REB : Reinforced classifier with prompt Engineering and Bert

Anonymous ACL submission

Abstract

In this study, we evaluated traditional models for binary classification along with the latest LLM models, and explored ways to improve them. Traditional models selected were logistic regression, random forest, naive Bayes, and multilayer perceptron (MLP), while the latest large language model chosen was Flan-T5-XXL. We trained using the GLUE-SST2 dataset and evaluated using test_dataset.csv. Among the traditional models, Logistic regression shows the best performance. Futhermore, BERT, one of the latest LLM models, was evaluated. The Flan-T5-XXL model was improved through prompt engineering and evaluated. Additionally, we propose a new model (Reinforced classifier with Prompt Engineering and Bert, REB), combining BERT inference and LLM sentiment clarification. In conclusion, while using LLMs alone yielded better results than using traditional models alone, the introduction of BERT and sentiment clarification in the new model, REB, showed better performance.

Keywords : Binary Classification, Traditional Model, Large Language Model, BERT, Sentiment Clarification, Model Fusion, REB

1 Introduction

In the field of supervised learning, binary classification problems are crucial research areas used in various fields of our society such as finance, healthcare, email filtering, machine fault detection, and more. Especially with the emergence of Large Language Models (LLMs) that excel at understanding subtle context and nuances in text data, previously challenging natural language-based classification problems can now be solved more accurately and extensively.(Gu et al., 2022) Among these natural language-based problems, this study focuses on the sentiment classification problem.

The objective of this research is to identify and improve the most effective AI models for binary sentiment classification. We compare traditional machine learning approaches with latest Large Language Models and apply various enhancement techniques. The GLUE-SST2 dataset is used for training, and model evaluations are conducted using the provided test_dataset.csv. We applied the TF-IDF format preprocessing to the given text dataset. and using them, we train and evaluate the traditonal models such as Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). Furthermore, for the latest Large Language Models, this study selects the BERT model and the Flan-T5-XXL model. Especially, for the Flan-T5-XXL model, different prompt methods like zero-shot, one-shot, few-shot, and chain-of-thought are applied, and their performances are measured. Additionally, We propose a new model that combines BERT and large language models(Flan-T5-XXL). This model is referred to as Reinforced Classifier with Prompt Engineering and Bert (REB). With the introduction of BERT and the new technique of sentiment clarification through prompt engineering, this study will lead to achieve improved performance and perspectives on binary sentiment classification problems.

2 Dataset

GLUE Benchmark: The GLUE (A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding)[1] is a project that provides benchmark datasets and

[1] https://github.com/nyu-mll/GLUE-baselines

evaluation metrics for assessing natural language processing models. The goal of the GLUE project is to compare and evaluate the performance of various models using diverse datasets and evaluation metrics for different natural language processing tasks.(Wang et al., 2018). (Anonymous, 2023)

Standford Sentiment Treebank V2: The SST-2 dataset[2] is a popular benchmark dataset used in natural language processing (NLP) for sentiment analysis tasks. SST-2 dataset is an extension of the Stanford Sentiment Treebank (SST), which originally had fine-grained sentiment labels (very positive, positive, neutral, negative, very negative). The SST-2 dataset contains movie review sentences, and the sentiment of each sentence are classified positively or negatively. This dataset is labeled as 0 for positive sentiment and 1 for negative sentiment.(Socher et al., 2013). (Anonymous, 2023)

3 Methodology

In this study, four traditional models and 4 number of prompt engineering methods were created to compare accuracy of each model.

3.1 Traditional Model

We utilized Scikit-learn library and Pytorch library to make traditioinal models.

