Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 8, No. 1, March 2026, Page. 9-17 e-ISSN 2685-5518

Classification of Article Types in the ITE Law using the KNN Algorithm with the Application of SMOTE, PCA, and GridSearchCV Hyperparameter Optimization

Alif Alpian Sahrul Muharom[1*], Aditiya Dwi Putro[2], Yohani Setya Rafika[3]
[1, 2,3] Program Studi Teknik Informatika, Direktorat Kampus Purwokerto, Telkom University
[*]Email: piansah@student.telkomuniversity.ac.id

Abstract

The advancement of information technology drives digital transformation, enhancing efficiency but also presenting challenges such as data management and privacy risks due to cybercrime. The Electronic Information and Transactions Law (UU ITE) serves as an essential legal foundation for protecting data and ensuring digital justice. This study employs the K-Nearest Neighbor (KNN) algorithm to classify UU ITE violations based on chronology texts, focusing on Articles 27 and 28 from 323 violation cases. The process includes text preprocessing, weighting, modeling, and evaluation. To address data imbalance, SMOTE (Synthetic Minority Oversampling Technique) and PCA (Principal Component Analysis) were applied. Hyperparameter optimization using GridSearchCV improved model performance. Initial accuracy of 57% increased to 75% after applying SMOTE and PCA, with a final result of 82.62%, a macro average F1-score of 0.82, and a weighted average F1-score of 0.83. The model showed the best performance on "Article 28 Paragraph 2" and the lowest on "Article 27 Paragraph 1". This study demonstrates the potential of Text Mining in supporting digital law enforcement.

Keywords: *Digital transformation, ITE Law, K-Nearest Neighbor (KNN), Text Mining*

INTRODUCTION

The advancement of information technology has induced substantial transformations in numerous facets of human existence, notably the transition from manual to digital data administration (Heliany, 2019). This transition has enhanced efficiency and precision across multiple sectors, including business, education, and government administration (Danuri, 2019). This progress has led to a surge of extensive, diverse, and intricate data, necessitating meticulous management and analysis to yield valuable insights (Wali, 2023). Additionally, new challenges have arisen, including issues of personal data protection and the rise in cybercrime, which has become a significant concern in the digital age (Saragih & Azis, 2020).

To address these difficulties, the Electronic Information and Transactions Law (UU ITE) was established as a legal framework regulating digital activities, encompassing offenses such as defamation, online gambling, and the dissemination of false information (Al Hadad, 2020; Rohmy & Suratman, 2021). The ITE Law provides the legal foundation for pursuing several forms of cybercrime, a trend whose significance is escalating as the number of internet users rises. According to data from the Indonesian Internet Service Providers Association (APJII) in 2023, internet users in Indonesia accounted for 78.19% of the total population, or around 215 million individuals (Winarno, 2021). Nonetheless, this elevated user rate correlates with a rise in criminal complaints, comprising 1,859 instances of bank account use for online gambling, up from 1,914 reports in 2023 (APJII, 2023).

Law enforcement regarding infractions of the ITE Law faces numerous challenges, including insufficient transparency and accountability, which engenders public scepticism about the legal system's effectiveness in Indonesia. In this regard, information technology and artificial intelligence (AI) serve as crucial tools to enhance law enforcement effectiveness, particularly in analyzing crime patterns. Machine learning methods, such as K-Nearest Neighbours (KNN), have demonstrated efficacy for clustering and analyzing data for predictive purposes (Nanda et al., 2022). KNN is a non-parametric technique commonly employed in text mining because of its exceptional capability for feature-based data categorization and proficiency in handling noisy data (Deolika and Taufiq Luthfi, 2019).

Additionally, the performance of KNN can be enhanced through methods such as SMOTE (Synthetic Minority Oversampling Technique), which mitigates data imbalance by creating synthetic samples for minority classes (Sulaiman, 2024), PCA, which diminishes data dimensionality while preserving critical information, thereby reducing complexity, enhancing efficiency, and lowering the likelihood of overfitting (Putra Pamungkas, 2019), and hyperparameter optimization via GridSearchCV to augment model accuracy (Toha, Purwono, and Gata, 2022).

This study investigates the use of the KNN algorithm to predict criminal acts under the ITE Law, specifically within the Indonesian law enforcement framework. This strategy aims to have technology-based systems deliver more precise, transparent, and efficient answers to the challenges of the digital world.

LITERATURE REVIEW

Investigations into the categorization of items under the ITE Law have been undertaken using diverse methodologies. The research titled "Application of Text Mining for Classifying Article Types in the ITE Law Using the Naive Bayes Algorithm" by Farhan (2021) examines the utilization of the Naive Bayes algorithm to categorize infringements of the ITE Law according to the textual chronology of incidents. This study examines Articles 27 and 28, using a dataset of 245 instances of violation chronology data, and applies text preprocessing, weighting, learning, and testing within the Python Flask framework.

