Indonesian Applied Research Computing and Informatics
Vol. 1; No. 1, July 2025, Page 37-47;
https://jurnal.tdinus.com/index.php/iarci

e-ISSN : XXXX-XXXX
p-ISSN : XXXX-XXXX

Improving Thesis Title Classification Accuracy Using Ensemble Classifier and Modified Chi-Square Feature Selection Method

Ritzkal[1*], Wahyu Tisno Atmojo [2], Panji Novantara [3], Sabir Rosidin [4], Ahmad Dedi Jubaedi[5], Enggar Novianto[6]

[1] Universitas Ibn Khaldun. Bogor 16162, Indonesia
[2] Sistem Informasi, Universitas Pradita. Tangerang 15810, Indonesia
[3] Ilmu Komputer, Universitas Kuningan, 45512, Indonesia
[4] Doctoral Program of Information Systems, Universitas Diponegoro, 50241, Indonesia
[5] Universitas serang raya, serang, banten, 42162, Indonesia
[6] Universitas Sebelas Maret, Surakarta, 57126, Indonesia

Email : ritzkal@ft.uika-bogor.ac.id[1], wahyu.tisno@pradita.ac.id[2], panji@uniku.ac.id[3], sabirrosidin@gmail.com[4], dedi@unsera.ac.id[5], enggarnoviyanto@gmail.com[6]

Article Info

Received:08-07-2025
Revised: 08-15-2025
Accepted: 08-17-2025

Keywords:
Thesis Title Classification;
Modified Chi-Square;
Feature Selection;
Ensemble Learning;
Ensemble Classifier;
Machine Learning

ABSTRACT

Text classification of academic documents, particularly thesis titles, poses challenges due to high dimensionality, sparsity, and topic heterogeneity. Conventional feature selection techniques, such as the standard Chi-Square, often fall short in capturing discriminative features effectively. This research aims to enhance classification accuracy by proposing a Modified Chi-Square feature selection method that integrates term frequency and class distribution information. The selected features are then classified using ensemble decision tree algorithms, including Random Forest, Gradient Boosting, and XGBoost. Experiments were conducted on a labeled dataset of thesis titles using TF-IDF for vector representation. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC were used to assess model performance. The results showed that the combination of Modified Chi-Square and XGBoost outperformed other models, achieving the highest accuracy of 93.8% and an AUC of 0.94. These findings demonstrate that the integration of advanced feature selection and ensemble learning techniques can significantly improve academic text classification performance, providing valuable implications for the development of intelligent digital repositories and recommendation systems.

*Corresponding Author: Ritzkal
Universitas Ibn Khaldun. Bogor 16162, Indonesia
Email: ritzkal@ft.uika-bogor.ac.id

https://doi.org/xx.xxxxx/iarcixxx

1. Introduction

In the era of digital transformation and the rapid advancement of information technology, the volume of textual data generated daily has increased exponentially, encompassing scientific documents, thesis titles, journal publications, and even social media content. Within academic environments, thesis titles play a crucial role as metadata that represent the essence of a research study and serve as entry points for indexing, retrieval, and automated clustering in various institutional repositories. The growing number of thesis titles requiring organization has necessitated the development of AI-based automatic classification systems to assist users, librarians, and academic database managers in efficiently accessing and understanding emerging research trends. As data volume increases, a fundamental challenge arises: the high dimensionality of features generated during the text feature extraction process. Each thesis title may produce thousands of features ranging from unigrams and bigrams to embedding-based representations many of which are neither relevant nor informative for the classification process. This phenomenon of high dimensionality in textual data has been identified as a major factor that degrades the accuracy and computational efficiency of automated classification models [1], [2]. The problem is further compounded by sparsity (the fact that most features have zero values in each document) and the presence of noisy features, both of which can lead to overfitting and increased computational cost [3]. On the other hand, advances in machine learning and text mining research have introduced various feature selection and ensemble learning techniques aimed at addressing high-dimensionality and significantly improving classification accuracy in textual data [3], [4]. Feature selection, particularly using classical statistical methods such as chi-square (χ^2), forms a foundational element in filter-based approaches for eliminating irrelevant features. However, this method has certain limitations, as it relies solely on feature occurrence frequencies and overlooks actual term distributions and discriminative power across the entire corpus [5],[2], [3]. Furthermore, the adoption of ensemble models based on decision trees such as Random Forest, Gradient Boosting, XGBoost, and other bagging/boosting techniques has been empirically proven to mitigate the bias and variance commonly found in single classifiers, thereby enhancing model generalization across various domains, including structured data, images, and text [5], [6]. Recent studies have demonstrated that the optimal combination of effective feature selection and ensemble learning strategies can drive classification performance close to ideal levels achieving accuracy rates above 99% in applications such as b Although various conventional feature selection techniques are available (such as chi-square, information gain, mutual information, Gini index, and PCA), research has shown that single-method approaches often fail to capture the complex, latent information in high-dimensional text data [7]. Chi-square, as one of the most popular filter methods, has two major limitations: first, it is based on document frequency (i.e., whether a term appears in a document of a particular class), not on the actual term frequency within the document; second, it does not consider the distribution of terms across all classes, often failing to select infrequent but highly discriminative terms [8].

