Effects of prompting on classification results

Term paper
Natural Language Processing - Computing Meaning
July 24th, 2025
Tanay Kewalramani, Matrikel 414724

Given problem statement:

Still in line with Pichler et al. (2025), we are interested in well-structured experiments to find out about the effects of the wording of prompts on classification results obtained from LLMs. Their experiments were run with prompts that were paraphrases of each other; it will be interesting to replicate such experiments on datasets you may choose. To control the paraphrases you may come up with, we suggest that you compare them by means of the BERT-Score (Zhang 2019). Additionally, you may test the LLM's sensibility to tipping ("I'll give you a 500 Dollar tip for a very good answer") or to threats ("Your career will be seriously affected by unsatisfactory answers").

We suggest that you opt for an interesting mixture of some of the following:

1. Get a text dataset (e.g. from kaggle) with gold classification data; - Split into training, (validation) and test subsets; other, e.g. by changing the order of elements, by using other words, by using different syntactic constructions; possibly also by using another langugage for your prompt than English – any good idea is welcome here;

2. Design a series of prompts that are paraphrases of each. Control the 'semantic' closeness of your prompt variants with the BERT score (); - Train e.g., a RoBERTa classifier or MISTRAL/MIXTRAL and/or FastText, to get a tool baseline result; - Run tests with the different prompts on different LLMs (e.g., Llamas, Qwen versions, etc.); - Maybe compare with online versions of (commercial) LLMs (GPTs, DeepSeek versions, etc.).

Reference: Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi: BERTScore: Evaluating Text Generation with BERT, https://doi.org/10.48550/arXiv.1904.09675

Literature Survey

1. BERTScore: Evaluating Text Generation with BERT (given reference)

This paper essentially gives us a robust method to compare the similarity between two pieces of text. It improves upon the earlier methods that rely on sentence structures and similar words by also including context. BERT-based embeddings help accomplish this.
It works by comparing all tokens of one piece of text with all tokens of the others, and selects the best candidate token for a reference token and vice versa thingy and there's IDF as well and stuff.

Pipeline

1. **Goal**: To understand how the wording of prompts (measured for similarity via BERTScore) affects classification outcomes in LLMs, and how incentive/threat cues further influence results.

2. **Dataset choice: SMS spam classification. 5572 datapoints.**
    a. Spam: 4825, Not spam: 747.
    b. Train dataset size: 3902
    c. Validation dataset size: 834
    d. Test dataset size: 836

3. **Establishing baseline performance on logistical regression and RoBERTa:**

Validation Set Performance (Logistic Regression)

```
              precision    recall  f1-score   support

         ham       0.98      0.98      0.98       724
        spam       0.90      0.88      0.89       112

    accuracy                           0.97       836
   macro avg       0.94      0.93      0.93       836
weighted avg       0.97      0.97      0.97       836
```

Test Set Performance (Logistic Regression)

```
              precision    recall  f1-score   support
```

```
         ham       0.97      0.98      0.98       724
        spam       0.89      0.83      0.86       112

    accuracy                           0.96       836
   macro avg       0.93      0.91      0.92       836
weighted avg       0.96      0.96      0.96       836
```

Classification Report (RoBERTa):

```
              precision    recall  f1-score   support

         ham       0.98      0.99      0.99       724
        spam       0.95      0.90      0.93       112

    accuracy                           0.98       836
   macro avg       0.97      0.95      0.96       836
weighted avg       0.98      0.98      0.98       836
```