentences from the dictionary, the focus was on how the emotional words are used in specific situations. Since the words are more 'the kind you would find in a dictionary,' there are words that are less frequently used in everyday conversations, and many sentences are non-conversational.
- Test Set-2: ChatGPT was tasked with generating sentences for each of the 24 emotions, producing twenty sentences per emotion for a total of 480 sentences. This data is unrelated to the *Multi-Label Korean Emotion Word Dictionary with 24 Emotions* and represents more general expressions. The purpose was to determine how well the model, trained on sentences derived from the emotion dictionary,

[2] http://aihumanities.org/en/archive/datasets/?vid=4

can classify different types of emotional expressions. The prompt used to generate Test Set-2 is outlined below. Although the original experiment was conducted in Korean, it has been translated into English for this paper.

Prompt:

I have created an emotion classification model and would like you to generate test data. The labels are as follows:

moved, fear, happiness, peace, jealousy, neutral, guilt, affection, fun, depression, passion, compassion, distress, sadness, shame, achievement, regret, excitement, love, anxiety, anger, surprise, tedium, and disgust.

Please create sentences that reflect each of these emotions. Generate 20 sentences for each emotion, ensuring that the name of the label does not directly appear in the sentences.

The examples generated by ChatGPT are provided below.

Distress

1. *My thoughts were tangled, and I couldn't organize anything in my mind.*
2. *I had no idea what decision to make.*
3. *A sigh escaped me without even realizing it.*
4. *I kept feeling like something was going wrong.*
5. *It felt like everything was falling apart.*

…Remaining sentences omitted…

We aimed to evaluate the model's performance from various perspectives using this dataset.

Model Training

In this study, we fine-tuned one of the LLMs, XLM-RoBERTa-base[3], to create an emotion classification model. XLM-RoBERTa-base is a transformer-based language model that belongs to the state-of-the-art BERT family and performs exceptionally well in multilingual environments, having been trained on over 100 languages, and can be applied to various natural language processing tasks. The model is pre-trained on a vast amount of text collected from the web, without specific labels or annotations, and excels in cross-lingual transfer learning and multilingual support. Since it can autonomously understand sentence structure, relationships between words, and context, it was chosen for its expected strong performance in emotion classification tasks. The model training was divided into two approaches: the basic approach (baseline) and the proposed method (SEL).

[3] https://huggingface.co/FacebookAI/xlm-roberta-base

- Baseline Model Training: First, the XLM-RoBERTa-base model was trained on 480,000 sentences with 24 emotion labels, split 80:20 for training and validation. This model served as the baseline to measure the basic performance of the study. We used the Hugging Face transformers module, and the optimizer was left as the default AdamW.
- Sequential Emotion Learning (SEL): This is the method proposed in this study, in which emotions are learned sequentially in two stages. First, the XLM-RoBERTa-base model is trained on data with seven emotion labels. Then, the output layer of the model is modified, and it is further trained on data with 24 emotion labels. The goal is to improve the model's ability to distinguish more complex emotions by first enhancing its understanding of simpler emotions. The seven-label emotion data was obtained from two datasets provided by the AI-hub website. These datasets consist of sentences expressing emotions from everyday life[4] and sentences extracted from literary works[5]. A total of 319,600 sentences were collected, and the labels consist of seven classes: *sadness, anger, anxiety, embarrassment, hurt, joy*, and *no emotion*. In the first phase of training, this dataset was split 80% for training and 20% for validation. In the second phase of training, the same dataset and methods as the baseline were used to complete the 24-label classification model.

Performance Evaluation

The performance of the model was measured using accuracy, precision, recall, and the F1-score, with the F1-score reported as the weighted F1. Additionally, top-3 accuracy was included in the evaluation for the test sets. Top-3 accuracy measures whether the correct label is among the top three predicted labels by the model. The baseline model and the proposed sequential learning model were compared in terms of their learning process, validation set, and test sets. The validation set consisted of 20% of the training data, representing similar sentences, while Test Set-1 and Test Set-2 were completely external datasets. This allowed us to analyze the performance differences between the models.

