Classifying Biomedical Natural Language Processing Domain-specific Papers

Rena Wu

**Dataset**

1. Manual Annotation (please check annotation\_agreement.ipynb for detail process):
   * A graph of a number of blue rectangular bars

     AI-generated content may be incorrect.I manually annotated the first 50 and last 30 papers from the raw data file based on the requirement from the instruction, resulting in a training set of 80 examples.
   * Annotation consistency was validated through cross-checking with two annotators (HK and QC).
   * The inter-annotator agreement, measured using Cohen’s Kappa, was 0.87, indicating strong agreement.
2. Test Data:

Figure 1Class distribution of the manual annotation, 58.75% for Non\_BioNLP and 41.25% for BioNLP, it maintains well class balance

* + The test set consists of entries from row 52 to row 302 (inclusive) from the same raw data file.
  + This ensures that the test data is separated and does not overlap with the annotated training set.

1. Extended Training Data:
   * To evaluate model performance with more training examples, we additionally used the provided annotated dataset from our GitHub repository [1] as an extended training set.
   * This extended training set was also evaluated on the same test data (rows 52–302) for direct comparability.

**Methods**

In this project, I implemented 3 different models, and comparison can be found in Table 1.

1. **Baseline Model (TF-IDF + Logistic Regression):** Our baseline approach uses a traditional machine learning pipeline involving TF-IDF vectorization for text representation and logistic regression for classification. Term frequency-inverse document frequency (TF-IDF) weights were calculated with a vocabulary size capped at 5000 features and n-gram ranges of 1 to 2, capturing both unigrams and bigrams to identify local word patterns effectively. The logistic regression classifier was trained with L2 regularization and balanced class weights to handle class imbalance. This setup ensures that no class dominates the learning process, providing a fair evaluation of model performance. The model was optimized with up to 1000 iterations for convergence. Performance metrics including accuracy, macro F1-score, precision, and recall were computed to evaluate this approach.
2. **Model 1: BERT Fine-Tuning Model:**This model is built based on the provided code on github [1], it use end-to-end fine-tuning of the BERT-base model for biomedical text classification. The data preprocessing involved combining the title and abstract of each entry, followed by tokenization with a maximum length of 512 tokens. The dataset was split into training and evaluation sets, maintaining the class distribution through stratified sampling. The fine-tuning process was optimized with weighted loss to address class imbalance, using a learning rate of 1e-4, batch sizes of 4 (due to device limitation) for both training and evaluation, and training for 3 epochs. A classification head was added to the BERT model, and the entire architecture was trained using the Trainer API from the Transformers library. The model utilized a binary classification setup with class weight balanced during training. The evaluation strategy included computing metrics such as accuracy, macro F1-score, precision, and recall, ensuring comprehensive performance assessment. The model was trained and evaluated on GPU when available, but as we don’t have GPU, we only train on CPU, which slows the progress.
3. **Model 2: Hybrid Model (TF-IDF + DistilBERT + Gradient Boosting):**We developed a hybrid architecture that combines multiple feature types. The pipeline includes three main components: TF-IDF vectors, DistilBERT embeddings, and manually engineered domain-specific features. The training data was split into training and test sets using stratified sampling to maintain class balance. The hybrid feature set was processed through a Gradient Boosting Classifier with 250 estimators, a learning rate of 0.05, and a max depth of 4. Class weights were computed and adjusted to handle class imbalance effectively. Performance was evaluated using accuracy, macro F1-score, precision, and recall, demonstrating the model's balanced capability in handling both BioNLP and Non\_BioNLP texts.

A white sheet with black text

AI-generated content may be incorrect.

**Experiment**

**Experimental Setup:**The experiments were designed to evaluate the performance of three classification models for biomedical text classification: a baseline model, Model 1 (BERT), and Model 2 (Hybrid). The dataset included Non-BioNLP and BioNLP categories, and the performance metrics calculated were accuracy, macro F1-score, precision, and recall. Initially, models were trained on a small, annotated training set (n=80 samples) and evaluated on a larger test set (n>300 samples) to ascertain each model's ability to use limited data for training and generalize new, unseen data. Subsequent experiments involved training the models with increasing sample sizes to provide a detailed comparison of their scalability and robustness.

**Training and Evaluation:**

1. Training Regime:
   * The models were initially trained on the manually annotated 80-sample dataset to evaluate their performance with a smaller dataset.
   * Models were then trained using various sample sizes (100, 200, 300, 400, and 500) to evaluate their scalability and robustness with increasing data.
   * An additional evaluation of model efficiency was conducted using a larger dataset of approximately 40,000 samples. Model 2 completed training in 1 hour and 40 minutes, whereas Model 1 processed only up to 400 samples within the same timeframe. This highlighted Model 1's slower processing speed due to its BERT-based architecture.
2. Performance Metrics:
   * Performance metrics evaluated included accuracy, macro F1-score, precision, and recall.
   * Evaluation was performed on unseen data to assess the generalization capability of the models.

**Result**:

A graph of different colored bars

AI-generated content may be incorrect.

Figure 2. Performance comparison of three classification models (training on 80 samples) on biomedical text classification.