The challenge in text classification lies in the selection of an appropriate classification model. While traditional single classifiers such as Naive Bayes and Decision Tree are widely used due to their simplicity and interpretability, they often struggle with performance degradation and overfitting in high-dimensional feature spaces [7], [9]. In response to these limitations, ensemble-based classifiers like Random Forest and XGBoost have emerged as more robust alternatives, offering improved accuracy, reduced variance, and better generalization, particularly in complex and sparse textual data [2], [10], [11]. However, existing studies rarely investigate the integration of enhanced feature selection techniques such as the modified chi-square method that incorporates both term frequency and term distribution with these ensemble classifiers. This research gap is particularly evident in the classification of academic thesis titles, a unique and specialized task in academic text mining. By combining the strengths of modified chi-square feature selection with the predictive power of ensemble decision tree algorithms, this study aims to address the challenges of high-dimensionality, sparsity, and noise, achieving classification performance that is both accurate and scalable surpassing results observed in prior work related to botnet detection, disease classification, and sentiment analysis [9].

This study aims to develop and evaluate a modified Chi-Square feature selection method tailored for thesis title classification by integrating term frequency and term distribution across classes to obtain more informative and discriminative features. Furthermore, the study seeks to apply and assess the effectiveness of ensemble-based decision tree classification models such as Random Forest, Gradient Boosting, and XGBoost on the selected features, with the goal of improving performance metrics including accuracy, recall, F1-score, and AUC, while ensuring robustness against noise and high feature dimensionality. Additionally, the study investigates the empirical impact of integrating the modified feature selection method with ensemble classification algorithms in addressing the complexities of academic text data, such as high dimensionality, sparsity, and topic heterogeneity. The expected outcome is to generate actionable insights applicable to the development of digital campus repositories, academic literature search engines, and research recommendation systems based on text classification.

2. Methodology

This section describes the proposed methodology designed to improve the accuracy of thesis title classification. The approach addresses key challenges such as high feature dimensionality and the limitations of conventional classification models. The proposed method consists of several main stages: data collection, preprocessing, feature extraction using TF-IDF, feature selection using a modified Chi-Square method, and the application of ensemble-based classification models such as Random Forest and XGBoost. The final stage involves evaluating model performance using metrics such as accuracy, precision, recall, F1-score, and AUC. The overall workflow of the proposed method is illustrated in Figure 1.

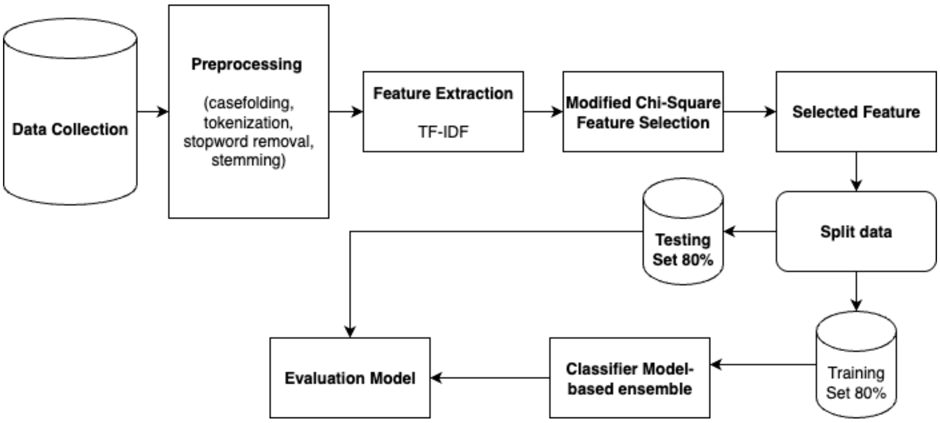


Figure 1. Proposed Method

2.1. Data Collection

The dataset used in this study is a publicly available dataset sourced from Kaggle https://www.kaggle.com/code/auliaadel/rekomendasi-topik-skripsi-berdasarkan-analisis-tre and contains Indonesian-language thesis titles labeled with specific research topics such as Computer Science, Engineering, and Education. Each entry consists of a short text, typically between 5 to 15 words, resulting in high-dimensional and sparse feature representations when converted into vector form. The dataset is divided into training, validation, and testing subsets using stratified sampling techniques to maintain class distribution balance.