3.1.1 TF-IDF

TF-IDF is a statistical method used to measure the importance of each word in text data. It calculates the weight value of a word w in a document d using the following formula

$$\mathrm{TFIDF}(w,d) = \mathrm{TF}(w,d) \cdot \log\left(\frac{n}{\mathrm{DF}(w)}\right) \quad (1)$$

In other words, the importance of a word is proportional to the number of times the word appears in the document and inversely proportional to the total number of documents containing that word. This method is used to adopt the word with the highest weight value in a document as the keyword of that sentence. The TF-IDF method is utilized to represent a sentence in a bag-of-words style vector, making it applicable for various models. (Ramos et al., 2003)

3.1.2 Logistic Regression

Logistic Regression is a model commonly used for binary classification. It employs an optimization algorithm that adjusts weights by minimizing the loss function. The accuracies of different solver values, including lbfgs and saga, were compared.

3.1.3 Random Forest

Random Forest is a algorithm widely used for classification and regression tasks. A random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, k), k = 1,...} where the k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. We adjusted n-estimators hyperparameter and found the optimal value. (Breiman, 2001)

3.1.4 Naive Bayes

Naive Bayes is a classification model that calculates conditional probabilities given an input under the assumption that all features are independent, and predicts the probability for each class. In this study, a Naive Bayes model was created using the sklearn library, and the degree of Laplace Smoothing was adjusted by controlling the alpha value to compare accuracy.

3.1.5 MutiLayer Perceptron

The MultiLayer Perceptron (MLP) model refers to an artificial neural network model composed of multiple layers. Each layer applies weights and activation functions to learn data and classifies classes in the final layer. This study involved varying the number of hidden layers and the number of epochs to find the optimal hyperparameters.

3.2 BERT

BERT is a language representation model released by Google in November 2018.(Devlin et al., 2018) Unlike previous language data preprocessing models like word2vec, which learn in a unidirectional structure, the BERT model learns by considering both left and right directions of the sequence.

[2] https://nlp.stanford.edu/sentiment/index.html

3.2.1 Preprocessing

The input part of BERT is a sequence of token which represents a single text sequence. Each sentence starts with a [CLS] token. The final hidden state output of the transformer corresponding to this token is the classification information of the sentence for classification tasks. The [SEP] token, added at the end of the sentence, serves to distinguish between two sentences when they exist in one task. Sentences are tokenized using the BERT Tokenizer. The tokenized sentences are converted to BERT token indices, set to a maximum length, and then either padded or trimmed. Finally, an attention mask for each sequence is created, which is used to prevent the BERT model from focusing attention on the padded parts.

3.2.2 Architecture

1. **Pre-trained**

   The training process of BERT includes MLM (Masked Language Model) and NSP (Next Sentence Prediction). MLM masks a certain percentage of input tokens and predicts the masked tokens. 15% of token are changed into [MASK] token, random word, and original state token to solve the mismatch problem that may occur between the pre-tuning and fine-tuning process. Through a bidirectional approach, the model is enabled to understand the entire meaning of a sentence and predict the [MASK] tokens. In the NSP process, given two sentences, the model predicts whether they are consecutive or not. This training allows the language model to learn context and sequence.

2. **Fine-tuning**

   Using the transformer library, a pre-trained BERT model is obtained for fine-tuning. Optimal parameters are found by adjusting various hyperparameters such as epoch, batch size, and learning rate. (Devlin et al., 2019).

3.3 LLM

LLM refers to models capable of understanding and generating natural language. By tuning billions of parameters through extensive data learning, they learn language. LLMs perform well in understanding various aspects of language, grasping context, and generating sentences. Such models are used in various natural language processing tasks like document creation, translation, and question answering. There are various LLM models, but this study used the 'Plan T5 XXL' model. The Plan T5 XXL model(Chung et al., 2022) is an LLM using decoder-only and aims to improve zero-shot prompting performance using instruction tuning.

3.3.1 Prompt Engineering

A prompt refers to natural language used to request a generative AI to perform a specific task. Prompt engineering is the process designing prompts to make generative AI systems generate accurate and relevant responses. Through this process, the efficiency and effectiveness of generative AI models can be enhanced, leading to more accurate answers.

3.3.2 Zero-shot

Prompts that provide only the task or query without examples are known as zero-shot prompts. Since generative AI models are already trained on a variety of languages, they can generate results by just being given the task.

