Individual Report of NLP Final Project

Yixi Liang

Department of Data Science, The George Washington University

DATS_6312_10: NLP for Data Science

Amir Jafari

Dec 12, 2022

Table of contents

Introduction

In this final project, we use the dataset contains argumentative essays written by U.S students in grades 6-12. The essays were annotated by expert raters for elements commonly found in argumentative writing. And we use this dataset to classify the different classes of the sentence.

Description of individual work

In this final project I generate summary column and train_balanced.csv, test_balanced.csv. Train them on different pretrained model and different head, and try to improve model performance.

Data preprocessing

In this section, I try to use transformer pipeline ‘Summarization’ to summarize text and use the result them to improve the performance of the model. The reason why I try to use the summarization is that we think summary can extract the core meaning and structure of the sentence, and I guess that might work for text classification. For instance, evidence is one of the seven classes in this dataset, and after using summarization the 200-length paragraph left only few sentences of 30-length may help model to classify them.

I use this code to initialize the summarizer ‘summarizer = pipeline("summarization", model="t5-base", tokenizer="t5-base")’ , then set the parameter like this ‘summarizer(text, min_length=5, max_length=30)[0]['summary_text']’, and run them on the ‘discourse_text’ column.

Figure 1

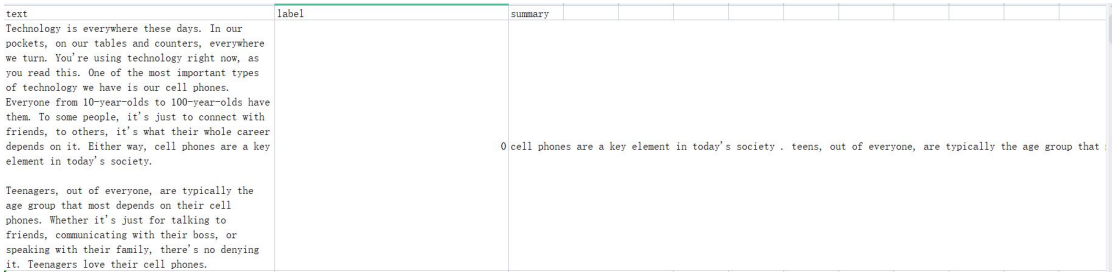
Script of generating the summary.

```
summarizer = pipeline("summarization",  model="t5-base", tokenizer="t5-base")
res = []
for i in tqdm(range(len(df_train))):
    text = df_train.iloc[i]['text']
    res.append(summarizer(text, min_length=5, max_length=30)[0]['summary_text'])

df_train['summary'] = res
```

Figure 2

Example of summary.



text	label	summary
Technology is everywhere these days. In our pockets, on our tables and counters, everywhere we turn. You're using technology right now, as you read this. One of the most important types of technology we have is our cell phones. Everyone from 10-year-olds to 100-year-olds have them. To some people, it's just to connect with friends, to others, it's what their whole career depends on it. Either way, cell phones are a key element in today's society. Teenagers, out of everyone, are typically the age group that most depends on their cell phones. Whether it's just for talking to friends, communicating with their boss, or speaking with their family, there's no denying it. Teenagers love their cell phones.	0	cell phones are a key element in today's society . teens, out of everyone, are typically the age group that

As we can see, summarization did work, it can reduce the length of the sentence, but it also has some problems of destructed the structure of whole paragraph and sometime the result of summarization is meaningless, since summarizer are not work very well, Last but not least, it cost plenty of time.

Training

After generated them, I try several combinations, such as only put the summary to train or add text and summary to train together. But I did not see much different between them. Finally, I chose summary and text to train, and test on text only. And I mostly try them on BERT + LSTM. Here is the table.

Results

Table 1

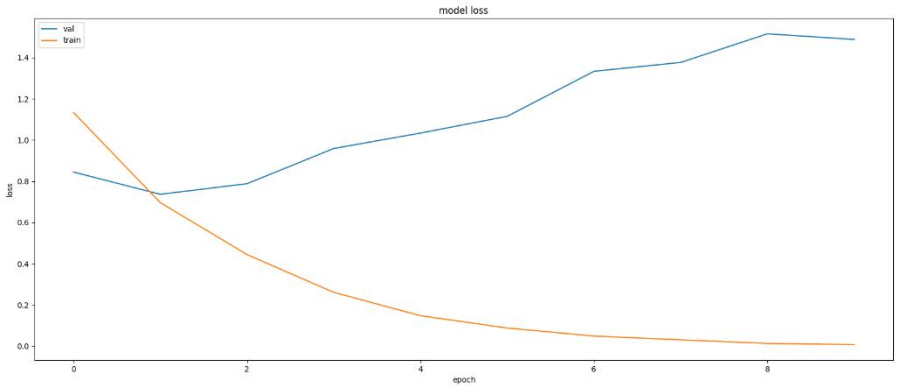
The result of training.

Pretrained	Model	Accuracy	F1 macro
bert-base-cased	BERT+LSTM	0.7820838627700127	0.7575876668982763
bert-base-uncased	BERT+LSTM	0.7916137229987293	0.7685482955944065
bert-base-uncased	BERT+LSTM 10000 datasets	0.7465057179161372	0.7174047084744588

And I also try some pretrained like 'bert-large-uncased-whole-word-masking', 'bert-large-cased', 'bert-large-uncased', but they did not have good performance.

Figure 3

Loss of train and validation.



And from Figure 3 we can find validation loss is increasing. We may not need to train so much epoch, since the model is overfitting.

Summary

In this final project, I try to use summarization to improve the result of the model. But there are still some problems remain. For instance, I do not know whether add original text and summary together is right or not to train the model. In addition, I the

summary of the text are sometime not good enough. Moreover, it costs plenty of time to use summary transformer. The percentage of the code is nearly 30%.

References

Feedback prize - evaluating student writing. Kaggle. (n.d.). Retrieved December 11, 2022, from https://www.kaggle.com/competitions/feedback-prize-2021/data

Sangani, R. (2022, January 26). Adding custom layers on top of a hugging face model. Medium. Retrieved December 11, 2022, from https://towardsdatascience.com/adding-custom-layers-on-top-of-a-hugging-face-model-f1ccdfc257bd