

UNIVERSITY OF DANANG
UNIVERSITY OF SCIENCE AND TECHNOLOGY

TRAN QUANG PHUC

OPINION MINING WITH ENSEMBLE METHODS

Major: Computer Science

Code: 9480101

SUMMARY OF DOCTORAL THESIS

Danang – 2026

The dissertation is completed at
UNIVERSITY OF SCIENCE AND TECHNOLOGY

Science instructor:

- 1. Prof.PhD. Huynh Xuan Hiep**
- 2. PhD. Le Thi My Hanh**

Reviewer 1:

Reviewer 2:

Reviewer 3:

The dissertation is defended before The Assessment Committee
at University of Science and Technology

At hour day month year

The dissertation can be found at:

- National Library;
- Center for Learning Information Resources and Communication,
The University of Da Nang

PREFACE

1. The urgency of the thesis

The rapid growth of the World Wide Web and social media platforms such as Twitter, Facebook, TripAdvisor, Yelp, and Amazon has generated enormous volumes of text data containing user opinions and reviews about products, services, and socioeconomic issues. Prominent examples include: Amazon Food Reviews [67] with more than 568,000 food reviews rich in product contexts; and TripAdvisor Hotel Reviews [2] containing 108,891 hotel reviews from 1,850 hotels covering 7 distinct aspects (value, rooms, location, cleanliness, check-in/front desk, service, business service). These online opinions increasingly play a pivotal role in supporting decision-making for both individuals and organizations.

However, current opinion mining systems still face three core challenges that remain insufficiently addressed: (1) Class imbalance: On Amazon Food Reviews, 5★ reviews account for 63.88% of all reviews, whereas 1★–2★ reviews represent only 14.42%; on Yelp, the positive class (4★–5★) accounts for about 63–65%, biasing models toward the majority class. (2) Non-linear relationships among opinion components: Sentiment depends simultaneously on the entity, aspect, holder, and time in complex non-linear ways that conventional vectorization methods fail to capture. (3) Lack of interpretability: Deep learning models such as BERT [81] and RoBERTa [83] achieve high accuracy but operate as black boxes, making it difficult to explain why a review is classified as positive or negative for a specific aspect, thereby limiting their applicability in domains that require transparency.

Critically, most existing studies treat opinion as a simple label, ignoring the rich structure of review data. In practice, users and businesses need to know: who wrote the review, about which aspect of which entity, with what sentiment, and at what time. This is the problem of opinion quintuple extraction and analysis that has not been systematically studied within an ensemble learning framework. Against this backdrop, the dissertation “Opinion Mining with Ensemble Methods” is conducted to propose ensemble learning-based models for opinion mining, addressing the identified limitations of existing approaches.

2. Objectives, Methods, Objects, and Scope of Research

2.1 Research Objectives

The general objective is to research and propose opinion mining models based on ensemble learning to overcome limitations of traditional models. Specific objectives include: (1) Surveying existing opinion mining research and systems, clarifying achievements and current limitations; (2) Studying the structured opinion quintuple model $o = (e, a, s, h, t)$ and techniques for extracting quintuples from text; (3) Proposing opinion mining models based on classification forests, density forests, and manifold forests suited to structured opinion data; (4) Building and converting opinion datasets into quintuple format from standard benchmarks; (5) Evaluating and comparing proposed models with existing approaches on real-world datasets.

2.2.2.2 Research Subjects

The research subjects of this dissertation include: (1) textual data containing user opinions; (2) aspect-level opinion mining, which represents the finest granularity of opinion analysis and provides the highest practical value; (3) techniques for extracting and representing opinion information as opinion quintuples; and (4) homogeneous ensemble learning methods based on decision trees.

2.3 Research Methods

This dissertation adopts a combined research methodology consisting of: (1) a literature review that analyzes and synthesizes more than 150 studies on opinion mining, ensemble learning, and natural language processing; and (2) an experimental approach implemented in R and Python, in which the proposed models are evaluated on six real-world review datasets (four English datasets and two Vietnamese VLSP-2018 datasets) to validate their effectiveness.

2.4.2.4 Research Scope

The scope of this dissertation is limited to the extraction and standardization of opinion quintuples from textual data, the investigation of ensemble learning methods based on the homogeneous base learner of decision trees, the application of the Opinion Classification Forest, Opinion Density Forest, and Opinion Manifold Forest to aspect-level opinion mining, and the evaluation of the proposed models using real-world service and product review datasets.

3. Research Problem and Contributions

The research question is: “Can decision tree-based ensemble learning methods exploit the quintuple structure $o = (e, a, s, h, t)$ to simultaneously improve classification accuracy, address class imbalance, and provide transparent model interpretability?”

The dissertation makes four main contributions:

(1) Opinion Classification Forest (F_{CF}). An Opinion Classification Forest model is proposed based on an $N \times 5$ opinion quintuple matrix and supervised Shannon entropy $H(O)$. The model addresses the research gap in structured sentiment representation for aspect-level opinion classification while providing interpretability through feature importance and partial dependence plots (Papers 2, 3).

(2) Opinion Density Forest (F_{DF}). An Opinion Density Forest model is developed by integrating DBSCAN clustering, Gaussian entropy $H_D(O)$ and SHAP-based interpretation. The model handles class imbalance through the density distribution of opinion clusters and uses SHAP values and Shapley interaction indices to interpret the importance and interactions of the quintuple components (Papers 4, 5).

(3) Opinion Manifold Forest (F_{MF}). An Opinion Manifold Forest model is introduced using a manifold relation matrix $W \in R^{N \times N}$, manifold distance $dist_{\mathcal{M}}$, and a stacking meta-learner $\mathbf{M}$. The manifold structure is learned directly from leaf co-occurrence in an unsupervised random forest, without solving a global eigenvalue problem as in ISOMAP or UMAP, in order to capture non-linear relationships among the quintuple components; the

experiments show the feasibility of this approach and identify the cost of the matrix W as the main barrier to scaling (Paper 6).

(4) Opinion Quintuple Datasets. A procedure for converting review data into the opinion quintuple structure is developed and applied to two English datasets published by Mukherjee et al., yelpHotelData (over 880,000 quintuples) and yelpResData (over 14 million quintuples), and to two Vietnamese datasets, VLSP-2018 Restaurant (over 15,000 review–aspect pairs) and VLSP-2018 Hotel (over 23,000 pairs), for training and evaluating the proposed models (Papers 4, 5, 6, 9); the quintuple extraction model ($F1 = 0.78$ on Twitter US Airline Sentiment) is published in Papers 7, 8.

Under the row-level random split of the quintuples used in the published papers (Table 4.4 of the full dissertation), the average accuracies on the two Yelp datasets are F_{CF} (0.77), F_{DF} (0.78) and F_{MF} (0.80). All proposed forest models run on standard CPUs; in particular, F_{MF} eliminates the need for GPU acceleration and provides transparent model interpretation through the manifold relation matrix W and the manifold information gain criterion (IG_M).

When re-evaluated with a review-level split (all quintuple rows of a review belong to the same subset; Appendix J of the full dissertation, 5 seeds $\times$ 10,000 reviews), the Balanced Accuracy of every model that uses only the quintuples drops to about 0.51–0.57: F_{CF} and F_{DF} are comparable to Random Forest (differences not statistically significant), while F_{MF} is about 0.02–0.03 below Random Forest because its manifold clustering step is learned on a subsample; the text-based models BERT and RoBERTa reach 0.76–0.82. A controlled check with the same Random Forest and the same 112,275 quintuples gives an Accuracy of 0.946 under the row-level split and 0.611 under the review-level split; the dissertation therefore recommends review-level splitting when evaluating models on quintuple data.

4. Dissertation Structure

In addition to the introduction, conclusions, and directions for further research, the dissertation comprises 4 chapters: Chapter 1 (Theoretical Foundations of Opinion Mining and Ensemble Learning); Chapter 2 (Opinion Mining with Classification Forest); Chapter 3 (Opinion Mining with Density Forest); Chapter 4 (Opinion Mining with Manifold Forest).

CHAPTER 1. THEORETICAL FOUNDATIONS OF OPINION MINING AND ENSEMBLE LEARNING

The main content encompasses the research results of Papers 1, 7, 8, and 9. This chapter establishes the theoretical foundation for the entire dissertation.