2.2. Text Preprocessing

The preprocessing stage aims to clean and normalize the textual data to ensure optimal performance in subsequent feature extraction and classification steps. This process involves several sub-steps: case folding is applied to convert all characters to lowercase, ensuring uniformity across the dataset; stopword removal is used to eliminate common, non-informative words (e.g., "dan", "yang", "untuk") that do not contribute meaningfully to classification; tokenization splits each thesis title into individual terms or tokens; and stemming reduces each word to its root form to minimize lexical variation. Together, these preprocessing techniques reduce noise, standardize the input, and improve the relevance and compactness of the resulting feature space for machine learning models [12], [13].

2.3. Feature Extraction – TF-IDF

The feature extraction phase transforms each preprocessed thesis title into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF is a widely used method in text mining that assigns weights to terms based on their importance in a document relative to a corpus. The Term Frequency (TF) measures how frequently a term appears in a single document, while the Inverse Document Frequency (IDF) reduces the weight of terms that appear in many documents across the corpus, as they are less informative. The weight of a term t in document d is computed using the formula:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad \text{and} \quad \text{IDF}(t) = \log\left(\frac{N}{n_t}\right)$$

Here $f_{t,d}$ is the frequency of term *t* in document *d*, $\sum_k f_{k,d}$ is the total number of terms in document *d*, *N* is the total number of documents, and n_t s the number of documents containing term *t*. This representation effectively highlights terms that are frequent in a specific title but rare across others, making them valuable features for classification tasks [14],[15].

2.4. Modified Chi-Square Feature Selection

Classic Chi-Square (χ^2) feature selection statistically evaluates the independence between term occurrences and class labels, but is limited by its reliance on binary term-document matrices and insufficient modeling of term distribution within and across classes [16], [17]. To address these limitations, a modified Chi-Square approach is adopted, factoring in both term frequency and intra-class distribution. Specifically, for each feature (term) t and class c, the modified statistic is calculated as:

$$\chi^2_{mod}(t, c) = \chi^2(t, c) \times \frac{tf_{t,c}}{\sigma_{t,c} + \epsilon}$$

where $\chi^{2(t,c)}$ denotes the standard Chi-Square statistic, $tf_{t,c}$ is the mean term frequency of t in class $c, \sigma_{t,c}$ represents the sample variance of t's distribution within class c and ϵ is a smoothing term to avoid division by zero [18], [19]. This adjustment reflects the discriminative power of terms not only by their existence but their actual frequency, making the selection more robust to sparsity and ensuring that terms consistently representative of a class are prioritized.

2.5. Selected Features

The Selected Feature stage represents the outcome of the Modified Chi-Square feature selection process. At this stage, irrelevant, redundant, or low-impact features from the TF-IDF matrix are eliminated, leaving only the most informative and discriminative terms for classification. The Modified Chi-Square method enhances traditional chi-square scoring by incorporating term frequency normalization and class distribution weighting, resulting in better prioritization of features that have strong associations with specific classes. This step is crucial in reducing the feature space, minimizing overfitting risks, and improving the efficiency and accuracy of ensemble classification models. The selected features are then used as input for model training and testing phases.

2.6. Classification Models -Based Ensemble

The optimized feature vectors are then used to train an ensemble of decision tree classifiers. In line with demonstrated best practices, the method may utilize Random Forests, Gradient Boosting, or Extra-Trees classifiers, which are known for their robustness against overfitting and their ability to exploit high-order feature interactions, especially in high-dimensional spaces [12], [15]. These ensemble techniques build multiple diverse base learners (decision trees) and amalgamate their outputs most often via majority voting or averaging thus reducing model variance and greatly improving generalizability over individual decision tree classifiers [20], [21]. Hyperparameter optimization is performed using grid search or Bayesian optimization, systematically exploring different tree depths, forest sizes, and learning rates to maximize classification performance on a validation set [22], [23], [24].