RESULTS AND ANALYSIS

The learning results of the baseline and SEL models were examined and compared.

[4] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=86

[5] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=485

Baseline Learning Results

Table 1 shows the learning results of the baseline model. The accuracy of the baseline model for classifying 24 emotions reached 0.958 at epoch 5, demonstrating very high performance. The F1-score was similarly high at 0.958. When fine-tuning large models like XLM-RoBERTa, it is common practice to set the number of epochs between 3 and 5 to prevent overfitting, as shown in studies such as Devlin et al., where 3 or 4 epochs were used during fine-tuning to achieve optimal performance depending on the dataset (Devlin, 2019). Similarly, Arslan and colleagues demonstrated that performance improvements in multi-class text classification tasks typically plateau after 5 epochs, suggesting that additional training offers diminishing returns and increases the risk of overfitting (Arslan et al., 2021). In this experiment, performance plateaued at epoch 5, leading us to stop training to maintain model generalization. This approach aligns with standard practices in optimizing performance while avoiding overfitting, especially for datasets where the training and validation sets share similar linguistic patterns. These results are based on the evaluation of the validation dataset, and since the training data and validation data contain very similar forms of emotional expression, the classification performance is exceptionally good. Despite the 24 classes being highly granular, the emotion classification performance, based on the contextual understanding capabilities of the LLM, is notably high.

Table 1. Baseline training: Results of 24-emotion classification

epoch	accuracy	precision	recall	F1-score
1	0.895760	0.896703	0.895760	0.895509
2	0.934385	0.934644	0.934385	0.934171
3	0.949239	0.949575	0.949239	0.949239
4	0.956822	0.956840	0.956822	0.956788
5	**0.958447**	**0.958466**	**0.958447**	**0.958424**

Sequential Emotion Learning (SEL) Results

The sequential learning proposed in this paper is divided into two steps. First, the results of training the seven-emotion classification are shown in Table 2. At epoch 7, the model reached an accuracy of 0.923 and an F1-score of 0.923, demonstrating excellent classification performance. It was determined that the model had sufficiently acquired basic emotion recognition capabilities, and this model was then used for transfer learning to perform the second phase of training, which involved classifying 24 emotions.

Table 2. SEL-step1 : Results of seven-emotion classification

epoch	accuracy	precision	recall	F1-score
1	0.671292	0.669688	0.671292	0.668243
2	0.877080	0.877042	0.877080	0.876621
3	0.908667	0.909143	0.908667	0.908729
4	0.902550	0.904122	0.902550	0.902767
5	0.917584	0.917975	0.917584	0.917573
6	0.920447	0.920905	0.920447	0.920332
7	**0.923545**	**0.923698**	**0.923545**	**0.923522**
8	0.923388	0.923584	0.923388	0.923407

The results of the second phase of training are shown in Table 3. At epoch 5, both the accuracy and F1-score reached 0.959, demonstrating excellent performance in the 24-emotion classification. As a result, it can be seen that the performance is similar to that of the baseline model.

Table 3 SEL-step2 : Results of 24-emotion classification

epoch	accuracy	precision	recall	F1-score
1	0.928604	0.928524	0.928604	0.928384
2	0.949375	0.949386	0.949375	0.949305
3	0.955489	0.955505	0.955489	0.955456
4	0.958145	0.958205	0.958145	0.958134
5	**0.959812**	**0.959836**	**0.959812**	**0.959795**

Comparative Analysis Using Test Sets

The comparison of learning results using the validation set showed that both the baseline and the proposed SEL method achieved a similar accuracy level, close to 0.96. However, for this model to be practically useful, it must also demonstrate good performance on data beyond the training set. In this study, we evaluated the performance of the emotion classification model using two test sets to measure the model's classification performance on example sentences from the emotion dictionary and more general sentences. First, Table 4 shows the performance of both models on Test Set-1.