Heliany (2019) emphasizes the freedom of expression under the ITE Law but does not implement a specific machine learning technique. The study by Saragih & Azis (2020) uses the K-Nearest Neighbours (KNN) algorithm for text classification. Still, it neglects the prevalent issue of data imbalance commonly encountered in legal classification scenarios. Nanda et al. (2022) employ Principal Component Analysis (PCA) for dimensionality reduction without integrating it with hyperparameter optimization.

Previous studies indicate a persistent research gap in the implementation of comprehensive, optimal classification systems. This work integrates KNN with the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance, employs PCA for dimensionality reduction, and uses GridSearchCV for hyperparameter optimization. This methodology is anticipated to yield superior classification performance compared to prior studies, while also enhancing the research's originality.

METHODS

This research technique employs methodical procedures to develop a classification system for items under the Electronic Information and Transactions Law (UU ITE) using the k-Nearest Neighbours (KNN) algorithm. The study commences with a literature review to identify a suitable methodology; this is followed by data collection from ITE Law documents, data preprocessing, TF-IDF text weighting, and the implementation of the KNN model. The evaluation assesses the model's performance, as depicted in the research flow in Figure 1.

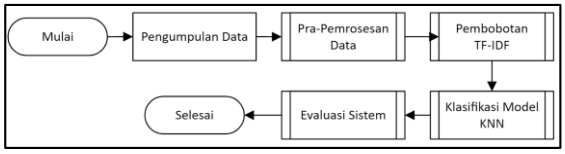


Figure 1. Research Framework

3.1 Data Collection

Chronological data on infractions of the Electronic Information and Transactions Law (UU ITE) under Articles 27 and 28 were obtained from the Supreme Court's rulings via the Supreme Court of the Republic of Indonesia Decision Directory website (www.putusan3.mahkamahagung.go.id; Registrar's Office of the Supreme Court of the Republic of Indonesia).

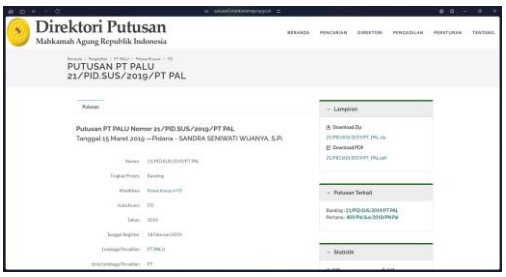
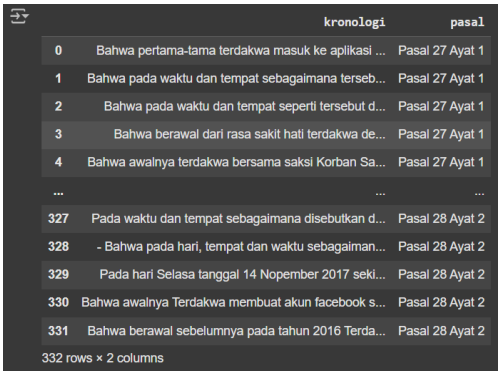


Figure 2. Directory of Decisions

PDF documents were downloaded and meticulously analyzed to determine the sequence of infractions, which were usually located within the legal factual descriptions or the judge's rationale. Irrelevant information, including

tangential legal arguments, was excluded to preserve the dataset's focus.



	kronologi	pasal
0	Bahwa pertama-tama terdakwa masuk ke aplikasi ...	Pasal 27 Ayat 1
1	Bahwa pada waktu dan tempat sebagaimana terseb...	Pasal 27 Ayat 1
2	Bahwa pada waktu dan tempat seperti tersebut d...	Pasal 27 Ayat 1
3	Bahwa berawal dari rasa sakit hati terdakwa de...	Pasal 27 Ayat 1
4	Bahwa awalnya terdakwa bersama saksi Korban Sa...	Pasal 27 Ayat 1
...	...	...
327	Pada waktu dan tempat sebagaimana disebutkan d...	Pasal 28 Ayat 2
328	- Bahwa pada hari, tempat dan waktu sebagaiman...	Pasal 28 Ayat 2
329	Pada hari Selasa tanggal 14 Nopember 2017 seki...	Pasal 28 Ayat 2
330	Bahwa awalnya Terdakwa membuat akun facebook s...	Pasal 28 Ayat 2
331	Bahwa berawal sebelumnya pada tahun 2016 Terda...	Pasal 28 Ayat 2

332 rows × 2 columns

Figure 3. Dataset

The outcomes of the verification and data extraction are preserved in CSV format, comprising two primary columns:

1. The Case Chronology encompasses a comprehensive sequence of events related to the defendant's conduct.
2. Breached Articles, specifically Articles 27 and 28 of the ITE Law, are pertinent to the infraction.

