Fair Political Debate Platform Using AI

Team A

Capstone Design Project

Abstract. This document is an online discussion platform development proposal. It is trying to develop a platform that enables smooth and fair discussion by implementing moderator AI. Moderator AI includes two functions: filtering offensive expressions and summarizing the text of debaters.

Keywords: moderator AI · Text Classification · Summarization.

1 Introduction

The candidates' conversations in this TV debate were far from elegant. The general atmosphere of discussion has been clouded by rough exchanges of negative things that have nothing to do with economic policy. If the debaters express such offensive expressions, the moderator must intervene and sanction them, but it is not easy. Because the moderator is a person. We are going to use the deep learning model to create a discussion platform that enables smooth and fair communication.

In our discussion platform, there is basically a moderator AI using deep learning, and it plays two roles and helps smooth and fair discussion. In addition, by introducing the national concept, people have a sense of solidarity and motivate them to participate in discussions. The first function of moderator AI is to filter out offensive expression. In the case of Kakao Talk open chat rooms, there are room robots. However, this bot only provides a simple function of automatically responding to messages according to the set items. Sentences that offend the other person, not severe swear words, can only be determined by understanding the context of the conversation. The second function summarizes the statements of both sides. In judging the outcome of the debate, it becomes easier to judge if someone summarizes the opinions of both sides without prejudice. In discussion, the moderator performs this function, but in our discussion platform, the moderator AI performs this function.

2 Motivation and Objective

2.1 Motivation

Before explaining the motive, I will mention two facts. The first is the importance of fundamental discussions. Unlike the conventional method of cramming,

discussion fosters the ability to think creatively and logically express one's opinions. Rather than direct teaching, it is more of a process of thinking and solving by oneself. Humans who live talking all their lives communicate their thoughts, achieve their goals, and solve problems through dialogue.

The second is the acceleration of digital transformation caused by COVID-19. COVID-19 has changed the it trend non-face-to-face, and this phenomenon is bound to continue even after COVID-19. Classes and lectures are now serviced online through platforms such as zoom, and it has become an era of untact where office workers also work from home. Naturally, the era has come when activities such as discussions are carried out online.

Ironically, however, there is no suitable online discussion platform. Therefore, we intend to create a discussion platform that provides convenience to proceed fairly and smoothly while conducting discussions non-face-to-face according to this untact situation.

2.2 Objective

There is no human moderator on the discussion platform we want to create. Therefore, moderator AI should take over the role of moderator. Our ultimate goal is to develop a moderator AI that enables discussion to proceed smoothly even without a human moderator. There are two intermediate goals for the development of moderator AI.

The first intermediate goal is to develop a model that effectively filters offensive expressions that can only be known by considering context as well as severe swear words. Discussions are offline-based and face-to-face conversations. The platform we want to create is discussed through non-face-to-face chat while being anonymous with each other. Since anonymity is guaranteed, there is a high possibility that various types of offensive expressions will be frequently used. Therefore, we intend to filter the offensive expression based on the Bert model that can understand the context in both directions.

The second intermediate goal is to develop a model that summarizes the statements of both sides without omission of important content and without bias. Since this is also the role of the moderator in traditional discussions, it is essential for the moderator AI to do so. Improving the performance of the summary model is the most important goal because omitting and summarizing important contents can hinder the smooth progress of discussion.

3 Background and Related Work

3.1 Background

Looking at the existing discussion, the moderator interferes subjectively for a fair discussion by blocking the attitudes of excessive commentators. However,

human intervention is not always fair and neutrality is not guaranteed. Therefore, people think of a fair and neutral debate as a good debate and want to see it. It's a very difficult topic for a person to maintain fairness in discussion, but what if it's a computer with no interest? The model created through machine learning cannot be said to be the answer, but at least the fact that it treats both commentators the same and is not entangled in interests will certainly serve as an advantage over people. For this reason, we intend to design a discussion platform so that fair discussion can proceed with artificial intelligence.

3.2 Realated Work

Bidirectional Encoder [1] BERT is at its core a transformer language model with a variable number of encoder layers and self-attention heads. We are going to use the model that pre-train the Korean dataset using this BERT for learning. Since BERT has a high understanding of the overall sentence, both the model for classifying offensive expressions and the conversation summary model are models to be used. Among them, we will use a distillation method that transfers knowledge from BERT with a large model size to a small model.

Knowledge distillation [3] Knowledge Distillation is the first concept to emerge from the NIPS 2014 workshop paper Distilling the Knowledge in a Neural Network. Deep learning models are generally wide and deep, so if there are many parameters and more computations, feature extensions will be better, and accordingly, performance such as Classification and Object detection, which are the purpose of the model, will also be improved. However, deep learning has gone beyond just saying, "Models with good purpose performance are good models." If a smaller model can deliver as much performance as a larger model, it can be said to be more efficient in terms of computing resources (GPU and CPU), energy (battery, etc.), and memory. For example, if you want to use an application that utilizes deep learning on your phone, but if you want to use a model that requires a few GB of memory, you have to connect to an online cloud server and use resources such as GPUs. Thus, Knowledge distillation aims to increase the performance of small networks by communicating knowledge of large networks to small networks in the learning process so that even small networks can perform similarly to large networks.

