Improving Unethical Dialogue Identification in Korean Text using CoMPM

Korea University

Sangwoo Ra
Department of Big Data
Team 30
2018380702

Ji Eom
Department of Digital Business
Team 30
2019390826

Abstract

This study applies the Context Modeling with Speaker's Pre-Trained Memory Tracking (CoMPM) model to the task of identifying unethical conversations in the digital domain, with a particular focus on Korean. Existing approaches rely heavily on external structured data, which is mainly available in English, which limits their applicability to other languages. We addressed these limitations by fine-tuning KcELECTRA, a model pre-trained on Korean text ethics verification data, and implementing a CoMPM approach that considers context and prior knowledge. Our results show that context plays an important role in text recognition, even in short digital exchanges. Using the F1 score to compare the performance of our model with a model that does not consider context, we found a significant improvement from 0.61 to 0.647. Despite limitations such as inaccurate predictions when context changes and possible overfitting due to limited data sets, these results provide promising directions for future research to improve model performance through improved context modeling, data expansion, and model simplification.

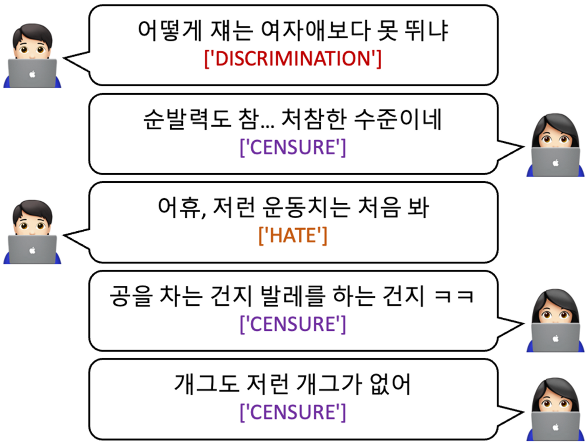
1 Introduction

In the rapidly evolving digital landscape, the proliferation of online platforms is creating countless opportunities for individuals to express their views and participate in discussions. This growth in communication, however, is not without its challenges, and a foremost among these is the identification and moderation of unethical or immoral utterances within digital conversations. Identifying such unethical remarks in text-based dialogues is a complex task, compounded by the intricate nature of human language, amplified by emotional inflection and subtle nuances of dialogue context.

Despite numerous advancements in the field of Emotion Recognition in Conversation (ERC), the identification of unethical remarks remains a formidable challenge. A substantial proportion of existing methods heavily rely on external structured knowledge, predominantly provided in English. This dependency poses significant difficulties when attempting to apply these methods to non-English languages.

As part of efforts to overcome these language-dependent limitations, CoMPM (Combination of CoM and PM)[1] was introduced. Being independent of external structured data, CoMPM can potentially serve as a viable option for application across a wider range of languages. This study aims to explore such possibilities by fine-tuning a pre-trained model with Korean text ethics verification data. The goal is to enable the model to distinguish between ethical and unethical dialogue content, utilizing CoMPM to reflect the context and prior knowledge in the classification process.

As an extension of previous research, this study seeks to expand the applicability of models previously used in ERC to facilitate multi-class classification of ethical dialogue types. Furthermore, by contrasting the results of context-reflecting models with those that do not consider context, this research aims to validate the hypothesis that context plays a pivotal role in text recognition even in relatively short digital exchanges, such as comments. As can be seen in Figure 1., the ethical characteristic of the last sentence alone is indeterminable. However, considering the overall context and the speaker's characteristics, it can be recognized as a censure. This research aims to validate these considerations.



어떻게 쟤는 여자애보다 못 뛰냐
['DISCRIMINATION']

순발력도 참... 처참한 수준이네
['CENSURE']

어휴, 저런 운동치는 처음 봐
['HATE']

공을 차는 건지 발레를 하는 건지 ㅋㅋ
['CENSURE']

개그도 저런 개그가 없어
['CENSURE']

Figure 1: Model Architecture

2 Related Work

Text classification is a crucial technique widely applied in the field of natural language processing, with various applications such as topic classification, sentiment analysis, spam detection, and intent detection. Various machine learning models and deep learning architectures have been employed in this field.

Kim (2014)[2] proposed the Text-CNN model. By applying convolutional techniques used in image processing to text classification, this method analyzes relationships between the central word and its surrounding words. It demonstrated excellent performance in various text classification tasks. However, the model had limitations in adequately reflecting prior context.

To overcome these limitations, several Emotion Recognition in Conversation (ERC) models that consider the context of a conversation have been proposed. These models model the state of each speaker participating in the conversation, and use this to reflect the context of the previous conversation. In our research, we utilize the CoMPM(Combination of CoM and PM) model[1]. CoMPM aims to enhance the predictive performance of ERC models by reflecting the context of previous conversations into a pre-trained memory.

