VOL. 23 (2) (2026), 105-115 p-ISSN: 693-7554, e-ISSN:2654-3990

A Multi-Ontology Approach for Deforestation Issue Analysis in Indonesian Digital Media

Dinar Munggaran Akhmad[1*], **Ersa Resita**[2], **Carli Apriansyah Hutagalung**[3], **Ryan Tsany Adelmar**[4], **Muhammad Dheki Akbar**[5]

[1*,2,3,4,5] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta, Indonesia

Abstract

The increasing volume of online news discussing deforestation has created challenges for automatically identifying and categorizing environmental issues because conventional text classification methods often fail to capture semantic relationships between terms. This study proposes a multi-ontology semantic feature extraction framework that integrates DBpedia, IndoWordNet, and a domain-specific keyword ontology to enrich document representation for Indonesian deforestation issue classification. News articles were preprocessed through case folding, tokenization, stopword removal, and stemming. Semantic scores generated from the three ontology resources were combined to produce semantic feature vectors and initial issue labels, which were subsequently classified using a Support Vector Machine (SVM). To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) and class-weight balancing were incorporated during model training. The proposed model was evaluated using 5-fold stratified cross-validation on a dataset of 567 labeled sentences collected from Indonesian online news portals. Experimental results achieved an accuracy of 73.24%, a macro precision of 61.39%, a macro recall of 47.63%, and a macro F1-score of 51.55%. These results indicate that the proposed multi-ontology framework provides richer semantic representations for deforestation issue classification and offers an interpretable approach for supporting large-scale analysis of Indonesian online news.

Keywords: deforestation, multi-ontology, semantic feature extraction, Support Vector Machine, text classification.

1. Introduction

Deforestation remains one of the most significant environmental challenges because it affects biodiversity, ecosystem sustainability, and climate stability. Indonesia, as one of the countries with the largest tropical forest areas, continues to experience forest degradation caused by agricultural expansion, plantation development, land conversion, and infrastructure projects. Besides reducing forest cover, these activities also contribute to environmental degradation and greenhouse gas emissions. Recent studies have highlighted that land-use change is one of the dominant drivers of deforestation, emphasizing the importance of effective environmental monitoring and sustainable forest management to support conservation policies [1], [2].

The rapid development of digital media has significantly increased the availability of online news discussing environmental issues. News portals continuously publish articles covering various aspects of deforestation, including its causes, impacts, government policies, and environmental events. This large volume of textual information provides valuable knowledge for researchers, policymakers, and environmental organizations. However, manually analyzing and categorizing news articles is inefficient and requires substantial effort. Consequently, automatic text classification has become an important task in Natural Language Processing (NLP) to organize textual information into meaningful categories.

*Corresponding author. *E-mail address:dinar.munggaran@unj.ac.id*
Received: 25 July 2026, Accepted: 30 July 2026 and available online 31 July 2026
DOI: https://doi.org/10.33751/komputasi.v23i2.118

Support Vector Machine (SVM) is one of the most widely adopted machine learning algorithms for text classification because of its effectiveness in handling high-dimensional textual data and its ability to produce reliable classification results. Previous studies have demonstrated that SVM performs well in various Indonesian text classification tasks, including sentence classification and fake news detection [3], [4]. Nevertheless, the effectiveness of SVM is strongly influenced by how documents are represented. Conventional feature extraction methods, such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), mainly rely on word frequencies and often ignore semantic relationships among words. Consequently, documents discussing the same concept using different terms may produce different feature representations, reducing classification performance.

To address this limitation, semantic enrichment has been introduced to improve document representation by incorporating semantic knowledge into textual features. Semantic similarity measures the relatedness between concepts rather than relying solely on lexical matching, allowing documents with similar meanings to be represented more consistently. Previous studies have shown that ontology-based semantic similarity can improve feature representation and support better machine learning performance in various text classification tasks [5], [6].

One of the most widely used ontology resources for semantic enrichment is DBpedia, which extracts structured knowledge from Wikipedia and publishes it as Linked Open Data. Through DBpedia Spotlight, textual entities can be automatically recognized and linked to their corresponding knowledge base entities, enabling documents to be enriched with semantic information beyond their lexical content [7], [8]. In addition, ontology-based document representation using DBpedia has been shown to improve short text classification by reducing data sparsity and providing more meaningful semantic features [9].