Table 4. Performance comparison on Test Set-1

	accuracy	**precision**	**recall**	**F1-score**	**top3 accuracy**
Baseline	0.686882	0.715876	0.686882	0.685668	0.8461
SEL	0.681627	0.711795	0.681627	0.680141	0.8421

In Test Set-1, the baseline model recorded an accuracy of 0.686 and an F1-score of 0.685, while the SEL model showed an accuracy of 0.681 and an F1-score of 0.680. The performance difference between the two models is minimal, and the difference in the third decimal place is not considered to be significant. No substantial differences were observed between the two models here as well. Although the performance is significantly lower compared to the 0.96 accuracy achieved during training, given that the task was to choose one out of 24 finely categorized emotions, a 68% success rate is considered high. This level of performance is likely due to the fact that Test Set-1 included the emotion words used during training. However, the sentences in Test Set-1 are diverse and not as well-structured as the training data, which may have contributed to the drop in classification performance. Nevertheless, the top-3 accuracy reached 0.84 for both models, indicating a high probability of correctly identifying the correct emotion among the top three choices out of 24.

On the other hand, Test Set-2 consists of general sentences that are unrelated to the emotion dictionary. Table 5 shows the comparative results for Test Set-2.

Table 5. Performance comparison on Test Set-2

	accuracy	**precision**	**recall**	**F1-score**	**top3 accuracy**
baseline	0.531250	0.591920	0.531250	0.531193	0.7750
SEL	**0.560417**	**0.598714**	**0.560417**	**0.552828**	**0.8021**

In both models, the accuracy decreased. However, a significant difference emerged between the two models, with the SEL model proposed in this paper outperforming the baseline model. The SEL model achieved an accuracy of 0.56, surpassing the baseline model's accuracy of 0.53. The F1-score for SEL was also higher at 0.55, compared to 0.53 for the baseline model. Additionally, the top-3 accuracy for the SEL model was 0.80, further demonstrating better performance.

These results suggest that the SEL method performs better on general sentences that are unrelated to the emotion dictionary. This indicates that the SEL approach, by learning emotions in stages, is better at classifying more generalized emotional expressions. Unlike the baseline model, which is specialized for the emotion dictionary, the SEL model enhances its understanding of emotions, resulting in less

performance degradation. In other words, the SEL model shows potential as a model that can provide more flexible emotion recognition in everyday situations.

In addition to the comparison with the baseline model, we conducted experiments using additional classification models, including Logistic Regression, SVM, Naïve Bayes, and Random Forest. Table 6 presents the accuracy of these models on the validation set, Test Set-1, and Test Set-2. The results show that SEL significantly outperforms traditional classification models, particularly in Test Set-1 and Test Set-2, which consist of more diverse and challenging data. While Naïve Bayes achieved slightly higher accuracy than SEL on Test Set-2, SEL consistently demonstrated superior performance across all metrics and datasets. These findings emphasize the effectiveness of the SEL approach in fine-grained emotion classification compared to existing methods.

Table 6. Accuracy comparison of SEL and traditional classification models on validation and test sets