Text Classification and Text summarization The Text Classification model is a model that classifies categories of documents using deep learning. Since each text has a label, it is an easy task to calculate and learn accuracy. In particular, since we classify non-anisotropic expressions, we will use comments left by people with different thoughts after seeing the same context as a dataset. This is similar to sentimental analysis rather than a general classification task. Since the understanding of sentences takes precedence over sentence generation, we will use the Bert-series model. Text summarization is a task that summarizes documents, and although there is no correct answer to the document summary,

it uses the reference summary to learn the standard and sample text of the summary. There are two representative Korean summary models that have been pre-trained about this. First, BART (Bidirectional and Auto-Regressive Transformers) learns in the form of an autoencoder that adds noise to some of the input text and restores it to the original text. Second, BertSum[4] is a structure in which inter-sentence Transformer 2-layers are placed on the BERT.

4 Problem Statement Proposed Solution

The most important points of discussion are fairness and neutrality. In other words, it is very important for the moderator to manage the discussion while looking at the topic without any prejudice or bias. Since the current discussion depends on the moderator's method and progress, it can cause unfairness depending on the moderator's competence. There are various ways of discussion, but the moderator's intervention inevitably occurs. Therefore, we intend to create an artificial intelligence moderator that is not bound by any interest that can replace the role of such a moderator. The data put into the model that plays the role of the moderator will be taught by collecting articles with fairness and neutrality as much as possible. Since the criteria for labeling data are based on principles and standards, it is revealed in advance that there will be no interest or bias.

5 Planning in Detail

5.1 Subtasks

Our project is largely divided into four parts of development. It is an offensive expression filtering model, a statement text summarization model, a front-end, and a back-end.

In the part of developing an offensive expression filtering model, the characteristics of the corresponding task must be identified first. Since it is not just a curse or not, we plan to examine the research trends of sentimental classification with the most similar characteristics and find or label a Korean dataset accordingly. In the offensive expression, there is a direct word expression and contextual offense, and since the model must understand the meaning of words by grasping the relationship between words in the sentence, two-way language models xlnet and bert are suitable. Since it is necessary to serve a model, it is planned to analyze the performance using not only kobert and kluebert but also the distilling version of large model, select an appropriate model, and fine-tuning.

In the part of developing the statement text summarization model, an existing text summarization Korean dataset and a pre-trained model will be used. Models currently available include Ala-kung-dala-kung, SKT-AI kobert, and uoneway-kobertsum. We plan to use each model to analyze the performance, and select and serve the model that best suits our task.

In the front-end development section, we plans to implement the overall UI and chat servers of iOS apps and introduce national concepts. The reason for introducing the concept of a country is not just a platform that ends with a single discussion, but a foundation to become a platform that can continue to discuss with people. Through the implementation of this system, people in the same country can engage in discussions with bonds, and the responsibilities assigned to each person in the discussion are catalysts that drive the discussion in a more democratic direction.

In the back-end development part, the main purpose is to serve the two developed models. The text message is received at the front end, the slanderous expression filtering model infers the received message in real time and delivers the result to the front end, and the text summarization model makes a summary inference based on text sentences accumulated on the server and sends the result to the front end. It plans to use torchserve to serve the model to the server and implement the part that processes the client's http requests.

It has not yet been decided whether to put a slanderous expression filtering model on the server or embed it into the app itself. This is because if the response is slow in terms of speed as it has to be inferred in real time, embedding a lightweight model into the app itself can be one way. Therefore, the slanderous expression masking model will be implemented in several ways, and after comparing performance, we will decide which method to use.

The work distribution of each team member is as follows.
JinWoo: Text summarization model development
ChaeYoon: Offensive expressions filtering model development
CheongSoo: Front-end development
Seok: Back-end development

Table 1. Detailed Project Schedule.

	Week5	Week6	Week7	Week8	Week9	Week10	Week11	Week12
JinWoo	Data search	summarization model develop			Model optimization			
ChaeYoon	Model performance test, model selection				Model optimization			
ChungSoo	Develop UI, chat function			Develop functions using response from servers				
Seok	Example model serving			Develop torchserve handler			Model serving	

References

1. Devlin, Jacob, et al. Article title."Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
2. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
3. Jiao, Xiaoqi, et al. "Tinybert: Distilling bert for natural language understanding." arXiv preprint arXiv:1909.10351 (2019).
4. Liu, Yang. "Fine-tune BERT for extractive summarization." arXiv preprint arXiv:1903.10318 (2019).
5. Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).