Chapter 1 at a glance – Problem: establishing the theoretical foundation and automatically extracting opinion quintuples (e, a, s, h, t) from text. Method: a review of opinion mining and ensemble learning; a probabilistic quintuple extraction framework combined with dependency parsing. Results: on Twitter US Airline Sentiment (4,500 annotated samples), the extraction model achieves $F1 = 0.78$ ($P = 0.81, R = 0.75$), 8.6 ms per tweet and 210 MB of memory. Contribution: the quintuple representation and data foundation for the three forest models in Chapters 2–4 (Papers 7, 8, 9).

1.1. Opinion Mining

1.1.1. The Opinion Quintuple Model

According to [1], an opinion can be formally represented as an opinion quintuple

$$o = (e, a, s, h, t),$$

where e denotes the target entity being evaluated, a represents the aspect (or attribute) of the entity, s indicates the sentiment orientation expressed toward aspect a of entity e , h identifies the opinion holder (i.e., the individual or organization expressing the opinion), and t denotes the time at which the opinion is expressed.

Given an opinion document

$$d_i,$$

the extracted opinion information is represented as a set of opinion quintuples

$$O = \{o_1, o_2, \dots, o_m\},$$

For sentiment classification, review ratings are transformed into sentiment labels using a three-level mapping strategy, where 1★–2★ reviews are assigned to the negative class, 4★–5★ reviews are assigned to the positive class, and 3★ reviews are excluded (for Amazon Food Reviews and TripAdvisor Hotel Reviews) because they generally represent neutral or ambiguous opinions, which will be addressed in future extensions of the dissertation; for the two Yelp datasets (Chapters 3 and 4), an asymmetric threshold is applied and 3★ reviews are merged into the negative class.

Compared with traditional sentiment representations, the opinion quintuple model provides a richer semantic description by simultaneously modeling the target entity, aspect, sentiment, opinion holder, and temporal information. However, the resulting five-dimensional representation introduces complex dependencies among opinion components. In particular, sentiment orientation is jointly influenced by multiple contextual factors rather than by individual components independently. These non-linear relationships cannot be effectively captured using conventional feature engineering or linear vector representations, thereby motivating the development of the three proposed ensemble learning models, namely the Opinion Classification Forest, Opinion Density Forest, and Opinion Manifold Forest.

1.1.2. Opinion Mining Tasks

Opinion mining consists of eight major tasks [1]: (1) entity extraction and resolution; (2) aspect extraction and resolution; (3) opinion expression extraction; (4) aspect-based sentiment classification; (5) opinion holder identification; (6) time identification; (7) opinion summarization; and (8) opinion comparison. This dissertation focuses on Aspect-Based Sentiment Analysis (ABSA) because it provides the finest-grained sentiment information and offers the highest practical value for e-commerce applications.

1.1.3. Levels of Opinion Mining

Opinion mining can be performed at three different levels of granularity.

The document level aims to determine the overall sentiment expressed in an entire document toward a particular entity.

The sentence level focuses on identifying the sentiment polarity of individual sentences by classifying each sentence as positive, negative, or neutral.

The aspect level, also known as Aspect-Based Sentiment Analysis (ABSA), performs fine-grained sentiment analysis by identifying opinion targets and determining the sentiment expressed toward each specific aspect of an entity.

This dissertation focuses on aspect-level opinion mining because it provides the most detailed and actionable sentiment information. Compared with document-level and sentence-level analysis, aspect-level opinion mining enables a more accurate understanding of customer preferences and dissatisfaction regarding individual product or service attributes. Consequently, it has become a fundamental technique in intelligent e-commerce platforms, recommender systems, customer feedback analysis, and business decision support.

1.2. Ensemble Learning and Forest Models

1.2.1. Decision Trees as Homogeneous Base Learners

A decision tree is a hierarchical predictive model consisting of internal decision nodes, branches, and leaf nodes. Each internal node represents a decision based on a feature, each branch corresponds to the outcome of the decision, and each leaf node represents the predicted class label. Owing to their simple structure and rule-based representation, decision trees provide high model interpretability through a set of human-readable IF-THEN decision rules.

For a classification problem with C classes, the optimal splitting criterion is determined using Shannon entropy

$$H(O) = -\sum_{c \in C} p(c) \log p(c)$$

where $p(c)$ denotes the probability of class c . The quality of a candidate split is evaluated using Information Gain

$$IG = H(O) - \sum_{i \in \{L, R\}} \frac{|O^i|}{|O|} H(O^i),$$

which measures the reduction in uncertainty after partitioning the training data.

Although decision trees are computationally efficient and highly interpretable, they are prone to overfitting because small variations in the training data may produce substantially different tree structures. Consequently, decision trees are commonly employed as homogeneous base learners within ensemble learning frameworks to improve generalization performance while preserving interpretability.

1.2.2. Bootstrap and Ensemble Learning Techniques

Bootstrap sampling [13] is a resampling technique that generates multiple training subsets by randomly sampling instances from the original dataset with replacement. Given an original dataset O , the Bootstrap procedure constructs B training subsets

$$O_1^*, O_2^*, O_B^*, \dots$$

each having the same size as the original dataset while containing approximately 63.2% unique samples.

Bootstrap serves as the foundation of several ensemble learning techniques. Bagging [62] trains multiple independent base learners on different Bootstrap subsets and combines their predictions using majority voting, thereby reducing model variance without significantly increasing bias. Boosting [94], in contrast, constructs base learners sequentially, where each subsequent learner focuses on samples that were incorrectly classified by previous learners, thereby reducing prediction bias. Stacking [63] adopts a different strategy by training a meta-learner

$$\mathbf{M},$$

which learns how to combine the predictions produced by multiple heterogeneous or homogeneous base learners using a validation dataset.

1.2.3. Opinion Classification, Opinion Density, and Opinion Manifold Forests

Forest-based ensemble learning methods have evolved from supervised classification toward unsupervised density estimation and manifold representation learning.

The Opinion Classification Forest [65][66] extends the conventional Classification Forest by integrating opinion quintuple representations into a Bagging framework composed of decision trees. During tree construction, each node is split using supervised Shannon entropy

$$H(O),$$

while random feature selection is employed to improve model diversity and generalization performance.

The Opinion Density Forest [66][68][69][70] further extends this framework to unsupervised learning by replacing Shannon entropy with Gaussian entropy

$$H_D(O),$$

thereby enabling the forest to capture the intrinsic density distribution of opinion data without requiring class labels. This property makes the model particularly suitable for handling class imbalance and discovering latent opinion clusters.

Building upon the previous models, the Opinion Manifold Forest [66][71] introduces manifold learning into the forest framework. The model constructs a manifold relation matrix

$$W \in R^{N \times N},$$

which represents the latent geometric relationships among opinion samples through a co-leaf occurrence mechanism. Based on this manifold representation, the proposed model estimates manifold distance

$dist_{\mathcal{M}}$, thereby preserving the intrinsic nonlinear structure of the opinion quintuple space and improving classification performance.

The three proposed forest models represent a progressive evolution of ensemble learning strategies. The Opinion Classification Forest employs supervised Shannon entropy

$$H(O),$$

to perform aspect-level opinion classification. The Opinion Density Forest replaces this criterion with Gaussian entropy

$$H_D(O),$$

to model the density distribution of unlabeled opinion data. Finally, the Opinion Manifold Forest utilizes manifold Gaussian entropy

$$H_{\mathcal{M}}(O),$$

to exploit the underlying manifold structure of opinion quintuples. This progression reflects the transition from conventional supervised learning to density-aware learning and ultimately to manifold-based representation learning for complex non-linear opinion data.

1.3. Ensemble Learning Approach

The dissertation proposes four selection criteria:

(1) Structured tabular data: when data has been extracted into structured features such as opinion quintuples, tree/forest methods remain a strong choice, particularly on medium/small datasets [141];

(2) Interpretability requirements: when applications require classification explanations (healthcare, finance, management), Forest + SHAP provides transparent interpretation that end-to-end deep learning models can hardly provide [144][145];

(3) Computational resources: when no dedicated GPU is available, random forests and density forests are a suitable choice because they do not require GPUs [143];

(4) Conversely, deep Transformer models are preferred for raw text without preprocessing, handling sarcasm/negation/implicit opinion, and when sufficiently large labeled training datasets are available.

Accordingly, this dissertation adopts an ensemble learning approach because the input data are first transformed into structured opinion quintuples. This representation enables efficient learning using tree-based ensemble models while satisfying two practical requirements: model interpretability and computational efficiency.

