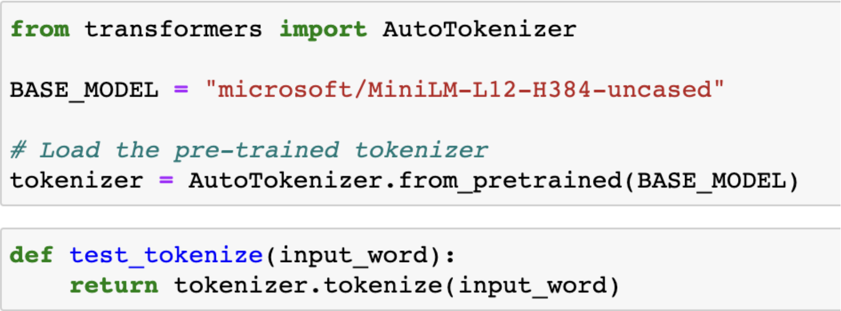
SI 630: Homework 4 – Prompt-based NLP

Yu Yan (kuminia)

Kaggle username: kkkyyy1

April 10, 2023

Problem 1



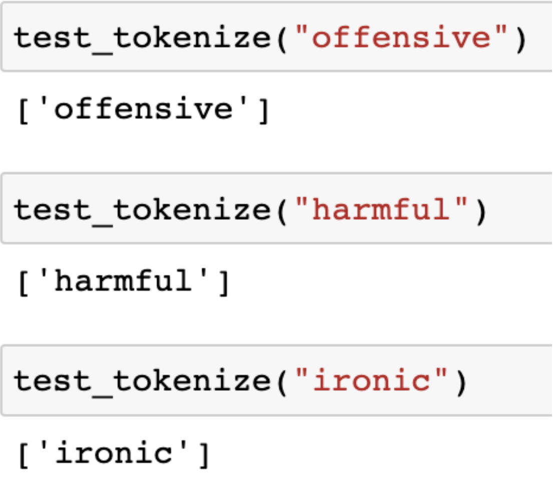
```python
from transformers import AutoTokenizer

BASE_MODEL = "microsoft/MiniLM-L12-H384-uncased"

# Load the pre-trained tokenizer
tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL)
```

```python
def test_tokenize(input_word):
    return tokenizer.tokenize(input_word)
```

I use this to check the single-token words.



```
test_tokenize("offensive")
```

```
['offensive']
```

```
test_tokenize("harmful")
```

```
['harmful']
```

```
test_tokenize("ironic")
```

```
['ironic']
```

Problem 2: Few-shot prompts

1. {"placeholder":"text_a"}. Overall, the sentence shows {"mask"} meaning.

toxic: ["offensive", "harmful","ironic"]

non-toxic: ["respectful", "objective", "calm"]

2. {"placeholder":"text_a"}. The sentence contains {"mask"} content.

toxic: ["offensive", "harmful", "ironic"]

non-toxic: ["respectful", "objective", "righteous"]

3. {"placeholder":"text_a"}. The sentence contents {"mask"} others.

toxic: ["dismiss", "blame", "discourage"]

non-toxic: ["respect", "praise"]

4. {"placeholder":"text_a"}. If I hear the words above I will be {"mask"}.

toxic: ["angry", "depressed", "discouraged"]

non-toxic: ["peaceful", "calm"])

5. {"placeholder":"text_a"}. The person who said the sentence above is {"mask"}.

toxic: ["angry", "ironic", "offensive"]

non-toxic: ["peaceful", "calm", "kind"])

6. {"placeholder":"text_a"}. The most likely context for the above conversation to occur is between {"mask"}.

toxic: ["enemies", "rivals"]

non-toxic: ["friends", "lovers"]

7. {"placeholder":"text_a"}. The above statement is most likely to occur in a movie belonging to the {"mask"} genre.

toxic: ["crime", "war"]

non-toxic: ["documentary", "children"]

8. {"placeholder":"text_a"}. Do you think a gentle and easygoing person would say the above statement? Answer: {"mask"}.

toxic: ["no", "unlikely"]

non-toxic: ["yes", "absolutely"]

9. {"placeholder":"text_a"}. Do you think a angry and offensive person would say the above statement? Answer: {"mask"}.

toxic: ["yes", "absolutely"]

non-toxic: ["no", "unlikely"]

10. {"placeholder":"text_a"}. Do you think the sentence above is a toxic sentence? Answer: {"mask"}.

toxic: ["yes", "absolutely"]

non-toxic: ["no", "unlikely"]

Problem 3: Zero-shot prompts

1. Please determine whether the sentence in <> is a offensive or hurtful sentence, If so, return 1; if not, return 0

F1: 0.178

2. Please determine whether the sentence in <> dismisses or invalidates someone's feelings or experiences, If so, return 1; if not, return 0

F1: 0.197

3. Please determine whether the sentence in <> is just a plain description without strong emotional color. If so, return 0; if not, return 1

F1: 0

4. Please determine whether the sentence in <> made by a person with criminal tendencies, If so, return 1; if not, return 0