Among the models used for text classification, BERT (Bidirectional Encoder Representations from Transformers)[3] is a representative example. BERT, based on the encoder of Transformer, understands context and learns relationships between sentences through MLM (Masked Language Model) and NSP (Next Sentence Prediction). However, since BERT only learns using masked data, there is a problem that considerable computing resources are required to obtain the desired performance.

To address this issue, Clark et al. (2020)[4] proposed ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately). ELECTRA increased the efficiency of learning by performing learning on all tokens. In our research, we use KcELECTRA[5], which is pre-trained using Korean comment data. Compared to KcBERT[6], which is a BERT trained with Korean comment data, it has a dataset about 70 million larger and 1.5 times larger vocabulary, and it demonstrated higher performance in downstream tasks.

Based on these studies, our research proposes a classification model that can learn more efficiently about ethical information by grafting KcELECTRA as pre-trained model and CoMPM. Through this, we aim to more accurately detect ethical violations in conversations.

3 Approach

Our model utilizes online conversation data, which includes speakers U_u and utterances txt_i, and is classified into 8 "Types". The conversation consists of $(U_A, txt_1), (U_B, txt_2), (U_A, txt_3), ..., (U_u, txt_i)$ with at least two participants in the conversation. In this study, given the previous conversation $h_t = \{(U_A, txt_1), (U_B, txt_2), ..., (U_{u-1}, txt_{t-1})\}$, we aim to predict the ethical type of the utterance txt_t of speaker U_u at the current time t.

3.1 Model Architecture

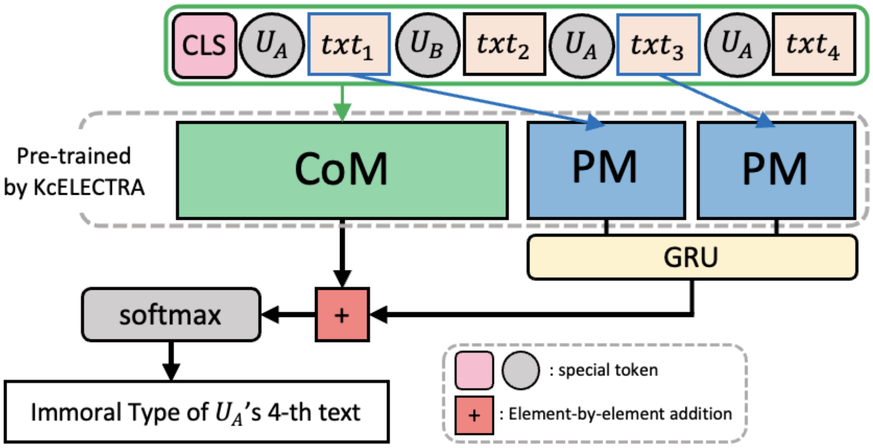


Figure 2: Model Architecture

Figure2. represents the architecture of our model. Conversation data are distinguished by special tokens for use in the CoM. The CoM accepts the conversation up to this point as an input and handles the relationships of previous utterances towards the current statement as context. Previous utterances (txt_1,txt_2) of the current speaker U_A are used in the PM. The PM uses only the previous utterances of the current speaker, which are processed as the speaker's prior knowledge. The prior knowledge processed through the PM passes through a unidirectional GRU and is then reflected in the results of the CoM. Subsequently, it is applied to the softmax function to predict the immoral type of the current utterance.

3.2 CoM (Context Embedding Module)

The CoM processes the conversation up to this point as context. The KcELECTRA, which is a pre-trained model tailored to unrefined Korean and trained on resources such as news comments, is utilized. A model using only the CoM, which does not reflect prior knowledge, was implemented. The output of the CoM is applied to the softmax function and returned as a predicted class.

3.3 PM (Pre-trained Memory Module)

The PM takes only the previous utterances of the current speaker as input and utilizes them as prior knowledge. The KcELECTRA, which is the same pre-trained model used in the CoM, is employed. A model using only the PM, which does not reflect the context of the conversation, such as the other person's emotions, was implemented. The output of the PM is applied to the softmax function and returned as a predicted class.

3.4 CoMPM (Combination of CoM and PM)

The CoMPM is a model that combines the CoM, which reflects context, and the PM, which reflects prior knowledge. The CoM and PM are pre-trained with KcELECTRA, allowing them to understand each other. The output of the PM is applied to a unidirectional GRU, which reduces the weight of older utterances. The outputs of the CoM and PM are combined using element-by-element addition, then applied to the softmax function and returned as a predicted class.

3.5 KcELECTRA

To compare with the results of our context-reflecting model, we implemented a model that does not reflect context. This is a classification model based on KcELECTRA, the pre-trained model of our model, which predicts the immoral type based solely on a single sentence.