3.2 Pre-Data Processing

1. Data Cleaning

The data purification phase aims to remove all non-alphabetic characters from the text, including symbols, punctuation, numerals, emojis, and URLs. This stage guarantees that the processed data comprises exclusively pure text formed of letters, devoid of any extraneous factors.

2. Case Folding

The case-folding stage standardizes text within a document by converting all characters to a consistent case format, thereby streamlining subsequent data processing.

3. Tokenization

The tokenization phase seeks to decompose normalized text, achieved through Case Folding, into smaller pieces known as tokens, including words and punctuation marks, thereby facilitating subsequent text analysis.

4. Stopword Removal

The stopword removal phase involves eliminating frequently occurring common words that lack substantial meaning, using the Case Folding method.

5. Stemming

The Stemming stage removes affixes from words, yielding their root forms and ensuring the dataset comprises only simpler, more prevalent terms.

3.3 TF-IDF Weighting

The preprocessed data subsequently undergoes a Feature Weighting phase utilizing the TF-IDF weighting method. This method assigns a weight to each word based on its significance within the document.

Weighting is executed via the TF-IDF methodology. The procedure for ascertaining the weight values is as follows:

1. Term frequency (TF) calculation

Enumerating the frequency of a term's occurrence inside the specified document. Terms may denote keywords pertinent to Articles 27 and 28 of the ITE Law. Each document detailing the chronology of cases under Articles 27 and 28 has its TF computed for each term. This is a quantitative depiction of the term's frequency in each document. The TF is computed by the formula provided below.

$$tf_{ij} = \frac{n_{ij}}{\sum n_{IJ}} \quad (3.1)$$

where:

tf_{ij} : number of term frequencies
N_{ij} : number of terms *i* in document *j*
$\sum n_{IJ}$: Total frequency of all terms in document *j*

2. Calculation of the number of documents containing the term

Document Frequency (DF) is calculated by summing the number of occurrences of a term across all documents. The greater a term's frequency across documents, the higher its Document Frequency (DF) value. This study illustrates the prevalence or specificity of a phrase in relation to Articles 27 and 28 of the ITE Law. DF is utilized to compute the Inverse Document Frequency (IDF) in the subsequent phase.

3. Calculation of number of documents (N)

The entire quantity of existing papers will be determined, typically denoted by the symbol "N".

4. Calculation of inverse document frequency (IDF)

The IDF quantifies a term's capacity to differentiate groupings. The IDF value is represented by the equation provided below.

$$idf_i = log \frac{N}{DF_i} \quad (3.2)$$

where:

idf_i : Inverse Document Frequency for term *i*

N : Number of total documents

idf_i : Number of documents that have term *i*

5. Weighting (W)

The TF-IDF weight can be calculated using the following formula.

$$W_{ij} = tf_{ij} \times idf_i \quad (3.3)$$

where:

W_{ij} : TF-IDF weight for term *i* in document *j*

tf_{ij} : Frequency of term *i* in document *j*

idf_i : Inverse Document Frequency for term *i*

3.4 KNN Model Classification

1. Determination of K Value

Establishing the K value in the K-Nearest Neighbours (KNN) method is an essential step in the classification process. The K value denotes the quantity of nearest neighbors employed to ascertain the classification of a test dataset. The KNN model identifies the K nearest neighbours in the test dataset and uses a voting mechanism to determine the most frequently occurring class among them.

2. Calculation of Euclidean Distance

This procedure entails calculating the Euclidean distance between every pair of data points. In text mining, this distance can be computed using the vector representation of the text, e.g., via TF-IDF (Term Frequency-Inverse Document Frequency).

3. Calculation of the Distance of the Nearest K Value

This stage entails identifying the K data points exhibiting the minimal Euclidean distance from the input data. It constitutes the nearest neighbor group.

4. Determination of Majors Based on the K-Score Approach

In accordance with the classification of Articles 27 and 28 of the ITE Law, the number of neighboring groups associated with Article 27 and the number associated with Article 28 will be determined. The predominant category of articles among these neighbors will be regarded as the categorization outcome.gga ini akan dianggap sebagai hasil klasifikasi.

3.4 Smote Oversampling

Data balancing was performed using the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in the target variable. The subsequent results pertain to the data analysis conducted before and after SMOTE was applied.