Besides entity knowledge, lexical ontology also plays an important role in representing semantic relationships among words. A lexical ontology organizes words into synonym sets (synsets) and defines semantic relationships between concepts, enabling different lexical forms that convey similar meanings to be mapped into the same semantic representation. Previous work on building an Indonesian lexical ontology demonstrated that semantic concepts can be constructed by mapping lexical entries into synset structures, providing a useful semantic resource for Natural Language Processing applications such as information retrieval, machine translation, and knowledge engineering [10].

Although previous studies have successfully applied ontology and semantic similarity in several domains, most of them employ a single ontology resource or focus on domains other than environmental news [11], [12]. Research integrating DBpedia with a lexical ontology through semantic mapping based on synonym-set structures for Indonesian deforestation news classification is still limited. Therefore, this study proposes a multi-ontology semantic enrichment approach that combines DBpedia and a lexical ontology through semantic mapping to compute semantic similarity before classification using the Support Vector Machine (SVM) algorithm. The proposed approach is expected to generate richer semantic feature representations and improve the classification of Indonesian online news into four issue categories: causes, impacts, policies, and events.

2. Methods

The research framework consists of five main stages, starting from data collection and ending with the evaluation of the classification results. The overall workflow of the proposed approach is illustrated in **Figure 1**.

Figure 1. Research Flow

This study follows a text classification workflow consisting of five stages: news article collection, text preprocessing, multi-ontology feature extraction, deforestation issue classification, and performance analysis and evaluation, as shown in Figure 1. The workflow begins with collecting and preparing the dataset, followed by semantic feature extraction, model training, and performance evaluation. Unlike conventional text classification methods, this study introduces a semantic enrichment stage by integrating multiple ontology resources before the classification process to produce richer document representations [13], [14].

2.1. Media Article Collection

News articles related to deforestation were collected from several Indonesian online news portals. The collection process was conducted using keywords associated with deforestation, including *deforestation*, *forest*, *illegal logging*, *forest fire*, and *land-use change*. After the articles had been collected, each document was manually assigned to one of four issue categories: Cause, Impact, Policy, and Event. The labeled dataset was then divided into training and testing data for the classification process.

2.2. Text Preprocessing

Before feature extraction, all collected documents undergo a preprocessing stage to improve text quality and reduce inconsistencies in word representation. Previous studies have shown that preprocessing helps reduce noise in textual data and improves the effectiveness of text classification models. Common preprocessing techniques include case folding, tokenization, stopword removal, and stemming [14], [15]. The preprocessing stage consists of four sequential steps.

a. **Case folding** converts all characters into lowercase and removes unnecessary characters, such as punctuation marks and numbers.
b. **Tokenization** splits each document into individual word tokens that can be processed independently.
c. **Stopword removal** eliminates common Indonesian words that carry little semantic information and have limited contribution to the classification process.
d. **Stemming** converts each token into its root form to reduce lexical variations among words with the same meaning.

The resulting normalized tokens are then used as input for the semantic feature extraction stage.

2.3. Multi-Ontology Feature Extraction

The proposed multi-ontology feature extraction framework integrates DBpedia, IndoWordNet, and a keyword ontology to generate semantic representations of deforestation-related texts. These complementary ontology resources provide entity-level knowledge, lexical semantic relationships, and domain-specific terminology, enabling richer semantic features than conventional lexical representations [16], [17].

2.3.1. Ontology Resources

Table 1 summarizes the ontology resources employed in this study.

Table 1. Ontology resources used in the proposed framework

Resource	Language	Structure	Purpose	Source
DBpedia Spotlight	Multilingual	RDF Knowledge Graph	Entity Linking & Categorization	https://www.dbpedia-spotlight.org/
IndoWordNet	Indonesian	Lexical Database (WordNet)	Semantic Similarity & Synonyms	https://wn-msa.sourceforge.net/

2.3.2. Multi-Ontology Architecture

The framework consists of three ontology components operating in parallel. DBpedia identifies domain entities through entity matching, IndoWordNet captures semantic relationships using synsets, while the keyword ontology serves as a complementary semantic resource. The outputs of these components are aggregated to produce ontology-based feature vectors.

