การประชุมทางวิชาการมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 64
The 64th Kasetsart University Annual Conference

Natural Language Processing with WangchanBERTa for Classifying Causes of Death Among Thai People Living with HIV

Sudawadee Chitlekasakul [a*], Sirinya Teeraananchaia [b]
[a] Master of Biomedical Data Science program, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
[b] Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
E-mail address: sudawadee.ch@ku.th

Abstract

This study developed a Natural Language Processing (NLP) model using WangchanBERTa embeddings to classify causes of death (COD) among Thai people living with HIV (PLHIV) into AIDS-related and non-AIDS-related categories. The dataset included free-text mortality records from individuals initiating antiretroviral therapy between 2021–2024 in the Universal Health Coverage (UHC) database. Model performance was evaluated using multiple metrics. The baseline model showed limited discrimination, with F1-score 57.10% and AUC 56.92%, while applying the Synthetic Minority Over-sampling Technique (SMOTE) substantially improved performance to F1-score 78.07% and AUC 90.63%, demonstrating the effectiveness of transformer-based Thai NLP combined with oversampling for automated mortality classification.

Keywords: HIV mortality, Natural Language Processing, WangchanBERTa, SMOTE

Introduction

- Accurate classification of COD among PLHIV is critical for epidemiological monitoring and national program evaluation, yet Thai mortality data are primarily recorded as unstructured free text, making standardized analysis difficult.
- Manual coding is time-consuming and prone to inconsistency, while mortality patterns are shifting toward non-AIDS chronic conditions, increasing the need for reliable classification systems.
- Transformer-based models, particularly WangchanBERTa, provide improved contextual representation for Thai clinical narratives and offer a scalable solution for automated mortality classification.

Objectives

- To develop and evaluate a WangchanBERTa-based NLP model to classify free-text COD among PLHIV into AIDS-related and non-AIDS-related categories.

Methods

Data Source and Study Population
- UHC database of PLHIV (n=5,159)
- PLHIV ≥15 years, ART initiation 2021–2024

Preprocessing
- Remove punctuation
- Remove symbols

Embedding
- WangchanBERTa Transformer

Imbalance Handing
- SMOTE on Training set

Model Training using WangchanBERTa
- Train/test split: 70% / 30%
- 5-Flod Cross-Validation

Model Evaluation
- Accuracy
- Sensitivity
- Specificity
- Precision
- F1-score
- AUC-ROC

Results

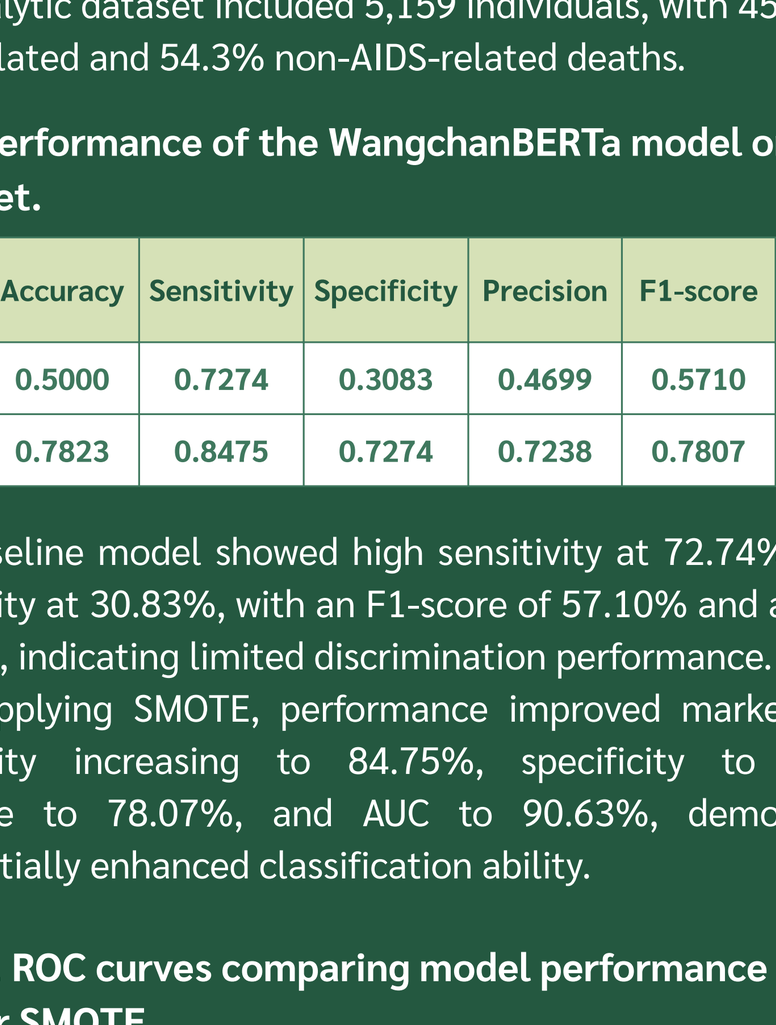
- The analytic dataset included 5,159 individuals, with 45.7% AIDS-related and 54.3% non-AIDS-related deaths.