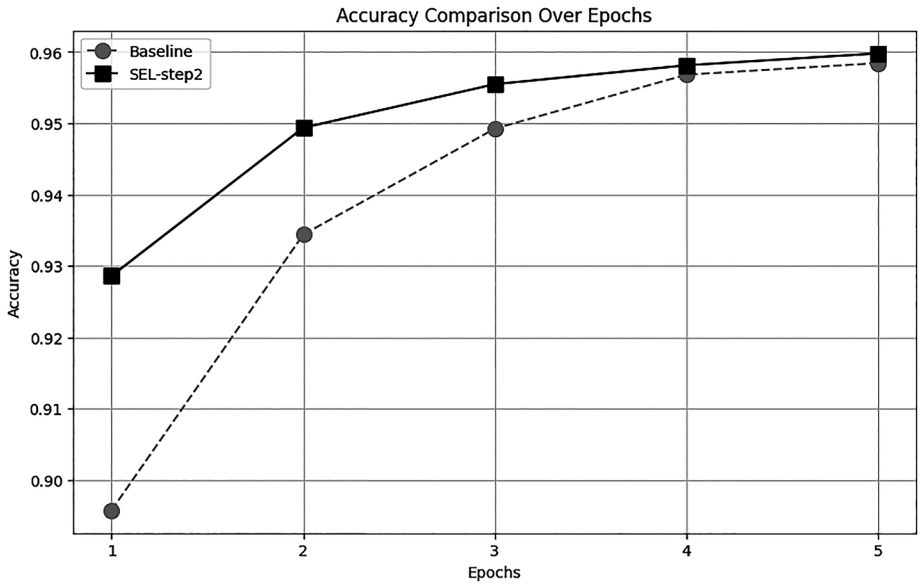
models	validation set	Test Set-1	Test Set-2
SEL	**0.9598**	**0.6816**	0.5604
Logistic Regression	0.7155	0.2748	0.5667
SVM	0.7644	0.2927	0.5542
Naïve Bayes	0.7051	0.317	**0.5979**
Random Forest	0.5266	0.1494	0.4688

DISCUSSION

In this study, we identified a significant performance difference by applying the Sequential Emotion Learning (SEL) method to the emotion classification task and comparing it with the baseline model. In this section, we analyze the advantages and implications of the SEL method based on the experimental results and discuss potential directions for future research.

First, Figure 1 compares the accuracy improvement trends by epoch during the training process of both the baseline and SEL in step 2.

Figure 1. Comparison of accuracy by epoch between the baseline and SEL-step2



As seen in the first epoch of the second phase of training in the SEL method, the accuracy starts at a considerably high level of 0.928. Compared to the baseline, which starts at 0.895, SEL demonstrates a better classification ability from the beginning of the training and reaches the peak more quickly than the baseline. Given that both the baseline and SEL-step2 models were trained on the same data and with the same settings, this difference is quite significant. It is likely that SEL already possessed a strong ability to differentiate between emotions because it had already learned the seven emotion labels in step 1. Therefore, SEL achieves good performance to the extent that it could be sufficiently utilized even if training is stopped at epoch 2 or 3.

The importance of sequential learning becomes particularly apparent in complex tasks such as multi-label problems, like emotion classification. By first learning the seven emotions, the model was able to distinguish more clearly between emotional differences, and as it expanded to 24 emotions, performance improved rapidly. This suggests that sequential learning methods may be advantageous in finely detailed tasks such as emotion classification.

To further evaluate the practical application and generalization capability of the proposed method, we tested the model on a set of arbitrary sentences reflecting various emotions. These sentences were carefully constructed to avoid directly using the emotion labels, ensuring thereby a more realistic evaluation of the model's ability to infer emotions from contextual cues. The results highlight the model's ability to predict emotions, as shown in Table 7.

Table 7. Emotion prediction results for arbitrary sentences

Sentence	**Predicted Emotions (Top 3)**
I couldn't stop crying all day after losing my beloved pet yesterday.	*sadness* (98.54%) *depression* (0.58%) *compassion* (0.19%)
The whole family cheered together after hearing the news of my acceptance today.	*happiness* (99.40%) *fun* (0.31%) *achievement* (0.12%)
I thought I had everything ready, but I kept feeling like something was missing, and my heart was racing.	*anxiety* (98.58%) *distress* (1.24%) *neutral* (0.04%)
After making a poor decision that disappointed a friend, I felt a sharp pain every time I saw their pained expression.	*sadness* (37.81%) *guilt* (20.50%) *compassion* (14.92%)
It feels like all the meaningful moments have passed, leaving my heart empty.	*depression* (98.74%) *tedium* (1.10%) *distress* (0.06%)
When someone distorted my intentions and criticized me, I felt an uncontrollable emotion surging within me.	*anger* (91.24%) *shame* (6.28%) *disgust* (0.93%)
I couldn't believe it when I bumped into a friend I hadn't seen in years.	*surprise* (97.90%) *neutral* (0.49%) *fear* (0.49%)
The smell in the messy room was unbearable.	*disgust* (52.16%) *tedium* (16.33%) *anxiety* (15.82%)
I was so excited about starting a new journey that I stayed up all night with anticipation.	*excitement* (80.18%) *surprise* (11.64%) *anxiety* (2.30%)
Everyone burst into laughter and had a wonderful time together.	*fun* (98.81%) *neutral* (0.98%) *happiness* (0.05%)
After persevering until the end, I finally achieved my goal.	*achievement* (98.24%) *passion* (0.84%) *neutral* (0.25%)
Seeing a small puppy shivering on the street made my heart ache.	*compassion* (96.83%) *sadness* (1.81%) *moved* (0.28%)