1. DBpedia Feature Extraction

DBpedia is an RDF-based knowledge graph that represents entities and their semantic relationships, making it suitable for ontology-driven feature extraction and semantic text representation [18], [19]. Each normalized token is matched against the DBpedia entity dictionary, and matched entities contribute to the corresponding category score.

$$Score_{\text{DBpedia}}(c) = \sum_{t \in T} \sum_{k \in K_c} M\,atch(t, k) \quad (1)$$

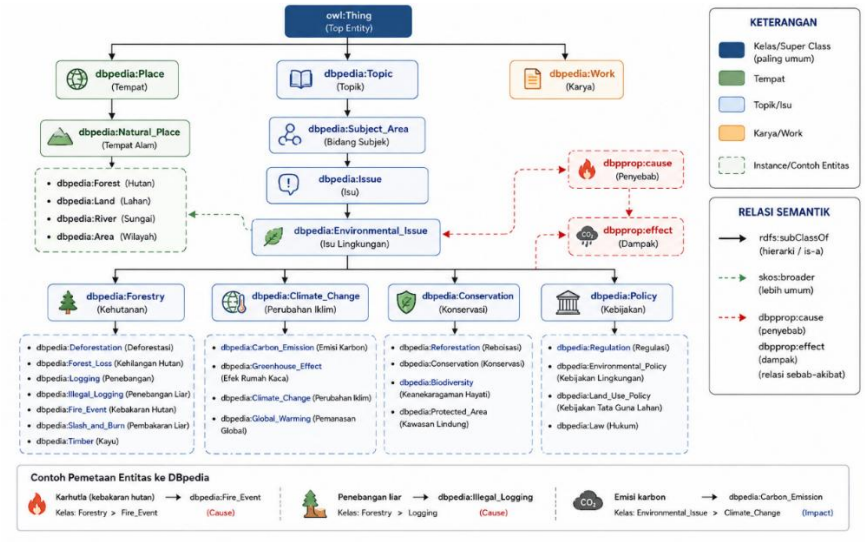

Figure 2. Simplified DBpedia ontology adapted for deforestation issue classification

Figure 2 illustrates the simplified DBpedia ontology adopted in this study. The ontology organizes entities hierarchically from the top-level concept (*owl:Thing*) into several domain-specific classes relevant to deforestation, including Forestry, Climate Change, Conservation, and Policy. Semantic relations such as *rdfs:subClassOf*, *skos:broader*, *dbpprop:cause*, and *dbpprop:effect* are used to model hierarchical and causal relationships among entities. During feature extraction, normalized tokens are matched against this ontology to identify corresponding entities and assign semantic scores to one of the four predefined issue categories: Cause, Impact, Policy, and Event.

2. IndoWordnet Feature Extraction

IndoWordNet provides lexical semantic knowledge by organizing words into synsets connected through semantic relations such as synonymy, hypernymy, and hyponymy. These semantic relationships enable the identification of conceptually related terms beyond exact lexical matching and improve semantic feature representation (Sikelis et al., 2021). Each normalized token is mapped to predefined deforestation synsets, and the corresponding semantic category score is accumulated based on successful semantic matches.

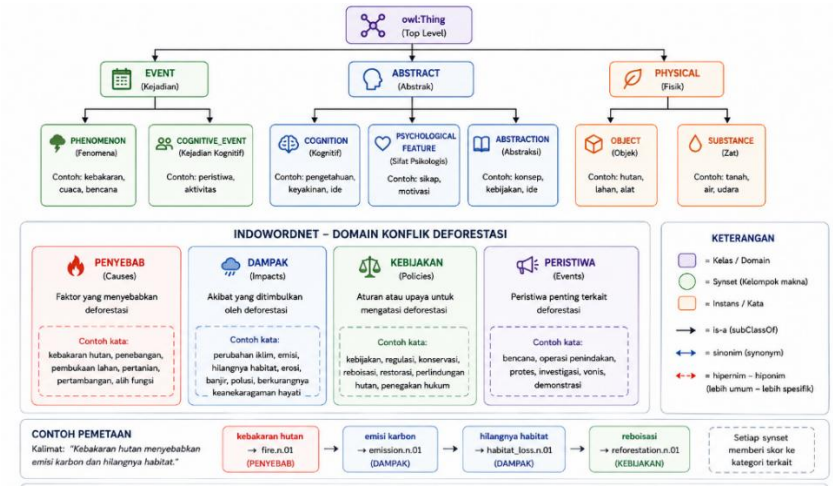

Figure 3. Simplified IndoWordNet lexical ontology adapted for deforestation issue classification