3.3.3 Few-shot

The purpose of few-shot prompting is to guide the model by giving some examples of answers. Examples of answers that include the task and reasoning process are provided to the model before the main question.

3.3.4 Chain of Thought

Chain of Thought prompting method provides the model with examples of answers that include both the task and the reasoning process. This allows the model to learn the process of reaching the answer so that model can provide a more accurate answer.(Wei et al., 2022)

3.3.5 Trigger

By adding the trigger sentence "Let's think step by step" to the existing prompt, the model is encouraged to derive results step by step.(Kojima et al., 2023)

4 Experiments

4.1 Overview

In this section, we conducted text semantic binary classification using traditional models, BERT and LLM (Large Language Model). We found the optimal hyperparameters and proper prompts for each model. Then, we will compare and analyze the results obtained. Through the analysis, we aim to identify limitations and in order to overcome them and achieve higher performance, we have developed a new model REB(Reinforced classifier with Prompt Engineering and BERT). The REB model performs text sentiment clarification on a LLM through prompt engineering and processes the transformed text through the pre-trained BERT model. Overall REB structure is shown in [Figure 2].

4.2 Traditional model

In this study, we found the optimal hyperparameters by adjusting various hyperparameters of the traditional model.

1. Logistic Regression :
   We adjusted the hyperparameters of the Logistic Regression model to find the optimal hyperparameters. We used lbfgs and saga as solver, adjusting the epochs to 1000, 5000, and 10000 to observe the accuracy of each model. The results showed that regardless of the number of epochs, saga displayed an accuracy of 74%, while lbfgs showed an accuracy of 76%.

2. Naive Bayes :
   We adjusted the alpha value in the Naive Bayes model to find the optimal model. By using a grid search algorithm with various alpha values(Laplace Smoothing)(Ji and Kwon, 2023), we found that the optimal value is 1.0. The results showed an accuracy of 75% when the alpha value was set to 1.0.

3. Random Forest :
   In Random Forest model, we adjusted the n-estimator parameter to control the number of trees and find the optimal value. The experimental results showed an accuracy of 73% when the n-estimator value was set to 500.

4.3 BERT

We used a pre-trained BERT model using the Hugging Face's Transformers. After importing the model, we fine-tuned it to achieve best results. Because of limitations in computing resources and performance, we could not conduct extensive fine-tuning for BERT, the large-scale model. we performed fine-tuning for 2 to 10 epochs, identifying the epoch value that yielded the highest performance and applied this value to the final BERT model. Also, we adjust various parameters such as learning rate and batch size. When using 2 epochs, a batch size of 32, and a learning rate of 2e-5, we achieved an accuracy of 94%.

4.4 LLM

We tried a lot of Prompts. [Figure 1] represents various evaluation criteria of each prompt we conducted.

4.5 REB

We utilized LLM to conduct sentiment clarification, modifying sentences to more clearly express emotions. By employing the Trigger sentence prompt(Kojima et al., 2023), we make LLM comprehend emotions better and transform them into easily understandable sentences. Using changed sentences, we created a new test dataset and classified them with the previously trained BERT model. The accuracy of REB model was 96%.

5 Results

The overall experimental results is shown in [Figure 3]. Among the traditional models, logistic regression has the highest accuracy at 76%, and Random Forest model and MLP model have the lowest accuracy at 73%. Among the LLM prompt engineering results, the result of one-shot prompting was the highest accuracy at 95%. Rather, the accuracy of prompting with trigger sentence, which was expected to show the highest result, was slightly lower at 90%. The trigger sentence we used does not appear to show high accuracy because trigger sentence like 'let's think step by step' does not fit into the classification of sentiment. Overall, it can be seen that the performance of the BERT or LLM model is about