1.4. Related Work

A survey of over 150 works across five major research groups is summarized in Table 1.1.

Table 1.1. Summary of Related Research Groups

Research Group	Representative Works	Strengths	Main Limitations
Traditional ML	SVM [86], NB [204], LR [203]	Effective on small data; Simple to implement	Cannot handle class imbalance; Lacks opinion quintuple representation
Ensemble Learning	RF [65], GBC [84], Stacking [63]	High accuracy; Stable performance	No density clustering integration; Not applied to quintuples
Deep Learning RNN/CNN	LSTM [85], BiLSTM [88]	Captures sequential context	Black-box; Requires large data; Difficult to interpret

Research Group	Representative Works	Strengths	Main Limitations
Transformer / LLM	BERT [81], RoBERTa [83], GPT	Highest accuracy	Black-box; Requires GPU; High cost; Not interpretable
Manifold / GCN	GCN [137] Manifold [118][129]	Exploits non-linear structure	Complex; Not applied to 5-dimensional quintuples

Three research gaps are identified: (1) No model integrates density clustering with random forests to simultaneously handle class imbalance and exploit opinion cluster structure; (2) Manifold learning has not been applied to 5-dimensional opinion quintuple representation; (3) State-of-the-art ABSA models lack transparent aspect-level interpretability. These three gaps are addressed by F_{CF} (Chapter 2), F_{DF} (Chapter 3), and F_{MF} (Chapter 4) respectively.

1.5. Model Evaluation

The effectiveness of the proposed models is evaluated using four standard performance metrics, as summarized in Table 1.2.

Table 1.2. Standard Performance Evaluation Metrics

Metric	Formula	Description
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Overall classification accuracy. This metric may become unreliable for highly imbalanced datasets ($\rho > 3$).
Precision (P)	$TP/(TP+FP)$	Measures the proportion of correctly predicted positive samples. It is particularly important when false positives incur high costs.
Recall (R)	$TP/(TP+FN)$	Measures the ability to correctly identify positive samples. It is particularly important when false negatives incur high costs.
F1-score	$2PR/(P+R)$	Harmonic mean of Precision and Recall. It is adopted as the primary evaluation metric for imbalanced classification tasks.

Additionally, Cohen's Kappa coefficient [79] $\kappa = (P_o - P_e)/(1 - P_e)$ is used to assess annotation agreement ($\kappa \geq 0.80$ indicates very good agreement). To provide a more comprehensive evaluation under severe class imbalance, Appendix F further reports F1-macro, F1-positive (F1-pos), and F1-negative (F1-neg) on the VLSP-2018 Restaurant and VLSP-2018 Hotel datasets, where the imbalance ratio is approximately $\rho \approx 10.5$ (Restaurant) and $\rho \approx 3.2$ (Hotel). These additional metrics provide a more balanced assessment of classification performance across both majority and minority classes.

1.6. Opinion Quintuple Extraction

1.6.1. Extraction Model

The automatic quintuple extraction model is based on a joint probability decomposition:

$$P(e, a, s, h, t|x) = P_e(e|x) \cdot P_a(a|e, x) \cdot P_s(s|e, a, x) \cdot P_h(h|s, G) \cdot P_t(t|x)$$

where P_e, P_a, P_s, P_h, P_t denote the conditional probabilities of extracting the entity, aspect, sentiment, opinion holder, and time, respectively. Here, $G=(V, E)$ represents the dependency graph obtained from syntactic dependency parsing.

The extraction framework consists of six sequential stages. First, the input text is preprocessed through tokenization, normalization, and linguistic preprocessing. Second, dependency parsing is performed to construct the syntactic dependency graph. Third, opinion entities are identified using Pointwise Mutual Information (PMI) together with domain-specific lexicons. Fourth, opinion aspects are extracted according to syntactic dependency relations between entities and candidate aspect terms. Fifth, sentiment polarity is determined by combining lexical resources, including VADER and SentiWordNet. Finally, the opinion holder and temporal information are extracted to complete the opinion quintuple representation. To improve the reliability of the extracted quintuples, each candidate quintuple is assigned a confidence score, and only those satisfying

$$Conf(o) \geq \tau,$$

are retained for subsequent learning. This confidence-based filtering reduces noisy annotations and ensures that the proposed Opinion Classification Forest, Opinion Density Forest, and Opinion Manifold Forest are trained using high-quality structured opinion data.

1.6.2. Quintuple Extraction Experimental Results

Dataset: Twitter US Airline Sentiment [80] with 14,640 tweets about 6 US airlines (United, US Airways, American, Southwest, Delta, Virgin America); 4,500 tweets manually annotated via Label Studio by three independent annotators.

Table 1.3. Experimental Results of Opinion Quintuple Extraction

Scenario	Description	Main Results
SC1: Annotation	4,500 tweets manually annotated by three independent annotators using Label Studio	Cohen's Kappa $\kappa > 0.79$ (entity, aspect, sentiment); $\kappa = 0.76$ (holder); $\kappa = 0.79$ (time)
SC2: Extraction	Quintuple extraction on 14,640 tweets; evaluation on 4,500 annotated tweets (70/15/15 split)	$F_1 = 0.78$ ($P=0.81, R=0.75$); Entity: $F_1=0.86$ (highest); Holder: $F_1=0.68$ (lowest)
SC3: Comparison	Comparison with rule-based method and spaCy	Proposed (0.78) > spaCy (0.70) > Rule-based (0.66); Processing time: 8.6 ms/tweet; memory: 210 MB

The proposed opinion quintuple extraction model achieved an overall $F_1 = 0.78$, compared with the rule-based approach (0.66) and spaCy (0.70), while maintaining an average processing time of 8.6 ms per tweet and a memory footprint of 210 MB, making it suitable for real-time applications.

Among the five opinion components, entity extraction achieved the highest performance with an $F_1 = 0.86$, mainly because it relies on domain-specific lexicons. In contrast, holder identification achieved the lowest performance ($F_1 = 0.68$) due to the opinion holder often being implicit and difficult to identify in social media text.

The Cohen's Kappa coefficient for the entity, aspect and sentiment components was greater than 0.79, indicating very good inter-annotator agreement [79].

1.7. Chapter conclusions

Chapter 1 establishes the theoretical foundation of the dissertation by introducing the opinion quintuple model, $o = (e, a, s, h, t)$ as the unified representation of opinion data, systematically reviewing ensemble learning methods based on decision trees, and clarifying the differences among Shannon Entropy, Gaussian Entropy, and Manifold Gaussian Entropy, which are employed in the three proposed models.

Furthermore, the chapter reviews related studies and identifies three research gaps, including handling class imbalance, exploiting the manifold structure of opinion data, and improving model interpretability.

In addition, the proposed opinion quintuple extraction framework achieved an overall F1-score = 0.78, providing high-quality input data and establishing the foundation for constructing opinion quintuple extraction models on other datasets.

CHAPTER 2. OPINION MINING WITH CLASSIFICATION FORESTS

The main content encompasses the research results of Papers 2 and 3.

2.1. Introduction

Chapter 2 proposes the Opinion Classification Forest model to address the first research gap, namely supervised sentiment representation and classification on opinion quintuple data with interpretability. The core contribution is the representation of opinion sentences as an $N \times 5$ opinion quintuple matrix, rather than a single feature vector, thereby preserving the multidimensional semantic structure of opinion information.

Chapter 2 at a glance – Problem: supervised opinion polarity classification on quintuple data with interpretable results. Method: the Opinion Classification Forest on the $N \times 5$ quintuple matrix (Bagging, Shannon entropy), interpreted through feature importance, minimal depth and PDP. Results: Accuracy 0.80–0.82 on Amazon Food Reviews; Accuracy 0.91 on TripAdvisor Hotel Reviews, with a negative-class Recall of only 0.50 due to label imbalance. Contribution: the first contribution of the dissertation (Papers 2, 3).

2.2. Opinion Quintuple Matrix

Given a set of opinion quintuples

$$O = (o_1, o_2, o_3, \dots, o_N),$$

where $o_i = (e_i, a_i, s_i, h_i, t_i), i = 1, 2, \dots, N$.

the opinion quintuple matrix O_{mat} of size $N \times 5$ is defined as

$$O_{mat} = \begin{bmatrix} e_1 & a_1 & s_1 & h_1 & t_1 \\ e_2 & a_2 & s_2 & h_2 & t_2 \\ e_3 & a_3 & s_3 & h_3 & t_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_N & a_N & s_N & h_N & t_N \end{bmatrix},$$

where each row represents an opinion quintuple, and each column corresponds to a feature component. All feature values are numerically encoded using label encoding before being input into the classification model. To the best of the author's knowledge, this is one of the few representations in Aspect-Based Sentiment Analysis (ABSA) that simultaneously employ all five components of the opinion quintuple as classification features.