The evaluation demonstrated that the model could accurately infer fine-grained emotions in diverse contexts, with particularly high performance on primary emotions such as *happiness*, *sadness*, and *anxiety*. However, an analysis of error patterns in SEL model predictions, as summarized in Table 8, provides insights into how the model handles overlapping or contextually similar emotions. For instance, *distressed* was often misclassified as *anxiety* (45.45%), and *disgust* as *anger* (50.41%), likely due to shared contextual or affective features. Similarly, *tedium* was confused with *depression* (61.17%), and *excitement* with *love* (59.57%), reflecting the subtle contextual nuances in these emotional expressions. These findings highlight specific areas where the SEL model, despite its high overall accuracy exceeding 95%, can be further refined to improve its handling of complex emotional overlaps. Future work could focus on augmenting the training dataset with examples that emphasize distinctions between overlapping emotions and exploring advanced architectures, such as multi-label learning, to better capture subtle emotional differences.

Table 8. Frequent error patterns in SEL model predictions

Actual Label	Most Frequent Error Label	Proportion of Misclassified Predictions (%)
distressed	*anxiety*	45.45
disgust	*anger*	50.41
moved	*affection*	40.19
tedium	*depression*	61.17
excitement	*love*	59.57

While this study focused on Korean datasets, the SEL method is not language-dependent and can be extended to other languages and datasets. Testing the method on multilingual and cross-cultural datasets would validate its generalizability and enhance its global applicability.

Overall, these results validate the robustness of the proposed method in practical scenarios and highlight its potential for applications in emotion-sensitive domains such as customer support, mental health, and social communication analysis. For example, in healthcare, the SEL model can be utilized for analyzing patient emotions to enhance psychological assessments or support therapy sessions. In education, the model could help measure student engagement by identifying emotional states during learning activities, enabling personalized learning strategies. In customer service, the SEL model can assist in developing emotion-based counseling systems that improve user satisfaction and personalize interactions. These applications demonstrate the versatility of the SEL model in addressing real-world challenges across various

domains. Future work could explore expanding the training dataset with more varied emotional expressions to further enhance the model's generalization capability.

Although this study focused on emotion classification, the SEL method presents the potential to be applied to other complex classification tasks. In domains beyond emotion, the strategy of first learning basic features with broad categories and then progressively expanding to more detailed classification tasks may prove effective. Of course, further evaluation of the generalizability of this method to other domains and various datasets is needed. Future research should explore the application of the SEL method to other classification tasks to assess its versatility.

Moreover, the quantity and quality of training data are critical factors that influence SEL model performance. While this study employed a balanced dataset carefully curated for fine-grained emotion classification, future research should examine how variations in data size, noise levels, and class imbalances affect learning outcomes. For example, noise reduction techniques, such as outlier detection and data cleaning, could improve the reliability of training data. Enhanced labeling processes, including expert annotation or consensus-based approaches, may also contribute to better model generalization. Additionally, increasing data quantity through data augmentation or by incorporating diverse datasets could help the SEL model generalize across broader scenarios. Such investigations will aid in deriving optimal strategies for data selection and preparation, enhancing the robustness and applicability of models.