Figure 3 illustrates the simplified IndoWordNet lexical ontology employed in this study. The upper layer presents the general semantic hierarchy derived from IndoWordNet, while the lower layer represents the domain adaptation for deforestation issue classification. Domain-specific concepts are organized into four issue categories, namely Cause, Impact, Policy, and Event. Each category contains representative synsets and example lexical terms that are used during semantic matching. The bottom section illustrates an example of semantic mapping, where words extracted from an input sentence are associated with the corresponding synsets before semantic scores are accumulated for each issue category.

3. Keyword Ontology Feature Extraction

The keyword ontology performs rule-based semantic matching using manually curated domain-specific keywords. Exact matches receive a higher weight than partial matches.

$$Score_{\text{Keyword}}(c) = \sum_{t \in T} \sum_{k \in K_c} w_{\text{match}} \; Match(t, k) \tag{2}$$

where

$$w_{\text{match}} = \begin{cases} 2, & \text{exact match} \\ 1, & \text{partial match} \end{cases}$$

4. Feature Combination

The scores obtained from the three ontology sources are combined by summing the scores of each component for every category. The combined score is calculated as follows:

$$Score_{combined}(c) = Score_{dbpedia}(c) + Score_{wordnet}(c) + Score_{keyword}(c) \tag{3}$$

Where c∈{CAUSE, IMPACT, POLICY, EVENT}

The combined semantic scores are used to automatically assign an initial issue category for each sentence. This automatically generated category is referred to as the **AI Label** or **pseudo-label**, because it is produced through ontology-based semantic matching rather than manual annotation. The pseudo-label provides an initial semantic categorization that is subsequently used during the supervised classification process.

The final pseudo-label is determined by selecting the category with the highest combined score:

$$Label = \arg\max_{c} Score_{combined}(c) \tag{4}$$

The confidence score is computed using proportional normalization:

$$Confidence = \frac{Score_{combined}(Label)}{\sum_{c} Score_{combined}(c)} \tag{5}$$

Where the confidence value ranges from 0 to 1. A confidence score closer to 1 indicates that the predicted category is substantially more dominant than the other candidate categories.

2.4. Deforestation Issue Classification

In this study, four issue categories are defined:

1. **CAUSE**, covering activities that trigger deforestation, such as forest fires, illegal logging, agricultural expansion, and mining.
2. **IMPACT**, covering environmental consequences, including climate change, biodiversity loss, ecosystem degradation, and air pollution.
3. **POLICY**, covering regulations, laws, moratoriums, and conservation initiatives.
4. **EVENT**, covering incidents related to deforestation, such as law enforcement operations, arrests, court decisions, protests, and land conflicts.

The ontology-based semantic feature extraction stage produces semantic scores and pseudo-labels for each sentence. Subsequently, each sentence is represented using TF-IDF features, while the ontology-derived semantic scores are incorporated as additional semantic features. The resulting feature vectors are then used to train a Support Vector Machine (SVM) classifier.

2.4.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is employed to classify deforestation issues into the four predefined categories. SVM identifies the optimal hyperplane that separates different classes in a high-dimensional feature space [20], [21].

1. Linear SVM

A linear SVM defines the decision boundary as $w^T x + b = 0$. Multi-class classification is implemented using the One-vs-Rest (OvR) strategy. The predicted class is computed as $\hat{y} = \arg\max_k (w_k^T x + b_k)$, where k represents the four issue categories.

2. Soft Margin Classification

To handle overlapping classes, soft margin classification is adopted. The optimization objective is

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i, \tag{6}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0, \tag{7}$$

where $\boldsymbol{C}$ is the regularization parameter that controls the trade-off between margin maximization and classification error.

2.4.2. TF-IDF Feature Extraction

Semantic feature extraction was performed by integrating statistical and semantic representations to generate a richer document representation for deforestation issue classification. Following the preprocessing stage, each news article was first transformed into a numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme. TF-IDF assigns a weight to each term according to its frequency within a document and its distribution across the document collection, enabling more representative terms to receive higher weights. The TF-IDF weight of term $t$in document $d$is calculated as [14], [15].