4 Experiments

4.1 Data

In this study, we employ the Text Ethics Verification Data provided by AI Hub[1]. This dataset comprises a compilation of digital dialogues in Korean, with each conversation being labeled with one of eight ethical types: "IMMORAL NONE", “CENSURE”, "HATE”, "DISCRIMINATION", “SEXUAL”, “ABUSE”, "VIOLENCE", "CRIME", as shown in Table 1. Each item in the dataset constitutes a conversation between at least two speakers, with specific individuals and products anonymized. The features utilized in this study include ’id’, ’sentence’, along with sub-features of ’sentence’ such as ’speaker’, ’text’, ’types’, ’is immoral’.

Label	Count
IMMORAL_NONE	15,566
CENSURE	10,876
HATE	5,830
DISCRIMINATION	3,369
SEXUAL	2,475
ABUSE	1,924
VIOLENCE	1,490
CRIME	587

Table 1: Count of each label in the training dataset.

While the original dataset is composed of six json files, we utilized only four due to the considerable size of the model. The dataset demonstrates an imbalance with a relatively larger count of "IMMORAL NONE" and "CENSURE". Therefore, preprocessing was performed in a manner that alleviated this imbalance while preserving the unique characteristics of the data. The result of this preprocessing is as depicted in Table 1, detailing the quantity of data per label. This preprocessing procedure was applied to only three of the json files, which were subsequently split into a training set and validation set at a ratio of 2:1. The remaining json file was used as a test set without preprocessing. The training set was utilized for model training, the validation set for fine-tuning model parameters and monitoring performance during training, and the test set for evaluating the final model’s performance. The number of data used in the study is detailed in Table 2.

Set	Type Count	Dialogue Sentence Count
Train	42,027	12,500
Valid	21,011	6,250
Test	45,215	13,297

Table 2: Count of dialogue sentences in the Train, Validation, and Test sets.

[1] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=558

4.2 Evaluation method

In this research, the evaluation during the training phase employed measures such as Accuracy, Precision, Recall, and F1-score. In the final phase, we used Accuracy and the F1-score, with the latter serving as the primary criterion for gauging performance superiority. These metrics allow for a comprehensive evaluation of the model's performance in Immoral Type Recognition in Online Conversation. Particularly, the F1-score, being the harmonic mean of precision and recall, provides a more balanced measure in the face of uneven class distribution. The evaluation made use of both the validation set (for model selection) and the test set (for final performance reporting).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{2}$$

4.3 Experimental details

All our experiments were conducted in a Google Colab Pro environment with 40GB of GPU RAM, utilizing an A100 GPU. To train the model, we used the AdamW optimizer, with a learning rate set at 1e-5 and a batch size of 16. We applied gradient clipping with a maximum gradient norm set to 10 to prevent the exploding gradient problem and trained the model over a total of 10 epochs. The training time was approximately 2.5 hours each for the CoM and PM, approximately 5 hours for the CoMPM, and about 3.5 hours for the KcELECTRA.

4.4 Results

We evaluated and compared the performance of KcELECTRA, CoM, PM, and CoMPM on our task of detecting ethical violations in text. The performance metrics for each model are presented in Table 3. Our results indicate that all three of our proposed models (CoM, PM, and CoMPM) demonstrated superior performance compared to the KcELECTRA baseline. This suggests that the integration of context (through CoM and PM) and pre-trained memory (through CoMPM) significantly improved the detection of ethical violations within the textual content.

However, it is noteworthy that while CoMPM outperformed the baseline, its performance was not as high as anticipated. A plausible explanation for this could be overfitting due to limited data. Overfitting typically arises when a model is excessively complex relative to the quantity and noise level of the training data. In such circumstances, the model tends to memorize the training data rather than generalize from it, which subsequently leads to decreased performance on unseen data. Therefore, it is essential to gather more data and further generalize the model to improve its performance and robustness.

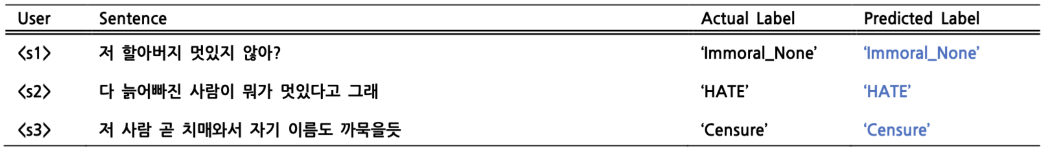
Model	Accuracy	F1 Score
KoELECTRA	0.59	0.61
CoM	0.643	0.653
PM	0.65	0.649
CoMPM	0.641	0.647

Table 3: Accuracy and F1 Score for different models.

5 Analysis

Our model considers previous conversations as context and utilizes the current speaker's previous utterances as prior knowledge to more accurately predict the ethical type of the current utterance.