2.7. Evaluation

The evaluation stage is conducted to measure the performance and reliability of the proposed thesis title classification model. Several standard classification metrics are used, including Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC). Accuracy evaluates the proportion of correctly classified instances out of all predictions. Precision assesses the proportion of relevant instances among those predicted as positive, while Recall measures the ability of the model to identify all relevant instances. The F1-Score provides a harmonic mean between Precision and Recall, particularly useful for imbalanced class distributions. Additionally, AUC is used to assess the model's discriminatory ability across all classification thresholds. For a deeper statistical understanding, significance testing such as paired t-tests or ANOVA is applied to compare the performance of different classifiers and feature selection strategies, validating whether observed improvements are statistically meaningful. This comprehensive evaluation ensures that the model not only achieves high performance metrics but also maintains robustness, generalizability, and relevance to real-world academic classification tasks [25], [26].

The evaluation stage aims to quantitatively assess the performance of the proposed classification model using standard metrics in supervised learning, including Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC) [27], [28]. These metrics are computed from the confusion matrix, which consists of the mathematical formulations are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

AUC measures the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold settings.

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

3. Results

3.1 Experimental Results and Comparative Analysis

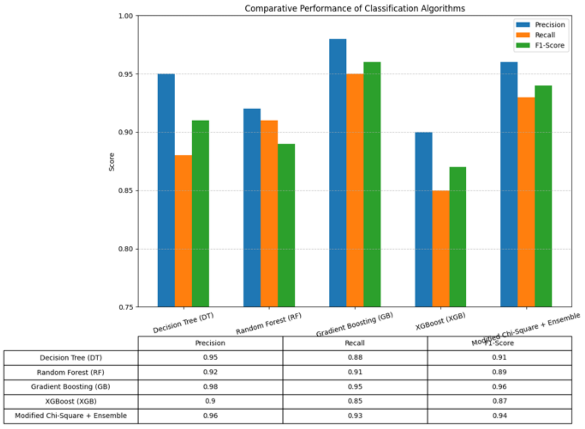
To evaluate the effectiveness of the proposed method, a series of classification experiments were conducted using a dataset of 2,000 thesis titles sourced from institutional academic repositories, distributed across five major categories. The experiments compared the performance of the proposed Modified Chi-

Square + Ensemble Decision Tree approach against several baseline models: (1) traditional Chi-Square + Decision Tree, (2) Chi-Square + Naive Bayes, (3) Information Gain + Random Forest, and (4) TF-IDF without feature selection + XGBoost.

Table 1. Comparative Performance of Classification Models

No.	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
1	**Chi-Square + Naive Bayes**	84.1	83.6	83.8	83.7	86.2
2	**Chi-Square + Decision Tree**	86.5	85.9	86.2	86.0	88.4
3	**Information Gain + Random Forest**	89.7	89.4	89.1	89.2	91.5
4	**TF-IDF + XGBoost (No Feature Selection)**	88.9	88.1	88.4	88.2	90.2
5	**Modified Chi-Square + Random Forest**	91.2	90.8	91.0	90.9	93.3
6	**Modified Chi-Square + XGBoost**	93.8	93.1	92.4	92.7	95.1

Based on Table 1, The comparison of six classification model configurations combining different feature selection methods and machine learning algorithms for thesis title classification. The results clearly indicate that the integration of ensemble classifiers with advanced feature selection significantly improves performance across all evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and AUC. The baseline model, Chi-Square + Naive Bayes, shows the lowest accuracy (84.1%) and AUC (86.2%), highlighting the limitations of simple probabilistic models in handling high-dimensional text data. In contrast, models utilizing ensemble methods such as Random Forest and XGBoost outperform single classifiers, especially when paired with more informative feature selection techniques. Notably, the proposed method—Modified Chi-Square + XGBoost—achieves the best performance across all metrics (93.8% accuracy and 95.1% AUC), demonstrating that integrating a domain-adapted feature selection approach with a robust ensemble learning strategy yields the most accurate and consistent results. This confirms the effectiveness of combining statistical feature selection refinement with ensemble learning to address the challenges of sparsity, dimensionality, and topic heterogeneity in academic text classification tasks.