Method	Prompt	Accuracy	F1-score (Weighted avg)	F1-score (Macro avg)
Zero shot	Classify whether the following sentence is positive or negative. \n If positive, print 1. If negative, print 0.	0.94	0.94	0.94
	considering this context, determine if the following sentence is more likely to be positive, print 1, if negative, print 0	0.92	0.93	0.93
One shot	Classify whether the following sentence is positive or negative. If positive, print 1. If negative, print 0. i give one example.[…]	0.95	0.95	0.95
Few shot – 3 examples	if sentence is positive, print 1, if negative, print 0. i give 3 test example. […]	0.92	0.92	0.92
Few shot – 9 examples	if sentence is positive, print 1, if negative, print 0. i give 9 test example. […]	0.91	0.91	0.91
Chain of Thought	Let's classify positive and negative sentences. If the sentence is positive, 1 is output, if it is negative, 0 is output. Q: the part where nothing's happening A: This sentence suggests boredom. Also there are no indicators indicating positivity: 0 Q: {} A:	.0.95	0.95	0.95
Trigger sentence	Let's classify positive(1) and negative(0) sentences. let's think step by step Q: the part where nothing's happening A: This sentence suggests boredom. Also there are no indicators indicating positivity: 0 Q: {} A:	0.90	0.90	0.90

Figure 1: Prompts.

