Project 2: Document Classification

Md Mohaiminul Islam
mmislam@iastate.edu
Kaggle Handle: Md Mohaiminul Islam/mahim05078

1 Methodology

1.1 Data Preprocessing

The preprocessing pipeline prepared documents for transformer-based classification , and also prepares to address severe class imbalance (10:1 ratio) in training data. The raw abstracts and titles underwent minimal normalization to preserve domain-specific terminology essential for domain specific language models. Documents were formatted by combining tile and abstracts as: "$Title : title \backslash n Abstract : abstract$". We then implemented k-fold cross-validation with two key modifications to address train-test distribution mismatch. First, validation folds were rebalanced through minority class oversampling to achieve 15% positive samples (vs. 9% in training), improving calibration for the estimated test distribution ($\approx$ 20% positive). Second, weighted random sampling assigned class weights inversely proportional to frequency during training, ensuring balanced batch compositions without data augmentation. To use curriculum learning in 1.4 and 1.5 we assigned difficulty rating to each sample based on previous prediction probability distribution.

1.2 Baseline: Fine-tuned BERT

As a baseline, we fine-tuned BERT-base-uncased (Devlin et al., 2019) for binary document classification. The model employed cross-entropy loss with class weights $w_0 = 0.55$ and $w_1 = 5.60$ to address the $10:1$ class imbalance. Training was conducted over 5 epochs using the AdamW optimizer with learning rate $\eta = 2 \times 10^{-5}$, weight decay $\lambda = 0.01$, batch size, $B = 16$, and maximum sequence length, $L = 512$ tokens. A linear warmup schedule spanning 10% of training steps was done before cosine annealing for learning rate decay.

1.3 DeBERTa: Addressing Class Imbalance

We selected *DeBERTa-v3-base* (He et al., 2021) as one of our improved architectures due to its disentangled attention mechanism and enhanced mask decoder, which have demonstrated superior performance on classification tasks with class imbalance. After hyperparameter optimization, the model was trained with focal loss (Lin et al., 2017) ($\gamma = 3.0$) to emphasize hard-to-classify minority samples, learning rate $\eta = 1 \times 10^{-5}$, weight decay $\lambda = 0.15$, and batch size $B = 8$ with gradient accumulation over 2 steps. Training was limited to 3 epochs with aggressive early stopping (patience=1) to prevent overfitting. This configuration achieved $F_1 = 0.86$ on public test data, substantially outperforming the BERT baseline.

1.4 PubmedBERT: A Domain Specific Approach

PubMedBERT (Gu et al., 2021) was another one of the improved model over the baseline, selected for its domain-specific pretraining on 14 million PubMed abstracts and full-text articles. This makes an expert in biomedical terminology and scientific discourse patterns suitable for our classification task. We implemented a two-stage training procedure: first, an initial model was trained without curriculum learning to generate confidence scores for all training samples. Sample difficulty was quantified as $d_i = 1 - \frac{|0.5 - p_i|}{0.5}$, where p_i represents the predicted probability for sample i. This metric assigns higher difficulty to samples near the decision boundary ($p_i \approx 0.5$). Using these scores, curriculum learning proceeded through three epochs with progressively expanding training subsets: easy samples (lowest 50% difficulty), medium samples (lowest 75%), and the complete dataset. We maintained identical hyperparameters as DeBERTa ($\eta = 1 \times 10^{-5}$, $\gamma = 3.0$), this approach achieved the highest public leaderboard score ($F_1 = 0.88$).

1.5 Ensembling Best Outcomes

The final ensemble combined *DeBERTa* and *PubMedBERT* predictions through probability averaging with a range of weights starting from equal wieghts ($w_{\text{DeBERTa}} = w_{\text{PubMed}} = 0.5$) to 20-80 ($w_{\text{DeBERTa}} = 0.2, w_{\text{PubMed}} = 0.8$) weights. This strategy tried to get the best of both worlds: DeBERTa's superior attention mechanism for general text understanding and PubMedBERT's domain-specific knowledge. The ensemble is also known to reduces model-specific biases and prediction variance. This approach performed on par with our last approach and achieved the highest score with given public test data.

