HALELab-NITK at #SMM4H-HeaRD2026: Inclusion of Feature Engineering for Detection of Patient Metadata in SARS-CoV2 Sequencing Articles

Aakarsh Bansal*, **Abhishek Srinivas*** and **Sowmya Kamath S.**
Healthcare Analytics and Language Engineering (HALE) Lab,
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, Srinivasnagar P.O.,
Mangaluru 575025, India
{aakarsh12bansal, sriabhi2407}@gmail.com, sowmyakamath@nitk.edu.in

Abstract

This article presents a system description for our work as part of Task 5 of the SMM4H-HeaRD 2026 workshop. We fine-tune pre-trained BERT and BiomedBERT models and further enhance them using custom feature augmentation techniques. Incorporating these engineered features results in improved performance, with the best model achieving a validation F1 score of 0.8419 and an evaluation phase F1 score of 0.753.

1 Introduction

In this work, we participate in Task 5: Detection of Patient Metadata in SARS-CoV-2 Sequencing Articles (Klein et al., 2025) of the SMM4H-HeaRD workshop co-located with ACL 2026, which aims to identify and extract key patient-level information from scientific publications. This task is particularly important for the field of genomic epidemiology, where linking viral genome sequences with patient metadata, such as demographics, clinical outcomes and geographic information is critical for understanding disease transmission, variant emergence and public health trends.

In real-world healthcare settings, such patient metadata is often embedded in unstructured text within research articles, making automated extraction both challenging and essential. Accurate detection of patient metadata enables the construction of richer, more complete datasets, improving downstream analyses and supporting data-driven decision-making in public health. By addressing this problem, this task contributes to bridging the gap between large-scale genomic data and the contextual clinical information needed for effective epidemiological modeling and response.

*Equal contribution.

2 Dataset

The dataset is provided as part of the SMM4H-HeaRD 2026 workshop's Task 5 (Klein et al., 2025), for detecting patient metadata in SARS-CoV-2 sequencing articles. It has been constructed from a filtered subset of the LitCovid (Chen et al., 2023) corpus, where articles likely to contain genomic sequencing information are selected using high-recall keyword-based filtering. The final version of this dataset consists of sentence-level entries labeled as either positive (containing patient metadata) negative (not containing patient metadata). The training set contains approximately 15,505 entries and the validation set consists of 2,215 entries.

3 Methodology

Our methodology encompasses four phases. A representative diagram summarizing our workflow is shown in Fig. 1.

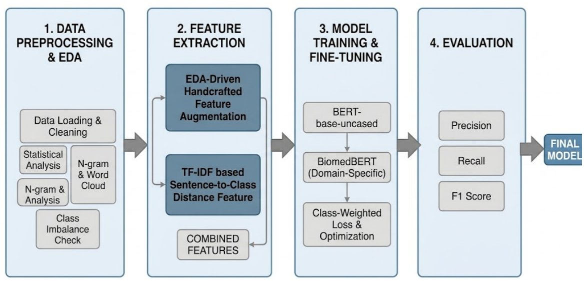

Figure 1: Proposed Methodology

3.1 Data Preprocessing

We apply a data cleaning pipeline to improve training data quality, which includes removal of missing values, filtering out extremely short sentences, elimination of duplicate entries and validation of label consistency. Exploratory Data Analysis (EDA) is performed to better understand the dataset characteristics. This includes analyzing class distribution,

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 222–224
July 3, 2026 ©2026 Association for Computational Linguistics

sentence length distributions and identifying discriminative n-grams and terms using TF-IDF (Term Frequency-Inverse Document Frequency) statistics. The analysis confirms severe class imbalance and highlights subtle patterns distinguishing the two classes.

3.2 Baseline Models

We fine tune transformer based language models for the sentence classification task. Specifically, we employ the following architectures listed below. Input sentences are tokenized with a maximum sequence length of 128. To address class imbalance, we use a class-weighted cross-entropy loss function during training.

- *BERT-base-uncased (Devlin et al., 2019)*: A widely used general purpose pretrained transformer model trained on large-scale English corpora.
- *BiomedBERT (Gu et al., 2021)*: A domain adapted variant pretrained on biomedical literature, enabling improved representation of medical and clinical terminology.

3.3 Proposed Enhancements

To improve model performance beyond standard fine-tuning, we incorporate additional features:

EDA-Guided Feature Augmentation. We extract patterns identified during EDA and encode them as a low-dimensional feature vector. These include age-related information, mentions of sex of a person and geographical information. Age and sex features are extracted using regular expression patterns. Geographical features are extracted using spaCy NER (Honnibal et al., 2020) to identify location-related entities and their occurrence statistics; we use the "FAC" (*Facility*), "GPE" (*Geopolitical Entity*) and "LOC" (*Location*) entity labels for this purpose. The final feature vector consists primarily of binary indicators capturing the presence of age, sex and geographical information, along with two normalized features representing the number of detected geographical entities and the count of unique geographical entities in a particular entry.