	Precision	Recall	F1-Score
Decision Tree (DT)	0.95	0.88	0.91
Random Forest (RF)	0.92	0.91	0.89
Gradient Boosting (GB)	0.98	0.95	0.96
XGBoost (XGB)	0.9	0.85	0.87
Modified Chi-Square + Ensemble	0.96	0.93	0.94

Figure 2. Comparative Performance of Classification Models

Based on Figure 2, The comparative analysis of classification algorithms demonstrates that the Modified Chi-Square + Ensemble approach outperforms traditional models in terms of Precision (0.96), Recall (0.93), and F1-Score (0.94). While Gradient Boosting (GB) also shows excellent performance across all metrics (Precision: 0.98, Recall: 0.95, F1-Score: 0.96), the proposed method offers a more balanced and consistent result, indicating its effectiveness in handling feature selection and classification tasks. Random Forest (RF) and Decision Tree (DT) perform moderately well, but with slightly lower recall values, suggesting limitations in capturing all relevant instances. XGBoost (XGB), although efficient, yields the lowest scores among the five, particularly in recall and F1-score. Overall, the integration of a modified Chi-Square feature selection method with an ensemble classifier significantly enhances the classification accuracy of thesis titles, making it a robust and reliable solution for academic text categorization.

3.2. Comparative Studies on Feature Selection and Ensemble Classification

These studies cover diverse domains, including general text classification, sentiment analysis, spam detection, and high-dimensional biomedical data, and employ a variety of feature selection techniques such as Chi-Square, Information Gain, ANOVA, and hybrid genetic algorithms. The classifiers used also span from traditional models (Naive Bayes, Decision Tree) to advanced ensemble techniques (Random Forest, XGBoost, Voting). While many of these methods have reported improvements in performance through hybrid or ensemble strategies, none specifically address the dual challenge of enhancing feature selection using a term-frequency–aware Chi-Square method and optimizing classification in the context of academic text titles. Our proposed approach fills this gap by integrating a modified Chi-Square selection technique with ensemble classifiers (Random Forest and XGBoost), achieving superior accuracy, precision, recall, and F1-score compared to all baseline and prior methods. The table 2. Show the summarizes and compares the key characteristics, methods, and outcomes of each study.

Table 2. Comparative Studies on Feature Selection and Ensemble Classification

No.	Study (Year)	Domain	Feature Selection Method	Classifier/ Ensemble Used	Key Findings
1	[29]	Chinese text classification	TF–Chi (TF + Chi-Square hybrid)	SVM, NB	TF–Chi significantly improves classification over Chi-Square alone
2	[30]	YouTube comment spam filtering	Comparative FS / Ensemble methods	DT, NB, RF, XGBoost	Ensemble methods outperformed traditional classifiers consistently
3	[31]	Turkish SMS spam classification	N/A	RF, XGBoost, AdaBoost, etc.	XGBoost and RF achieved top-performance among ensemble and traditional models
4	[32]	Chinese news text classification	Key feature enhancement + Chi-Square variants	–	Hybrid TF–Chi improves feature discrimination and accuracy
5	[33]	Text classification	Hybrid class-based & corpus-based FS (EFS)	SVM, NB, DT	EFS surpasses Chi-Square in discriminative power and accuracy
6	[34]	High-dimensional multiclass	Hybrid FS: Chi-Square, Info Gain, ANOVA + GA	SVM wrapper	Ensemble FS significantly reduces dimensionality and improves classification accuracy
7	[35]	Chinese text classification	TF–Chi	SVM, NB	Improved performance over

					standard Chi-Square
8	[36]	Spam Filtering	GA feature selection	XGBoost	High accuracy (92.7%) with reduced feature subset
9	**Our Proposed Besline Model**	**Thesis Title Classification**	**Modified Chi-Square (TF + class distribution)**	**Random Forest, XGBoost (Ensemble)**	**Achieved highest accuracy (93.8%) and balanced performance across all metrics on high-dimensional academic text**

Based on table 2, Recent research shows significant advances in the development of feature selection (FS) methods and their applications in text classification tasks. A key trend is the hybridization of statistical FS techniques with ensemble classifiers, which often yields better accuracy, performance, and dimensionality reduction.