F1: 0.182

5. Please determine whether the sentence in <> would be said by a genteel person., If so, return 0; if not, return 1

F1: 0

6. Please determine whether the sentence in <> will blame someone and make them unhappy. If so, return 1; if not, return 0

F1: 0.161

7. Please determine whether people would become angry or discouraged after hearing the sentence in <>. If so, return 1; if not, return 0

F1: 0.172

8. Please determine whether the content in <> is more likely to be said by a person with a bad temper. If so, return 1; if not, return 0

F1: 0.18

9. Please determine whether the sentence in <> is a toxic sentence, If so, return 1; if not, return 0

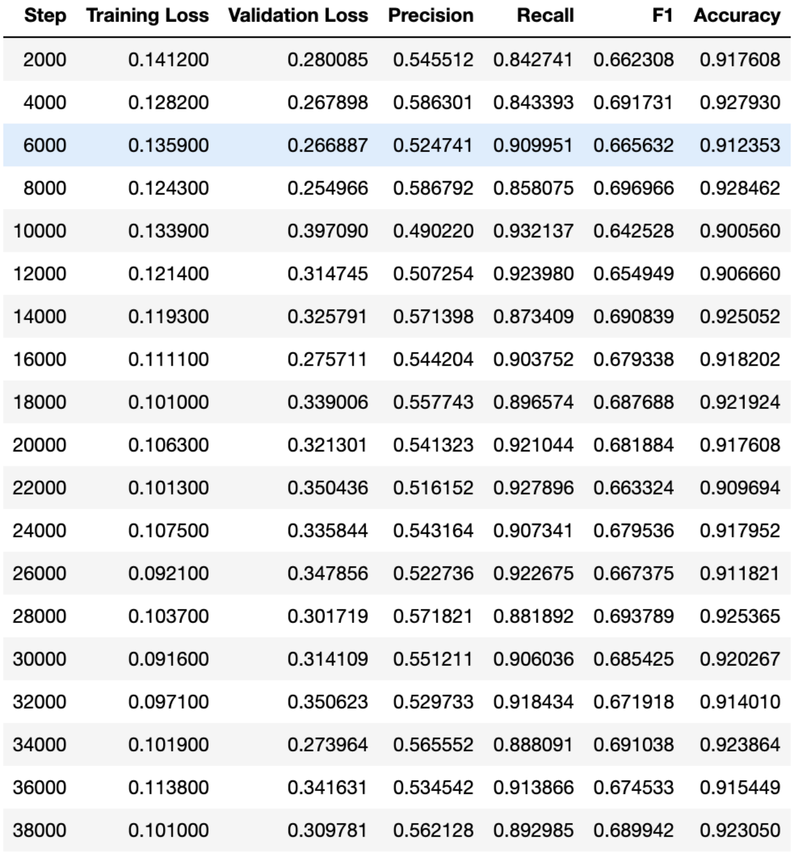
F1: 0.203

10. Please determine whether the sentence in <> would cause tension in the relationship between people., If so, return 1; if not, return 0

F1: 0.154

Problem 4

Here is the result for the regular classifier



Step	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
2000	0.141200	0.280085	0.545512	0.842741	0.662308	0.917608
4000	0.128200	0.267898	0.586301	0.843393	0.691731	0.927930
6000	0.135900	0.266887	0.524741	0.909951	0.665632	0.912353
8000	0.124300	0.254966	0.586792	0.858075	0.696966	0.928462
10000	0.133900	0.397090	0.490220	0.932137	0.642528	0.900560
12000	0.121400	0.314745	0.507254	0.923980	0.654949	0.906660
14000	0.119300	0.325791	0.571398	0.873409	0.690839	0.925052
16000	0.111100	0.275711	0.544204	0.903752	0.679338	0.918202
18000	0.101000	0.339006	0.557743	0.896574	0.687688	0.921924
20000	0.106300	0.321301	0.541323	0.921044	0.681884	0.917608
22000	0.101300	0.350436	0.516152	0.927896	0.663324	0.909694
24000	0.107500	0.335844	0.543164	0.907341	0.679536	0.917952
26000	0.092100	0.347856	0.522736	0.922675	0.667375	0.911821
28000	0.103700	0.301719	0.571821	0.881892	0.693789	0.925365
30000	0.091600	0.314109	0.551211	0.906036	0.685425	0.920267
32000	0.097100	0.350623	0.529733	0.918434	0.671918	0.914010
34000	0.101900	0.273964	0.565552	0.888091	0.691038	0.923864
36000	0.113800	0.341631	0.534542	0.913866	0.674533	0.915449
38000	0.101000	0.309781	0.562128	0.892985	0.689942	0.923050