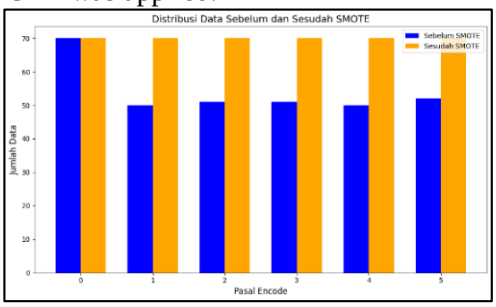


Figure 2. Smote Oversampling

3.5 Dimensionality Reduction – PCA

Principal Component Analysis (PCA) is a dimensionality reduction method employed to decrease the number of features in a dataset while preserving essential information. PCA identifies the principal components that account for the greatest variance in the data, thereby streamlining analysis and enhancing model performance.

The PCA dimensionality reduction approach effectively decreased more than 420 characteristics to 257 principal components, preserving 95% of the cumulative variance. It guarantees the retention of the majority of critical information from the original data, while substantially reducing data dimensionality.

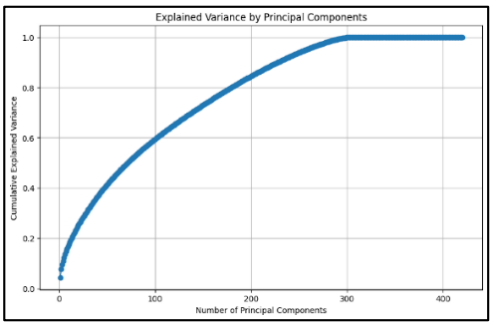


Figure 3. Explained Variance by Principal

The Explained Variance by Principal Components graph indicates that after roughly 257 principal components, subsequent components contribute minimally to the cumulative variance. It illustrates the efficacy of PCA in condensing information from numerous features into a reduced dimension.

3.6 Hyperparameter Tuning KNN

In the K-Nearest Neighbours (KNN) algorithm, hyperparameters include the K value (the number of nearest neighbours evaluated), weights (the method for assigning weights to neighbours), and metrics (the method for quantifying the distance between data points). Selecting appropriate hyperparameter values can influence the accuracy and performance of the KNN model.

3.7 System Evaluation

This study employed systematic testing and analysis during the evaluation phase, using a confusion matrix to assess the effectiveness of a text classification model based on the K-Nearest Neighbours (KNN) method for classifying infractions of the ITE Law. The assessment matrix evaluated the model's accuracy and performance, incorporating TF-IDF weighting and KNN classification. This procedure quantified parameters such as accuracy, precision, recall, and F1-score, enabling a thorough evaluation of the model's ability to categorize infractions of the ITE Law using the training data.

RESULTS AND DISCUSSION

4.1 Modeling K-Nearest Neighbors (KNN)

The KNN model achieved an overall accuracy of 57%, indicating a substantial categorization error rate. Performance differed across classes: Article 27, Paragraph 1, had a precision of 0.56 and a recall of 0.41, indicating the model often struggled to identify this class. Article 27, Paragraph 2 exhibited exceptional performance, achieving an accuracy and recall of 0.96, whereas Article 27, Paragraph 3 demonstrated low precision at 0.18 but strong recall at 0.82. Article 27, Paragraph 4 had moderate precision (0.63) and low recall (0.31), whereas Article 28,

Hasil precission, recall, F1-Score, dan akurasi dari model KNN adalah:

	precision	recall	f1-score	support
Pasal 27 Ayat 1	0.56	0.41	0.47	96
Pasal 27 Ayat 2	0.96	0.96	0.96	50
Pasal 27 Ayat 3	0.18	0.82	0.29	11
Pasal 27 Ayat 4	0.63	0.31	0.42	103
Pasal 28 Ayat 1	0.38	0.95	0.54	20
Pasal 28 Ayat 2	0.73	0.86	0.79	44
accuracy			0.57	324
macro avg	0.57	0.72	0.58	324
weighted avg	0.64	0.57	0.57	324

Figure 4. Result of KNN Model Evaluation

Paragraph 1 demonstrated poor precision (0.38) but excellent recall (0.95). Article 28, Paragraph 2, exhibited a precision of 0.73 and a recall of 0.86. The model demonstrated superior efficacy in classifying the majority class compared to the minority class, as indicated by the macro- and weighted-average scores.

The researchers used SMOTE to address data imbalance by augmenting the representation of the minority class, thereby enhancing the model's ability to identify rare classes. PCA was used to eliminate irrelevant or redundant features, thereby enhancing model efficiency. Furthermore, GridSearchCV was employed for hyperparameter optimization, including the identification of the optimal number of neighbours (k) and the most effective measure to enhance KNN performance.

