**Observations on the Dataset of Korean Threatening Conversations**

Lee Jin1

1Aiffel

**Abstract**

This research explores methodologies for classifying threatening conversations in the Korean language, an effort initiated as part of the DLthon competition organized for the 2021 AI Grand Challenge. Leveraging the TUNiB dataset, we aim to classify conversations into four distinct threat categories, including general conversation as a non-threatening control. Our approach utilizes data augmentation techniques, synthetic data generation for general conversation samples, and an extensive comparison of various model architectures to optimize classification performance. The final model is evaluated by F1 score, with findings showing that our synthetic data approach significantly enhances non-threatening class identification.

1. **Introduction**

The detection and classification of threatening language in Korean present unique challenges due to the language's morphology and contextual nuances. Threat recognition is crucial for applications in online monitoring, safety, and conversational AI, especially for distinguishing subtle linguistic cues across diverse threat types. The DLthon competition task requires categorizing dialogues into four threat classes: extortion, workplace harassment, other harassment, and a general conversation class, the latter generated as synthetic data to balance the dataset. This study aims to achieve high classification performance on the TUNiB dataset and assess which data augmentation techniques contribute most to accuracy.

**1.1 Background**

As online interactions become an integral part of modern life, the need for effective classification of threatening language has intensified. Natural Language Understanding (NLU) techniques are widely applied in tasks such as sentiment analysis and AI-driven conversational agents, helping systems interpret user intent accurately. However, identifying and categorizing various types of threats—such as extortion, workplace harassment, and intimidation—remains a challenging task, especially in the Korean language, where context and nuances are critical for accurate classification.

One major hurdle is that existing models often fail to differentiate between types of threatening language, treating all forms of threats as a single category. Furthermore, in many Korean datasets, non-threatening or "general" conversation examples are either absent or underrepresented, resulting in data imbalances that can hinder model performance.

**1.2 Related Works**

Research in harmful content detection has progressed substantially in recent years, with models like BERT and its variants achieving significant success in English language toxic language detection. Studies on Korean-specific models, such as KoBERT and KoELECTRA, have also shown promise in tasks related to sentiment analysis and hate speech detection. However, these applications generally approach harmful language classification broadly, without distinguishing between specific threat types.

For Korean language processing, existing approaches often focus on single-label classification of harmful content. This approach limits the model's effectiveness in handling diverse threat types, which is particularly relevant for applications requiring accurate identification of workplace harassment, extortion, and other forms of intimidation.

**1.3 Limitations of Previous Research**

While advancements have been made in toxic language classification, current approaches exhibit several limitations:

1. **Lack of Granular Classification**: Many models treat all types of threats as a single class, leading to misclassifications and limiting applicability in situations that require precise threat identification.
2. **Data Imbalance**: Existing datasets often lack adequate non-threatening conversation examples, which hampers the model’s ability to distinguish between threats and general conversations. This imbalance tends to cause models to overfit on threat classes.
3. **Limited Use of Non-Threatening Data**: Few studies explore strategies to supplement datasets with general conversation data to balance training samples and improve the model's ability to handle non-threatening language.
4. **Methodology**

**2.1 Data Preparation**

**2.1.1 Data Analysis and Tokenization**

An initial analysis of conversation lengths revealed a wide distribution, with a maximum length of 839 tokens and a mean length of 214 tokens. Based on this, we determined an optimal sequence length for tokenization to retain essential context while avoiding unnecessary padding.

**Tokenization Process** : Using KoBERT’s get\_tokenizer utility, we tokenized the dataset by segmenting each conversation into subword units. This approach utilizes a word-piece tokenization strategy, which allows frequently occurring words to remain intact as single tokens while less common words are split into subword components. This method ensures effective handling of complex Korean morphology, preserving the semantic and syntactic integrity of conversations.

**Stopword Handling**: Unlike traditional approaches, we opted not to remove stopwords, as even filler words can carry contextual meaning critical for threat detection. Through experiments, we confirmed that stopwords did not negatively impact model performance and, in some cases, aided in better interpretation of user intent.