Text is represented using TF-IDF features. The term frequency (TF) is defined as

$$TF(t, d) = 1 + \log\left(f_{t,d}\right), \tag{8}$$

where $f_{t,d}$denotes the frequency of term $t$in document d. The inverse document frequency (IDF) is computed as

$$IDF(t) = \log\frac{N}{df(t)}, \tag{9}$$

where $N$is the total number of documents and $df(t)$is the document frequency of term t. The TF-IDF weight is calculated as

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t). \tag{10}$$

The TF-IDF vectorizer is configured with max_features = 1000, ngram_range = (1,2), and sublinear_tf = True.

2.4.3. Hyperparameter Tuning

Hyperparameter tuning is performed using GridSearchCV with 5-fold stratified cross-validation. The regularization parameter $\boldsymbol{C}$ is searched over $\{\mathbf{1.0}, \mathbf{2.0}, \mathbf{5.0}, \mathbf{10.0}\}$, with the macro F1-score used as the evaluation metric.

2.4.4. Class Imbalance Handling

To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples for minority classes, while class_weight = "balanced" is employed to automatically adjust class weights during SVM training. SMOTE is one of the most widely adopted oversampling techniques for handling imbalanced datasets because it generates new minority-class instances through interpolation rather than simple duplication, thereby reducing the risk of overfitting and improving classifier generalization [22], [23]. In this study, each minority-class instance is paired with one of its k nearest minority-class neighbors, and a synthetic sample is generated according to where denotes the current minority-class sample, is one of its k nearest minority-class neighbors, and is a uniformly distributed random number.

The generated synthetic sample is located along the line segment connecting the two minority-class instances, thereby increasing the diversity of minority samples while preserving their underlying data distribution [22], [24]. In addition, the parameter class_weight = "balanced" automatically adjusts the penalty assigned to each class according to its frequency in the training data, allowing the Support Vector Machine (SVM) classifier to reduce bias toward majority classes and improve classification performance on imbalanced datasets [25].

2.5. Analysis and Evaluation

The classification results are evaluated using four commonly adopted performance metrics: Accuracy, Precision, Recall, and F1-score. These metrics are calculated from the confusion matrix by comparing the predicted labels with the manually assigned labels. They provide complementary information about the classifier's ability to correctly identify each issue category rather than relying on accuracy alone [13], [14]. The evaluation metrics are defined as follows:

Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

where **TP** (True Positive) and **TN** (True Negative) represent correctly classified instances, while **FP** (False Positive) and **FN** (False Negative) denote incorrectly classified instances.

Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

which measures the proportion of correctly predicted positive instances among all positive predictions.

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

which measures the proportion of actual positive instances that are correctly identified by the classifier.

F1-score

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{14}$$

which represents the harmonic mean of Precision and Recall, providing a balanced evaluation of the classification performance, particularly when class distributions are unequal.

The proposed multi-ontology SVM model was evaluated using these four metrics to assess its effectiveness in classifying Indonesian online news articles into the predefined issue categories (Cause, Impact, Policy, and Event).

3. Result and Discussion

This section presents the implementation results of the proposed system. The discussion begins with the preprocessing results, followed by ontology mapping, semantic feature extraction, and the final classification output. The findings are then discussed by highlighting the contribution of the multi-ontology approach to document representation and comparing them with related studies.

3.1. Dataset

The dataset used in this study consists of Indonesian online news articles related to deforestation collected from three national news portals, namely Mongabay Indonesia, Kompas Lestari, and Kompas Regional. The articles were published between September 2025 and August 2026. Each article was processed through a sentence-splitting stage, where every sentence was treated as an individual instance for the classification task. Subsequently, each sentence was manually assigned to one of four predefined issue categories: Cause, Impact, Policy, and Event. The general characteristics of the dataset are presented in Table 2.

Table 2. General Characteristics of the Dataset

Characteristic	Value
News sources	Mongabay Indonesia, Kompas Lestari, Kompas Regional
Publication period	September 2025 – August 2026
Number of articles	14
Number of sentences	567
Number of words	100,010*
Number of characters	74,647*
Issue categories	Cause, Impact, Policy, Event

As shown in Table 1, the dataset comprises 14 news articles, resulting in a total of 567 labeled sentences used as the input for the classification model. The corpus contains approximately 100,010 words and was divided into training and testing sets using a 75:25 split ratio. To improve the robustness and reproducibility of the experimental results, model hyperparameters were optimized using 5-fold Stratified Cross Validation during the training phase. The distribution of the dataset across the four issue categories is presented in Table 3.