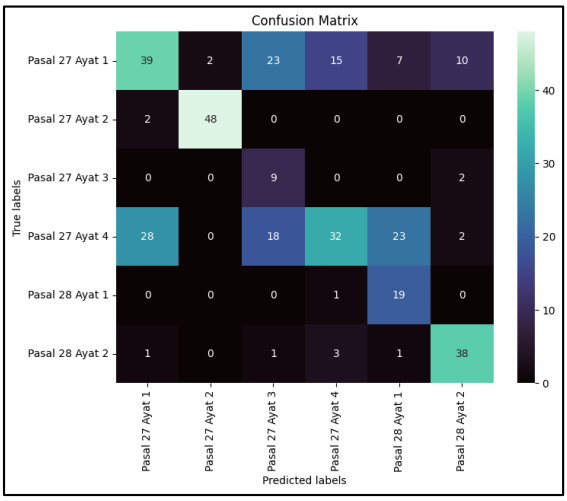


Figure 5. Confusion Matrix of KNN Model

4.2 Smote Oversampling & PCA

The KNN model's total accuracy improved to 75%, up from 57% before applying SMOTE and PCA. The improvement in accuracy indicates that the data preprocessing measures effectively enhanced the model's ability to discern patterns across all categories.

Hasil precission, recall, F1-Score, dan akurasi dari model KNN adalah:

	precision	recall	f1-score	support
Pasal 27 Ayat 1	0.64	0.53	0.58	85
Pasal 27 Ayat 2	1.00	0.97	0.99	72
Pasal 27 Ayat 3	0.60	0.71	0.65	59
Pasal 27 Ayat 4	0.70	0.69	0.70	71
Pasal 28 Ayat 1	0.70	0.82	0.75	60
Pasal 28 Ayat 2	0.84	0.81	0.83	73
accuracy			0.75	420
macro avg	0.75	0.75	0.75	420
weighted avg	0.75	0.75	0.75	420

Figure 6. Result of Smote & PCA

The model achieved optimal performance in the "Article 27 Paragraph 2" category, with precision (1.00) and recall (0.97) both at 1.00. Conversely, it demonstrated the poorest performance in the "Article 27 Paragraph 1" category, with precision (0.64) and recall (0.53), indicating a persistent failure to identify samples in that class accurately. The macro-average F1-score of 0.75 indicates balanced performance across the courses. The assessment demonstrates the importance of SMOTE and PCA in improving the distribution of minority data to optimize the model. Additional parameter adjustment is required to enhance overall performance.

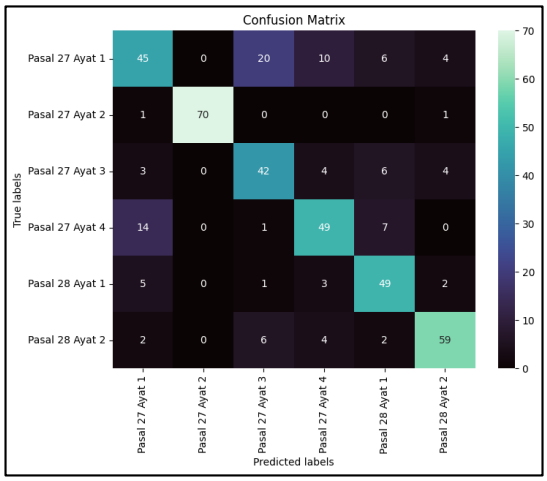


Figure 7. Confusion Matrix Smote & PCA

4.3 Hyperparameter KNN - K Value, Weight, and Metrics

Hyperparameter optimization was conducted for the K-Nearest Neighbours (KNN) model to identify the optimal parameter combination for improved performance. The assessed parameters comprised the K value (2–100), neighbor weights (uniform or distance), and distance measurement metrics (Minkowski, Euclidean, Manhattan).

	parameter_K	parameter_weights	parameter_metric	evaluasi_accuracy	evaluasi_precision	evaluasi_recall	evaluasi_f1_score
0	2	uniform	minkowski	0.747619	0.747619	0.780422	0.752416
1	2	uniform	euclidean	0.747619	0.747619	0.780422	0.752416
2	2	uniform	manhattan	0.578571	0.578571	0.737587	0.581066
3	2	distance	minkowski	0.811905	0.811905	0.813353	0.812072
4	2	distance	euclidean	0.811905	0.811905	0.813353	0.812072
...	...	...	...	...	...	...	...
589	100	uniform	euclidean	0.295238	0.295238	0.660211	0.216740
590	100	uniform	manhattan	0.273810	0.273810	0.407731	0.153212
591	100	distance	minkowski	0.471429	0.471429	0.777420	0.458466
592	100	distance	euclidean	0.471429	0.471429	0.777420	0.458466
593	100	distance	manhattan	0.442857	0.442857	0.824779	0.408599

Figure 8. Iteration Table

The evaluation process employed Stratified K-Fold Cross Validation to maintain balanced data distribution across folds, with outcomes evaluated using accuracy, precision, recall, and F1-score metrics.