**Distribution Adjustment**: After tokenization, the maximum sequence length was reduced to 458, with a mean of 123 tokens. This adjustment balanced preserving conversational content and reducing padding, optimizing input for KoBERT.

**Embedding Components**:

* **Token Embeddings** capture the individual word-pieces, allowing the model to process both high-frequency and rare terms.
* **Segment Embeddings** are used to differentiate parts of the conversation, facilitating better handling of conversational exchanges as a coherent sequence.
* **Position Embeddings** retain the order of tokens, helping the model understand sentence structure, which is crucial for interpreting intent.

**2.1.2 Model Input Preparation**

To further optimize model input, we set the max sequence length parameter to 200 based on insights from the distribution of tokenized sentence lengths. This value balances capturing full conversational context while minimizing excessive padding for shorter conversations. Additionally, we maintained special tokens, such as [CLS] and [SEP], to signal the start and end of sequences, respectively, enhancing KoBERT’s attention to conversation boundaries.

**2.2 Model Selection**

For the task of threat classification, this study employs KoBERT as the primary model. KoBERT is a transformer-based model specifically optimized for Korean language understanding, utilizing a bidirectional encoder architecture. This structure enables the model to capture context from both preceding and following tokens within a sentence, an essential feature for accurately interpreting the nuanced intent inherent in conversational data.

The selection of KoBERT is based on its proven efficacy in Korean natural language understanding tasks. The model leverages BERT's encoder architecture, which is well-suited to capturing sentence-level context and meaning. This capability makes KoBERT particularly appropriate for tasks that require a deep understanding of user intent, as is necessary for accurately classifying threatening language in conversations.

KoBERT's architecture comprises multiple transformer layers that process a combination of token, segment, and position embeddings. These embeddings collectively generate a rich contextual representation of the input conversation, enabling the model to discern subtle semantic and syntactic variations critical to effective threat classification.

**2.3 Outlier Detection and Classification**

To manage uncertain predictions, we implemented an outlier detection mechanism based on prediction confidence. During inference, the model computes the probability of each class for a given input. When the confidence for a specific class falls below a predefined threshold, the input is classified as "general conversation." This approach addresses the absence of explicit general conversation data and allows the model to handle non-threatening conversations with reasonable accuracy.

1. **Results**

**3.1 Overall Performance**

Our model achieved an optimal F1-score of 0.90443 with a max\_len of 200 and a classification threshold of 0.96. This high F1-score indicates the model's robust ability to distinguish between threatening and general conversations, achieving balanced precision and recall across all classes. Detailed performance metrics for each threat class (extortion, workplace harassment, other harassment) demonstrate the model’s effectiveness in capturing distinct linguistic patterns associated with each type of threat.

**3.2 Impact of Synthetic Data**

The inclusion of synthetic general conversation data significantly improved the model's performance in identifying non-threatening conversations. Comparative results showed enhanced precision and recall for the general conversation class, with synthetic data yielding a clear advantage over models trained solely on threatening conversations. This finding supports the hypothesis that data augmentation can mitigate data imbalance, enhancing the model's generalization capabilities.

**3.3 Ablation Study**

An ablation study was conducted to evaluate the effect of each hyperparameter (max\_len, epoch, and threshold) on model performance. Results indicate that tuning sequence length and epoch count significantly impacts the model’s ability to capture conversational nuances, while threshold adjustments improve classification confidence for uncertain predictions. The optimal values identified through this process contributed to a more nuanced and reliable classification model.

1. **Conclusion**

This study presents a novel approach for classifying threatening conversations in Korean, addressing the specific challenges posed by language morphology and context sensitivity. Leveraging the KoBERT model and synthetic general conversation data, we successfully developed a classification model capable of distinguishing between four distinct threat types and a general conversation class. Our experiments demonstrate that synthetic data generation plays a crucial role in mitigating data imbalance and improving classification accuracy for non-threatening dialogues.

Our findings indicate that careful hyperparameter tuning and the inclusion of general conversation data significantly enhance the model's performance, providing a viable framework for practical applications in online safety, content moderation, and conversational AI. Future research should continue to refine data augmentation techniques and explore multi-label classification models to further advance the field of Korean-language threat detection.