2 Results and Discussion

Table 1 summarizes the performance metrics for each model in terms of F1 score. We use F1 as the evaluation metric because it is the criterion on which the Kaggle leaderboard is based.

Our first approach was to use 5-fold cross-validation on the baseline BERT, which overfit severely. Investigating the results showed that the main reasons were class imbalance in the training data and a discrepancy between the class distributions of the training and test data. This was evident from the fact that the training data contained about 9% class 1 samples, but predicting a similar percentage yielded a very poor score on the test data. Hence, we inferred that the test data had a substantially higher percentage of class 1 samples and relatively less class imbalance. Another observation was that the best public score achieved with the baseline required setting the decision threshold to $p(y{=}1) \geq 0.05$, which indicates very low predicted probabilities for class 1, even though the training process used weighted sampling for the minority classes based on the label distributions per fold. This suggests poor probability calibration and limited separation for the positive class.

For the next approach, we considered DeBERTa for its stronger architecture with disentangled attention, which we hypothesized could generalize better under imbalance. Moreover, some of the five folds performed poorly because the already scarce class 1 training data were not incorporated equally across folds. To address this and reduce training time, we reduced the number of folds to 3 so that each fold contained more positive examples. We also reduced the number of epochs from five to three and halved the learning rate. All of these changes were made to reduce overfitting. For this approach, continuously tuning the decision threshold yielded the best result at $p(y{=}1) \geq 0.185$, which was better than the previous setting but still quite low. The model's average confidence for both classes did not show strong separation, indicating under-confident predictions and limited class polarity.

In our third attempt, we chose PubMedBERT due to its domain-specific expertise in analyzing biomedical text from PubMed articles. It also has fewer parameters than DeBERTa, which can encourage less memorization and more reasoning. Keeping all hyperparameters the same as before, this model achieved the most significant improvement among all experiments, obtaining a best private score of 0.86 and a public score of 0.87. The optimal prediction threshold in this case was ≥ 0.775, marking a substantial improvement over previous approaches in terms of stability, generalization, and confidence in class 1 predictions.

Subsequently, we implemented a curriculum learning approach during training. To achieve this, we introduced an additional preprocessing step in which we used PubMedBERT's prediction scores to estimate a relative difficulty rating for the training samples. Samples whose average (across all validation folds) prediction probabilities were closest to 0.5 were considered the hardest, as the model was most uncertain about them. Although this approach achieved a higher public leaderboard score (0.888), it resulted in a lower private score, which unfortunately ended up being our final submission. But, we can argue that, it did increase the robustness of the model since optimal prediction threshold in this case was better (≥ 0.8375) signifying more confident predictions. Moreover without curriculum learning it yielded best private test results(0.865).

Finally, we experimented with two ensemble configurations combining PubMedBERT and DeBERTa predictions in $50-50$ and $80-20$ ratios. The goal was to use the complementary strengths of both models—PubMedBERT's domain-specific reasoning and DeBERTa's contextual understanding. While the $50-50$ ensemble improved stability, the $80-20$ mix resulted in a higher private score (0.855). This is mostly due to PubMEdBERTs confidence dominating the ensemble and pulling through whereas DeBERTa's prediction scores were less confident.

Model	k	Best Fold (F1 score)	Avg. Fold (F1 score)	Test Score(Public)	Test Score(Private)
Baseline BERT	5	0.868	0.816	0.733	0.660
DeBERTa	3	0.806	0.785	0.838	0.868
PubmedBERT	3	0.699	0.683	0.876	**0.865**
PubmedBERT(Curriculum)	3	0.700	0.683	**0.888**	0.847
Ensemble$(50-50)$	3	0.685	0.591	0.852	0.836
Ensemble$(20-80)$	3	0.693	0.652	0.868	0.855

Table 1: Performance comparison of all the models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.