The machine learning model evaluation produced 594 parameter combinations, each corresponding to a distinct evaluation result. Parameter combinations, namely parameter_weights = distance and parameter_metric = euclidean, showed higher accuracy, precision, recall, and F1-score than other combinations.

	parameter_K	parameter_weights	parameter_metric	evaluasi_accuracy	evaluasi_precision	evaluasi_recall	evaluasi_f1_score
46	9	distance	euclidean	0.826190	0.826190	0.825653	0.824766
45	9	distance	minkowski	0.826190	0.826190	0.825653	0.824766
63	12	distance	minkowski	0.821429	0.821429	0.820983	0.819812
64	12	distance	euclidean	0.821429	0.821429	0.820983	0.819812
52	10	distance	euclidean	0.816667	0.816667	0.815438	0.814430
	parameter_K	parameter_weights	parameter_metric	evaluasi_accuracy	evaluasi_precision	evaluasi_recall	evaluasi_f1_score
46	9	distance	euclidean	0.826190	0.826190	0.825653	0.824766
45	9	distance	minkowski	0.826190	0.826190	0.825653	0.824766
63	12	distance	minkowski	0.821429	0.821429	0.820983	0.819812
64	12	distance	euclidean	0.821429	0.821429	0.820983	0.819812
51	10	distance	minkowski	0.816667	0.816667	0.815438	0.814430
	parameter_K	parameter_weights	parameter_metric	evaluasi_accuracy	evaluasi_precision	evaluasi_recall	evaluasi_f1_score
233	40	distance	manhattan	0.480952	0.480952	0.858327	0.469996
239	41	distance	manhattan	0.480952	0.480952	0.858017	0.468841
221	38	distance	manhattan	0.485714	0.485714	0.857288	0.475117
251	43	distance	manhattan	0.478571	0.478571	0.857161	0.462713
227	39	distance	manhattan	0.483333	0.483333	0.856599	0.471601
	parameter_K	parameter_weights	parameter_metric	evaluasi_accuracy	evaluasi_precision	evaluasi_recall	evaluasi_f1_score

Figure 9. Iteration Table of the Highest K Value

The iteration table presents the evaluation of the KNN model across essential parameters: the number of nearest neighbours (K), the weight type, and the employed distance metric. This assessment was performed to identify the parameter combination that produced optimal performance. The iteration results indicated that the combination of K = 9, distance weights, and the Euclidean distance metric achieved the greatest accuracy of 0.826190. Distance weights substantially enhanced accuracy by accounting for the proximity of test data to its neighbours. Proximal facts have greater significance, hence rendering classification decisions more susceptible to the influence of pertinent neighbors.

	Pasal 27 Ayat 1	Pasal 27 Ayat 2	Pasal 27 Ayat 3	Pasal 27 Ayat 4
precision	0.657143	1.0	0.828571	0.742857
recall	0.741935	1.0	0.852941	0.702703
f1-score	0.696970	1.0	0.840580	0.722222
support	62.000000	70.0	68.000000	74.000000

Pasal 28 Ayat 1	Pasal 28 Ayat 2	accuracy	macro avg	weighted avg
0.785714	0.942857	0.82619	0.826190	0.831395
0.820896	0.835443	0.82619	0.825653	0.826190
0.802920	0.885906	0.82619	0.824766	0.827615
67.000000	79.000000	0.82619	420.000000	420.000000

Figure 10. Evaluation Table of K = 9, Weight: Distance, Metric: Euclidean

The Euclidean distance metric is significant as it is typically appropriate for data with reduced dimensionality, such as in Principal Component Analysis (PCA). By employing PCA to diminish data dimensionality, less relevant features are discarded, leading to enhanced accuracy in distance computations.

Figure 12 presents the assessment findings of the KNN model utilizing optimal parameters: k = 9, weights = distance, and metric = Euclidean, regarding the classification performance of each class (Article 27 Paragraph 1 to Article 28 Paragraph 2). This model achieved an overall accuracy of 0.826190, indicating superior classification performance. The macro-average F1-score of 0.824766 and the weighted-average F1-score of 0.827615 demonstrate the model's proficiency in handling imbalanced class distributions.

Performance varied across classes, with the highest precision recorded in Article 28 Paragraph 2 at 0.942857, indicating that the model infrequently misclassifies data in this category. Moreover, the greatest F1-score of 0.885906 was observed in the same class, signifying an ideal equilibrium between precision and recall for that category.