Table 3. Distribution of Sentences Across Issue Categories

Category	Number of Sentences	Percentage
Cause	381	67.2%
Impact	80	14.1%
Policy	84	14.8%
Event	22	3.9%

Total	**567**	**100%**

Table 2 indicates that the dataset is class-imbalanced, with the Cause category representing the majority class (381 sentences, 67.2%), while the Event category represents the minority class (22 sentences, 3.9%). Such an imbalance may bias the classifier toward the majority class. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during the training stage to generate a more balanced class distribution before training the Support Vector Machine (SVM) classifier.

3.2. Pre-Processing Results

The preprocessing stage transforms the original news text into a normalized representation before ontology mapping and classification. The generated output at each preprocessing stage is presented in Table 4, illustrating how the text is progressively processed through cleaning, tokenization, stopword removal, and stemming. The final preprocessed tokens provide a consistent textual representation for the subsequent semantic analysis.

Table 4. Example of preprocessing results.

Step	Description	Example Output
Cleaning	Remove URLs, numbers, punctuation and convert text to lowercase	kebakaran hutan di kalimantan menyebabkan kerusakan ekosistem
Tokenization	Split text into individual tokens	["kebakaran", "hutan", ...]
Stopword Removal	Remove Indonesian stopwords	["kebakaran", "hutan", "kalimantan", ...]
Stemming	Convert tokens into root forms using Sastrawi	["bakar", "hutan", "kalimantan", "sebab", "rusak", "ekosistem"]

3.3. Multi-Ontology Feature Extraction

The proposed multi-ontology framework extracts semantic features by integrating three complementary ontology resources, namely DBpedia, IndoWordNet, and a keyword-based ontology. This semantic enrichment process aims to generate richer document representations before the classification stage. The extracted features include keyword matches, ontology entities, semantic synsets, combined semantic scores, and confidence values.

3.3.1. Keyword Matching Results

Keyword matching was performed on all 567 sentences in the dataset to identify domain-specific terms associated with the four predefined issue categories. Table 1 presents the distribution of keyword matches. The results indicate that the **Cause (PENYEBAB)** category obtained the highest number of keyword matches, with a total of **1,847 matches** (average **3.26 matches per sentence**) appearing in **423 sentences**. Meanwhile, the **Event (PERISTIWA)** category produced the fewest matches, with only **156 matches** distributed across **89 sentences**. This distribution reflects the dominance of causal information in Indonesian deforestation news and the relatively limited occurrence of event-specific terminology.

3.3.2. DBpedia Feature Extraction

DBpedia feature extraction was carried out using keyword matching as a fallback mechanism because the DBpedia Spotlight API was unavailable during the experiment. Despite this limitation, the extraction process successfully identified an average of **0.89 entities per sentence**.

A total of **504 entities** were recognized from the corpus, covering 312 sentences (55%), while 89 entities could not be assigned to any predefined issue category. Among the four issue categories, the Cause category obtained the highest semantic contribution, followed by Impact, Policy, and Event, indicating that most recognized entities are strongly associated with activities causing deforestation.

3.3.3. IndoWordNet Feature Extraction

The IndoWordNet feature extraction process employed manually constructed deforestation-related synsets. On average, 1.24 semantic synsets were identified for each sentence.

The fire synset appeared most frequently (287 matches), followed by deforestation (156), climate_change (98), biodiversity (67), policy (45), conservation (34), disaster (23), and enforcement (12). These findings indicate that wildfire-related terminology dominates the collected news articles, which is consistent with the characteristics of the dataset.

3.3.4. Feature Combination

The semantic scores obtained from DBpedia, IndoWordNet, and the keyword-based ontology were combined into a unified semantic feature vector for each sentence. The resulting feature vectors were normalized to ensure comparable feature scales.

The combined scores reveal a substantial difference between issue categories. The Cause category achieved the highest average semantic score (4.69), followed by Impact (1.58), Policy (0.90), and Event (0.35). This result suggests that the proposed semantic enrichment framework effectively differentiates between the four deforestation issue categories by generating distinctive semantic representations.

