ANLP : Sentence topic classification

Hugo PLOTTU[*] **Alexandre PETIT**[†] **Axel VOYER**[‡] **Aymeric PALARIC**[§]

Abstract

In this project, we addressed the challenge of sentence classification using a hybrid approach combining deep learning models and traditional machine learning techniques. Initially constrained by limited training data, we expanded our dataset through semi-supervised methods, employing pseudo-labeling with a pre-trained BERT classifier and creating additional training sets derived from Wikipedia articles. We implemented a variety of classification strategies, including zero-shot classification with DeBERTa, fine-tuning a BERT model with an additional classification layer, and utilizing embeddings with traditional classifiers like SVM and KNN. Furthermore, we explored traditional information retrieval techniques, applying TF-IDF and BM25 algorithms to categorize sentences based on keyword relevance and document similarity. These diverse methods allowed us to robustly enhance our model's performance by leveraging both the semantic understanding capabilities of deep learning and the precision of conventional algorithms.

1 Introduction

The task of sentence classification presents numerous challenges, particularly when the sentences involved are complex and pertain to niche subjects within broader categories. This complexity is further exacerbated by the limited availability of labeled training data. In our project, we were tasked with classifying 1140 sentences into 12 distinct categories based on their subject matter. Each category represented a specific domain such as politics, technology, or health, among others. The primary challenge lay in the dataset's skewed distribution; the training set provided only three examples per category, significantly fewer than what was available in the test set. This stark discrepancy posed a significant hurdle in training effective classification models.

Given the intricacies of the sentences and the scarcity of training examples, our research focused on three core areas: training or fine-tuning models with minimal data, deciphering the complex semantics of niche sentences, and strategically expanding our training dataset with accurately labeled new data. To address these challenges, we explored a variety of approaches that combined the cutting-edge capabilities of deep learning models and the robustness of traditional machine learning techniques.

2 Solution

Naive Method: Keywords and TF-IDF on Given Training Data : The naive approach began with the application of a keyword-based classification system, which utilized predefined keywords linked to specific categories to quickly classify sentences. In parallel, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) technique on the initial training dataset to highlight the importance of specific terms in relation to their document frequency across texts. This method served as a foundational layer for more complex algorithms by providing a preliminary sorting of sentences into potential categories based on keyword relevance.

BERT (Devlin et al., 2019) Use and Fine-Tuning on Given Training Data : We harnessed the power of BERT, a pre-trained model known for its deep understanding of language nuances. The BERT model was fine-tuned using our limited dataset, which involved adjusting the model to better align with the specific linguistic patterns and classification needs of our data. By fine-tuning BERT on our training set, we enhanced its capability to generate embeddings that are more tailored to the context of our classification tasks.

BERT Use for Semi-Supervised Learning to Create New Training Data : Leveraging the semi-

[*]hugo.plottu@student-cs.fr
[†]alexandre.petit@student-cs.fr
[‡]axel.voyer@student-cs.fr
[§]aymeric.palaric@student-cs.fr

supervised capabilities of BERT, we implemented a strategy to expand our training data. By using BERT to predict labels on the initial dataset, we identified sentences that the model classified with high confidence (over 99%). These sentences were then added back to the training set as pseudo-labeled data. This process of generating pseudo-labels allowed us to significantly enlarge our training corpus without requiring additional labeled data from human annotators.

Scraping Wikipedia to Create Labeled Training Data : To further augment our training data, we turned to Wikipedia as a resource for extracting content related to our categories. By selecting Wikipedia pages relevant to each category, we were able to scrape and filter text to match the specific characteristics—such as sentence length and entropy—of our training set. This method provided us with a rich source of diverse sentences that broadened the scope and variety of our training dataset.

Applying Search Engines Methods on Augmented Training Data : Upon augmenting our training data, we applied advanced search engine methods such as TF-IDF and BM25 to organize and classify this new data. These information retrieval techniques were particularly useful in sorting and categorizing large volumes of text based on their relevance to specific topics, allowing us to refine the accuracy of our dataset categorization.

Applying ML Methods on Augmented Training Data : With an enriched dataset in place, we applied traditional machine learning models to further refine our classification process. Utilizing models like SVM, XGBoost, and KNN on the embeddings generated from the augmented data, we were able to leverage the nuanced understanding of sentence embeddings provided by these algorithms to achieve higher classification accuracy.

Applying Pre-trained Zero-Shot Classifier Model DeBERTa (He et al., 2021) : In addition to the methods described above, we utilized the DeBERTa model for zero-shot classification tasks. This advanced model, which requires no specific training on the target labels, was able to predict sentence categories based on its pre-trained understanding of language. This approach was particularly valuable for handling categories with very few training examples, as it bypassed the need for extensive labeled data.

Mixing Methods to Improve Results : Finally, to capitalize on the strengths of each individual method, we employed a mixed-method approach for final predictions. This included merging the results from different models—such as DeBERTa and our naive keyword method—and using a voting system that incorporated outcomes from various classification strategies. By combining these methods, we were able to harness the best features of each, from semantic understanding and keyword detection to vector proximity, leading to a robust and accurate classification system.

3 Results and Analysis

Our experimentation yielded compelling results, with the DeBerta Large model demonstrating remarkable efficacy in zero-shot classification, achieving an accuracy of 84.5%. Notably, our test dataset frequently comprised sentences spanning multiple topics, rendering the attainment of this accuracy particularly noteworthy.

To further assess the robustness of our model, we conducted an in-depth analysis focusing on the second-highest confidence predictions for each sentence. Our findings revealed that 11.5% of accurately identified topics were relegated to secondary choices by the model. This observation suggests a high degree of proficiency, with over 95% of predictions correctly identifying the primary or secondary topics.

Given the importance of scrutinizing borderline cases, we meticulously examined instances where the confidence margin between the first and second predicted topics was narrow, varying from 5% to 15%. The rationale behind this investigation was to identify potential areas of uncertainty within our model's predictions. However, contrary to our initial hypothesis, our evaluation revealed a more evenly distributed pattern of errors across the dataset. This implies that the model's inaccuracies are not predominantly clustered within instances of ambiguous topic classification.

Method	Accuracy
Naive Method (Keywords and TF-IDF on given training data)	0.24298
BERT fine-tuned on given training data	0.52719
BERT fine-tuned with frozen layers, classification layer only	0.41403
BERT used for semi-supervised learning to create new training data	0.79824
BERT fine-tuned with additional data (semi-supervised)	0.79824
Scraping Wikipedia to create labeled training data (TF-IDF on augmented data)	0.46929
Scraping Wikipedia to create labeled training data (BM25 on augmented data)	0.54385
Applying ML methods on augmented training data (SVM, XGBoost, KNN)	0.79824, 0.79824, 0.70877
Applying pre-trained zero-shot classifier model DeBERTa	**0.84561**
Mixing methods to improve results (ML and search engine methods voting)	0.81491
Naive method combined with BERT	0.78771
Naive method combined with DeBERTa	0.75350

Table 1: Accuracy results for various sentence classification methods.

References

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.