Figure 1: result for the regular classifier

Problem 6

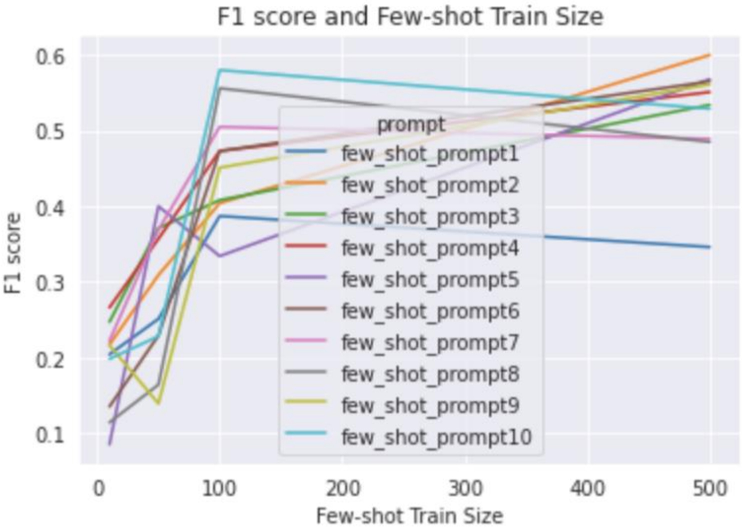


Figure 2: F1 score for few shot learning

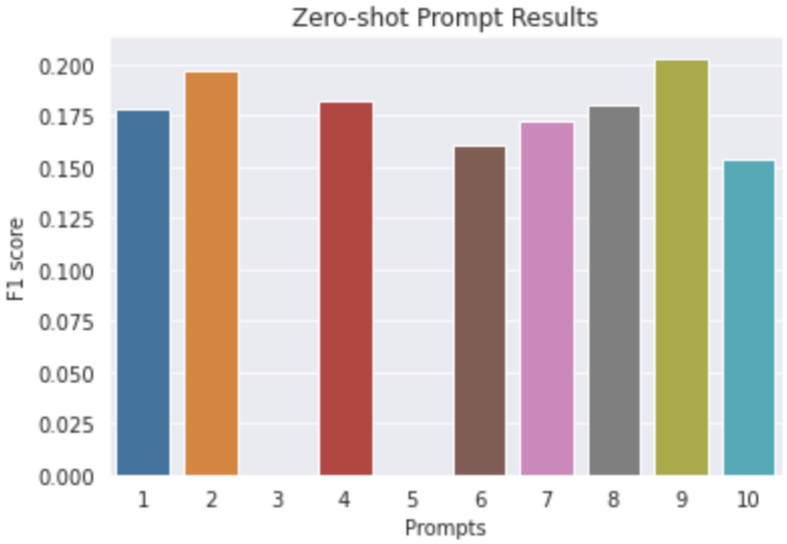


Figure 3: F1 score for zero shoot learning

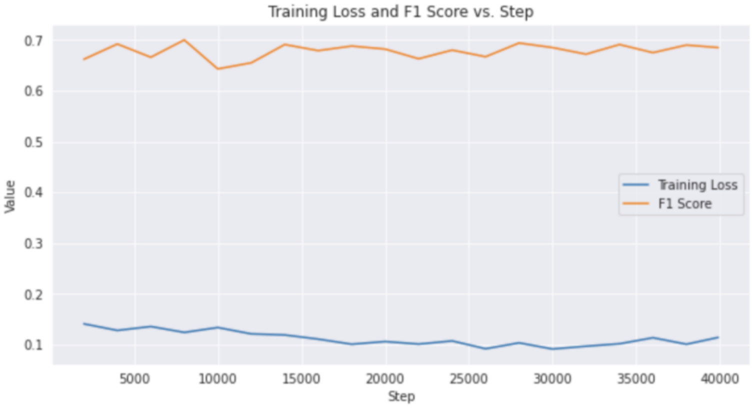


Figure 4: F1 score and training loss for regular classification

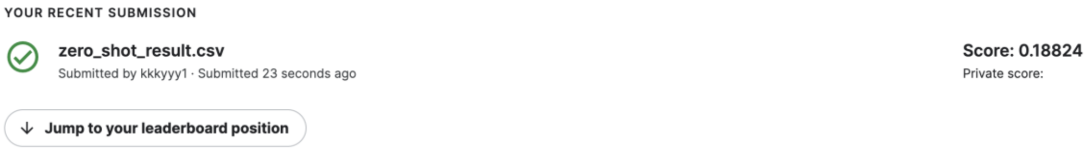
Write your guess on how many instances you think you need to train a prompt-based learning model that will reach the performance of a MiniLM model trained on all the data:

The F1 score for few shot learning usually increases as the training size increase from 10 to 500, we can see from Figure 2 that with 10 training size, the F1 score is about 0.2, and with 500 training size the F1 score will reach about 0.55. Moreover, the upward trend is getting slower and slower. And The F1 score for the MiniLM model trained on all the data is about 0.69.

Therefore, I guess the F1 score will reach about 0.69 on 1000 to 1500 training instances.

Problem 7

I use the prompt that I reach the best F1 score with the development dataset to predict the test label and my final result on Kaggle is 0.18824.



YOUR RECENT SUBMISSION

zero_shot_result.csv

Submitted by kkkyyy1 · Submitted 23 seconds ago

Score: 0.18824

Private score: