Automating ICD Code Prediction with Natural Language Processing Techniques

Rozita Haghighi, Yongqi Hao

Abstract

This study addresses transforming unstructured Electronic Health Record (EHR) data into structured ICD codes through Automated Medical Coding (AMC) to reduce workload, improve consistency, and enhance patient care. Using NLP models such as BERT, BioBERT, and Naive Bayes, we focus on improving coding accuracy and efficiency. BioBERT-FFNN, a domain-specific transformer, outperforms BERT-FFNN due to its specialized biomedical training, while Naive Bayes achieves high performance. Future work involves refining models with RoBERTa PM for better ICD prediction.

Introduction and Research Goal

Electronic health record (EHR) data contain diverse patient information, with over 80% being unstructured text like clinical notes, making it challenging for predictive analysis. This project addresses these challenges through Automated Medical Coding (AMC), focusing on multilabel classification for ICD code prediction. By employing Natural Language Processing (NLP) techniques to convert unstructured data into structured ICD codes, AMC aims to enhance coding accuracy, reduce errors, streamline billing, and lessen the workload of healthcare professionals, ultimately improving efficiency and quality of patient care. [6].

Dataset Selection

- The dataset includes 20,000 texts from MIMIC-IV and MIMIC-IV-Note, containing hospital and ICU records, as well as discharge summaries from 2020 to 2022. [4, 3]
- The data is split into 80% for training and 20% for testing. Preprocessing involves extracting relevant text, removing stopwords and special characters, retaining critical numbers, and using the top 10 most frequent ICD codes.

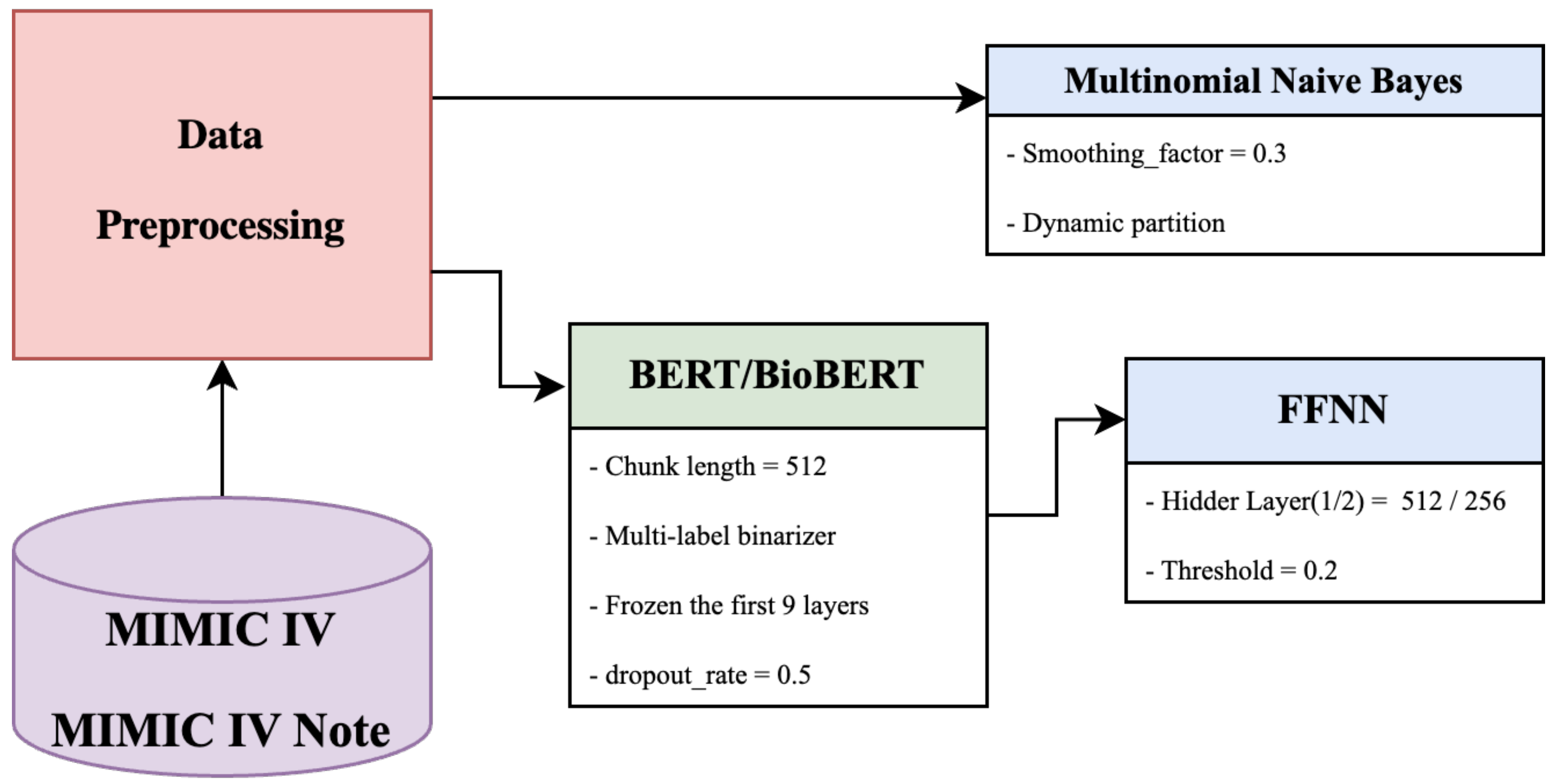
Methods

Our method involves first fine-tuning **BERT**, which is pre-trained on general corpora (English Wikipedia and BooksCorpus), on our clinical text corpus to generate contextual embeddings. We also fine-tune **BioBERT**, which is initialized with BERT's weights and then pre-trained further on biomedical domain corpora (PubMed abstracts and PMC full-text articles), to extract domain-specific embeddings from clinical data. These embeddings are then fed into a feedforward neural network (**FFNN**) for multilabel classification of ICD codes, leveraging both general and domain-specific information to improve accuracy. Additionally, we implement **Naive Bayes** to extract the top 10 probabilities of ICD codes as predictions, and compare all models to evaluate their performance in ICD code prediction. [5, 7, 1, 2]

