Enhancing Self-Attention via Knowledge Fusion: Deriving Sentiment Lexical Attention from Semantic-Polarity Scores

Dongjun Jang
Department of Linguistics
Seoul National University
qwer4107@snu.ac.kr

Jinwoong Kim
Graduate School of Data Science
Seoul National University
kjw900106@snu.ac.kr

Hyopil Shin
Department of Linguistics
Seoul National University
hpshin@snu.ac.kr

Abstract

In recent years, pre-trained language models have demonstrated exceptional performance across various natural language processing (NLP) tasks. One fundamental component of these models is the self-attention mechanism, which has played a vital role in capturing meaningful relationships between tokens. However, a question still remains as to whether injecting lexical features into the self-attention mechanism can further enhance the understanding and performance of language models. This paper presents a novel approach for injecting semantic-polarity knowledge, referred to as Sentiment Lexical Attention, directly into the self-attention mechanism of Transformer-based models. The primary goal is to improve performance on sentiment classification task. Our approach involves consistently injecting Sentiment Lexical Attention derived from the lexicon corpus into the attention scores throughout the training process. We have conducted empirical analysis on our approach using the NSMC, a benchmark for Korean sentiment classification, where it demonstrated substantial performance enhancements and secured state-of-the-art achievements. Furthermore, our approach demonstrates robustness and effectiveness in out-of-domain tasks, indicating its potential for broad applicability. Additionally, we analyze the impact of Sentiment Lexical Attention on the view of the CLS token's attention distribution. Our method offers a fresh perspective on synergizing lexical features and attention scores, thereby encouraging further investigations in the realm of knowledge injection utilizing the lexical features.

1 Introduction

In recent years, pre-trained language models such as BERT (Devlin et al., 2018), XLNet (Yang et al., 2019b), BART (Lewis et al., 2019), and GPT-3 (Brown et al., 2020) have demonstrated remarkable performance across various downstream tasks in NLP. These language models (LMs) are characterized by a vast number of trainable parameters, which many researchers believe encode valuable knowledge during the processing of contextualized token embeddings (Wang et al., 2020; Incitti et al., 2023). Among these parameters, self-attention mechanisms play a crucial role and are widely considered as the foundation of nearly all language models. Many studies have led to performance improvements by attempting to inject knowledge into self-attention, based on the understanding that it learns based on relationships between tokens (Hu et al., 2023; Kaddari and Bouchentouf, 2023; Xie et al., 2023; Zhao et al., 2024). However, an important question remains: Can we leverage the sentiment lexical features to enhance the self-attention mechanism and gain a deeper understanding of the relationships between semantically meaningful tokens?

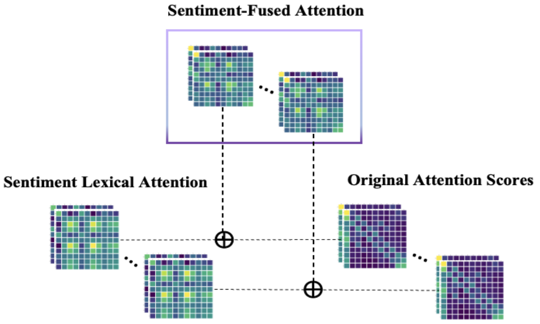

Sentiment-Fused Attention
Sentiment Lexical Attention
Original Attention Scores

Figure 1: The Sentiment-Fused Attention, induced by forming a linear combination of the Sentiment Lexical Attention and the Original Attention Scores.

Many studies have investigated the methods of knowledge fusion on LMs to enhance performance in natural language understanding tasks (Sun et al., 2019; Liu et al., 2020; Wang et al., 2023). Knowledge injection techniques can be applied to any part of the LMs (Colon-Hernandez et al., 2021; Wei et al., 2021). Among the many methods, we introduce a method to convert lexical sentiment features

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 334 - 344
June 20-21, 2024 ©2024 Association for Computational Linguistics

into a computable matrix (Sentiment Lexical Attention), which is then induced into a linear combination with the Self-Attention. We denote this fused attention mechanism as Sentiment-Fused Attention (Figure 1). For injecting Sentiment-Fused Attention well on attention score, we partially follow Xia et al. (2021)'s way, which proposes a method of guiding the attention output by injecting similarity knowledge into the attention score.

In Section 3, we suggest our novel approach of injection, pointing out Xia et al. (2021)'s injection methods might have the potential to distort the relationships of tokens. Furthermore, the process of extracting Sentiment Lexical Features from the Polarity Score of the Lexicon Corpus and deriving them into a fusible matrix is elaborately described in Section 3. We believe our method of injection is meticulously formulated to augment the weights between tokens with similar sentiment features. In Section 4, we substantiate the effectiveness of our injection method through experiments, simultaneously empirically demonstrating that it yields stable performance improvements, unlike Xia et al. (2021)'s approach. To the best of our knowledge, our results represent the state-of-the-art on the Naver Sentiment Movie Corpus (NSMC)[1], which is widely regarded as the leading benchmark for sentiment analysis in the Korean language. Additionally, experimental observations reveal the effectiveness of sentiment lexical features in out-of-domain tasks. In Section 6, we investigate the impact of Sentiment-Fused Attention on attention by statistically examining the attention dynamics of the CLS token (serving as the classifier) and demonstrate through analysis that it exerts significant influence.