As shown in Figure 2, when trained on a small annotated dataset of 80 samples, Model 2 demonstrated better performance among the models with an accuracy of 75.53%, an F1-score of 74.42%, a precision of 74.19%, and a recall of 74.75%. Notably, Model 2 achieved balanced F1 scores for the Non-BioNLP (79.76%) and BioNLP (69.07%) classes. The baseline model showed an accuracy of 79.68%, an F1-score of 78.77%, but suffered from low precision (64.35%) despite its high recall (88.10%), indicating a tendency for false positives. Model 1, on the other hand, showed instability with an accuracy of 72.45%, and a significant disparity between the F1 scores for Non-BioNLP (80.47%) and BioNLP (53.23%). The performance on the larger test set (n>300) suggests good generalization capability, though the precision-recall tradeoffs in Model 1 and the baseline may reflect challenges in learning decision boundaries from minimal examples.

A graph of different types of lines

AI-generated content may be incorrect.

Figure 3. Performance comparison when evaluated on unseen data between Model 1 (BERT, blue) and Model 2 (Hybrid, red) across different training sample sizes

The comparative analysis between Model 1 and Model 2 as illustrated in Figure 3, reveals several key performance distinctions across different training sample sizes. Model 1 demonstrates some instability, with erratic performance particularly evident at sample sizes of 100, 200, 300, and 500. But it seems to perform better and better as the training size goes up. This instability is most pronounced in the BioNLP class, where Model 1 fails to predict correctly in multiple instances, resulting in an F1 score of 0. Conversely, Model 2 exhibits a more stable and consistent improvement as the training sample size increases. This consistency is particularly notable in the BioNLP class, where Model 2 maintains a superior performance, indicating a better balance between classes. The overall accuracy and macro F1-score metrics further emphasize Model 2's robustness and data efficiency, suggesting more effective feature engineering and a better handling of class imbalance. These findings suggest potential issues with Model 1 related to class imbalance handling, optimization instability, and feature representation limitations, whereas Model 2's approach seems to achieve better generalization and stability from limited data.

Furthermore, given a training dataset of approximately 40,000 samples, Model 2 completed training in 1 hour and 40 minutes, and the evaluation result of unseen data is shown in Figure 4. In contrast, Model 1 only processed up to 400 samples within the same time frame (and I don’t have enough resources like GPUs and time to wait till it get finish). This considerable difference in training speed highlights that Model 1, which utilizes a BERT-based approach, is significantly slower compared to Model 2, which employs a hybrid approach combining BERT with TF-IDF techniques. These findings underscore not only the superior performance and robustness of Model 2 but also its efficiency and practicality for large-scale datasets, making it a more effective choice for real-world applications.A graph of different colored rectangular shapes

AI-generated content may be incorrect.

Figure 4. Performance metrics of Model 2 (Hybrid) when classified the same unseen data with model trained on a dataset of approximately 40,000 samples

**Discussion and Conclusion:**

This study evaluated the performance of three distinct models for the task of biomedical text classification: a baseline model utilizing TF-IDF and logistic regression, Model 1 employing a BERT-based architecture, and Model 2 implementing a hybrid approach combining TF-IDF, DistilBERT embeddings, and domain-specific features.

When trained on a small annotated dataset of 80 samples, Model 2 demonstrated a balanced and robust performance with an accuracy of 75.53%, an F1-score of 74.42%, a precision of 74.19%, and a recall of 74.75%. It outperformed Model 1, which has some instability, especially in predicting the BioNLP class, and slightly lagged behind the baseline model in terms of recall but maintained better precision.

The experimental results on a larger dataset of approximately 40,000 samples further underscored the superior performance of Model 2 across all evaluated metrics, achieving an accuracy of 85.75%, a macro F1-score of 85.08%, a precision of 84.77%, and a recall of 85.51%. Additionally, Model 2 maintained balanced performance between the Non-BioNLP (F1 score: 88.24%) and BioNLP (F1 score: 81.93%) classes, demonstrating effective handling of class imbalances. In contrast, Model 1 showed significant performance instability, particularly at smaller sample sizes, with pronounced difficulties in predicting the BioNLP class. The baseline model, despite its simplicity, achieved reasonable recall but suffered from low precision, indicating a propensity for generating false positives.

Moreover, the efficiency analysis revealed that Model 2 completed training on the 40,000 sample dataset in 1 hour and 40 minutes, while Model 1, constrained by its BERT-based approach, managed to process only up to 400 samples within the same timeframe. This difference in processing speed further underscores Model 2's practicality and efficiency for large-scale datasets.

In summary, the hybrid architecture of Model 2, which combines the strengths of TF-IDF vectors, DistilBERT embeddings, and domain-specific features, proved to be a powerful and efficient solution for biomedical text classification. Its balanced performance, scalability, and efficiency make it a promising choice for real-world applications where large and diverse datasets are common. On the other hand, the slower processing speed and instability of Model 1 highlights the need for improvements in class imbalance handling and optimization stability when utilizing BERT-based models. Future work could focus on enhancing the use of BERT models and exploring additional hybrid approaches to further improve classification performance and efficiency.

**Reference:**

[1] [YinYuBB/BioNLP\_BERT](https://github.com/YinYuBB/BioNLP_BERT)

Require file for submission:

1. Github link and readme: [renawuq/BioNLP\_BERT](https://github.com/renawuq/BioNLP_BERT)
2. My annotation result: [annotations - Google Drive](https://drive.google.com/drive/u/1/folders/1kjcZ470gxs-j0B4XFc4YuhPBQd1B92PI)
3. My model: [model\_result - Google Drive](https://drive.google.com/drive/u/1/folders/1oo4OLEpEyXFiAeP-9PIJUkuuyuLvPO08)
4. If you need my other model, please contact me directly, currently drive can’t store all my model…