Additionally, a more detailed analysis of the impact of the quantity and quality of training data on model performance could help in deriving optimal learning strategies, which could be proposed as a future research direction. Adjusting hyperparameters such as batch size and learning rate during step 1 and step 2 of SEL may further improve performance. In this study, these hyperparameters were carefully selected based on the computational resources available, including an Intel i7 CPU and an NVIDIA RTX 3090 GPU. These hardware constraints necessitated balancing model performance with memory and processing capacity limitations, influencing the choice of hyperparameters to ensure stable training. Future research could systematically explore automated hyperparameter optimization techniques, such as grid search or Bayesian optimization, to further enhance model performance. Adopting resource-efficient learning methods, such as knowledge distillation and lightweight architectures, could also address these constraints and improve scalability. In addition, we are considering leveraging state-of-the-art language models, such as T5 or GPT-based architectures, to further refine the SEL method. These advanced models represent the cutting edge of natural language processing and could significantly enhance the SEL model's effectiveness in complex and nuanced classification tasks.

CONCLUSION

This study introduced a Sequential Emotion Learning (SEL) method for fine-grained emotion classification and demonstrated its performance compared to simpler learning methods. The SEL approach, particularly, showed high accuracy from the early stages of training and quickly converged to its peak. It also proved to be effective in improving the classification of emotions in general, unseen sentences, demonstrating the model's ability to generalize beyond the training data. These findings suggest that sequential learning can be an effective strategy for tackling complex emotion classification tasks, offering a promising approach that could be applied across various domains.

Looking ahead, it will be essential to apply the SEL approach to other classification tasks to fully evaluate its versatility and generalizability across different datasets and domains. Beyond emotion classification, SEL shows potential for broader applications in complex classification tasks, such as multi-label classification, intent detection, and topic categorization. Exploring these possibilities will help validate its utility and expand its practical applications. Future studies should also explore the applicability of SEL to multilingual and culturally diverse datasets, as this would enhance its practical value and establish its robustness across languages. Investigating how classification patterns evolve when the SEL method is trained on multilingual datasets will provide insights into its performance across various languages and cultural contexts. These studies could not only validate the generalizability of SEL but also identify the most effective ways to implement sequential learning in diverse fields.

Further research is also needed to explore the optimal learning strategies, including fine-tuning hyperparameters and analyzing the impact of training data size and quality. Such efforts will contribute to the development of more robust and efficient models for complex classification tasks.

REFERENCES

Ahanin, Z., Ismail, M. A., Singh, N. S. S., & AL-Ashmori, A. (2023). Hybrid feature extraction for multi-label emotion classification in English text messages. *Sustainability*, *15*(16), 12539.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In R. Mooney, C. Brew, L.F. Chien & K. Kirchhoff (Eds.), *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 579-586).

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue* (pp. 196-205). Berlin, Heidelberg: Springer Berlin Heidelberg.

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., & Goujon, A. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. In J. Leskovec, M. Grobelnik, M. Najork, J. Tang & L. Zia (Eds.), *Companion Proceedings of the Web Conference 2021* (pp. 260-268).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, *1*.

Buechel, S., & Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Conneau, A. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Desai, S., Kshirsagar, A., Sidnerlikar, A., Khodake, N., & Marathe, M. (2022). Leveraging emotion-specific features to improve transformer performance for emotion classification. *arXiv preprint arXiv:2205.00283*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186).

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual review of psychology*. *30*(1), 527-554.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hwang, S., & Kim, D. (2020). BERT-based classification model for Korean documents. *Journal of Society for e-Business Studies*, *25*(1).

Khairnar, J., & Kinikar, M. (2013). Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, *3*(6), 1-6.

Kim, T., & Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Kusal, S., Patil, S., Kotecha, K., Aluvalu, R., & Varadarajan, V. (2021). AI based emotion detection for textual big data: Techniques and contribution. *Big Data and Cognitive Computing*, *5*(3), 43.

