Sentiment Analysis About Hybrid Learning by BERT(Bidirectional Encoder Representations from Transformer)

Seho Jeong

1. Introduction

Since the outbreak of COVID-19 in 2019, the structure and delivery of higher education have changed significantly. Online lectures—either pre-recorded or conducted live via platforms such as Zoom—have become widespread. Assignments, quizzes, and exams have also shifted to online formats in many institutions.

As societies began to stabilize post-COVID, many universities adopted hybrid-learning models, where in-person classes are combined with online elements. This form of education, often referred to as hybrid lectures or hybrid learning, reflects a broader transition in educational infrastructure and pedagogy.

In response to these changes, this study aims to explore how students, especially those from various universities, perceive this shift toward hybrid education. Using sentiment analysis techniques based on Hugging Face's BERT model, I analyze social media data to understand students' attitudes, opinions, and emotional reactions to hybrid learning.

2. Data

2.1. Twitter Data

I collected Twitter data over a six-month period starting in 2020 using the Twitter API. Tweets were filtered using specific keywords related to the topic, such as "hybrid", "hybrid learning", and "online learning". After collection, we removed advertisements and tweets unrelated to the context of education. Stopwords were also removed as part of preprocessing.

For model training, each tweet was manually labeled into one of three sentiment categories: positive, negative, or neutral.

2.2. Reddit Data

Reddit data was also collected over the same six-month period in 2020 using the Reddit API. We used the same keyword set as for Twitter and focused on posts from student communities of 20 universities (e.g., University of Washington, UC San Diego, UC Berkeley, etc.).

Similar to the Twitter dataset, irrelevant posts and advertisements were removed, and stopwords were filtered out during preprocessing.

I manually labeled each Reddit post as positive, negative, or neutral to be used for training the sentiment classification model.

2.3. Data Augmentation

After collecting data over a six-month period, we filtered out irrelevant advertisements and off-topic posts, resulting in approximately 4,000 usable tweets and Reddit posts. However, the overall dataset was relatively small, and a significant imbalance was observed: the number of neutral samples greatly exceeded that of positive and negative samples.

To address both the class imbalance and the limited size of the training data, we performed data augmentation focused on the positive and negative classes. We generated between 30,000 and 40,000 augmented samples using a synonym replacement approach. Specifically, a predefined number of words in each sentence were replaced with appropriate synonyms to create variant sentences while preserving the original meaning.

We also utilized Weights & Biases to optimize augmentation parameters, including the number of word substitutions per sentence and the total number of augmented sentences. These tuning efforts helped maximize model performance while minimizing semantic drift in the generated data.

3. Model / Experiment

I used a BERT-based sequence classification model for the sentiment analysis task. To adapt the model to the specific language and context of hybrid learning discussions, I first performed domain-specific pretraining on our collected dataset (Twitter + Reddit) based on the pre-trained model (bert-based-cased). After this, I fine-tuned the model for three-way sentiment classification: **positive**, **neutral**, and **negative**.

During training, I used the **AdamW optimizer**, and to improve the model's performance, we conducted **hyperparameter tuning** focused on maximizing the F1-score. I applied a range search method by testing multiple combinations of upper and lower bounds for key hyperparameters.

The following values were tuned:

- Learning rate : 1e-5 ~ 4e-5
- Batch : 4, 6, 8, 12
- Epochs : 6, 7, 8, 9, 10
- Data_augmented_set : 30000 ~ 40000
- Word_change : 5, 6, 7, 8, 9, 10

We selected the best configuration based on F1-score performance on a validation set. The final model achieved an **F1-score of 0.86** on the held-out test set which is shown in Figure 1.

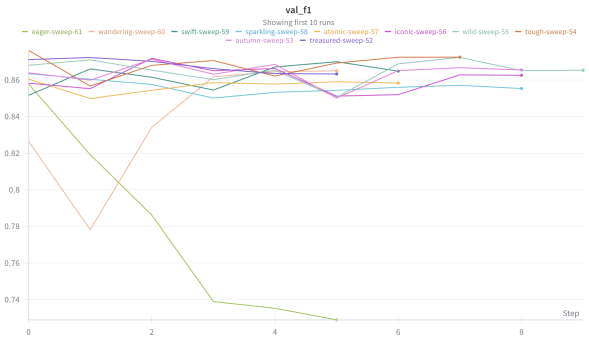

Figure 1: Graph of f1-scores for validation set

4. Limitation

One major limitation was the lack of a clearly defined term. Although this paper and some universities use the expression "Hybrid learning," there is no universally agreed-upon terminology for this concept, and various terms were used inconsistently across contexts. Furthermore, even when the same expressions such as "Hybrid learning" or "Hybrid class" were used, they often referred to different concepts depending on context. Due to this ambiguity, a considerable portion of the collected data turned out to be irrelevant, and there was a high possibility that the model would process and analyze unrelated content. As a result, the model could not be reliably applied for real-world analysis.

In future research, this issue should be addressed by identifying and selecting keywords that are both semantically consistent and unambiguous, in order to improve the quality of data collection.

Additionally, due to the lack of advanced linguistic preprocessing techniques, I focused on hyperparameter tuning to improve model performance. However, this approach requires strictly controlled training conditions and high-quality data, and even under such conditions, the improvement in F1-score was limited to around 0.01–0.03. Therefore, future work should place more emphasis on natural language preprocessing and semantic filtering, which are likely to contribute more substantially to model performance than tuning alone