2.3. Opinion Classification Forest Model

2.3.1. Opinion Classification Tree

An Opinion Classification Tree is constructed by recursively splitting each node v into two child nodes, L and R . The splitting criterion is based on Shannon Entropy (Equation (1.4)) and the information gain (Equation (1.6)), where

$$s \in \{\text{positive, negative}\}.$$

The resulting decision tree can be represented as a set of IF–THEN decision rules, where each path from the root node to a leaf node corresponds to one classification rule. For example: “IF Entity=Hotel AND Aspect=Room AND Score $\geq$ 4 THEN sentiment=positive”.

2.3.2. Opinion Classification Forest

The Opinion Classification Forest, denoted by F_{CF} , is constructed through the following four steps:

Step 1 – Bootstrap Sampling.

Generate B bootstrap subsets,

$$O^*_1, O^*_2, \dots, O^*_B$$

by sampling with replacement from the opinion quintuple matrix O_{mat} . Each subset contains N samples (with duplicates), corresponding to approximately 63.2% unique samples.

Step 2 – Random Feature Selection.

At each split node, randomly select two features (the square root of the number of input features) from the four candidate features of the opinion quintuple instead of considering all features, thereby increasing the diversity among decision trees.

$$m_{try} = \lfloor \sqrt{4} \rfloor = 2$$

Step 3 – Opinion Classification Tree Construction.

For each bootstrap subset O^*_b , an opinion classification tree T_b is constructed using the randomly selected m_{try} features at every split. The tree is recursively grown by maximizing the information gain until the stopping criterion is satisfied.

Step 4 – Prediction Aggregation.

The predictions of $n=100$ decision trees are combined using the majority voting strategy:

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}} \sum_{i=1}^n \mathbb{1}[T_i(o) = c]$$

2.3.3. Computational Complexity

Training Complexity.

Each decision tree is trained on a bootstrap subset containing approximately $N' \approx 0.63N$ samples. At each split node, m_{try} features are randomly selected, and the optimal splitting threshold is determined. For a balanced decision tree, the training complexity is given by

$$\Theta(n \times N' \times m_{try} \times h) = O(n \times N \times m_{try} \times \log N)$$

where h denotes the average depth of the tree.

Prediction Complexity.

Each unseen sample is classified by traversing all n decision trees. Since each tree has an average depth of $O(\log N)$, the overall prediction complexity for a single sample is

$$\Theta(n \times \log N)$$

2.3.4. Overall Framework

The Opinion Classification Forest model consists of three stages:

1. Opinion Modeling: Construct the opinion quintuple matrix O_{mat} and generate bootstrap subsets through bootstrap sampling.
2. Opinion Discovery and Representation: Build the Opinion Classification Forest F_{CF}
3. Opinion Orientation: Predict the sentiment label using the majority voting strategy, and interpret the classification results through feature importance, Partial Dependence Plots (PDPs), and feature interaction analysis.

2.4. Interpretation of the Opinion Classification Forest

The proposed Opinion Classification Forest is interpreted at two levels:

1. *Tree Level*: Decision rules are represented as IF–THEN rules along the path from the root node to a leaf node. For example:

IF Entity = Hotel AND Aspect = Room AND Score ≥ 4 THEN Sentiment = Positive.

In addition, the minimum tree depth and feature importance at each split node are analyzed.

2. *Forest Level*: Model interpretation is performed through feature importance, feature interaction analysis, and Partial Dependence Plots (PDPs), which illustrate the nonlinear relationships between individual features and the predicted sentiment probability.

The analysis shows that Time has the smallest average minimum depth (1.222) and appears as the root node 146 times, indicating that the time of review is the most influential feature for sentiment classification. Holder ranks second, with an average minimum depth of 1.286 and 119 occurrences as the root node. Strong negative interactions between Entity and Aspect (−1.728) and between Entity and Time (−0.898) reveal the nonlinear structure of the five-dimensional opinion quintuple space, providing the experimental basis for the subsequent models presented in Chapters 3 and 4.

Table 2.1. Feature Importance of the F_{CF} Model on the TripAdvisor Hotel Reviews Dataset (Source: Table 2.11 of the full dissertation)

Feature	Average Minimum Depth	Times as Root Node	Information Gain (IG) Importance	Permutation Importance	Role in Classification
Time (t)	1.2220	146	89.3295	0.0000	Most influential feature; captures seasonal effects, temporal trends, and review timing.
Holder (h)	1.2860	119	92.1901	0.0021	Reflects the consistent evaluation behavior of individual users.
Aspect (a)	1.4820	110	28.6544	0.0043	Determines aspect-specific sentiment polarity.
Entity (e)	1.2900	125	—	0.0122	Represents the target entity and exhibits strong

Feature	Average Minimum Depth	Times as Root Node	Information Gain (IG) Importance	Permutation Importance	Role in Classification
					interaction with the Aspect feature (−1.728).

The Partial Dependence Plots (PDPs) reveal that: (1) in the {Time, Holder} space of TripAdvisor Hotel Reviews (Figures 2.11–2.14 of the full dissertation), Holder has a strong influence, with some users showing stable sentiment over time while others change markedly with Time; (2) in the {Entity, Time} space, some entities keep a consistently high positive probability regardless of time, while others fluctuate strongly; and (3) Time:Holder (0.321) is the only positive interaction (Table 2.12 of the full dissertation): when Time is at the root, Holder tends to appear earlier in the tree.

2.5. Experimental Results

2.5.1. Datasets

Table 2.2. Dataset Information for Chapter 2 Experiments

Dataset	Original Size	After Preprocessing	Split Ratio
Amazon Food Reviews [67]	568,454 reviews	8,020 quintuple samples (stratified sampling 1:1)	80% train/20% test
TripAdvisor Hotel Reviews [2]	108,891 reviews	45,706 quintuple samples (neutral 3★ removed)	80% train/20% test

2.5.2. Experimental Results on Amazon Food Reviews

Chapter 2 conducts three experimental scenarios across two datasets. The Amazon Food Reviews dataset [67] comprises 568,454 reviews with an imbalanced distribution (5-star reviews account for 63.88%, while 1–2-star reviews account for only 14.42%); after applying balanced sampling, the resulting experimental set achieves an approximate positive:negative ratio of 1:1. After preprocessing, the TripAdvisor Hotel Reviews dataset yields 45,706 opinion quintuple samples.

Scenario 1 – Aspect-based food classification (Table 2.8 of the full dissertation): the proposed model achieves Accuracy = 0.80, Precision = 0.80, Recall = 0.76, and F1 = 0.78. The small Precision–Recall gap ($0.80 - 0.76 = 0.04$) indicates balanced classification between the two classes, a result of the balanced sampling strategy.

Scenario 2 – Classification based on time and reviewer identity (Table 2.9 of the full dissertation): Accuracy increases to 0.82 and Precision reaches 0.85, while Recall decreases to 0.67. The larger Precision–Recall gap ($0.85 - 0.67 = 0.18$) indicates that the Time and Holder features are effective at identifying the positive class but miss more instances of the negative class. Feature importance analysis (Table 2.11 of the full dissertation) shows that Time has the smallest average minimum depth (1.2220) and serves as the root node 146 times, making it the strongest classification feature, while Holder ranks second (1.2860, 119 times), reflecting the consistent evaluation style of individual reviewers. The strongest positive

interaction is Time:Holder (0.321), while the strongly negative Entity:Aspect interaction (−1.728) (Table 2.12 of the full dissertation) reveals a nonlinear relationship among the quintuple components and directly motivates the Opinion Density Forest (Chapter 3) and the Opinion Manifold Forest (Chapter 4).