The main contribution of our work are as follows:

- We propose the way of inducing Sentiment Lexical Attention from the semantic-polarity score, which means that any corpus containing the polarity information could follow our work for the enhancement on downstream tasks.
- We establish a mathematical formula that combines two different attention matrix. The theoretical underpinnings and empirical evidence supporting this approach are demonstrated through experiments.
- We achieve a state-of-the-art performance on NSMC benchmark.

2 Related Work

Previous research has extensively investigated the injection of knowledge into self-attention based language models to augment its language representation prowess (Wang et al., 2023). In this chapter, we introduce prior research on knowledge graph-based approaches, which are most commonly utilized for Knowledge Injection, and discuss how knowledge integration has been approached from the perspective of Lexical Semantics. Finally, we justify the validity of our research by introducing prior studies related to self-attention distributions.

Infusing Knowledge Graph into the Self-Attention Mechanism Zhang et al. (2019) pioneered the development of the ERNIE model, an innovative approach that employs knowledge integration to enhance language representation. Liu et al. (2020) propose K-BERT with knowledge graphs, in which triples are injected into the sentences as domain knowledge. Peters et al. (2019) present KnowBERT, a model that integrates knowledge bases (KBs) into the pre-trained BERT model. Xu et al. (2020) utilize external entity descriptions to provide contextual information for knowledge understanding task. Yu et al. (2022) propose JAKET, the framework to model both the knowledge graph and language model. Ostendorff et al. (2019) combine text representations with metadata and knowledge graph embeddings to enhance BERT performance for document classification tasks.

Lexical Semantics Approach Xia et al. (2021) induce Word Similarity Matrix based on the similarity of lexical pair from the semantics role in WordNet. They inject Word Similarity Matrix directly into BERT's attention. Zhang et al. (2020) propose SemBERT, which integrates explicit contextual semantics from pre-trained semantic role labeling. Wu et al. (2021) also introduce SIFT, which incorporate explicit semantic structures into the training paradigm. Yin et al. (2020) propose SentiBERT, which incorporates contextualized representation with binary constituency parse tree to capture semantic composition.

Distribution of Self-Attention Several studies explore the characteristics of self-attention distribution and their implications for enhancing transformer-based Pre-trained Language Models (PLMs). Gong et al. (2019) investigate the

[1] https://github.com/e9t/nsmc

self-attention distribution within BERT models, demonstrating that the distribution tends to be focused around the token's position and the start-of-sentence token. They also find striking similarities in the attention distributions across the lower and upper layers. Kovaleva et al. (2019) propose that selectively disabling attention in certain heads can actually enhance the performance of fine-tuned BERT models. This discovery suggests the potential redundancy and over-complexity in the current attention mechanism. Additionally, Shi et al. (2021) present empirical evidence that the diagonal elements of the self-attention matrix, representing the attention of each token to itself, can be removed without compromising model performance. This finding further emphasizes the importance of inter-token attention over self-attention in PLMs.

3 Direct Injection of Sentiment Lexical Attention into Self-Attention

In this investigation, we enhance the existing Self-Attention mechanism by embedding Sentiment Lexical Attention within its attention matrix, thereby integrating sentiment-related connections among tokens. Sentiment Lexical Attention is conceptualized through the quantification of semantic-polarity similarity among token pairs, established via the dot product computation of their context polarity vectors. This process engenders a semantic-polarity similarity matrix that meticulously delineates the sentiment linkages inherent in tokens within a specific input sequence, ensuring a nuanced comprehension of these interrelations. Notably, a pronounced amplification of polarity similarities is observed among tokens sharing analogous sentiment properties, with the similarity values delineated within a spectrum ranging from 0 to 1.

By leveraging this semantic-polarity similarity as Sentiment Lexical Attention, we could directly inject this information into attention scores. This methodology enables us to potentially refine the attention mechanism by injecting sentiment-associated values between tokens. Consequently, this process facilitates the generation of a more informed representation of the sentiment relationships within sentences.

3.1 Knowledge-Guided Attention Approach Proposed by Xia et al. (2021)

Xia et al. (2021) proposed a methodology for directly incorporating knowledge into the self-attention mechanism by utilizing a Word Similarity Matrix. Their main objective is to enhance the focus of BERT on word pairs that demonstrate semantic similarity. To calculate the Attention Weight, they utilize the Similarity Matrix, which allows the model to assign higher weights to tokens with high similarity. The conventional definition of Self Attention can be described as follows:

$$\textit{Self Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q represents the query matrix, K represents the key matrix, V represents the value matrix, and d_k represents the dimensionality of the key matrix. Xia (2021)'s Knowledge-Guided Attention calculates the Hadamard product of the QK^T using the similarity matrix S:

$$\textit{score} = QK^T \odot S$$

$$\textit{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{\textit{score}}{\sqrt{d_k}}\right)V.$$

However, the Hadamard product of S and QK^T in the self-attention mechanism can potentially lead to issues, particularly when negative values are present in the attention score output. The nature of the Hadamard product can cause the loss of significance of certain elements if negative attention scores exist. This can dampen the importance of positive similarity values and result in an unintentional representation of token relationships. To address this, non-linear transformations or the addition of bias terms to the attention scores is necessary to ensure more reliable and stable attention distributions.

