
# Interpretable Machine Learning

Hendra Bunyamin Maranatha Christian University September 24, 2023

# Table of Content

# Table of Content

# What is Interpretability?

- Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019).

- Interpretability is the degree to which a human can consistently predict the model's result (Kim et al., 2016).

# Topic Clustering

- Bunyamin and Sulistiani (2017) explore the **application of a topic clustering model on students' final project abstracts**.
- The topic clustering algorithm follows Latent Dirichlet Allocation (LDA) which is an expansion model from probabilistic latent semantic analysis (PLSA) (Hofmann, 1999).
- The results of clustering are obscure and difficult to interpret. We believe the clustering model is not suitable for this problem.

# Text Classification

- Bunyamin et al. (2019) study **text classification problem** from students' final projects. Specifically, the features are extracted from abstracts, chapter 1, chapter 2, and chapter 3.

- Ridge classifier, passive-aggressive, linear Support Vector Classifier (SVC), and Stochastic Gradient Descent (SGD) classifier gain the highest accuracy among all 12 text classification models. Surprisingly, the deep learning models are unable to attain results as comparable as the ones of linear models.

# Language Model $\Rightarrow$ Text Classification

- Bunyamin (2021) investigates an adoption of the transfer learning technique named Universal Language Model for Fine-tuning (ULMFiT) (Howard and Ruder, 2018).

- The experiments show that ULMFiT achieves high accuracy ($\approx$ 93%); additionally, this high performance can mitigate the overfitting symptom.

# Table of Content

# Classical vs. Deep Learning

- Bunyamin and Meyliana (2021) study the utilization of classical and deep learning time series prediction models in the case of Indonesian economic growth.
- The dataset comprises World Development Indicators of Indonesia from 1962 to 2016.
- Through experiments, Seasonal Autoregressive Integrated Average (SARIMA) and Convolutional LSTM give the best performance based on Root-Mean-Square Error (RMSE).

# Table of Content

# Breast Cancer Histopathological Image Classification

- Bunyamin. et al. (2022) implement **Progressive Resizing Approach** (PRA). The idea is very simple, that is to start training using small images and finally end the training using large images.
- Another highlight of the research is the use of **Vahadane technique** (Vahadane et al., 2016) which solves both stain separation and color normalization problems.
- The result shows that PRA achieves a promising performance measured by $F_1$ score.

# Building Damage Image Segmentation

- Toba et al. (2023) develop **transfer learning models** using low-cost masking pre-processing in the experimental building damage (xBD) dataset, a large-scale dataset for advancing building damage assessment.

- The experiments show that ResNet-34 is the best with an F1 score of 71.93%, and an intersection over union (IoU) of 66.72%.

# References I


Bunyamin, H. (2021). Utilizing indonesian universal language model fine-tuning for text classification. *Journal of Information Technology and Computer Science*, 5(3):325–337.

Bunyamin, H., Heriyanto, Novianti, S., and Sulistiani, L. (2019). Topic clustering and classification on final project reports: a comparison of traditional and modern approaches. *IAENG International Journal of Computer Science*, 46(3):506–511.

Bunyamin, H. and Meyliana (2021). Classical and deep learning time series prediction techniques in the case of indonesian economic growth. *IOP Conference Series: Materials Science and Engineering*, 1077(1):012014.

Bunyamin, H. and Sulistiani, L. (2017). Automatic topic clustering using latent dirichlet allocation with skip-gram model on final project abstracts. In *2017 21st International Computer Science and Engineering Conference (ICSEC)*, pages 1–5.

Bunyamin., H., Toba., H., Meyliana., and Wahyudianingsih., R. (2022). Breast cancer histopathological image classification using progressive resizing approach. In *Proceedings of the 1st International Conference on Emerging Issues in Technology, Engineering and Science - ICE-TES,*, pages 351–357. INSTICC, SciTePress.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

# References II


Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Toba, H., Bunyamin, H., Widyaya, J. E., Wibisono, C., and Haryadi, L. S. (2023). Masking preprocessing in transfer learning for damage building detection. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 12(2).

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971.