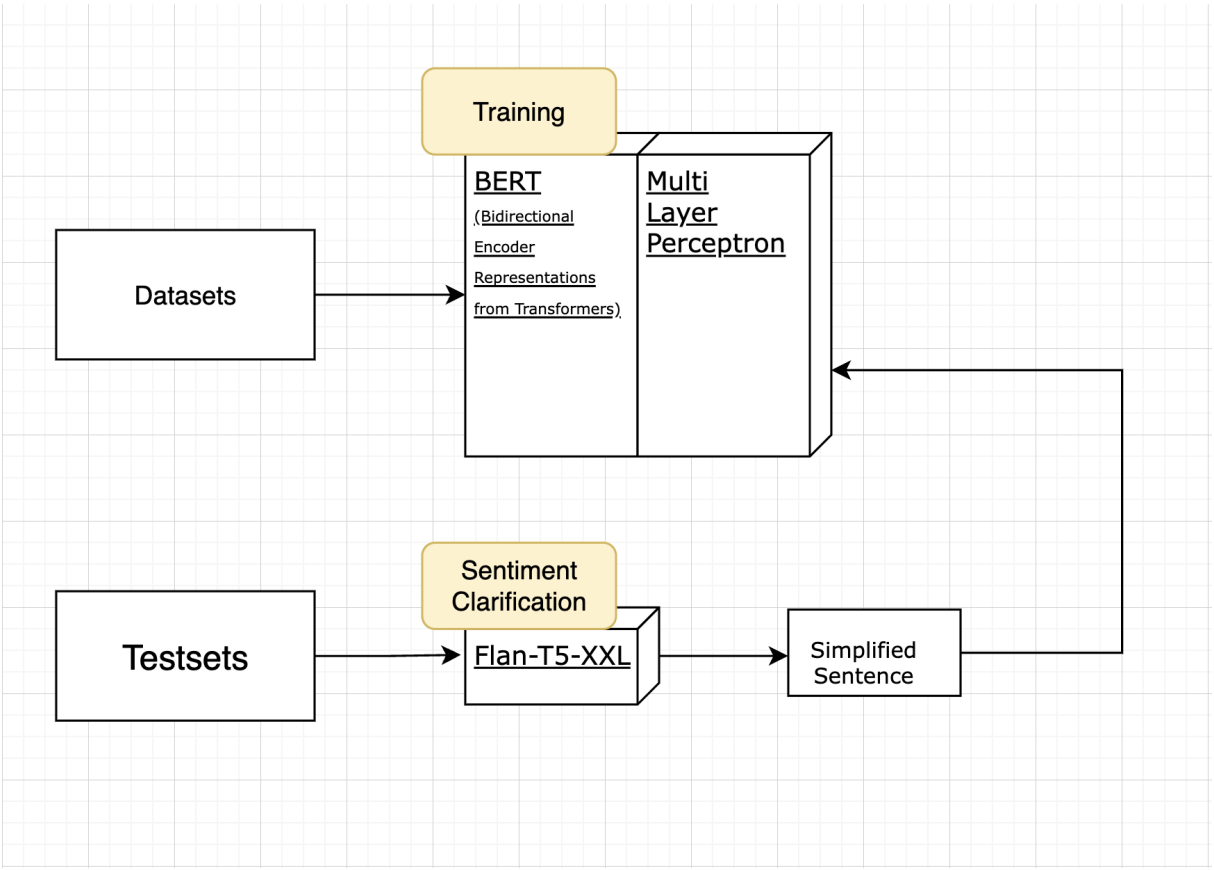


Figure 2: REB Structure.

20%p better than the traditional model. This can be seen as a limitation of the TF-IDF preprocessing method. TF-IDF is a simple formula based on the frequency of appearance of words and the frequency of documents, so it may not understand the semantic content of sentence. In addition, it is difficult to classify sentences that must be interpreted differently according to the order because they ignore the order of words. On the other hand, since the

BERT model performs bidirectional learning, it shows good performance in grasping the context of sentences and understanding the meaning of words. The LLM model also has a structure that includes multiple attention layers, so it performs well to understand the context and the meaning of each word. Therefore, there is a large difference in accuracy between the traditional model learned by TF-IDF preprocessing and the BERT and LLM models.

In addition to simple BERT and LLM models, REB proceeds with sentimental clarification that makes the sentiment of the sentence more clearly revealed. Through prompt engineering with Flan-T5-XXL model, we modify the given sentences and apply them to the BERT model to distinguish sentiment more clearly. The REB model also shows a performance improvement of about 1-2%p compared to the LLM or BERT models.

An example of the clarification of the sentence is shown below.

1. Before : **'manages to show life in all of its banality when the intention is quite the opposite.'**

2. After : **'The film makes life look boring, when the intention is quite the opposite.'**

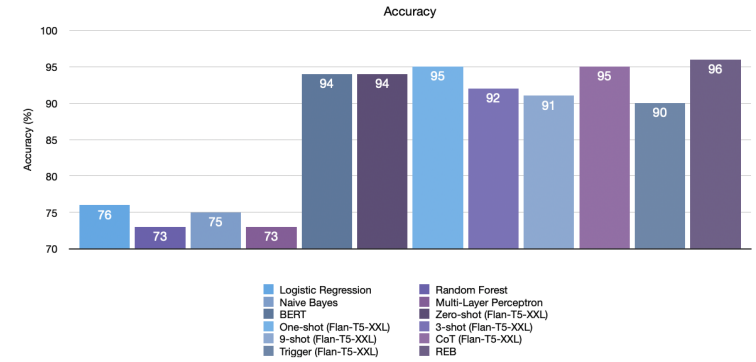
Sentences before sentimental clarification are somewhat difficult to understand even for humans and have difficult words such as 'banality'. Looking at the sentences after Clarification, 'banality' was changed to words that are easy for anyone to understand, such as 'boring', making the sentimental more clearly. Thus, we can observe an improvement in accuracy by employing a Large Language Model (LLM) to modify the emotional tone of a sentence for better clarity and subsequently applying this to the BERT model.

6 Discussions

In this study, we proposed the REB (Reinforced classifier with prompt engineering and Bert) model, a new approach to improving the performance of binary sentiment classification. Existing studies have mainly relied on traditional machine learning models for classification or focused on large language model prompt engineering for classification. Although these methods can be effective, they have often had limitations in fully understanding and classifying complex sentence structures or various expressions. We proposed REB to solve these existing problems. REB converts a given sentence into a more understandable form using the Large Language Model (LLM), and performs classification by inputting the converted sentence to the BERT model. This process simplifies complex language structures, making it easier for BERT to analyze the sentiment contained in sentences more effectively. Unlike conventional approaches, this is an creative way to improve performance by combining the strengths of the two models. REB achieved a slight performance improvement over the existing method. This means that our REB model can better understand and classify complex sentence structures and nuances. In the real world, it is expected that our model will have a significant impact in areas where text classification is important (sns, advertising, and spam blocking). However, our model has a disadvantage in terms of latency. Since it has to go through the converting sentence step through LLM and the inference step of the Bert model, it has a disadvantage that the latency increases compared to the existing classification models. In addition, loss of important information to sentimental judgment may be made in the transformation step. Future research can be expanded in the direction of reducing latency by using a technique (quantization(Yang et al., 2019), pruning(Wang et al., 2020)) that lowers the complexity of the model while maintaining accuracy as much as possible. Furthermore, we can extend the study to minimize emotional information loss by applying new prompt engineering techniques in the sentence conversion step through LLM.