3.2 Sentiment-Fused Attention

We propose a novel concept called **Sentiment-Fused Attention**, which presents an advanced formulation for incorporating sentiment knowledge into the self-attention mechanism. Building upon the work of Xia et al. (2021), we modify the injection of knowledge to mitigate the risks associated with the Hadamard product. Instead of using the Hadamard product, we employ a summation operation to combine the Sentiment Lexical Attention with the attention scores. This alteration effectively integrates the knowledge without distorting token relationships. By using summation, we retain the positive characteristics of the original model while addressing the issues related to negative attention

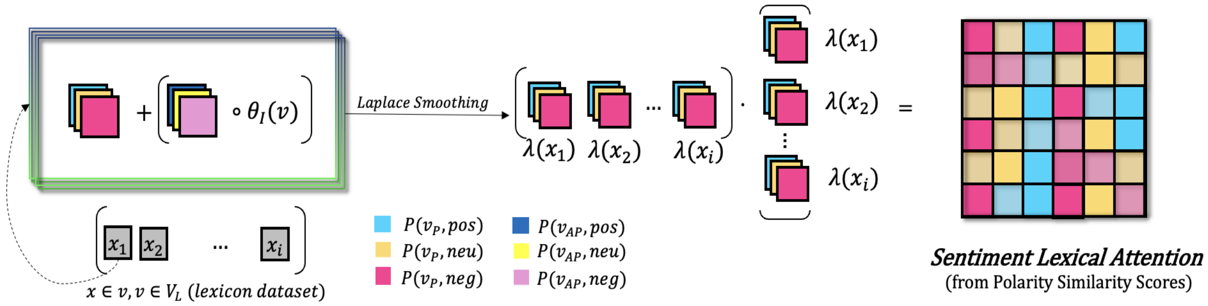


Figure 2: Pictorial Illustration of Sentiment Lexical Attention Induction from Input Variables Using Sentiment Information Sourced from the Lexicon Dataset

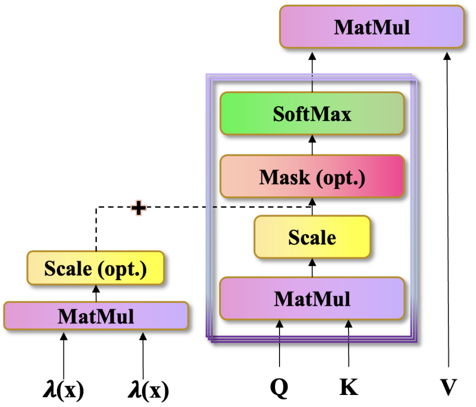


Figure 3: Visualizing the Linear Combination of Sentiment Lexical Attention and Original Attention Score Matrix

scores. This approach ensures a more accurate and stable representation of token relationships, resulting in more reliable attention distributions.

3.2.1 Leveraging Semantic-Polarity Similarity Score as Sentiment Lexical Attention

To leverage semantic-polarity similarity, we introduce the notion of Total Sentiment, denoted as $\boldsymbol{\lambda}(x) \in \mathbb{R}^3_{(0,1)}$, for each token x in an input sentence. The context polarity $\boldsymbol{\lambda}(x)$ is defined as a combination of aspect polarity vectors $\boldsymbol{\theta}_{AP}$, aspect-agnostic polarity vectors $\boldsymbol{\theta}_P$, and intensity values θ_I of words or phrases containing x. We compute Total Sentiment by Laplace smoothing the aggregate of aspect polarity vectors $\boldsymbol{\theta}_{AP}$ and taking the product of aspect-agnostic polarity vectors $\boldsymbol{\theta}_P$ and intensity values θ_I. We denote $V_{L,x}(\subset V_L)$ as a set of words or phrases containing token x, then Total Sentiment of the token is represented as follows,

$$\boldsymbol{\lambda}(x) = LS(\sum_{v \in S_L} (\boldsymbol{\theta}_P(v) + \theta_I(v)\boldsymbol{\theta}_{AP}(v)) I_{V_{L,x}}(v))$$

These modifications allow us to incorporate Sentiment Lexical Attention effectively, leading to improved attention mechanisms that provide a more accurate and stable representation of token relationships.