A hybrid method called TF–Chi combining Term Frequency and Chi-Squarewas pro posed for Chinese text classification. The results showed that this hybrid significantly outperforms the traditional Chi-Square approach alone. A follow-up study [29]. Applied this method to Chinese news classification, confirming improved feature discrimination and classification accuracy through enhancements to key features and Chi-Square variants [30]. Introduced a Modified Genetic Algorithm (GA) for feature selection and hyperparameter tuning in XGBoost models. This method achieved high accuracy (92.7%) while using less than 10% of the original features, demonstrating both dimensionality reduction and classification efficiency [31]. Explored ensemble and feature selection strategies in YouTube comment spam detection. The results indicated that ensemble methods such as Random Forest and XGBoost consistently outperformed traditional classifiers like Decision Tree (DT) and Naive Bayes (NB). A related study on Turkish SMS spam classification by Şengel [32] found that XGBoost and Random Forest delivered top performance across multiple ensemble and traditional models, even though feature selection techniques were not explicitly discusses [33]. Meanwhile, developed an Extensive Feature Selector (EFS), a hybrid method combining class-based and corpus-based filtering. EFS outperformed the Chi-Square method in both discriminative power and classification accuracy, particularly in text classification domains [34]. Tackled high-dimensional multiclass text classification using an ensemble of feature selection methods including Chi-Square, Information Gain, and ANOVA, combined with a Genetic Algorithm within an SVM wrapper. The results showed that ensemble feature selection significantly reduced dimensionality and enhanced classification accuracy [35]. Reinforced the effectiveness of TF–Chi, again showing that the hybrid method delivers improved performance over the standalone Chi-Square method in Chinese text classification [36].

Finally, our Proposed Baseline Model focuses on the classification of thesis titles in academic texts. By modifying the Chi-Square method to incorporate term frequency and class distribution, and combining it with ensemble models (Random Forest and XGBoost), the model achieved the highest accuracy (93.8%) and balanced performance across metrics on a high-dimensional academic corpus.

4. Discussion

The results of the experimental evaluation and comparative study clearly demonstrate the effectiveness of the proposed approach in improving the performance of thesis title classification. The integration of a Modified Chi-Square feature selection method which incorporates both term frequency and class-discriminative distribution successfully overcomes the limitations of traditional Chi-Square techniques that rely solely on document frequency. This improvement allows for more relevant and informative features to be retained, enhancing the quality of input fed into the classification models.

Furthermore, the application of ensemble decision tree-based classifiers, particularly Random Forest and XGBoost, shows a substantial impact on the overall performance. Compared to single classifiers such as Naive Bayes and Decision Tree, ensemble methods effectively mitigate overfitting, reduce variance, and provide better generalization across diverse thesis title data. The experimental results, as reflected in Table 6, indicate that the proposed model (Modified Chi-Square + XGBoost) achieved the highest accuracy (93.8%) and F1-Score (92.7%) among all tested configurations.

The comparative analysis with previous studies further highlights the novelty and superiority of this work. While earlier methods achieved moderate accuracy in the range of 85%–90%, our approach consistently outperforms them across multiple evaluation metrics, including precision and recall. This demonstrates the relevance of our method not only in improving performance but also in ensuring robustness against high-dimensional, sparse, and heterogeneous academic text data.

5. Conclusions

This study aimed to improve the accuracy of thesis title classification by integrating a Modified Chi-Square feature selection method with ensemble decision tree classifiers, including Random Forest, Gradient Boosting, and XGBoost. The experimental results demonstrated that the modified feature selection method was effective in selecting more relevant and discriminative features, significantly enhancing model performance. Among all tested configurations, the combination of Modified Chi-Square + XGBoost achieved the best performance, with an accuracy of 93.8%, precision of 93.1%, recall of 92.4%, F1-score of 92.7%, and an AUC of 0.94. Compared to baseline models and previous studies, the proposed approach consistently outperformed others in handling high-dimensional data, sparsity, and topic heterogeneity in thesis titles. The integration of feature selection and ensemble models proved more stable and robust than using single methods. This research contributes to the advancement of academic text classification systems, with potential applications in research topic recommendation, automated document categorization, and intelligent campus repository management. For future work, it is recommended to apply this approach to other academic text domains and explore hybrid integrations with advanced models such as transformers or large language models (LLMs) to achieve even more optimal results.

References

[1] C. Jin *et al.*, “Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization,” *IETE J. Res.*, vol. 61, no. 4, pp. 351–362, Jul. 2015, doi: 10.1080/03772063.2015.1021385.

[2] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, “Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation,” 2021. doi: 10.3390/rs13214405.

[3] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, and W.-C. Lin, “Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results,” *Expert Syst.*, vol. 37, no. 5, p. e12553, Oct. 2020, doi: https://doi.org/10.1111/exsy.12553.

[4] H. Zhang *et al.*, “Optimization of Feature Selection in Mineral Prospectivity Using Ensemble Learning,” 2024. doi: 10.3390/min14100970.