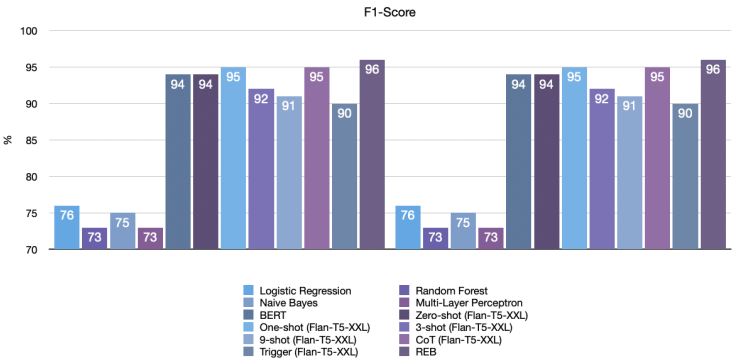
7 Conclusion

We have trained and tested various models for text sentiment binary classification, identifying their limitations along the way. As a solution to these limitations, we propose our new framework called REB. Our new model shows approximately a 20%p higher accuracy compared to traditional machine learning models



Accuracy

	Accuracy
Logistic Regression	76
Random Forest	73
Naive Bayes	75
Multi-Layer Perceptron	73
BERT	94
Zero-shot (Flan-T5-XXL)	94
One-shot (Flan-T5-XXL)	95
3-shot (Flan-T5-XXL)	92
9-shot (Flan-T5-XXL)	91
CoT (Flan-T5-XXL)	95
Trigger (Flan-T5-XXL)	90
REB	96



F1-score

	F1-score macro Average	F1-score weighted Average
Logistic Regression	76	76
Random Forest	73	73
Naive Bayes	75	75
Multi-Layer Perceptron	73	73
BERT	94	94
Zero-shot (Flan-T5-XXL)	94	94
One-shot (Flan-T5-XXL)	95	95
3-shot (Flan-T5-XXL)	92	92
9-shot (Flan-T5-XXL)	91	91
CoT (Flan-T5-XXL)	95	95
Trigger (Flan-T5-XXL)	90	90
REB	96	96

Figure 3: Accuracy and F1 scores.

and a 1-2%p improvement compared to the latest language models. To enhance classification performance, we introduce the new model REB by combining two existing models. This research can be seen as a novel perspective in the field of natural language processing, as it goes beyond simply altering the structure or parameters of the model to improve performance.

References

Anonymous. 2023. *2023 Fall SWE301141 - Homework Description.*

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mose Gu, Junhee Kwon, Jaehoon Paul Jeong, and Sanghee Kwon. 2022. An emotion classification scheme for english text using natural language processing. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1941–1946. IEEE.

Keungyeup Ji and Youngmi Kwon. 2023. Implementation of malicious mail filtering system through refinement of multinomialnb technique. *The Journal of Korean Institute of Information Technology*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

OpenAI. 2023. ChatGPT (Feb 13 version). https://chat.openai.com. Large language model.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Huan Wang, Can Qin, Yulun Zhang, and Yun Fu. 2020. Neural pruning via growing regularization. *arXiv preprint arXiv:2012.09243*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. 2019. Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7316.

This paper is written in Korean and translated English by chatGPT-3.5 (OpenAI, 2023)