Table 2.3. F_{CF} model results (Sources: Tables 2.8–2.10 of the full dissertation)

Scenario	Dataset	Accuracy	Precision	Recall	F_1
SC1. Aspect-based food classification	Amazon Food Reviews	0.80	0.80	0.76	0.78
SC2. Time/Reviewer classification	Amazon Food Reviews	0.82	0.85	0.67	0.75
SC3. TripAdvisor Hotel (imbalanced 4:1)	TripAdvisor Hotel Reviews	0.91	0.96	0.50	0.66

2.5.3. Experimental Results on TripAdvisor Hotel Reviews

On TripAdvisor Hotel (45,706 quintuple samples, 80/20 split): Accuracy = 0.91, but the negative-class Recall is only 0.50 and $F_1 = 0.66$, which illustrates the accuracy paradox. The 4:1 class imbalance biases the model towards the positive class to maximize Accuracy, so that half of the negative samples are missed, confirming deeper imbalance challenges in hotel review data compared to Amazon.

The experimental results show that the proposed Opinion Classification Forest achieves an Accuracy of 0.80–0.82 on the Amazon Food Reviews dataset and 0.91 on the TripAdvisor Hotel Reviews dataset. However, the Accuracy of 0.91 on the TripAdvisor dataset reflects the accuracy paradox caused by the 4:1 class imbalance, where the Recall is only 0.50 and the F_1 -score is 0.66. A comprehensive comparison between the proposed F_{CF} model and other machine learning models, including Random Forest (RF), Gradient Boosting (GBM), BERT, RoBERTa, BiLSTM-Attention, LSTM, and SVM, is presented in Chapters 3 and 4 using the Yelp opinion quintuple datasets.

The proposed Opinion Classification Forest has a training complexity of $\mathcal{O}(n \times N \times m_{try} \times \log N)$ and a prediction complexity of $\mathcal{O}(n \times \log N)$ which are equivalent to those of the standard Random Forest [65]. The computational complexity of the F_{CF} model is low, requiring only conventional CPU resources. Therefore, the F_{CF} model is computationally efficient and serves as an appropriate foundation for the more advanced models proposed in the subsequent chapters.

2.6. Chapter Conclusion

Chapter 2 proposes the Opinion Classification Forest model, whose core contribution is the representation of opinion quintuples as an $N \times 5$ matrix. The proposed F_{CF} achieves an Accuracy of 0.80–0.82 on the Amazon Food Reviews dataset and an Accuracy of 0.91 on the TripAdvisor Hotel Reviews dataset, although the Recall is only 0.50 and the F_1 -score is 0.66 owing to the 4:1 label imbalance, while still providing rich model interpretability.

However, the proposed F_{CF} does not effectively address class-imbalanced data and does not exploit the nonlinear cluster structure of opinion data. These two limitations are targeted in Chapters 3 and 4, respectively. The results presented in Table 2.10 of the full dissertation demonstrate the limitations of the FCF model on real-world imbalanced datasets and explain the motivation for developing the Opinion Density Forest presented in Chapter 3.

Furthermore, the strong negative interactions between Entity and Aspect (-1.728) and between Entity and Time (-0.898) (Table 2.12 of the full dissertation) reveal the nonlinear structure of the five-dimensional opinion quintuple space, providing empirical motivation for the development of the Opinion Manifold Forest proposed in Chapter 4.

Therefore, Chapter 2 not only proposes the Opinion Classification Forest but also provides quantitative evidence for the research gaps addressed in Chapters 3 and 4.

CHAPTER 3. OPINION MINING WITH DENSITY FORESTS

The main content of this chapter is based on the research results reported in Papers 4 and 5.

3.1. Introduction

Chapter 3 proposes the Opinion Density Forest model, which integrates DBSCAN clustering with SHAP-based model interpretation. The proposed model automatically discovers opinion clusters ($K = 7 - 9$ clusters), constructs an Opinion Density Forest using Gaussian Entropy $H_D(O)$, and provides model interpretability through SHAP values ϕ_j and Shapley interaction values $\phi_{i,j}$. The proposed model achieves an average Accuracy of 0.78 on the two Yelp datasets (published figures, row-level split) and runs on standard CPUs without GPU acceleration.

Chapter 3 at a glance – Problem: class imbalance and density-based cluster structure in quintuple data. Method: the Opinion Density Forest – DBSCAN clustering (7–9 clusters), one forest per cluster with Gaussian entropy, interpreted through SHAP values and Shapley interaction indices. Results: Accuracy 0.75 (Hotel) and 0.80 (Restaurant), 0.78 on average (published figures, row-level split), average F1 0.68; Entity and Time have the largest mean |SHAP|. Under the review-level split (Appendix J of the full dissertation), the density forest reaches a Balanced Accuracy comparable to Random Forest. Contribution: the second contribution of the dissertation (Papers 4, 5).

3.2. Quintuple Representation and Normalization

Each quintuple $o_i = (e_i, a_i, s_i, h_i, t_i)$ is mapped to a numerical vector in $\mathbb{R}^5$ via label encoding. Z-score normalization $\hat{x} = (x - \mu)/\sigma$ ensures uniform scaling before DBSCAN clustering.

3.3. Opinion Density Forest Model

3.3.1. Density-Based Opinion Clustering Using DBSCAN

DBSCAN [72] detects clusters based on ε (neighborhood radius) and $MinPts$. Point o is a core point if $|N_\varepsilon(o)| \geq MinPts$. Settings: $\varepsilon = 0.5$, $MinPts = 5$; results in $K = 7$ clusters (yelpHotelData) and $K = 9$ clusters (yelpResData).

3.3.2. Opinion Density Tree

The density tree is constructed in an unsupervised manner using Gaussian Entropy:

$$H(O) = \frac{1}{2} \log((2\pi e)^d |\Lambda(O)|),$$

where $d = 4$ is the dimensionality of the input features (e, a, h, t) and $|\Lambda(O)|$ is the determinant of the $d \times d$ covariance matrix at the node.

The density information gain criterion is:

$$IG_D(O_j, O_{jL}, O_{jR}) = H_D(O_j) - \frac{|O_{jL}|}{|O_j|} H_D(O_{jL}) - \frac{|O_{jR}|}{|O_j|} H_D(O_{jR}),$$

where the optimal split is obtained by maximizing IG_D , thereby partitioning the data into two subsets with the highest density.

3.3.3. Opinion Density Forest

The Opinion Density Forest consists of $n=100$ density trees, each trained independently on the density clusters C_k :

$$p_{DF}(o) = \frac{1}{n} \sum_{k=1}^n p_k(o).$$

where $p_{DF}(o)$ denotes the aggregated forest density, i.e. the average of the densities estimated by the individual density trees.

The sentiment label of an opinion sample is determined according to the cluster with the highest estimated density.

3.4. Interpretability with SHAP and Shapley Interaction

The SHAP value [125] of feature j is defined as

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(o_{S \cup \{j\}}) - f_S(o_S)]$$

where φ_j represents the contribution of feature j to the prediction of the proposed model.

The Shapley interaction value, denoted by $\varphi_{i,j}$, measures the joint contribution of the feature pair (i, j) and is computed according to Equation (3.28):

$$I_{i,j}(o) = \sum_{S \subseteq F \setminus \{i,j\}} \frac{|S|! (|F| - |S| - 2)!}{2(|F| - 1)!} \delta_{ij}(o, S)$$

3.5. Computational Complexity

Training Complexity.

The training complexity of the proposed Opinion Density Forest is determined by the combined computational cost of constructing the density trees and performing DBSCAN clustering at the leaf nodes. The overall training complexity is given by

$$\Theta(n \times m_{try} \times N \times \log N) + \Theta(n \times L \times N \times \log N)$$

where n denotes the number of density trees, m_{try} is the number of randomly selected features at each split, N is the number of training samples, and L represents the average number of leaf nodes per tree (approximately $N/\text{min_split}$).

Prediction Complexity.

The computational complexity of sentiment prediction and opinion orientation consists of three independent execution stages. Given n density trees, K opinion aspects to be classified, and N training samples, the overall prediction complexity is

$$\Theta(n \times \log N) + \Theta(K \times n \times \log N) + \Theta(n \times L \times d^2)$$

3.6. Experiments

3.6.1. Datasets

Two Yelp benchmark datasets were used for the experiments: yelpHotelData (hotel reviews) and yelpResData (restaurant reviews). After binarizing the labels ($4\star-5\star \rightarrow$ positive; $1\star-3\star \rightarrow$ negative), yelpHotelData contains 560,769 positive and 321,705 negative samples, whereas yelpResData contains 9,650,537 positive and 5,140,963 negative

samples. Both datasets were partitioned into 80% training and 20% testing sets using stratified sampling.