If there is no word or phrase containing a token x in S_L, we set $\boldsymbol{\lambda}(x)$ to a neutral sentiment vector, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This can be expressed as follows:

$$\begin{cases} LS(\sum\limits_{v \in V_{L,x}} \boldsymbol{\lambda}(v)) & \text{if } V_{L,x} \neq \emptyset, \\ (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) & \text{if otherwise.} \end{cases}$$

Additionally, we define the semantic-polarity similarity σ_{ij} (represented as $\sigma(x_i, x_j)$) as the product of the context polarities $\boldsymbol{\lambda}(x_i)$ and $\boldsymbol{\lambda}(x_j)$ for tokens x_i and x_j, respectively (Figure 2).

$$\sigma_{ij} = \sigma(x_i, x_j) = \boldsymbol{\lambda}(x_i) \cdot \boldsymbol{\lambda}(x_j)$$

3.2.2 Injection of Sentiment Lexical Attention into Attention Scores for Sentiment-Fused Attention

The formulation of Sentiment-Fused Attention involves a linear combination of Sentiment Lexical Attention, denoted as σ_{ij}, and the initial attention score, represented by $\frac{QK^T}{\sqrt{d_k}}$. This combination takes place during the forward propagation of $\frac{QK^T}{\sqrt{d_k}}$, prior to its non-linear activation through the $softmax$ function (Figure 3). The formula for the attention distribution is as follows:

$$softmax\left(\frac{QK^T}{\sqrt{d_k}} + \sigma_{ij}\right) \cdot V.$$

To stabilize the range of σ_{ij}, we include a scaling factor, $\sqrt{d_k}$, in the denominator. This architectural design ensures that the distribution of the original attention score outputs is preserved while mitigating the impact of Sentiment Lexical Attention.

$$softmax\left(\frac{QK^T}{\sqrt{d_k}} + \frac{\sigma_{ij}}{\sqrt{d_k}}\right) \cdot V$$

Throughout the training process, σ_{ij} consistently promotes similar tokens to have higher values. To ensure this consistency, we set σ_{ij} as a constant, thereby providing the model with a unidirectional stream of information. By doing so, σ_{ij} continues to provide consistent information about the relationships among similar tokens to the model. During the training procedure, σ_{ij} consistently induces high values for similar tokens, maintaining a reliable signal throughout the training process.

3.3 CARBD-ko dataset

CARBD-ko (A Korean Contextually Annotated Review Benchmark) dataset (Jang et al., 2024) is a comprehensive dataset ($S_L = \{X_i, \theta_i\}$) that includes reviews (X_i) paired with corresponding sets of sentiment factors (θ_i) for words or phrases (v_j) in the reviews. These sets consist of aspect-agnostic polarity attribute vectors ($\boldsymbol{\theta}_P(v_j)$), aspect polarity attribute vectors ($\boldsymbol{\theta}_{AP}(v_j)$), and associated intensity information of polarity values ($\theta_I(v_j)$) for each word or phrase (v_j) within the reviews. Both the aspect-agnostic and aspect polarity vectors are represented as three-dimensional one-hot vectors, which correspond to the polarity values of -1, 0, or 1. For example, a polarity value of 1 is represented by the vector $(1, 0, 0)$. The set of all words or phrases having sentiment factors is denoted as V_L. Leveraging the CARBD-ko dataset, our approach focuses on the extraction of context polarity vectors.

4 Experiments

4.1 Sentiment Classification Task

Our study focuses on conducting experiments in the domain of Sentiment Analysis, which provides a natural application for leveraging pre-existing knowledge in the field of natural language understanding. Sentiment classification tasks typically involve binary classification, distinguishing between positive and negative sentiments. Transformer-based models have shown high performance on such tasks. In our case, we evaluate the language model's performance on Sentiment Classification using the NSMC (Naver Sentiment Movie Corpus) benchmark dataset, which is widely used in Korean sentiment analysis work. The dataset consists of 200K reviews, with 150K reviews for the training set and 50K reviews for the test set. To assess the broader implications of our approach, we examine the effectiveness of Sentiment-Fused Attention in tasks that extend beyond sentiment classification.

4.2 Out-of-Domain Tasks

In addition to sentiment classification, we conduct experiments on diverse out-of-domain downstream tasks, including KorNLI (Ham et al., 2020), PAWS-ko (Yang et al., 2019a), Hate Speech Detection (Moon et al., 2020), and Question-Pair benchmark[2]. These tasks are commonly used to evaluate the overall performance of Korean language models. By evaluating our approach on these tasks, we aim to determine the generalizability and applicability of the Sentiment Lexical Attention and understand its impact on performance across various out-of-domain tasks.

4.3 Scaling Factor and Attention Head Configuration

We design a suite of experiments consisting of four distinct scenarios on NSMC benchmark. These scenarios involve different configurations of the scaling factor $\sqrt{d_k}$ and the injection scope of the Sentiment Lexical Attention. The objective is to quantify the influence of the Sentiment Lexical Attention on attention mechanisms.

The first setting involves the direct injection of values from the Sentiment Lexical Attention into all attention heads across all layers without normalization by $\sqrt{d_k}$. The second setting modifies the approach by normalizing the Sentiment Lexical Attention values σ_{ij} with $\sqrt{d_k}$ to constrain their range. The third and fourth scenarios exclusively inject the Sentiment Lexical Attention σ_{ij} into the final attention head (Att_{last}) across all layers. The fourth scenario further reduces the range of the external knowledge values through the application of $\sqrt{d_k}$.

