CSC 495/791 Natural Language Processing

Project Assignment 2

Vikram Pande

Introduction

Task: News Classification

News classification is of importance in the field of information retrieval and media analysis. It plays a pivotal role in efficiently organizing and categorizing the vast volume of news articles available online, facilitating easier access to relevant information for both individuals and automated systems. By accurately classifying news content into categories such as politics, sports, health, and more, it enables users to filter and consume information that aligns with their interests and preferences, thus enhancing their overall information consumption experience. Additionally, news classification has significant implications in the realm of media monitoring, sentiment analysis, and trend prediction, making it a critical research area with broad applications in today's data-driven society.

Dataset: AG News (AG's News Corpus)

AG is a collection of more than 1 million news articles. These news articles have been gathered from more than 2000 news sources. The dataset is provided by the academic community for research purposes in data mining, information retrieval, data compression and data streaming. [1]

The dataset is constructed in the following manner: It contains 4 largest classes from the original corpus. Each class contains 30,000 training samples and 1900 testing samples and the total number of training samples is 120,000 and testing samples is 7,600. The classes are divided into categories: world, sports, business, and science/technology. [1]

The evaluation metric here is the testing errors, which are essentially a straightforward accuracy comparison represented as a percentage. Additional metrics include precision, recall, and F1-measure.

Related work

Text classification is a classic task in Natural Language Processing. There are many sequential model-based approaches are already employed for text classification. In the work [2] Xian Zhan et al, implemented character-level convolution network for text classification using several datasets and various models. Zhen wang et al [3] used prompt engineering-based approach to classify news articles of AG dataset by providing different patterns in prompt and comparing different models. Schick and Schütze [4] propose to reformulate input examples into cloze-style phrases and show superiority in few-shot text classification and natural language inference.

State-of-the-art performance on AG's dataset for news classification is the approach of ASCM+SL on 4 training sizes of 10,50,100 and 1000. ASCM that transforms token embeddings to a semantic-clustered embedding space and categorizes all answer tokens embeddings in that space. Besides, to exploit massive unlabeled data, we propose a semi-supervised method called stair learning (SL) which transfers knowledge orderly and further increases the performance. [4]

Prompting Strategy

1. **Baseline**

   Large Language Models (LLMs) are typically trained on extensive datasets, allowing them to work effectively with direct inputs and outputs. When prompted directly, an LLM can often generate a reasonably accurate response in a single attempt. This approach is referred to as "zero-shot prompting" and is considered a baseline technique in prompt engineering. In simple terms, the model is presented with a straightforward question without any surrounding context or background information, and it is expected to provide a corresponding answer. The extended version would be "few-shot-prompting".

2. **Few-shot prompting**

   In few-shot prompting, only a small number of examples/shots are also provided in prompt. This helps model in decision making for new data and useful when annotated data is limited to guide its understanding of a particular task or generate desired responses.

3. **Chain-of-thought (CoT)**

   Chain-of-thought (CoT) prompting enables complex reasoning capabilities through intermediate reasoning steps. To maintain a coherent flow of conversation. Instead of asking isolated or individual questions, this approach involves generating prompts that build upon the previous responses or questions, forming a logical and connected chain of thought in the conversation. [6]

4. **Zero Shot Chain-of-thought (CoT)** *[used for error analysis]*

   A novel concept, the zero-shot CoT (Kojima et al. 2022)[7], introduces a distinct approach by incorporating the phrase "Let's think step by step" into the original prompt. This strategy aims to enhance model performance. This approach empowers the model with the capacity to engage in systematic and logical reasoning. By introducing the directive to "think step by step," it encourages the model to break down complex problems into manageable components, facilitating more coherent and reasoned responses.

Results and Analysis

Experiments with different prompting strategies were done and stored in a file.
The main implementation of three different strategies was done on the News AG dataset and the results and analysis are: ***Accuracy Table***

Samples	Baseline	Few shot	COT
T = 100	67	69	65
T = 500	69.2	68.6	64.8
T = 100 (Error Analysis)	ZSCOT 83		

The result analysis was done by taking 100 and 500 samples because of API limitation. The results for 100 samples for the three approaches are shown below and it was seen that few-shot was better in comparison with others. For 500 samples, baseline model yielded better results.

Error Analysis

For error analysis, attempted the ZSCOT method. In the zero-shot Chain of Thoughts Method, an additional element can be incorporated into the prompt, such as "Let's think step by step." This addition prompts the model to carefully consider the given input, providing insights into the reasoning behind the model's predictions.

When utilizing prompt engineering with ZSCOT to analyse 100 samples, the model achieved an accuracy of 83%, which was the highest among all.

Based on the observations,

1. The model was mainly confused between classes like Business and Science/Technology. The potential reason might be interdisciplinary nature; both business and technology are becoming interconnected, technology advancements drive business decisions. Another reason might be shared vocabulary; For instance, terms like "analytics," "strategy," "innovation," and "disruption" are commonly used in both business and technology contexts.
2. Another observation indicates that when using zscot, the model was able to think step by step. If a description did not align with any of the four classes, the model simply refrained from classifying that particular sentence. In such scenarios, a manual assignment of -1 is made to the classes.
3. There was also ambiguity between classes such as World and Business may be due to contextual overlap. Both fields frequently address topics such as global affairs, economics leading to shared subject matter. Vocabulary used in these fields can also be common.
4. The model was not able to classify a sentence belonging to Sports class even if the vocabulary was related to sports class. This might be because of limited range of sport-specific context or some ambiguity.

Error analysis is shown in excel file: - `error_analysis.xlsx`

Challenges faced:

1. Limited requests per minute in OpenAI API (3 rpm), increased requests per minute by purchasing credits.
2. API Timeout: API response timeout after RPM issue.
3. Internal server – bad gateway error.
4. Tried async call to API, and different libraries such as backoff and tenacity.
5. Solution: Tried retry logic.

Code structure:

- `prompt_play.ipynb`
  - `Prompting experiments`
- `news_classification_100.ipynb`
  - `Prompting on AG news dataset for 100 samples.`
- `news_classification_500.ipynb`
  - `Prompting on AG news dataset for 500 samples.`
- `error_analysis.ipynb`
  - `Error analysis on 100 samples.`
- `error_analysis.xlsx`
  - `Error analysis of 100 samples.(ZSCOT method)`

References

1. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
2. Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015).
3. Zhen Wang, Yating Yang, Zhou Xi, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. 2022. ASCM: An Answer Space Clustered Prompting Method without Answer Engineering. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2455–2469, Dublin, Ireland. Association for Computational Linguistics.
4. Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 255–269, Online. Association for Computational Linguistics.
5. Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, Thomas Brightwell 2023. Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification.
6. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models NeurIPS 2023.
7. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, Large Language Models are Zero-Shot Reasoners, 2023.