3.6.2. Experimental Results

Table 3.1. F_{DF} Results on the Two Yelp Datasets

Dataset	Acc (Train)	Acc (Test)	Precision	Recall	F ₁
yelpHotelData (K=7 clusters)	0.84	0.75	0.71	0.62	0.67
yelpResData (K=9 clusters)	0.85	0.80	0.72	0.67	0.69

The SHAP analysis indicates that Entity and Time are the most influential features (mean $|\text{SHAP}| \approx 0.087$ and 0.085), Holder ranks third (≈ 0.077), and Aspect is the least influential (≈ 0.023) because the small number of distinct aspects limits its discriminative power (Figure 3.5 of the full dissertation).

The SHAP interaction analysis (Figure 3.6 of the full dissertation) shows that self-interactions and the Entity–Time pair appear closest to the root of the trees, whereas interactions involving Aspect appear only in deep branches.

The proposed Opinion Density Forest achieves a Testing Accuracy of 0.80 on the yelpResData dataset, which is higher than the 0.75 obtained on the yelpHotelData dataset. This difference is attributed to the greater diversity and clearer distinction of restaurant entities compared with hotel entities.

3.6.3. Comparison with Representative Methods

Table 3.2. Comparison of F_{DF} with 7 Representative Methods

Model	Struct. Model.	Avg. Acc.	Avg. F ₁	Interpretability	Computational Complexity
F_{DF} (Proposed)	Yes	0.78	0.68	Very High	$\Theta(n \times m_{\text{try}} \times N \times \log N)$ + $\Theta(n \times L \times N \times \log N)$
RoBERTa [83]	No	0.79	0.70	Very Low	$\Theta(L^2 \times d)$ per token (Very High - GPU)
BERT [81]	No	0.78	0.69	Very Low	$\Theta(L^2 \times d)$ per token (Very High - GPU)
BiLSTM-Attention [88]	No	0.77	0.68	Low	$\Theta(n \times L \times d)$ (Medium)
Gradient Boosting [84]	No	0.76	0.67	High	$\Theta(n \times N \times \log N)$ (Medium)
Random Forest [65]	No	0.76	0.66	High	$\Theta(n \times N \times m_{\text{try}} \times \log N)$ (Low)
LSTM [85]	No	0.74	0.65	Low	$\Theta(n \times L \times d^2)$ (Medium)
SVM [86]	No	0.75	0.66	Medium	$\Theta(N^2 \sim N^3)$ (High)

The proposed Opinion Density Forest achieves an average Accuracy of 0.78, higher than the standard Random Forest (0.76) under the row-level split of the published papers; it

requires no GPU and provides interpretation through the cluster structure and SHAP values. The deep learning and Transformer models take the raw review text as input and reach between 0.74 (LSTM) and 0.79 (RoBERTa); they are reported as reference points. The figures follow Paper 5; results re-run under a unified protocol are reconciled in Appendix J of the full dissertation: under the review-level split, the density forest is comparable to Random Forest (difference not statistically significant).

The proposed Opinion Density Forest has a higher training complexity than the standard Random Forest because of the additional DBSCAN component,

$$\Theta(n \times L \times N \times \log N)$$

used during density clustering. However, the practical computational cost is lower than the theoretical worst-case complexity because DBSCAN is applied only to the subsets of samples associated with individual leaf nodes. In contrast, BERT and RoBERTa, with a computational complexity of

$$\Theta(L^2 \times d) \text{ per token}$$

require dedicated GPU resources, whereas the proposed F_{DF} can be efficiently executed on a conventional CPU with a reasonable computational cost.

3.7. Chapter Conclusion

Chapter 3 proposes the Opinion Density Forest model by integrating DBSCAN clustering with SHAP-based interpretation for opinion mining. The proposed model automatically partitions opinion data into 7–9 density-based clusters, constructs an unsupervised Opinion Density Forest using Gaussian Entropy $H_D(O)$, and explains the prediction results through SHAP values and Shapley interaction values. The proposed F_{DF} achieves an average Accuracy of 0.78 (published figures, row-level split; comparable to Random Forest under the review-level split) and runs on standard CPUs at low computational cost. However, the proposed F_{DF} relies on the Euclidean distance in DBSCAN and therefore cannot fully exploit the underlying nonlinear geometric structure of opinion data. This limitation is targeted in Chapter 4.

CHAPTER 4. OPINION MINING WITH MANIFOLD FORESTS

The main content encompasses the research results of Paper 6.

4.1. Introduction

Chapter 4 proposes the Opinion Manifold Forest (F_{MF}) model to address deeper nonlinear relationships within the opinion quintuple space. Instead of relying on the conventional Euclidean distance, the proposed F_{MF} learns a manifold distance, denoted by $dist_{\mathcal{M}}$, directly from the data through a relationship matrix W . This enables the model to exploit the latent d' -dimensional geometric structure ($d' < 5$) underlying the Yelp opinion quintuple space.

Chapter 4 at a glance – Problem: nonlinear relationships among quintuple components that the Euclidean distance does not capture. Method: the Opinion Manifold Forest – an affinity matrix W learned from an unsupervised forest, the manifold distance $\sqrt{1 - W}$, DBSCAN on the manifold (3–4 clusters), and Stacking over the K forests. Results: Accuracy 0.78 (Hotel) and 0.82 (Restaurant), 0.80 on average (published figures, row-level split; about 0.02–0.03 Balanced Accuracy below Random Forest under the review-level split); estimated intrinsic dimension $d' \approx 2.78$ (Hotel) and 3.66 (Restaurant), below the observed dimension $D = 4$ (on Restaurant, DANCo gives $d' = 4$, so the evidence is not consistent). Contribution: the third contribution of the dissertation (Paper 6).

4.2. Opinion Clustering on the Manifold Space

4.2.1. Feature Representation and the Manifold Hypothesis

The proposed model is based on the manifold hypothesis, which assumes that the Yelp opinion quintuple data are distributed near a low-dimensional manifold of dimension d' , where $d' < 5$. This assumption is motivated by three observations:

1. The nonlinear relationship between entities and aspects;
2. The consistent reviewing behavior of individual users (holders);
3. The concentration of review ratings at the two extremes of the five-star scale.

The validity of this hypothesis is indirectly verified through the experimental results presented in Section 4.8 of the full dissertation and directly examined by intrinsic dimension estimation (Appendix I of the full dissertation), with clear support on Hotel but inconsistent evidence on Restaurant.

4.2.2. DBSCAN Clustering on the Manifold Space

The DBSCAN algorithm is performed in the manifold space using the manifold distance

$$dist_{\mathcal{M}}(o_i, o_j) = \sqrt{1 - W_{ij}},$$

instead of the conventional Euclidean distance.

Under this formulation, two opinion samples are regarded as semantically close if they frequently appear in the same leaf node of the manifold forest.

4.3. Opinion Manifold Tree

The Opinion Manifold Tree is constructed using the Manifold Gaussian Entropy, defined as

$$H_{\mathcal{M}}(O) = \frac{1}{2} \log[(2\pi e)^{d'} |\Lambda_{\mathcal{M}}(O)|]$$

where d' denotes the intrinsic dimensionality of the manifold ($d' < d = 5$), and $\Lambda_{\mathcal{M}}(O)$ is the weighted covariance matrix computed as

$$\Lambda_{\mathcal{M}}(O) = \frac{\sum_{i,j \in O} W_{ij} (o_i - \mu(O))(o_j - \mu(O))^T}{\sum_{i,j \in O} W_{ij}},$$

where W_{ij} represents the manifold relationship between opinion samples o_i and o_j .

4.4. Opinion Manifold Forest

4.4.1. Manifold Affinity Matrix

The manifold affinity matrix

$$W \in \mathbb{R}^{N \times N},$$

is constructed from n_W unsupervised trees using a co-leaf counting mechanism

$$W_{ij} = \frac{1}{n_W} \sum_{k=1}^{n_W} W_{ij}^{(k)} = \frac{1}{n_W} \sum_{k=1}^{n_W} \mathbf{1}[l_{T_k^{(w)}}(o_i) = l_{T_k^{(w)}}(o_j)],$$

where $L(o_i, T_k)$ denotes the leaf node reached by opinion sample o_i in the k -th manifold tree.

The affinity value satisfies

$$W_{ij} \in [0,1],$$

where

- $W_{ij} = 1$ indicates maximum affinity;
- $W_{ij} = 0$ indicates no affinity.