3.3.5. Confidence Score Distribution

Confidence scores were calculated using proportional normalization of semantic scores. The distribution demonstrates the reliability of the ontology-based pseudo-labeling process.

Approximately 27.5% of the sentences achieved confidence values between 0.90 and 1.00, while 33.3% fell within the 0.70–0.89 range. Overall, 60.8% of all sentences obtained confidence scores greater than 0.70, indicating that the ontology-based semantic mapping can assign issue labels with relatively high confidence for the majority of the dataset. Conversely, only 15.5% of the sentences obtained confidence values below 0.50, suggesting that these instances contain ambiguous semantic information or insufficient ontology evidence.

3.4. Deforestation Issue Classification

The final stage of the proposed framework is deforestation issue classification. Each sentence is represented using the semantic features extracted from the proposed multi-ontology framework and then classified into one of four issue categories: Cause, Impact, Policy, or Event. The semantic feature vectors are used as input to the Support Vector Machine (SVM) classifier to produce the final prediction.

3.4.1. Overall Classification Performance

The proposed Support Vector Machine classifier with a linear kernel was evaluated using 5-fold stratified cross-validation. The model achieved an accuracy of 73.24%, macro precision of 61.39%, macro recall of 47.63%, and macro F1-score of 51.55%. These results demonstrate that the proposed classification framework is capable of automatically categorizing Indonesian deforestation news into four issue categories with acceptable performance.

3.4.2. Performance by Category

Performance varies considerably across issue categories. The Cause category achieved the best performance, with 85% precision, 85% recall, and an F1-score of 85%. Conversely, the Event category obtained the lowest performance, with an F1-score of only 25%. The lower performance can be attributed to the limited number of training instances and the relatively small amount of event-related semantic evidence available in the dataset.

3.4.3. Confusion Matrix Analysis

The confusion matrix shows that the classifier tends to predict samples as Cause, reflecting the dominance of this category in the original dataset. The largest classification error occurred when Policy instances were incorrectly classified as Cause, indicating that these categories often share overlapping vocabulary and semantic characteristics. This observation suggests that although SMOTE improves class balance numerically, semantic overlap between issue categories remains a challenge for automatic classification.

3.4.4. Comparison with Other Methods

The proposed SVM model combined with SMOTE outperformed several baseline classifiers. Compared with Random Forest, Logistic Regression, Naïve Bayes, and SVM without SMOTE, the proposed configuration achieved the highest macro F1-score (51.55%). In particular, incorporating SMOTE improved the macro F1-score by approximately 3% compared with the same classifier trained without class balancing.

3.4.5. Ablation Study

An ablation study was conducted to evaluate the contribution of each component within the proposed classification pipeline. The results indicate that SMOTE contributes most significantly to improving classification performance, whereas data augmentation alone slightly decreases performance. This finding suggests that handling class imbalance is more influential than simple augmentation for the current dataset.

4. Conclusion

Based on the results of this study, several conclusions can be drawn. First, deforestation issue classification into four predefined categories, namely Cause, Impact, Policy, and Event, can be effectively automated using a linear Support Vector Machine (SVM) classifier. Second, the combination of TF-IDF feature extraction, SMOTE, and class weighting improves the classification performance on imbalanced datasets by reducing the bias toward majority classes. Third, the classification performance varies considerably across issue categories, with categories containing a larger number of training samples generally achieving higher F1-scores. Finally, the proposed ontology-based pseudo-labeling approach provides an efficient mechanism for generating initial labels for deforestation-related text. Nevertheless, expert validation remains necessary to ensure label quality and improve the reliability of the annotated dataset.

5. Acknowledgement

The authors would like to express their sincere gratitude to the Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, for providing financial support through its internal research funding program.

References

[1] D. Florini, I. Darmatama, T. Wulandari, M. Devianri, and Reflis, "Analisis deforestasi dan dinamika tutupan lahan berbasis data KLHK terhadap kesesuaian rencana tata ruang wilayah di Provinsi Bengkulu," Integrative Perspectives of Social and Science Journal, 2026.

[2] R. S. Nasution, S. A. Lubis, L. Immanuel, M. I. Parinduri, and M. R. Rizqullah, "Kebijakan hukum pengendalian laju deforestasi terhadap lingkungan di Indonesia," IURIS STUDIA: Jurnal Kajian Hukum, 2025.