Lee, S. J., Lim, J., Paas, L., & Ahn, H. S. (2023). Transformer transfer learning emotion detection model: synchronizing socially agreed and self-reported emotions in big data. *Neural Computing and Applications*, *35*(15), 10945-10956.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Liu, X., & Wangperawong, A. (2019b). Transfer learning robustness in multi-class categorization by fine-tuning pre-trained contextualized language models. *arXiv preprint arXiv:1909.03564*.

Liu, Y. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luz de Araujo, P. H., de Almeida, A. P. G., Ataides Braz, F., Correia da Silva, N., de Barros Vidal, F., & de Campos, T. E. (2023). Sequence-aware multimodal page classification of Brazilian legal documents. *International Journal on Document Analysis and Recognition (IJDAR)*, *26*(1), 33-49.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In D. Lin, Y. Matsumoto & R. Mihalcea (Eds.), *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural language engineering*, *20*(1), 1-28.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, *51*(4), 480-499.

Oberländer, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In E. M. Bender, L. Derczynski & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 2104-2119).

Ortony, A., Clore, G. L., & Collins, A. (2022). *The cognitive structure of emotions*. Cambridge university press.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)* (pp. 1320-1326).

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (eds.), *Theories of emotion* (pp. 3-33). Academic press.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in facebook posts. In A. Balahur, E. van der Goot, P. Vossen & A. Montoyo (Eds.), *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 9-15).

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, *66*(2), 310.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In R. Baeza-Yates, M. Lalmas, A. Moffat & B. A. Ribeiro-Neto (Eds.), *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 959-962).

Singh, G., Brahma, D., Rai, P., & Modi, A. (2021). Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-8). IEEE.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In E. Agirre, L. Màrquez & R. Wicentowski (Eds.), *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 70-74).

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Yang, Z. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Yi, Y., Park, J., & Kim, B. (2020). The construction of digital ontology for Korean emotions. *Korean Semantics*, *68*, 131-162.

Unaprjeđenje precizne klasifikacije emocija korištenjem velikih jezičnih modela putem sekvencijalnog učenja emocija

SAŽETAK

Ovim istraživanjem predlaže se pristup za poboljšanje učinkovitosti klasifikacije emocija uvođenjem metode sekvencijalnog učenja emocija (*Sequential Emotion Learning*, SEL). Konvencionalne metode učenja često imaju poteškoća s preciznim klasificiranjem emocija. Kako bi se to prevladalo, pristupom SEL prvo se trenira model na sedam osnovnih emocija, koje je relativno lakše klasificirati zbog njihove jasne međusobne razlike. Potom se model dodatno prilagođava pomoću 24 specifičnije emocionalne oznake, čime se poboljšava njegova sposobnost rješavanja složenih zadataka klasifikacije emocija. Eksperimentalni rezultati pokazuju da metoda SEL nadmašuje osnovni model, postižući veću točnost već u ranim fazama treniranja. Model SEL vrlo brzo postiže svoju maksimalnu učinkovitost te pokazuje poboljšane sposobnosti klasifikacije na neviđenim, općenitim rečenicama, što ukazuje na njegovu robusnost u različitim tekstualnim kontekstima. Dobiveni rezultati sugeriraju da metoda SEL može učinkovito poboljšati klasifikaciju emocija, osobito u zadacima koji zahtijevaju razlikovanje složenih emocija. Ovaj sekvencijalni pristup učenju nudi potencijalnu prednost u odnosu na tradicionalne metode i može se primijeniti na druga područja koja uključuju složene zadatke klasifikacije. Budućim istraživanjima mogla bi se istražiti opća primjenjivost ove metode na druge probleme klasifikacije kako bi se dodatno povećala njezina korisnost.

Ključne riječi: klasifikacija emocija, sekvencijalno učenje, precizna klasifikacija, veliki jezični model, afektivno računarstvo.