Our working hypothesis suggests that if the use of $\sqrt{d_k}$ leads to superior results compared to alternative approaches, there may be a positive correlation between the efficacy of σ_{ij} and overall model performance. On the other hand, if enhanced performance is observed when solely activating the last attention head, it could indicate that a more targeted application of σ_{ij} yields outputs that are more representative of the context, contributing to more effective convergence of the model's objective loss.

[2] https://github.com/songys/Question_pair

NSMC	ko-electra	kr-electra	kc-electra	xlm-roberta-base	kr-bert
Baseline	90.63	91.17	91.97	89.49	90.1
Xia et al. (2021)	90.06(-0.57)	91.11(-0.06)	92.10(+0.13)	89.17(-0.32)	89.35(-0.75)
$\sum Att + \sqrt{d_k}$	91.18(+0.55)	91.73(+0.56)	92.6(+0.63)	90.42(+0.93)	90.19(+0.09)
$\sum Att$	**91.32**(+0.69)	**91.82**(+0.65)	92.56(+0.59)	90.47(+0.98)	90.17(+0.07)
$Att_{last} + \sqrt{d_k}$	91.17(+0.54)	91.78(+0.61)	92.56(+0.59)	**90.55**(+1.06)	**90.3**(+0.2)
Att_{last}	91.24(+0.61)	**91.82**(+0.65)	**92.65**(+0.68)	90.33(+0.84)	90.23(+0.13)

Table 1: Accuracy of Performance on NSMC dataset

Model	μ	σ	σ^2
$\sum Att + \sqrt{d_k}$	0.596	0.109	0.33
$\sum Att$	0.552	0.091	0.301
$Att_{last} + \sqrt{d_k}$	0.582	**0.071**	**0.267**
Att_{last}	**0.600**	0.094	0.306

Table 2: Analysis of Performance Variations via Statistical Configuration

We will employ statistical analysis to identify the scenarios that yield acceptable performance. Subsequently, we intend to assess the performance of these optimized scenarios in other out-of-domain contexts.

4.4 Models and Hyper-Parameters

To conduct our experiments, we utilize four prominent Korean Transformer Encoder-based pretrained language models (ko-electra (Park, 2020), kr-electra (Lee and Shin, 2022), kc-electra (Lee, 2021), kr-bert (Lee et al., 2020)), as well as a multilingual model (Conneau et al., 2019). The baseline performance of each model on the NSMC task is shown in Table 1, which serves as our initial reference point for comparison. To further improve the performance of our models, we engage in hyperparameter tuning. This involves adjusting the learning rate within a range of 1e-5 to 5e-5 and extending the number of training epochs from 3 to 10. By employing this rigorous setup, we aim to ensure that our experimental results accurately capture the potential benefits of our proposed approach.

4.5 $\boldsymbol{\lambda}(x)$ Initialization

In our experimental setup, we extract the context polarity $\boldsymbol{\lambda}(v)$ from the CARBD-ko dataset to initialize the context polarity $\boldsymbol{\lambda}(x)$ for individual tokens x_i, aligned with the appropriate tokenizer for each language model. However, in real-world datasets, it is possible for previously unseen tokens x_i to appear. For such cases, we initialize all $\boldsymbol{\lambda}(x_i)$ to $\frac{1}{3}$, as described in Section 3.3.1.

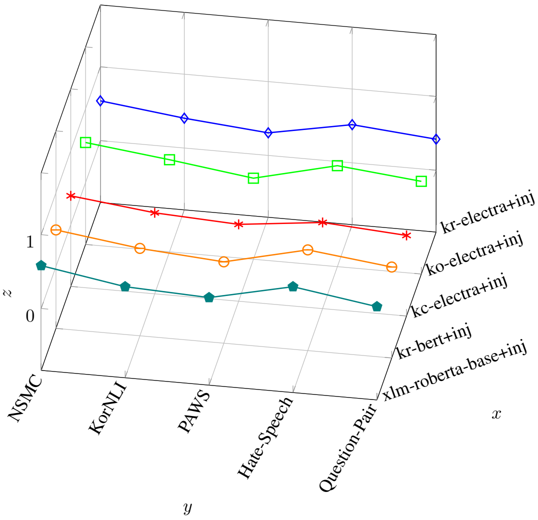


Figure 4: Appearance Rates of Initialized Tokens across 5 Downstream Tasks

When a significant number of tokens are initialized with $\frac{1}{3}$, it becomes challenging to establish a clear correlation between improved model performance and the use of σ_{ij}. Figure 4 provides insights into the appearance rates of tokens that have been initialized by σ_{ij} across five different tasks. As depicted in Figure 4, there are minimal variations observed between tasks and models, with most of the rates centered around 50%. Notably, the results on the NSMC dataset exhibit consistent stability. This finding underscores the significance of Sentiment Lexical Attention on attention, emphasizing that its impact cannot be disregarded.

