Report for the NLP - CS - 2024 Competition

Ayman MOUMEN Francisco GARCIA Ibrahim RAMDANE
Marouane MAAMAR Samer LAHOUD

Team ASMF submission

Abstract

This report outlines the application of several text classification strategies for the NLP-CS-2024 competition, aimed at developing an effective classifier under data scarcity. We employed a zero-shot BART MNLI model, hybrid models combining zero-shot classification with KNN and cosine similarity, and traditional models like BERT and DeBERTa enhanced by data augmentation. Our findings indicate that the DeBERTa model with data augmentation achieved the highest accuracy, providing a promising direction for future research in text classification. This work evaluates all these methodologies in a low-data scenario, providing insights into their efficacy and limitations.

1 Introduction

Text classification is a fundamental task in Natural Language Processing (NLP), crucial for tackling a wide array of applications such as sentiment analysis, spam detection, and beyond. The NLP-CS-2024 competition poses a unique challenge: to build accurate and robust text classifiers with very limited data and without the aid of third-party APIs. Tasked with categorizing 1140 sentences across twelve diverse categories—Politics, Health, Finance, Travel, Food, Education, Environment, Fashion, Science, Sports, Technology, and Entertainment—we explore a variety of methodologies to address this issue.

Our approach in the competition involves a meticulous combination of models and techniques designed to optimize performance under these constraints. Specifically, we have implemented a zero-shot BART MNLI model, hybrid strategies combining zero-shot classifications with KNN and cosine similarity, and traditional classifiers like BERT and DeBERTa augmented with additional training data. This report will delve into these strategies, detailing their implementation, effectiveness, and the insights gained through their application, with a particular focus on how the DeBERTa model enhanced with LLM data augmentation outperformed other methods.

2 Preliminary Approaches

In the first methodology, we used a pre-trained BERT model to transform training and test texts into semantic embeddings. The embeddings of the training were averaged to create centroids for each category, representing the semantic core of each group. For classification, we used cosine similarity between test text embeddings and these centroids determined the closest category. This approach resulted in a low accuracy of 0.45, indicating that while conceptually sound, the method was not effective in practice.

We tested a similar methodology, but instead of calculating centroids for each category, we directly predicted the label of new texts by identifying the closest existing text label in the training set. The result show a low performance of 0.46. This weak result warns of the need to apply different methods to deal with the problem.

3 Solution

Our methodology was designed to effectively leverage the limited data available in the NLP-CS-2024 competition. We utilized a combination of zero-shot learning, hybrid methods, and fine-tuning strategies to explore various ways of text classification under constraints.

1. **Zero-Shot Classification with BART MNLI:** Initially, we applied the BART MNLI model in a zero-shot framework, which provided a baseline accuracy of 0.76. This model was chosen for its ability to perform classification without needing labeled training data, an advantage given our dataset size limitations.

2. **Hybrid Approach for Data Augmentation:**

   - **Data Selection:** We employed the BART MNLI zero-shot model to selectively augment our dataset. Instances where a single class exhibited a confidence score exceeding 0.7 were exclusively included in the augmented dataset. This stringent criterion ensured that the appended data maintained a high confidence level, thereby enhancing the reliability of the augmented dataset.
   - **Secondary Classification Techniques:**
     - **K-Nearest Neighbors (KNN):** For sentences that did not meet the augmentation threshold, we applied the KNN algorithm, which resulted in a 0.66 accuracy. KNN was chosen for its simplicity and effectiveness in handling smaller, less complex data scenarios.
     - **Cosine Similarity:** We then employed cosine similarity as an alternative to classify the remaining non-augmented data, achieving a 0.67 accuracy. This method was used to measure the textual similarity, providing a different angle of classification based on text context.

3. **Advanced Model Implementation:**

   - **BERT and DeBERTa Zero-Shot:** Further explorations with BERT (Devlin et al., 2019) (Pietro, 2023) and DeBERTa in a zero-shot (PradipNichite) configuration yielded accuracies of 0.74 and 0.76, respectively, showing significant potential in zero-shot applications for future research directions.
   - **DeBERTa Fine-Tuning:**(Gugge) Applying fine-tuning to the DeBERTa model using the augmented data, we achieved an accuracy of 0.78.
   - **BERT Fine-Tuning:** Finally, we utilized a fine-tuned BERT model on the augmented datasets. It resulted in a 0.8 accuracy.
   - **BERT with LLM augmented data** We tried to use LLM, specifically Mistral-7b, to improve the data augmentation but it resulted in an accuracy of 0.65. Which is rather disappointing.

4 Results and Analysis

Table 1 present a summary of the accuracies achieved by each method.

Method	Accuracy
BART MNLI Zero-Shot	0.76
Hybrid (KNN)	0.66
Hybrid (Cosine Similarity)	0.67
BERT Zero-Shot	0.74
DeBERTa Fine-Tuned	0.75
DeBERTa FT with frozen layers	**0.81**
DeBERTa Zero-Shot	0.76
Deberta MisralAug	0.66

Table 1: Summary of classification accuracies by method

Analysis:

- The **hybrid models** exhibited slightly lower performance compared to the baseline, indicating possible issues with data selection or the suitability of secondary classification methods under stringent confidence thresholds.

- The **BERT Fine-Tuned** model outperformed other strategies, suggesting that extensive fine-tuning on a sufficiently augmented dataset can significantly enhance performance, especially in a scenario with limited initial data.

- The **DeBERTa Fine-Tuned** model achieved our highest score of 0.81 by freezing 6 to 8 layers, representing a significant improvement from the previous score of 0.75, which helps stabilize training, prevents overfitting emphasizing the effectiveness of fine-tuning on a well-prepared dataset. It surpassed BERT by introducing improvements in handling long-range dependencies and capturing contextual information more effectively. Specifically, DeBERTa's disentangled attention mechanism allows for better isolation of different aspects of input, reducing interference between tokens. Additionally, its enhanced relative positional encoding improves the model's ability to capture word order and position-sensitive information

This analysis demonstrates the critical importance of model selection and data preparation in achieving high accuracy in text classification tasks, particularly when data is scarce. Further exploration into hybrid models and alternative data augmentation strategies could potentially yield even better results.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sylvain Gugge. Fine-tune a pretrained model with hugging face.

Mauro Di Pietro. 2023. Bert for text classification with no model training.

PradipNichite. Youtube-tutorials/youtube_zero_shot_learning.ipynb at main · pradipnichite/youtube-tutorials.