

MAIN CONTRIBUTIONS OF THE DOCTORAL DISSERTATION

Thesis Topic: Opinion Mining with Ensemble Methods

Major: Computer Science

Code: 9480101

PhD Candidate: Tran Quang Phuc

Supervisors:

1. Prof. Dr. Huynh Xuan Hiep
2. Dr. Le Thi My Hanh

Institution: The University of Danang - University of Science and Technology

CONTRIBUTIONS OF THE DISSERTATION

- First, the dissertation proposes an Opinion Classification Forest (F_{CF}) model based on an $N \times 5$ opinion-quintuple matrix representation (entity, aspect, sentiment, holder, time) combined with supervised Shannon entropy as the node-splitting criterion. The semantic structure of the opinion quintuple is preserved throughout tree construction, and the model integrates Feature Importance and Partial Dependence Plot (PDP)-based interpretation. Under the row-level random split used in the published papers, the model achieves an accuracy of 0.77 on the yelpHotelData and yelpResData datasets.
- Second, the dissertation proposes an Opinion Density Forest (F_{DF}) model that integrates DBSCAN clustering at leaf nodes and Gaussian entropy as the splitting criterion to handle class imbalance through the density distribution of opinion clusters, together with SHapley Additive exPlanations (SHAP) and Shapley interaction indices to analyze the importance of and interactions among quintuple components. Under the published figures, the model achieves an average accuracy of 0.78 without requiring a GPU, with SHAP identifying Entity as the dominant feature (mean $|SHAP| \approx 0.087$).
- Third, the dissertation proposes an Opinion Manifold Forest (F_{MF}) model that learns a low-dimensional manifold structure directly from leaf co-occurrence statistics of an unsupervised random forest, via a relation matrix $W \in \mathbb{R}^{(N \times N)}$ and a manifold distance $dist_M$, avoiding the global eigen-decomposition required by ISOMAP or UMAP; the data are clustered with DBSCAN in the manifold space and predictions are aggregated through a stacking meta-learner. Under the published figures, the model achieves the highest average accuracy among the three proposed models (0.80) and is interpretable through the relation matrix W ; the storage cost of W is the main barrier to scaling.
- Fourth, the dissertation builds a pipeline that converts four opinion review datasets into the standardized opinion-quintuple structure (e, a, s, h, t) used to train and evaluate the proposed forest models, comprising two English datasets, yelpHotelData and yelpResData (publicly released data from Mukherjee et al.), and two Vietnamese

datasets, VLSP-2018 Restaurant and VLSP-2018 Hotel (community-released VLSP-2018 data); the quintuple extraction model achieves $F1 = 0.78$ on the Twitter US Airline Sentiment dataset.

- Finally, the three models are re-evaluated with a review-level split ($5 \text{ seeds} \times 10,000$ reviews): the Balanced Accuracy of all quintuple-only models is about 0.51–0.57, with F_{CF} and F_{DF} comparable to Random Forest and F_{MF} about 0.02–0.03 below Random Forest; this quantifies the leakage of the row-level split (Accuracy 0.946 versus 0.611 with the same Random Forest) and leads to the recommendation of review-level splitting when evaluating models on quintuple data.

Danang, ... 2026

Supervisors' Confirmation

Supervisor 1

Supervisor 2

Doctoral Candidate

Prof. Dr. Huynh Xuan Hiep

Dr. Le Thi My Hanh

Tran Quang Phuc