5 Result

5.1 NSMC

The evidence in Table 1 emphatically underscores the advantage of injecting Sentiment Lexical Attention during fine-tuning, leading to a consistent improvement in performance across all four scenarios enumerated in Section 4.3, as compared to the baseline models and Xia et al. (2021)'s way. An intriguing observation lies in the fact that Xia et al.

Downstream Tasks		**ko-electra**	**kr-electra**	**kc-electra**	**xlm-roberta-base**	**kr-bert**
KorNLI	baseline	82.24	82.51	82.12	79.92	77.13
	+injection	**+83.25**(+1.01)	82.48(-0.03)	82.07(-0.05)	80.07(+0.15)	79.3(+2.17)
PAWS	baseline	84.45	82.05	76.5	82.95	80.35
	+injection	**85.35**(+0.9)	81.3(-0.75)	76.9(+0.4)	83.1(+0.15)	80.65(+0.3)
Hate-Speech	baseline	67.45	73.2	**73.67**	64.06	66.45
	+injection	67.73(+0.28)	73.04(-0.16)	73.46(-0.21)	66.02(+1.96)	66.67(+0.22)
Question-Pair	baseline	95.25	95.51	95.12	93.8	94.591
	+injection	95.51(+0.26)	95.51(0)	**96.04**(+0.92)	94.06(+0.26)	94.591(0)

Table 3: Accuracy of Performance Evaluation of Models on Four Out-of-Domain Tasks. We inject σ_{ij} exclusively into the last attention head of each layer with scaling factor $\sqrt{d_k}$

(2021)'s injection method and our approach yield entirely distinct outcomes. As previously noted in Section 3.2, we pointed out the potential for distortion in the mathematical derivations of Xia et al. (2021)'s method, and this has manifested in empirical results (Table 1).

Among the models, the xlm-roberta-base model illustrates the most substantial performance enhancement, whereas the kr-bert model exhibits a modest performance gain. The remaining three models demonstrate performance amplifications exceeding 0.5% across all investigated scenarios. When considering the magnitude of the NSMC benchmark's test dataset (50K), these improvements are of considerable significance, indicating a potential escalation in the number of correct predictions varying from an average of 250 to almost 500 sentences.

Notably, the kc-electra model, upon the injection of Sentiment Lexical Attention into ATT_{last} without the utilization of $\sqrt{d_k}$, achieves an accuracy metric of 92.65%. To the best of our knowledge, this represents a state-of-the-art (SoTA) result for the NSMC benchmark. These findings highlight the effectiveness of directly injecting sentiment knowledge into the attention mechanism during the training phase, leading to improved model performance.

5.2 Other Downstream Task

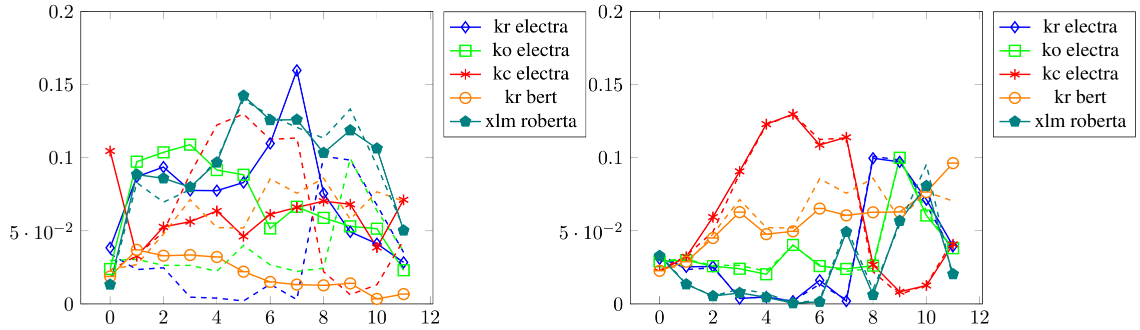
Table 2 indicates that, on average, the Att_{last} scenario results in the most significant performance improvements. The configuration of Att_{last}+$\sqrt{d_k}$ demonstrates the smallest standard deviation and variance, indicating its stability across a diverse range of models. Therefore, we adopt the Att_{last}+$\sqrt{d_k}$ configuration to inject knowledge into out-of-domain tasks.

In Table 3, out of the 20 cases examined, 13 show an increase in performance, 5 show a decrease, and 2 maintain their performance. Interestingly, these performance increases in different domains occur despite the absence of a direct correlation between the domain and the σ_{ij} values established in our experiments. This suggests that the similarity between tokens can facilitate a model's decision-making processes. However, the lack of consistent performance gains across all models, as seen in the NSMC benchmark, highlights the need for task-specific knowledge development. One notable aspect of our results is the variability and model-dependency observed in the injection of the sentiment knowledge. Performance decreases are exclusively observed in the kr-electra and kc-electra models, while other models either maintain or improve their performance. It is worth mentioning that both the kr-electra and kc-electra models consistently exhibit stable performance enhancements on the NSMC task.