Important Result

BioBERT-FFNN outperforms **BERT-FFNN**, showing domain-specific advantages, while **Naive Bayes** achieves best performance.

Methodology pipeline



Data Preprocessing

Multinomial Naive Bayes
- Smoothing_factor = 0.3
- Dynamic partition

BERT/BioBERT
- Chunk length = 512
- Multi-label binarizer
- Frozen the first 9 layers
- dropout_rate = 0.5

FFNN
- Hidder Layer(1/2) = 512 / 256
- Threshold = 0.2

MIMIC IV

MIMIC IV Note

Figure 1:Methodology pipeline and Parameters

Results

Metric	BERT-FFNN	BioBERT-FFNN	Naive Bayes
F1 Micro	0.1989	0.2165	0.4619
F1 Macro	0.1244	0.1356	0.1419
AUC-ROC Micro	0.6934	0.6922	0.6714
AUC-ROC Macro	0.6462	0.6471	0.5668
Precision@5	0.1820	0.1801	0.3391

Table 1:Performance Metrics for Methods

Sample Predictions

Method	Actual ICD	Predicted ICD
BERT	2724, 53081	2724, 4019, 53081
BioBERT	4019	2724, 4019
MNB	5859, 25000, 42731, 2724	4280, 42731, 25000, 2724

Table 2:Sample Prediction Results for Methods

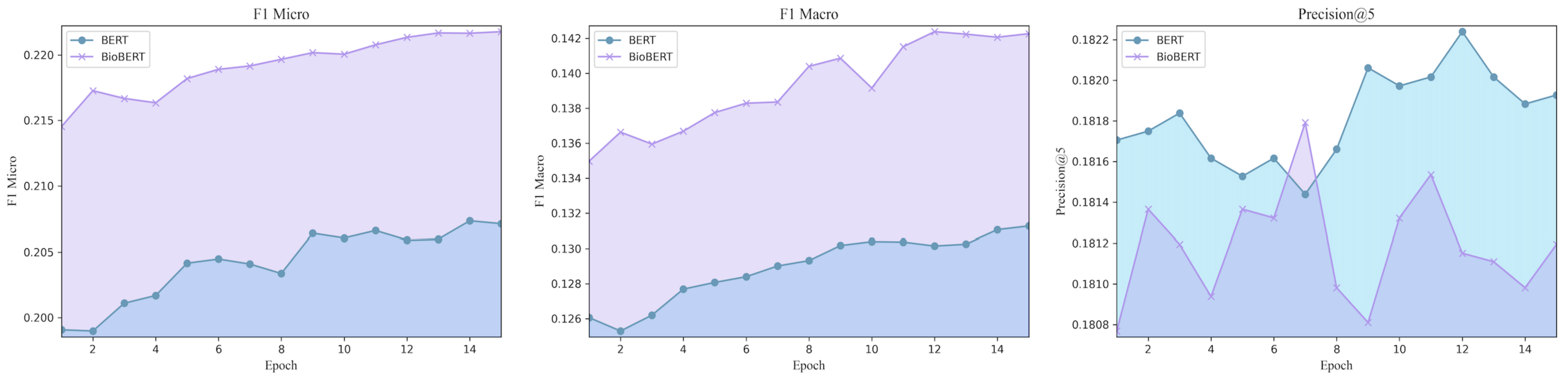


Figure 2:Metrics

Analysis and Conclusion

BioBERT-FFNN outperforms **BERT-FFNN**, suggesting that domain-specific contextual embeddings provide an advantage over general embeddings for biomedical text mining.

Naive Bayes performs surprisingly well, with the highest F1 Micro and Precision@5, likely due to its ability to focus on high-confidence predictions. However, its lower AUC-ROC Macro indicates a potential lack of generalizability across all classes.

To achieve optimal performance from transformer-based models, it is recommended to use larger datasets, leveraging their full potential for capturing intricate, context-specific relationships.

Future Work and Limitations

Future Work: We plan to implement **RoBERTa PM**, refined using biomedical corpora such as PubMed abstracts, PMC articles, and MIMIC-III, to enhance ICD code prediction. We will compare model on the full dataset after removing rare codes (fewer than ten instances).

Limitations: Due to limitations in the computing units, we are unable to implement more powerful classifiers(BART/XGBoost).

References

[1] Bert model on hugging face.

[2] Biobert model on hugging face.

[3] Mimic-iv-note (version 2.2), 2022. Accessed: 2024-10-06.

[4] Mimic-iv (version 3.0), 2024. Accessed: 2024-10-06.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics, 2019.

[6] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*, 177:141–153, 2019.

[7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.