Table 1 Performance of the WangchanBERTa model on testing set.

Data set	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
Original	0.5000	0.7274	0.3083	0.4699	0.5710	0.5692
SMOTE	0.7823	0.8475	0.7274	0.7238	0.7807	0.9063

- The baseline model showed high sensitivity at 72.74% but low specificity at 30.83%, with an F1-score of 57.10% and an AUC of 56.92%, indicating limited discrimination performance.
- After applying SMOTE, performance improved markedly, with sensitivity increasing to 84.75%, specificity to 72.74%, F1-score to 78.07%, and AUC to 90.63%, demonstrating substantially enhanced classification ability.

Figure. 1 ROC curves comparing model performance before and after SMOTE.



- The ROC curve shows that the baseline model had weak discrimination with an AUC of 0.569, while SMOTE improved performance markedly, increasing AUC to 0.906.
- The clear separation between the two curves confirms that SMOTE substantially improved the model's classification capability.

Conclusion

- WangchanBERTa effectively captures complex semantic patterns from Thai free-text COD narratives, enabling reliable mortality classification.
- Although baseline performance was constrained by residual imbalance and linguistic heterogeneity, SMOTE markedly enhanced discrimination and overall predictive performance.
- The integration of transformer-based embeddings with oversampling provides a practical and deployable framework that can be applied to real-world automated HIV mortality surveillance and routine public health monitoring.

References

1. Coutinho I, Martins B. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. J Biomed Inform. 2022;136:104232. doi:10.1016/j.jbi.2022.104232
2. Lertpiriyasuwat C, Kerr SJ, Noknoy S, Namahoot P, Punsuwan N, Apornpong T, et al. Cause of death in people living with HIV who initiated antiretroviral therapy after enrolling to the Thai National AIDS Program from 2008 to 2021. Lancet Reg Health Southeast Asia. 2025;36:100576. doi:10.1016/j.lansea.2025.100576
3. Lowphansirikul L, Polpanumas C, Phatthiyaphaibun W, et al. WangchanBERTa: Pretraining transformer-based Thai language models. arXiv preprint. 2021;arXiv:2101.09635.
4. Smith C, Duda SN, et al. Time trends in causes of death in people with HIV: Insights from the Swiss HIV Cohort Study. HIV Med. 2010;11(9):621-631. doi:10.1111/j.1468-1293.2010.00834.x
5. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc. 2020;27(11):1741-1748. doi:10.1093/jamia/ocaa189

Acknowledgements: This study was made possible through the guidance and support of Assoc. Prof. Dr. Sirinya Teeraananchai and the Co-Advisors of the Department of Biomedical Data Science, Faculty of Science, Kasetsart University. The authors express their sincere gratitude for their valuable assistance. This study was approved by the Institutional Review Board of Kasetsart University Research Ethics Committee (KUREC-HSR67/043), Thailand. A waiver of informed consent was granted for this secondary data analysis. All data were de-identified by the National Health Security Office prior to analysis.