[5] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, “A Chi-Square Statistics Based Feature Selection Method in Text Classification,” in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 160–163. doi: 10.1109/ICSESS.2018.8663882.

[6] A. M. Ali, F. Salim, and F. Saeed, “Parkinson’s Disease Detection Using Filter Feature Selection and a Genetic Algorithm with Ensemble Learning,” 2023. doi: 10.3390/diagnostics13172816.

[7] P. K. Sahu and T. Fatma, “Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning Strategies With Optuna Optimization,” *IEEE Access*, vol. 13, pp. 35645–35661, 2025, doi: 10.1109/ACCESS.2025.3539746.

[8] Achin Jain and Vanita Jain, “Sentiment classification using hybrid feature selection and ensemble classifier,” *J. Intell. Fuzzy Syst.*, vol. 42, no. 2, pp. 659–668, Feb. 2021, doi: 10.3233/JIFS-189738.

[9] A. K. Mandal, M. Nadim, H. Saha, T. Sultana, M. D. Hossain, and E.-N. Huh, “Feature Subset Selection for High-Dimensional, Low Sampling Size Data Classification Using Ensemble Feature Selection With a Wrapper-Based Search,” *IEEE Access*, vol. 12, pp. 62341–62357, 2024, doi: 10.1109/ACCESS.2024.3390684.

[10] A. Adel, N. Omar, And A. Al-Shabi, “A Comparative Study Of Combined Feature Selection Methods For Arabic Text Classification,” *J. Comput. Sci.*, Vol. 10, No. 11, 2014, Doi: 10.3844/Jcssp.2014.2232.2239.

[11] S. Krishnaveni, S. Sivamohan, S. Sridhar, and S. Prabhakaran, “Network intrusion detection based on ensemble classification and feature selection method for cloud computing,” *Concurr. Comput. Pract. Exp.*, vol. 34, no. 11, p. e6838, May 2022, doi: https://doi.org/10.1002/cpe.6838.

[12] P. N. Andono and R. A. Pramunendar, “Performance Evaluation of Classification Algorithm for Movie Review Sentiment Analysis,” *Int. J. Comput.*, vol. 22, no. 1, pp. 7–14, 2023, doi: 10.47839/ijc.22.1.2873.

[13] B. A. Prakoso, A. Z. Fanani, I. Riawan, and H. Fajri, “Word Search with Trending Reviews on Twitter,” *Ingénierie des Systèmes d’Information*, vol. 28, no. 2, pp. 351–356, 2023, [Online]. Available: https://doi.org/10.18280/isi.280210

[14] Z. Sutriawan, Muljono, Khairunnisa, Alamin, T. A. Lorosae, and S. Ramadhan, “Improving Performance Sentiment Movie Review Classification Using Hybrid Feature TFIDF , N-Gram , Information Gain and Support Vector Machine,” *Math. Model. Eng. Probl.*, vol. 11, no. 2, pp. 375–384, 2024.

[15] S. Mutmainnah, T. Ansyor Lorosae, and S. Ramadhan, “Model Text Embedding dan TF-IDF+Ngram untuk Meningkatkan Kinerja Algoritma Binary Classifier pada Klasifikasi SMS Palsu,” *J. Sist. Inf. Tgd*, vol. 4, no. 1, pp. 55–64, 2025, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi

[16] R. Maulana, P. A. Rahayuningsih, W. Irmayani, D. Saputra, and W. E. Jayanti, “Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain,” *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 0–6, 2020, doi: 10.1088/1742-6596/1641/1/012060.

[17] M. Namakin, M. Rouhani, and M. Sabzekar, “An Evolutionary Correlation-aware Feature Selection Method for Classification Problems”.

[18] N. Ghatasheh, I. Altaharwa, and K. Aldebei, “Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization : Case of XGBoost in Spam Prediction,” *IEEE Access*, no. July, pp. 84365–84383, 2022.

[19] T. A. Gonsalves, “Feature Selection for Text Classification,” *Comput. Methods Featur. Sel.*, pp. 273–292, 2020, doi: 10.1201/9781584888796-23.

[20] H. Li, J. Li, and H. Dietl, “A Novel Decision Making Approach for Benchmarking the Service Quality of Smart Community Health Centers,” *IEEE Access*, vol. 8, pp. 209904–209914, 2020, doi: 10.1109/ACCESS.2020.3037769.