Based on these findings, we conclude that directly injecting sentiment knowledge into the training process may lead to varying performance outcomes depending upon the specific model. If the knowledge, however, is logically structured and has a direct causal link with the task, it has the potential to yield stable performance improvements.

6 Dissecting the Impact of σ_{ij} on Attention Dynamics: An In-depth Analysis Centered on the CLS Token

In this section, we investigate the differences in standard deviation between the baseline model and the Att_{last} + $\sqrt{d_k}$ model concerning the CLS token at each layer. Our approach involves the direct injection of Sentiment Lexical Attention into the attention scores. We hypothesize that this injection of knowledge will lead to alterations in the relationship centered on the CLS token, which serves as



(a) Standard Deviation (σ^2) of CLS Token's Attention towards Other Tokens

(b) Standard Deviation (σ^2) of the Other Token's Attention towards CLS Tokens

Figure 5: Layer-wise Distributional Differences in Five Baseline Models and the $Att_{last} + \sqrt{d_k}$ Models, Centered on CLS Tokens. The dashed line represents the baseline models, while the solid line corresponds to the $Att_{last} + \sqrt{d_k}$ model.

the representative vector for the subsequent classifier. To test this hypothesis, we conduct an analysis of the standard deviation of attention scores surrounding the CLS token at each layer, aiming to understand the impact of σ_{ij}.

For statistical analysis, we examine the standard deviation of attention scores between the CLS token and other tokens in both the baseline model and the Att_{last} + $\sqrt{d_k}$ model. We conduct this analysis layer-by-layer while processing the 50K test dataset from the NSMC dataset. By comparing the standard deviation of attention scores, we aim to understand how the attention patterns of the CLS token change when Sentiment Lexical Attention is incorporated into the model. This analysis provides insights into the impact of knowledge injection on the attention mechanism and its effect on the relationship between the CLS token and other tokens.

Figure 5 demonstrates that the deviation between the baseline models and the Att_{last} + $\sqrt{d_k}$ model, specifically regarding the CLS tokens, primarily manifests in the alterations in the distribution of attention scores between the CLS token and other tokens. The presence of such disparities between models that differ solely based on the injection of σ_{ij} in their training processes strongly suggests a significant influence of σ_{ij} on the dispersion of attention scores. Interestingly, the distribution of attention scores from other tokens towards the CLS token remains mostly unchanged.

These findings can be attributed to the fact that the CLS token does not derive its context polarity $\boldsymbol{\lambda}(x)$ from $\boldsymbol{\lambda}(v)$, resulting in minimal differences in the attention weights towards the CLS token compared to the baseline models. On the other hand, tokens other than the CLS token, influenced by $\boldsymbol{\lambda}(v)$, consistently induce modifications in the attention score distribution throughout the training process, which likely affects the final attention distribution of the model. Through this analysis, we propose that these shifts in attention distribution serve as the primary catalyst for the performance alterations depicted in Tables 1 and 2.

7 Discussion

In this paper, we have introduced a novel approach for enhancing the self-attention mechanism of Transformer-based models through the injection of Sentiment Lexical Attention, derived from semantic-polarity scores. Our results demonstrate significant improvements in sentiment classification, particularly in the Korean language context. However, the applicability and challenges of this method across different tasks and languages, as well as its technical novelty, warrant further discussion.

7.1 Applicability to Other Languages

Our approach's effectiveness in the Korean language context opens up intriguing prospects for its applicability to other languages. Firstly, the fundamental principle of leveraging semantic-polarity scores for Sentiment Lexical Attention is language-agnostic and can be adapted to any language with available sentiment lexicons. However, the adaptation process requires careful consideration of linguistic nuances and sentiment expression in target languages. It involves meticulous curation of sentiment lexicons that accurately reflect the sentiment polarity in diverse linguistic contexts. Future work will explore the cross-linguistic applicability of our method, focusing on curating high-quality sentiment lexicons and adjusting the model to account for language-specific sentiment expression patterns.

7.2 Addressing Tasks Beyond Sentiment Analysis

The current study focuses on sentiment classification, leveraging semantic-polarity scores. While this is a direct application of Sentiment Lexical Attention, extending our approach to tasks without a clear relevance to sentiment poses challenges. To enhance the versatility and scalability of our approach, we are exploring strategies to generalize the concept of lexical feature-based attention. Future research could investigate domain-specific knowledge injection, where domain-related lexical features are derived and injected similarly to sentiment features. Additionally, integrating multiple types of lexical knowledge simultaneously could lead to a more robust and versatile model applicable across a wider range of tasks.

7.3 Technical Novelty and Contribution

While our approach builds upon existing work by Xia et al. (2021), it introduces significant innovations that extend beyond their framework. Specifically, the method of deriving Sentiment Lexical Attention from semantic-polarity scores and integrating it into the self-attention mechanism represents a novel contribution to the field. Our approach also presents a comprehensive empirical analysis across multiple architectures and tasks, establishing the effectiveness and robustness of our method. The novelty of our work lies in the specific application of lexical sentiment knowledge in enhancing the attention mechanism.