Figure 3. displays one of the test results from our model. As expected, we observed accurate label prediction. When predicting the label for the statement <s3>, the change in the context of the previous statement from 'Immoral_None' to 'HATE' was reflected, leading to a 'Censure' prediction. On



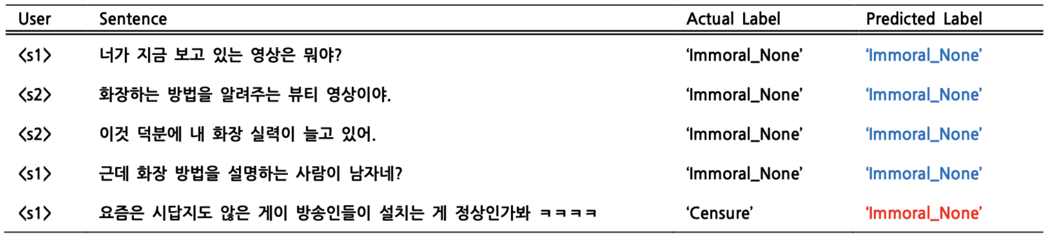
User	Sentence	Actual Label	Predicted Label
<s1>	저 할아버지 멋있지 않아?	‘Immoral_None’	‘Immoral_None’
<s2>	다 늙어빠진 사람이 뭐가 멋있다고 그래	‘HATE’	‘HATE’
<s3>	저 사람 곧 치매와서 자기 이름도 까묵을듯	‘Censure’	‘Censure’

Figure 3: True predictions due to context

the contrary, KcELECTRA predicted the statement <s3> as ’Immoral_None’ for the same dialogue. Thus, this test case underscores the significance of context in ensuring precise predictions.

Figure 4. illustrates a situation where our model fails to predict a change in ethical types in the current statement when the previous ethical types in the context remain the same. Given the CoMPM model’s dependency on context and prior knowledge, it sometimes over-relies on previous dialogues, hindering its optimal performance when the context transitions. This suggests that the model’s bias towards the previous context could lead to erroneous predictions.

Similar to many Natural Language Processing (NLP) models, our model also struggled with accurately classifying sentences employing satire or indirect speech. This highlights the ongoing challenge in NLP models to comprehend the subtle nuances of language and the inherent meanings encapsulated within specific sentences.



User	Sentence	Actual Label	Predicted Label
<s1>	너가 지금 보고 있는 영상은 뭐야?	‘Immoral_None’	‘Immoral_None’
<s2>	화장하는 방법을 알려주는 뷰티 영상이야.	‘Immoral_None’	‘Immoral_None’
<s2>	이것 덕분에 내 화장 실력이 늘고 있어.	‘Immoral_None’	‘Immoral_None’
<s1>	근데 화장 방법을 설명하는 사람이 남자네?	‘Immoral_None’	‘Immoral_None’
<s1>	요즘은 시답지도 않은 게이 방송인들이 설치는 게 정상인가봐 ㅋㅋㅋㅋ	‘Censure’	‘Immoral_None’

Figure 4: False predictions due to context

6 Conclusion

In our study, we applied the CoMPM model, an ERC model, not for emotion recognition but for recognizing immoral types in digital dialogues online. Our model employed KcELECTRA, a pre-trained model learned from unrefined Korean news comments, and used CoMPM, which reflects the context and prior knowledge. Consequently, our approach proved effective, showing a 3.7%p. higher F1-score compared to the conventional text classification model implemented with KcELECTRA. Therefore, it implies that our approach can be extended for not only immoral types but also various other types of classification tasks.

In our experiments, the CoM model exhibited superior performance to the PM model. Based on this, we discerned that context might be more crucial than prior knowledge, suggesting that future research could explore more granular approaches to context modeling. On the other hand, we identified a limitation where the model fails to perform accurate predictions when a conversation shifts from one context to another. Therefore, it is essential to explore solutions for this issue. Conversely, we noticed a slightly lower performance of the CoMPM model in comparison to the CoM and PM models, which could be attributed to a degree of overfitting due to limited datasets. Future research should aim to collect and integrate more data and seek strategies for generalization, such as regularization techniques or model simplification, to resolve these issues.

References

[1] Wooin Lee Joosung Lee. CoMPM: Context Modeling with Speaker’s Pre-trained Memory Tracking for Emotion Recognition in Conversation, 2022.

[2] Yoon Kim. Convolutional Neural Networks for Sentence Classification, 2014.

[3] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[4] Quoc V. Le Christopher D. Manning Kevin Clark, Minh-Thang Luong. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, 2019.

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

[6] Junbum Lee. Kcbert: Korean comments bert. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pages 437–440, 2020.

A Appendix: Team contributions

Both team members, Sangwoo Ra, and Ji Eom, contributed equally to all aspects of this research project.