Nonetheless, these findings also reveal performance discrepancies among classes. Classes such as Article 27, Paragraph 1, performed worse, with precision and F1 Scores that were not as high as those of Article 28, Paragraph 2. This is probably attributable to an imbalanced data distribution or less different feature patterns among those classes.

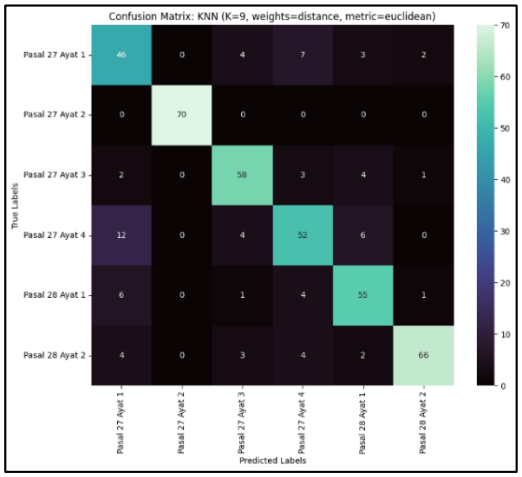


Figure 11. Matrix Confusion of KNN (K=9)

	Pasal 27 Ayat 1	Pasal 27 Ayat 2	Pasal 27 Ayat 3	Pasal 27 Ayat 4
precision	0.085714	0.928571	0.285714	1.000000
recall	1.000000	0.970149	1.000000	0.246479
f1-score	0.157895	0.948905	0.444444	0.395480
support	6.000000	67.000000	20.000000	284.000000

Pasal 28 Ayat 1	Pasal 28 Ayat 2	accuracy	macro avg	weighted avg
0.400000	0.185714	0.480952	0.480952	0.873469
0.933333	1.000000	0.480952	0.858327	0.480952
0.560000	0.313253	0.480952	0.469996	0.491909
30.000000	13.000000	0.480952	420.000000	420.000000

Figure 12. Evaluation Table of K = 40, Weight: Distance, Metric: Manhattan

Figure 14 compares performance across different parameter combinations in the KNN model. The combination of k = 40, weights = distance, and metric = Manhattan had markedly inferior performance compared to the other combinations, particularly compared to k = 9. The model's accuracy was merely 0.480952, considerably lower than the 0.826190 achieved with k = 9. It suggests that excessively large k values can diminish the model's sensitivity to local data patterns.

Moreover, the global average F1-score for this combination was merely 0.469996, signifying a substantial imbalance in classification among the classes. The disparity is apparent in the performance fluctuations of the evaluation indicators between classes, especially in Article 27, Paragraph 1. The precision for this class was merely 0.085714, signifying that the model often misclassified data into this category, despite its comparatively high recall. It may result from insufficient data representation in this class in the nearest-neighbour distribution.

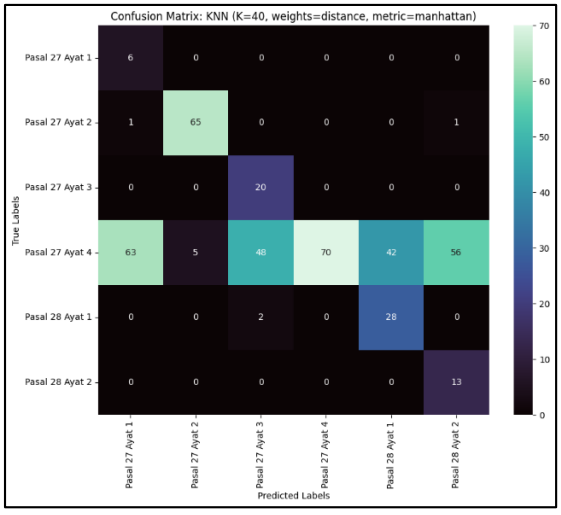


Figure 13. Matrix Confusion of KNN (K=40)

The utilization of the Manhattan distance metric in this combination is another aspect contributing to performance loss. This metric is less appropriate when data has been subjected to dimensionality reduction via PCA, as the Manhattan distance is more sensitive to absolute changes in each dimension, whereas PCA data is better suited to distance metrics such as the Euclidean distance.

CONCLUSION

A text-mining classification model using the KNN algorithm successfully categorized articles in the ITE Law into six distinct categories. This procedure encompassed critical stages, including data preprocessing, addressing data imbalance with the SMOTE technique, dimensionality reduction via PCA, and parameter optimization with GridSearchCV. The initial model achieved 57% accuracy, which rose to 75% after implementing SMOTE and PCA, indicating a notable performance improvement.