8 Conclusion

In our study, we introduced a new approach to inject sentiment knowledge into the self-attention mechanism of Transformer-based models. This approach yielded significant improvements, particularly in Korean sentiment classification benchmark, where we achieved a new state-of-the-art performance. Moreover, the promising results obtained across various out-of-domain tasks highlighted the general applicability of our method. Although the observed performance variations were task- and model-dependent, they underscored the substantial potential of incorporating human-derived knowledge into Transformer-based language models. Furthermore, in our examination of the CLS token, we could ascertain the direct impact of knowledge injection on the layer-wise attention distribution. The approach presented in this study opens the door for further exploration of effective techniques for injecting human knowledge into language models.

Limitations

Despite the promising results obtained in our study, it is important to acknowledge several limitations that should be addressed. Firstly, the application of the Sentiment Lexical Attention in our method assumes a direct relevance of the semantic-polarity scores to the specific task being addressed. This assumption limits the versatility and scalability of our approach, as the selection and application of relevant knowledge may require careful consideration and may not be readily available for all tasks. Secondly, the variation in performance observed across different models indicates that the efficacy of our approach may not be uniform across all types of Transformer-based models. It is necessary to conduct preliminary tests to assess the compatibility and effectiveness of our method with a given model before deploying it in real-world scenarios.

Thirdly, the success of our approach relies heavily on the quality and accuracy of the Sentiment Lexical Attention being employed. Tasks that require high-precision or complex human knowledge can be challenging, as even small inaccuracies in the knowledge may lead to significant deviations in performance. Careful attention should be given to the selection and curation of the Sentiment Lexical Attention to ensure its reliability and relevance to the task at hand.

Lastly, while we have made progress in understanding how to integrate pre-annotated sentiment values into Transformer models, there is still much to explore and understand about the precise influence of this knowledge on the model's training and decision-making processes. Further research and analysis are needed to gain a comprehensive understanding of these dynamics, particularly in complex real-world applications. Future work could focus on addressing these limitations by developing more adaptable knowledge injection mechanisms or conducting a more comprehensive analysis of how sentiment information influences model behavior.

By addressing these limitations, we can further enhance the effectiveness and applicability of integrating Sentiment Lexical Attention into Transformer-based models, opening up new avenues for advancements in NLP and related fields.

Ethics Statement

This research study follows ethical guidelines for conducting experiments following ACL rules. It utilizes publicly available datasets and sentiment lexicons, ensuring user privacy and avoiding any ethical concerns. The focus is on enhancing language models through the injection of the semantic-polarity scores, without manipulation or deception. The research does not involve human subjects or human-generated data. The study acknowledges potential biases and takes steps to mitigate them. Transparency and ethical considerations are paramount in the research process.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Efficient training of bert by progressively stacking. In *International conference on machine learning*, pages 2337–2346. PMLR.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Francesca Incitti, Federico Urli, and Lauro Snidaro. 2023. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436.

Dongjun Jang, Jean Seo, Sungjoo Byun, Taekyoung Kim, Minseok Kim, and Hyopil Shin. 2024. Carbd-ko: A contextually annotated review benchmark dataset for aspect-level sentiment classification in korean.

Zakaria Kaddari and Toumi Bouchentouf. 2023. A novel self-attention enriching mechanism for biomedical question answering. *Expert Systems with Applications*, 225:120210.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Junbum Lee. 2021. Kcelectra: Korean comments electra. `https://github.com/Beomi/KcELECTRA`.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *ArXiv*, abs/2008.03979.

Sangah Lee and Hyopil Shin. 2022. Kr-electra: a korean-based electra model. `https://github.com/snunlp/KR-ELECTRA`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

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

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification. *arXiv preprint arXiv:1909.08402*.

Jangwon Park. 2020. Koelectra: Pretrained electra model for korean. `https://github.com/monologg/KoELECTRA`.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and James Tin-Yau Kwok. 2021. Sparsebert: Rethinking the importance analysis in self-attention. In *International Conference on Machine Learning*, pages 9547–9557. PMLR.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing*, 102:717–740.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. 2023. Fusing external knowledge resources for natural language understanding techniques: A survey. *Information Fusion*, 92:190–204.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey. *arXiv preprint arXiv:2110.08455*.

Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. 2021. Using prior knowledge to guide bert's attention in semantic textual matching tasks. In *Proceedings of the Web Conference 2021*, pages 2466–2475.

Yifeng Xie, Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, and Dongsheng Chen. 2023. Syntax matters: Towards spoken language understanding via syntax-aware attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11858–11864.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing context into knowledge graph for commonsense question answering. *arXiv preprint arXiv:2012.04808*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2022. Jaket: Joint pre-training of knowledge graph and language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11630–11638.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Ying Zhao, Tingyu Xia, Yunqi Jiang, and Yuan Tian. 2024. Enhancing inter-sentence attention for semantic textual similarity. *Information Processing & Management*, 61(1):103535.