

ABSTRACT OF THE THESIS

Thesis Title: “OPINION MINING WITH ENSEMBLE METHODS”

Author: Tran Quang Phuc

Major: Computer Science

Code: 9480101

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

1. Objectives and Scope of Research

1.1. Objectives of the Thesis

The overall objective of the dissertation is to investigate and propose opinion mining models based on ensemble methods to overcome the limitations of traditional opinion mining models. The specific objectives are:

- To conduct a comprehensive review of research and systems in opinion mining, clarifying current achievements and limitations.
- To investigate an opinion model structured as a quintuple (entity, aspect, sentiment, holder, time) and techniques for extracting quintuples from text data.
- To propose opinion mining models based on combined influence to improve accuracy, generalization capability, and interpretability.
- To build opinion mining models based on classification forests, density forests, and manifold forests, suited to the structured nature of opinion data.
- To evaluate and compare the proposed models with existing opinion mining methods on real-world datasets.

1.2. Research Subjects

The research subjects of the dissertation include:

- Text data containing user opinions.
- Opinions at different levels of granularity, with a focus on the aspect level.
- Techniques for extracting and representing opinion data in quintuple form.
- Ensemble learning methods, particularly forest-based models, applied to opinion mining.

2. Research Methods

- Literature review method: analyzing and synthesizing research related to opinion mining, ensemble learning, and natural language processing.
- Experimental method: building, implementing (in R and Python), and evaluating the proposed models on real opinion data to verify the effectiveness of the approach.

3. Key Contributions

3.1. Scientific Contributions

- Proposes an Opinion Classification Forest (F_{CF}) model based on an opinion-quintuple matrix representation and supervised Shannon entropy, preserving semantic structure during tree construction and integrating Feature Importance and PDP-based interpretation.
- Proposes an Opinion Density Forest (F_{DF}) model integrating DBSCAN, Gaussian entropy and SHAP to handle class imbalance through the density distribution and to interpret the importance of and interactions among quintuple components.
- Proposes an Opinion Manifold Forest (F_{MF}) model that learns manifold structure directly from leaf co-occurrence statistics of an unsupervised random forest (relation matrix W , manifold distance $dist_M$), avoiding the global eigen-decomposition required by ISOMAP/UMAP, with DBSCAN clustering in the manifold space and results aggregated via stacking.

3.2. Practical Contributions

- Builds a pipeline converting four original review datasets (two English datasets, yelpHotelData and yelpResData — publicly released data from Mukherjee et al.; two Vietnamese datasets, VLSP-2018 Restaurant and VLSP-2018 Hotel — community-released VLSP-2018 data) into the standardized opinion-quintuple structure for model training and evaluation; the quintuple extraction model achieves $F1 = 0.78$ on the Twitter US Airline Sentiment dataset.
- Provides an empirical basis for comparing the three proposed forest models with existing machine learning, deep learning, and Transformer models (Logistic Regression, SVM, Random Forest, Gradient Boosting, LSTM, BiLSTM-Attention, BERT, RoBERTa, PhoBERT), supporting model selection for opinion mining systems with limited computational resources and high interpretability requirements.

4. Main Results

- The Opinion Classification Forest (F_{CF}) achieves an accuracy of 0.77 on yelpHotelData and yelpResData (published figures, row-level random split), with good interpretability via Feature Importance and PDP; further validation on Amazon Food Reviews and TripAdvisor Hotel Reviews achieves Accuracy of 0.80–0.82 and 0.91, respectively (with a Recall limitation of 0.50 on TripAdvisor due to label imbalance).
- The Opinion Density Forest (F_{DF}) achieves an average accuracy of 0.78 (published figures) without requiring a GPU; the SHAP analysis identifies Entity as the dominant feature (mean $|SHAP| \approx 0.087$).

- The Opinion Manifold Forest (F_{MF}) achieves the highest average accuracy among the three proposed models (0.80, published figures), runs on CPU and is interpretable through the relation matrix W and the manifold information gain IG_M .
- Successfully builds a pipeline converting original review data into the opinion-quintuple structure (e, a, s, h, t) across four datasets (yelpHotelData, yelpResData, VLSP-2018 Restaurant, VLSP-2018 Hotel); on VLSP-2018, F_{DF} reaches a Balanced Accuracy of 0.724 (Hotel) and 0.660 (Restaurant), while the text-based PhoBERT reference reaches 0.832 and 0.791.
- When re-evaluated with a review-level split (5 seeds $\times$ 10,000 reviews), the Balanced Accuracy of every quintuple-only model drops to about 0.51–0.57: F_{CF} and F_{DF} are comparable to Random Forest, while F_{MF} is about 0.02–0.03 below Random Forest; the text-based models (BERT, RoBERTa) reach 0.76–0.82. The dissertation quantifies the leakage of the row-level split (Accuracy 0.946 versus 0.611 with the same Random Forest) and recommends review-level splitting.

Under the published figures, the progressive accuracy gains from the Opinion Classification Forest (0.77) to the Opinion Density Forest (0.78) and the Opinion Manifold Forest (0.80) reflect increasingly refined exploitation of the data structure, while all three models retain interpretability through their tree, cluster and relation-matrix structures. Hence, the contribution of the three forest models lies mainly in the structured quintuple representation and in interpretability; their accuracy is bounded by the information carried by the four components (e, a, h, t).

Danang, ... 2026

Supervisors' Confirmation

Supervisor 1

Supervisor 2

Doctoral Candidate

Prof. Dr. Huynh Xuan Hiep

Dr. Le Thi My Hanh

Tran Quang Phuc