TF-IDF Class-Centroid Similarity. We compute TF-IDF representations of sentences using unigram and bigram features and construct centroid vectors for both positive and negative classes from the training set. For each input sentence, cosine similarity scores with respect to both centroids are calculated, along with their difference, and both are used as additional features. This provides a class-level similarity signal that complements contextual model embeddings. All extracted features are used to build a feature vector which is then concatenated with the [CLS] embedding from the transformer model, allowing the classifier to use both learned representations and explicit dataset-specific patterns.

3.4 Training Setup

Models are trained using the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 0.01 for 4 epochs with a batch size of 32. A linear learning rate scheduler with a warmup ratio of 10% of the total training steps is employed to gradually increase the learning rate during the initial phase of training. Gradient clipping with a maximum gradient norm of 1.0 is applied to stabilize optimization and prevent exploding gradients. To address class imbalance, a weighted cross-entropy loss function is used with class-specific weights computed from the training distribution. Evaluation is conducted using precision, recall and F1-score.

4 Experimental Results

We evaluate the performance of our models on the validation set provided in the dataset. The models are assessed using standard classification metrics, such as precision, recall and F1 score, which are more useful in cases of class imbalance. Table 1 summarizes the performance of our models on the validation set, while Table 2 shows the performance in the evaluation phase of the shared task.

Effect of Domain-Specific Model. BiomedBERT consistently does better the generic BERT-base model across all evaluation metrics. This improvement can be explained because of its pretraining on biomedical corpora, which allows it to better capture domain-specific terminology and nuances present in scientific literature.

Impact of Feature Augmentation. The usage of EDA-guided features and TF-IDF class-centroid similarity further improves performance over the base BiomedBERT model. These additional features explicitly provide complementary information that are not captured by the learned model embeddings alone. In particular, EDA-based features add dataset-specific patterns related to personal identification cues, while, TF-IDF similarity

Model	Precision	Recall	F1-score
BERT-base (no augmentation)	0.7397	0.7684	0.7538
BiomedBERT (no augmentation)	0.7613	0.8571	0.8064
BiomedBERT (+ feature augmentation)	**0.7948**	**0.8947**	**0.8419**

Table 1: Comparison on the Validation set

Model	Precision	Recall	F1-score
Mean	0.712	0.756	0.729
Median	0.741	0.772	0.754
Ours (BiomedBERT + feature augmentation)	0.740	0.767	0.753

Table 2: Comparison with Evaluation Phase (Test Set) Submissions

captures global class-level distance. This approach shows that combining extracted and learned representations can lead to improved classification performance.

Handling Class Imbalance. Due to the severe class imbalance (approximately 13% positive samples), recall for the positive class remains a challenging metric to optimize. The use of class-weighted loss helps improve sensitivity toward the minority class.

Validation vs. Evaluation Performance. Although the augmented BiomedBERT model achieves a strong F1-score of 0.8419 on the validation set, the performance on the hidden evaluation set decreases to 0.753. The proposed feature augmentation strategy incorporates dataset-specific patterns observed during EDA from the training data, which may align more strongly with the validation distribution than with unseen evaluation samples. Since the hidden evaluation set may contain differences in language and contextual patterns, we hypothesize that the constructed features may not have generalized equally across both distributions.

5 Conclusion

In this work, we fine-tuned BERT and BiomedBERT for the shared task and observed that domain-specific pretraining plays an important role in improving performance, with BiomedBERT outperforming the generic BERT-base model. Based on our EDA observations, we further incorporated constructed features through regular-expression matching and geographical NER and TF-IDF class-centroid similarity to provide complementary guidance to improve performance. However, the disparity between our validation and the evaluation phase scores suggests that some of these observed and extracted patterns may be specific in nature and may not generalize well to unseen data distributions.

References

Qingyu Chen, Alexis Allot, Robert Leaman, Chih-Hsuan Wei, Elaheh Aghaarabi, John J Guerrerio, Lilly Xu, and Zhiyong Lu. 2023. Litcovid in 2022: an information resource for the covid-19 literature. *Nucleic Acids Research*, 51(D1):D1512–D1518.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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 Trans. Comput. Healthcare*, 3(1).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ari Z. Klein, Davy Weissenbacher, Karen O'Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models. *medRxiv*.