The corresponding manifold distance is

$$\text{dist}_{\mathcal{M}}(o_i, o_j) = \sqrt{1 - W_{ij}}$$

Compared with the Euclidean distance, the relationship matrix W better preserves the semantic structure of opinion data. Consequently, opinion samples sharing the same Entity and Aspect generally have larger relationship values even when their Euclidean distance is relatively large.

4.4.2. Opinion Manifold Forest

Following the manifold clustering procedure described in Section 4.2.2, a specialized Opinion Manifold Forest

$$\mathcal{F}_k,$$

is constructed for each manifold cluster.

A Stacking meta-learner,

$$M,$$

implemented using a Random Forest, combines the predictions of the K specialized manifold forests on the validation set (20%). The final prediction is obtained as

$$\hat{y} = M(RF_1(o), RF_2(o), \dots, RF_K(o)).$$

4.5. Computational Complexity

Computational complexity of constructing a single Opinion Manifold Forest. The training complexity is divided into two stages: (1) constructing the manifold affinity matrix W using n_W unsupervised trees $T_k^{(w)}$; and (2) constructing the Opinion Manifold Forest based on the learned manifold representation. The overall training complexity is given by

$$\Theta(N \times \log N \times (n_W + n \times d'))$$

Computational complexity of manifold forest prediction. The prediction complexity consists of two steps: (1) estimating the affinity value $\hat{W}(o, o_i)$ for k_{rep} pairs between a new opinion sample o and the reference samples o_i ; and (2) traversing the n manifold trees for classification. The overall prediction complexity is

$$\Theta((k_{rep} + 1) \times n \times \log N)$$

Computational complexity of the ensemble manifold forest. The overall training complexity of the ensemble manifold forest consists of four independent stages executed sequentially: (1) constructing the manifold affinity matrix W ; (2) performing DBSCAN clustering in the manifold space; (3) building the K specialized manifold forests $\mathcal{F}_k$; and (4) training the Stacking meta-learner. The overall computational complexity is expressed as

$$\Theta(N \times \log N \times (n_W + n \times d')) + \Theta(K \times N_{val})$$

4.6. Experimental Results

4.6.1. Results on the Two Yelp Datasets

The clustering parameters are configured as follows:

- $\varepsilon = 0.3$, which is smaller than the value ($\varepsilon = 0.5$) used in Chapter 3 because the manifold distance $dist_{\mathcal{M}} \in [0,1]$ exhibits a denser distribution than the Euclidean distance. Using the same value of ε would therefore produce an excessive number of clusters.

- $MinPts = 5$, which is kept identical to Chapter 3 to ensure a fair comparison.

The value of ε is determined using the k-distance graph constructed in the manifold space by identifying the elbow point of the curve of distances to the $MinPts$ -th nearest neighbor.

The clustering results produce $K = 3$ clusters for the yelpHotelData dataset (corresponding to the component forests RF1, RF2 and RF3) and $K = 4$ clusters for the yelpResData dataset (RF1–RF4). The clusters are characterized by the actual statistics of the features (e, a, h, t) within each cluster rather than by semantic topic labels, since yelpHotelData contains only three aspects: AcceptsCreditCards, PriceRange and WiFi.

The resulting numbers of clusters ($K=3$ and $K=4$) are substantially smaller than those obtained in Chapter 3 ($K=7$ and $K=9$ using Euclidean distance). This is because the manifold space compresses semantically similar opinion samples into more compact clusters. Consequently, each specialized manifold forest $\mathcal{F}_k$ learns a more homogeneous decision boundary, which may help explain why, in the published figures, the proposed Opinion Manifold Forest achieves higher classification accuracy than the Opinion Density Forest.

Table 4.1. F_{MF} Results on yelpHotelData and yelpResData (Sources: Tables 4.2 and 4.3 of the full dissertation)

Dataset	Split	Accuracy	Precision	Recall	F_1
yelpHotelData (K=3 clusters)	Train	0.94	—	—	—
yelpHotelData (K=3 clusters)	Test	0.78	0.74	0.62	0.67
yelpResData (K=4 clusters)	Train	0.95	—	—	—
yelpResData (K=4 clusters)	Test	0.82	0.75	0.67	0.71

Note: Only test-set metrics are reported, as they form the basis for comparison with other models (Table 4.4 of the full dissertation); training Accuracy is shown in Figures 4.7 and 4.8 to assess overfitting. Results of re-running the three forest models with the current code are given in Appendix J (Table J.7).

On the test sets, the proposed Opinion Manifold Forest achieves Accuracy scores of 0.78 (Hotel) and 0.82 (Restaurant), the highest among the three proposed models under the row-level split of the published papers. The Train–Test accuracy gap is 0.16 for the Hotel dataset ($0.94 \rightarrow 0.78$) and 0.13 for the Restaurant dataset ($0.95 \rightarrow 0.82$); this gap is larger than that of the Opinion Density Forest (0.09/0.05) and calls for caution regarding overfitting; under the review-level split (Appendix J of the full dissertation), the manifold forest falls about 0.02–0.03 Balanced Accuracy below Random Forest. The Opinion Manifold Forest requires no GPU and provides transparent interpretation through the manifold affinity matrix W and the manifold Information Gain criterion.

Progression F_{CF} (0.77) $\rightarrow$ F_{DF} (0.78) $\rightarrow$ F_{MF} (0.80): under the row-level split of the published papers, +1% from the classification forest to the density forest and +2% on average from the density forest to the manifold forest (Hotel +3%: 0.78 vs 0.75; Restaurant +2%: 0.82 vs 0.80). This improvement is not reproduced under the review-level split (Appendix J of the full dissertation). Possible reasons for the published improvement: (1) $dist_{\mathcal{M}}$ better reflects semantic meaning; (2) more accurate clustering on manifold space (K=3/4 vs. 7/9 Euclidean clusters); (3) Stacking learns optimal weights rather than simple averaging. Consistently low test-set Recall (0.62–0.67) across the forest models reflects the shared challenge of the Yelp class imbalance (about 1.7:1 and 1.9:1); the Recall of the manifold forest equals that of the density forest (Restaurant 0.67; Hotel 0.62). A comprehensive model comparison is given in Table 4.2.

4.6.2. Comprehensive Comparison of All Models

As in the methodology described in Section 3.6.3, the traditional machine learning methods based on structured features (Random Forest, Gradient Boosting, SVM, and the three proposed forest models) use the numerically encoded opinion quintuple as input, whereas the deep learning and Transformer models (LSTM, BiLSTM-Attention, BERT, RoBERTa) use the original review text as input, following the standard operating procedure of these

architectures; this ensures a fair comparison on the same samples, the same sentiment labels, and the same training/testing split ratio.

Table 4.2. Comprehensive Model Comparison on yelpHotelData and yelpResData

Model	Acc Hotel	Acc Rest.	Avg. Acc.	Interpretability	Computational Complexity
Opinion Manifold Forest (F_{MF})	0.78	0.82	0.80	Very High	$\Theta(N \times \log N \times (nW + n \times d'))$ + $\Theta(K \times N_{val})$ (Two-stage combination)
Opinion Density Forest (F_{DF})	0.75	0.80	0.78	Very High	$\Theta(n \times m_{try} \times N \times \log N)$ + $\Theta(n \times L \times N \times \log N)$ (Medium)
Opinion Classification Forest (F_{CF})	0.76	0.77	0.77	Very High	$\Theta(n \times N \times m_{try} \times \log N)$ (Low)
BiLSTM-Attention [88]	0.76	0.78	0.77	Low	$\Theta(n \times L \times d)$ (Medium)
Random Forest [65]	0.74	0.77	0.76	High	$\Theta(n \times N \times m_{try} \times \log N)$ (Low)
Gradient Boosting [84]	0.74	0.78	0.76	High	$\Theta(n \times N \times \log N)$ (Medium)
SVM [86]	0.70	0.80	0.75	Medium	$\Theta(N^2 \sim N^3)$ (High)
LSTM [85]	0.71	0.76	0.74	Low	$\Theta(n \times L \times d^2)$ (Medium)
RoBERTa [83]	0.74	0.83	0.79	Very Low	$\Theta(L^2 \times d)$ per token (Very High - GPU)
BERT [81]	0.74	0.82	0.78	Very Low	$\Theta(L^2 \times d)$ per token (Very High - GPU)

The proposed Opinion Manifold Forest achieves an average Accuracy of 0.80 under the row-level split of the published papers, the highest among the three proposed forest models, 2% higher than the Opinion Density Forest and 3% higher than the Opinion Classification Forest. Possible reasons for this improvement are: (1) exploiting the nonlinear structure of the data through manifold learning; (2) more accurate clustering in the manifold space than in the Euclidean space; and (3) the manifold affinity matrix W better reflecting the semantic meaning of opinions than the Euclidean distance. The deep learning and Transformer models take the raw review text as input and reach between 0.74 (LSTM) and 0.79 (RoBERTa); they are reported as reference points (figures from Papers 5 and 6; re-run results are reconciled in Appendix J of the full dissertation). The Opinion Manifold Forest requires no GPU and provides transparent interpretation through the manifold affinity matrix W and the manifold Information Gain criterion.