[3] R. Rianto, E. R. Rianto, E. S. Humanika, and I. H. T. Untoro, "Peningkatan sensitivitas Support Vector Machine pada klasifikasi kalimat baku dan tidak baku bahasa Indonesia," Jurnal Teknologi Informasi dan Ilmu Komputer, 2026.

[4] D. A. E. Aritonang, P. Maharani, N. Ramadhani, K. D. Tania, and A. Meiriza, "Klasifikasi berita hoaks dan fakta berdasarkan representasi TF-IDF menggunakan Support Vector Machine," JATI (Jurnal Mahasiswa Teknik Informatika), 2026.

[5] M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, "Semantic similarity and machine learning with ontologies," Briefings in Bioinformatics, vol. 22, no. 4, 2021.

[6] M. D. Pratama, R. Sarno, and R. Abdullah, "Sentiment analysis user regarding hotel reviews by aspect based using latent Dirichlet allocation, semantic similarity, and Support Vector Machine method," International Journal of Intelligent Engineering and Systems, vol. 15, no. 4, pp. 415–426, 2022.

[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in Proc. 6th Int. Semantic Web Conf. (ISWC/ASWC 2007), Busan, South Korea, 2007, pp. 722–735.

[8] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia Spotlight: Shedding light on the web of documents," in Proc. 7th Int. Conf. Semantic Systems (I-SEMANTICS), Graz, Austria, 2011, pp. 1–8.

[9] J. Flisar and V. Podgorelec, "Improving short text classification using information from DBpedia ontology," Fundamenta Informaticae, vol. 172, no. 3, pp. 261–297, 2020.

[10] D. D. Putra, A. Arfan, and R. Manurung, "Building an Indonesian WordNet," in *Proceedings of the 2nd International MALINDO Workshop*, Jakarta, Indonesia, Jun. 2008, pp. 12–13

[11] D. M. Akhmad, Y. Herdiyeni, and E. Sudarnika, "Similarities of antimalarial resistance genes in

Plasmodium falciparum based on ontology," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 1, pp. 415–423, 2018.

[12] D. M. Akhmad and S. Andriani, "Pengukuran semantic similarity pada gen PFCRT dan DHFR berbasis ontology," *Jurnal Teknoinfo*, 2023.

[13] N. I. Raharko and Y. Yamasari, "Klasifikasi Emosi Ulasan Produk E-Commerce Menggunakan Support Vector Machine," Journal of Informatics and Computer Science (JINACS), vol. 6, no. 2, pp. 606–616, 2024.

[14] H. Ma'rifah, A. P. Wibawa, and M. Iqbal Akbar, "Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing," Sains, Aplikasi, Komputasi dan Teknologi Informasi (SAKTI), vol. 2, no. 2, pp. 70–78, Apr. 2020.

[15] I. A. Fadilla, "Classification of Hate Speech Against Cak Nun on Twitter Multinomial Naive Bayes," Journal of Computing and Data Science (JODA), vol. 1, no. 1, pp. 47–53, 2023.

[16] S. Sikelis, A. Antoniou, and G. Paliouras, "Ontology-based semantic feature extraction for text classification," Expert Systems with Applications, vol. 185, p. 115642, 2021.

[17] M. Bouchiha, A. Dahbi, and M. El Yadari, "Ontology-driven semantic representation for document classification: A survey," Knowledge-Based Systems, vol. 272, p. 110650, 2023.

[18] X. Chen, Y. Wang, and H. Zhang, "Ontology-based semantic feature extraction for text classification: A survey," IEEE Access, vol. 9, pp. 112345–112360, 2021.

[19] C. Brockmeier, M. Galkin, J. Lehmann, and A.-C. Ngonga Ngomo, "Knowledge Graph Embeddings for Entity Alignment: A Survey," Semantic Web, vol. 12, no. 4, pp. 1–26, 2021.

[20] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[21] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[22] H. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance," Information Sciences, vol. 505, pp. 32–64, 2020.

[23] M. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 20th Anniversary," Journal of Artificial Intelligence Research, vol. 61, 2021.

[24] Y. Douzas and F. Bacao, "Effective Data Generation for Imbalanced Learning Using SMOTE Variants," Expert Systems with Applications, vol. 193, 2022.

[25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python—User Guide," Version 1.6 Documentation, 2025.