The hyperparameter assessment identified the optimal configuration as K = 9, weights = "distance," and metric = "euclidean/minkowski," resulting in an accuracy of 82.62%, a macro F1-score of 0.82, and a weighted F1-score of 0.83. The model excelled on "Article 28 Paragraph 2," but "Article 27 Paragraph 1" requires improvement in precision and recall. The preprocessing and optimisation procedures improved model performance.

Model development should be conducted by:

1. Implement alternative algorithms, including Random Forests, Support Vector Machines (SVMs), and Deep Learning, for performance evaluation.
2. Employ larger, more diverse datasets to enhance the generalizability and robustness of categorisation outcomes.
3. Incorporate modern natural language processing (NLP) methodologies, including word embeddings and transformer models, to enhance the quality of legal text representation.

REFERENCES

Al Hadad, A. (2020). *POLITIK HUKUM DALAM PENERAPAN UNDANG-UNDANG ITE UNTUK MENGHADAPI DAMPAK REVOLUSI INDUSTRI 4.0. Khazanah Hukum*, 2(2), 65–72. https://doi.org/10.15575/kh.v2i2

APJII. (2023, March 10*). Survei APJII Pengguna Internet di Indonesia Tembus 215 Juta Orang. APJII.*

Danuri, M. (2019*). PERKEMBANGAN DAN TRANSFORMASI TEKNOLOGI DIGITAL. INFOKAM*, 15(2), 116–123. https://doi.org/10.53845/infokam.v15i2.178

Deolika, A., & Taufiq Luthfi, E. (2019). *ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING. Jurnal Teknologi Informasi*, 3(2).

Farhan. (2021). *PENERAPAN TEXT MINING UNTUK KLASIFIKASI JENIS PASAL UU ITE MENGGUNAKAN ALGORITMA NAIVE BAYES.*

Heliany, I. (2019). *Wonderful Digital Tourism Indonesia Dan Peran Revolusi Industri Dalam Menghadapi Era Ekonomi Digital 5.0. Destinesia Jurnal Hospitaliti Dan Pariwisata*, 1(1), 21–35. https://doi.org/10.31334/jd.v1i1.551

Kepaniteraan Mahkamah Agung Republik Indonesia. (n.d.). Direktori Puturan Mahkamah Agung. Publikasi Dokumen Elektronik Putusan Seluruh Pengadilan Di Indonesia. Retrieved December 29, 2024, from www.putusan3.mahkamahagung.go.id.

Nanda, R., Haerani, E., Gusti, S. K., & Ramadhani, S. (2022). *Klasifikasi Berita Menggunakan Metode Support Vector Machine. Jurnal Nasional Komputasi Dan Teknologi Informasi*, 5(2).

Putra Pamungkas, D. (2019). *Ekstraksi Citra menggunakan Metode GLCM dan KNN untuk Indentifikasi Jenis Anggrek* (Orchidaceae). INNOVATICS, 1(2), 51–56. http://innovatics.unsil.ac.id

Rohmy, A. M., & Suratman, T. (2021). *UU ITE Dalam Perspektif Perkembangan Teknologi Informasi dan Komunikasi. DAKWATUNA Jurnal Dakwah Dan Komunikasi Islam*, 7(2), 309–339.

Saragih, Y. M., & Azis, D. A. (2020). *Perlindungan Data Elektronik Dalam Formulasi Kebijakan Kriminal Di Era Globalisasi. SOUMATERA LAW REVIEW*, 3(2), 265–279. https://doi.org/10.22216/soumlaw.v3i1.4125

Sulaiman, K. (2024). *Dampak Pengambilan Sampel Data untuk Optimalisasi Data Tidak Seimbang pada Klasifikasi Penipuan Transaksi E-Commerce. Indonesian Journal of Computer Science Attribution*, 13(2), 3070.

Toha, A., Purwono, P., & Gata, W. (2022). *Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV*. *Buletin Ilmiah Sarjana Teknik Elektro*, 4(1), 12–21. https://doi.org/10.12928/biste.v4i1.6079

Wali, M. (2023). *Penerapan & Implementasi Big Data di Berbagai Sektor (Pembangunan Berkelanjutan Era Industri 4.0 dan Society 5.0)* (Efitra, A. Juansa, & Sepriano, Eds.; 1st ed.). PT. Sonpedia Publishing Indonesia.

Winarno, W. A. (2021). *SEBUAH KAJIAN PADA UNDANG-UNDANG INFORMASI DAN TRANSAKSI ELEKTRONIK (UU ITE)*. *Jurnal Ekonomi Akuntansi Dan Managemen*, 10(1), 43–48.