The proposed $\mathbf{F}_{MF}$ model has a training complexity of $\Theta(N \times \log N \times (nW + n \times d')) + \Theta(K \times N_{val})$ which consists of two stages: constructing the manifold affinity matrix W and building the Opinion Manifold Forest. However, the manifold affinity matrix W is constructed only once and subsequently shared throughout the training process. The proposed $\mathbf{F}_{DF}$ model incurs additional computational cost due to the DBSCAN clustering performed at the leaf nodes, whereas $\mathbf{F}_{CF}$ is the simplest model in terms of computational complexity. All three proposed models operate efficiently on standard CPUs without requiring GPU acceleration, whereas Transformer-based models such as BERT and RoBERTa have a computational complexity of $\Theta(L^2 \times d)$ per token and are typically trained on dedicated GPU hardware.

4.7. Chapter Conclusion

Chapter 4 proposed the Opinion Manifold Forest (F_{MF}) model, which exploits the manifold hypothesis in opinion quintuple data as an approach for Aspect-Based Sentiment Analysis (ABSA). The model constructs a manifold affinity matrix W from an unsupervised forest, performs DBSCAN clustering in the manifold space, and integrates the specialized manifold forests using a Stacking framework. The average Accuracy of 0.80 (published figures, row-level split) and the intrinsic dimension estimates indicate that the manifold hypothesis with intrinsic dimension $d' < 5$ is feasible for Yelp opinion data (clear on Hotel, inconsistent on Restaurant). Under the review-level split, the manifold forest falls about 0.02–0.03 Balanced Accuracy below Random Forest because the manifold clustering step is learned on a subsample, and the cost of the matrix W remains a barrier to scaling. The proposed Opinion Manifold Forest contributes to the development of ensemble learning methods for natural language processing and is well aligned with the growing demand for more responsible and interpretable AI models.

CONCLUSION AND FUTURE DIRECTIONS

1. Conclusion

Opinion mining has become an important research area with broad applications in business, entertainment, healthcare, education, and social management. This dissertation, entitled “*Opinion Mining with Ensemble Methods*”, has proposed and evaluated three ensemble learning models for opinion mining, all of which are built upon the homogeneous base learner of decision trees.

The main results of this dissertation are summarized as follows (the accuracy values below are published figures obtained with a row-level random split of the quintuples; results under the review-level split are summarized at the end of this section):

- **Opinion Classification Forest:** The proposed model achieved an average Accuracy of 0.77 on the yelpHotelData and yelpResData datasets. In addition, the model provides strong interpretability through feature importance analysis and Partial Dependence Plots (PDPs).

- **Opinion Density Forest:** The proposed model achieved an average Accuracy of 0.78 on the yelpHotelData and yelpResData datasets, without requiring GPU acceleration. Furthermore, SHAP analysis identified Entity as the most influential feature with a mean |SHAP| of about 0.087 (Figure 3.5 of the full dissertation).

- **Opinion Manifold Forest:** The proposed model exploits the manifold hypothesis in opinion quintuple data from the yelpHotelData and yelpResData datasets. The model constructs a manifold affinity matrix W , performs DBSCAN clustering in the manifold space ($d' < 5$), and integrates specialized manifold forests using a Stacking framework. The experimental results achieved an average Accuracy of 0.80, the highest among the three proposed models.

- **Opinion quintuple datasets:** The dissertation developed a procedure for converting review data into opinion quintuples and applied it to yelpHotelData (more than 880 thousand quintuples), yelpResData (more than 14 million quintuples) and two Vietnamese datasets, VLSP-2018 Restaurant (more than 15 thousand review–aspect pairs) and VLSP-2018 Hotel (more than 23 thousand pairs); on VLSP-2018, the results of the three forest models are reported together with baseline models and the PhoBERT reference point (Appendix F of the full dissertation).

- **Opinion quintuple extraction model:** In addition, the dissertation proposed an automatic opinion quintuple extraction model, achieving an F_1 -score of 0.78 (Precision = 0.81, Recall = 0.75) on the Twitter US Airline Sentiment dataset. This model provides a practical foundation for constructing opinion quintuple datasets in other application domains.

The progressive improvement in classification performance from F_{CF} (Accuracy = 0.77) to F_{DF} (Average Accuracy = 0.78) and finally F_{MF} (Average Accuracy = 0.80) reflects, in the published figures, increasingly sophisticated exploitation of the underlying data structure, while all three models retain interpretability through their tree, cluster and affinity-matrix structures.

When re-evaluated with a review-level split (Appendix J of the full dissertation, 5 seeds $\times$ 10,000 reviews), the Balanced Accuracy of every quintuple-based model drops to about 0.51–0.57 on both datasets: F_{CF} and F_{DF} are comparable to Random Forest (differences not statistically significant) and about 0.01 below linear SVM; F_{MF} is about 0.02–0.03 below Random Forest because its manifold clustering step is learned on a subsample. The text-based models (BERT, RoBERTa) reach 0.76–0.82, but when every model uses only the quintuples, F_{CF} and F_{DF} are comparable to BiLSTM-Attention and higher on average than LSTM, BERT and RoBERTa. On VLSP-2018, F_{DF} reaches a Balanced Accuracy of 0.724 on Hotel (not significantly different from Logistic Regression, 0.730) and 0.660 on Restaurant (the best among the quintuple-based models), while the text-based PhoBERT reference reaches 0.832 (Hotel) and 0.791 (Restaurant).

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). With respect to the three research gaps: interpretability at the level of each opinion component is addressed; class imbalance is addressed only partly (with a clear effect under strong imbalance, as in VLSP-2018 Restaurant); and exploiting the manifold structure is shown to be feasible, but the size of the matrix W remains a barrier to scaling. In addition, the dissertation quantifies the leakage caused by the row-level split (with the same Random Forest and 112,275 quintuples, Accuracy is 0.946 under the row-level split versus 0.611 under the review-level split) and recommends review-level splitting when evaluating models on quintuple data.

Nevertheless, the dissertation also acknowledges several limitations. The proposed forest-based models operate on pre-extracted opinion quintuple representations and therefore depend on the quality of the preceding extraction process. In contrast, end-to-end Transformer models learn representations directly from raw text and are capable of capturing sarcasm, negation, and implicit opinions.

2. Future Directions

- Expand yelpHotelData and yelpResData into multi-domain, multilingual datasets; incorporate Vietnamese from VLSP-2018 and other Asian languages.
- Integrate semantic representations from deep learning (BERT embeddings, PhoBERT for Vietnamese) with forest structures to improve quintuple extraction quality and handle sarcasm, negation, and implicit opinions.
- Develop opinion mining models toward deeper Explainable AI (XAI), increasing effectiveness, trustworthiness, and alignment with Responsible AI - an increasingly important research direction.
- Extend to real-time opinion analysis on streaming data and multilingual ABSA tasks.
- Reduce the cost of computing and storing the affinity matrix W (subsampling, sparse leaf-index representations, local W within clusters) to scale the Opinion Manifold Forest to large datasets.
- Combine model explanation with causal inference (counterfactual analysis, structural causal models) to distinguish the factors that genuinely drive sentiment polarity from correlations in observational data.

LIST OF PUBLISHED SCIENTIFIC WORKS

No.	Work Title	Type	Year Published
1	Paper No. 1	International Journal	2020
2	Paper No. 2	Conference Proceedings	2021
3	Paper No. 3	Conference Proceedings	2023
4	Paper No. 4	International Journal	2023
5	Paper No. 5	Conference Proceedings	2024
6	Paper No. 6	Conference Proceedings	2024
7	Paper No. 7	Conference Proceedings	2025
8	Paper No. 8	Conference Proceedings	2026
9	Paper No. 9	International Journal (Accepted)	2026