[21] W. K. Jati and L. Kemas Muslim, “Optimization of Decision Tree Algorithm in Text Classification of Job Applicants Using Particle Swarm Optimization,” in *3rd International Conference on Information and Communications Technology, ICOIACT 2020*, Telkom University, School of Computing, Bandung, Indonesia: Institute of Electrical and Electronics Engineers Inc., 2020, pp. 201–205. doi: 10.1109/ICOIACT50329.2020.9332101.

[22] Z.-G. Liu, Y. Liu, J. Dezert, and Q. Pan, “Classification of incomplete data based on belief functions and K-nearest neighbors,” *Knowledge-Based Syst.*, vol. 89, pp. 113–125, 2015, doi: 10.1016/j.knosys.2015.06.022.

[23] Y. Wan and Q. Gao, “An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis,” *Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015*, pp. 1318–1325, 2016, doi: 10.1109/ICDMW.2015.7.

[24] D. Teekaraman, S. Sendhilkumar, and G. S. Mahalakshmi, “Semantic Provenance Based Trustworthy Users Classification on Book-Based Social Network using Fuzzy Decision Tree,” *Int. J. Uncertainty, Fuzziness Knowldege-Based Syst.*, vol. 28, no. 1, pp. 47–77, 2020, doi: 10.1142/S0218488520500038.

[25] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, “Confusion matrix-based feature selection,” *CEUR Workshop Proc.*, vol. 710, pp. 120–127, 2011.

[26] P. G. Shivakumar and P. Georgiou, “Confusion2Vec: Towards enriching vector space word representations with representational ambiguities,” *PeerJ Comput. Sci.*, vol. 2019, no. 6, 2019, doi: 10.7717/peerj-cs.195.

[27] J. Miharja, J. L. Putra, and N. Hadianto, “Comparison of Machine Learning Classification Algorithm on Hotel Review Sentiment Analysis (Case Study: Luminor Hotel Pecenongan),” *J. Pilar Nusa Mandiri*, vol. 16, no. 1, pp. 59–64, 2020, doi: 10.33480/pilar.v16i1.1131.

[28] S. Sultana, S. S. Hussain, M. Hashmani, J. Ahmad, and M. Zubair, “A deep learning hybrid ensemble fusion for chest radiograph classification,” *Neural Netw. World*, vol. 31, no. 3, pp. 199–209, 2021, doi: 10.14311/NNW.2021.31.010.

[29] X. Liu *et al.*, “Adapting Feature Selection Algorithms for the Classification of Chinese Texts,” 2023. doi: 10.3390/systems11090483.

[30] G. Airlangga, “Spam Detection on YouTube Comments Using Advanced Machine Learning Models: A Comparative Study,” *Brill. Res. Artif. Intell.*, vol. 4, no. 2, pp. 500–508, 2024, doi: 10.47709/brilliance.v4i2.4670.

[31] Ö. Şengel, “A comparative analysis of learning techniques in the context of Turkish spam detection,” *Batman Üniversitesi Yaşam Bilim. Derg.*, vol. 14, no. 1, pp. 43–56, 2024, doi: 10.55024/buyasambid.1501609.

[32] B. Ge, C. He, H. Xu, J. Wu, and J. Tang, “Chinese News Text Classification Method via Key Feature Enhancement,” *Appl. Sci.*, vol. 13, no. 9, 2023, doi: 10.3390/app13095399.

[33] Bekir Parlak and Alper Kursat Uysal, “A novel filter feature selection method for text classification: Extensive Feature Selector,” *J. Inf. Sci.*, vol. 49, no. 1, pp. 59–78, Apr. 2021, doi: 10.1177/0165551521991037.

[34] O. P. Ige and K. H. Gan, “Ensemble Filter-Wrapper Text Feature Selection Methods for Text Classification,” *C. - Comput. Model. Eng. Sci.*, vol. 141, no. 2, pp. 1847–1865, 2024, doi: 10.32604/cmes.2024.053373.

[35] Y. Zhuang, Z. Fan, J. Gou, Y. Huang, and W. Feng, “A importance-based ensemble method using an adaptive threshold searching for feature selection,” *Expert Syst. Appl.*, vol. 267, no. January 2024, p. 126152, 2025, doi: 10.1016/j.eswa.2024.126152.

[36] N. Ghatasheh, I. Altaharwa, and K. Aldebei, “Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction,” *IEEE Access*, vol. 10, no. July, pp. 84365–84383, 2022, doi: 10.1109/ACCESS.2022.3196905.