Digital Zone
Jurnal Teknologi Informasi dan Komunikasi

Volume 16 Issue 2 Year 2025 | Page 72-83 | e-ISSN: 2477-3255 | ISSN: 2086-4884
Received: 13-06-2025 | Revised: 14-07-2025 | Accepted: 16-09-2025

Integrated Named Entity Recognition and Identical-Entity Detection for Extracting Unique Information Sources in News Articles

Adi Surya Suwardi Ansyah[1], Daniel Oranova Siahaan[2], Brian Rizqi Paradisiaca Darnoto[3]

[1,2]Departemen Teknik Informatika, Fakultas Teknik Elektro dan Informatika Cerdas – Institut Teknologi Sepuluh Nopember
[1,2]Jl. Raya ITS, Sukolilo Surabaya, Jawa Timur, 60111, Telp. 031-5994251-54
[3]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Jember
[1,2]Jl. Krajan Timur, Sumbersari, Kec. Sumbersari, Kabupaten Jember, Jawa Timur, Telp. 031-5994251-54
e-mail: [1]adisur.sa@gmail.com, [2] daniel@if.its.ac.id , [3]brianrizqi@unej.ac.id

*Corespondence: adisur.sa@gmail.com,

Abstract: *Native advertising is often difficult to detect because it resembles regular news articles. One indicator is the absence of diverse information sources or the reliance on a single perspective. Therefore, it is necessary to employ an extraction technique capable of consolidating various forms of identical entity mentions. This study integrates an NER model based on XLNet+BiLSTM+CRF with identical entity classification using Levenshtein distance features and static and contextual vector representations. The results show an F1-score of 93.71% at the entity level and 92.84% for identical entity identification, along with a list of unique citation sources. These findings demonstrate that this unique list can be an additional feature in detecting native advertising, which often relies on a single source. With an average unique entity coverage of 97.40%, the proposed architecture can extract unique entities within news articles.*

Keywords: *Named Entity Recognition, Identical-Entity Detection, Source Information*

1. Introduction

A news article generally refers to a publication that presents news or information regarding events, facts, or occurrences in our surroundings. With the advent of the internet, the distribution of news articles has expanded publishers' reach and facilitated readers' access to news content. A survey involving 98,000 respondents revealed that 88% of Indonesians prefer online news articles as their primary source of information [1]. This shift has significantly impacted the business model of news publishing, particularly since advertising remains one of the primary revenue pillars. In line with this, native advertising is designed to resemble non-commercial content in style and layout, making its commercial features less conspicuous and more acceptable to readers. This strategy not only aims to capture readers' attention but has also significantly contributed to the revenue streams of media organizations [2].

Native advertisements have become a preferred approach in advertising, as studies have shown that very few readers can recognize native ads embedded within news articles [3]. As a consequence, readers may misinterpret such articles as credible sources of information when, in fact, they are advertisements with unverified content [4]. Several factors can influence readers' trust in native advertising articles, including persuasive language, the perspective presented, and the extent to which information sources are utilized to support the content. Pasandaran [5] notes that using only one information source may indicate native ads, as it can reflect a singular viewpoint without contradiction or comparison. Li [6] further highlights that the absence of

information sources is another characteristic of native ads, which can lead to reader deception due to a lack of transparency. Therefore, it is necessary to develop natural language processing (NLP) methods to extract the number of information sources in news articles, which can serve as one of the criteria for identifying news articles that are, in fact, native advertisements. Darnoto et al. [7] have studied the detection of native ads in Indonesian news articles by utilizing the entire article text as the main feature in native ad classification. Furthermore, they have explored the classification of persuasive sentences as a characteristic for detecting native ads [8]. However, there remains potential for improvement, as current models have yet to fully exploit the distinctive characteristics of native ads in their classification processes.

The main challenge in identifying information sources lies in effectively locating and extracting them from news articles. Information sources typically include individuals, governments, organizations, and other entities. Previous studies addressing this challenge have utilized Named Entity Recognition (NER) approaches, such as Static Word Embeddings (SWE) combined with Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) for Indonesian datasets [9]. However, SWE cannot capture contextual meaning or represent unseen words. Koto et al. [10] showed that BERT outperforms SWE on an Indonesian news dataset, whereas Yan et al. [11] reported that XLNet + BiLSTM + CRF surpasses BERT, reaching 97.02 % accuracy on the English CoNLL-2003 benchmark.

However, prior research by Yan et al. [11] has not yet addressed the recognition of specific entities as information sources or identifying unique entities. Therefore, this study seeks to address these limitations by enhancing the XLNet+BiLSTM+CRF architecture to identify unique information sources effectively. This approach employs NER with the XLNet+BiLSTM+CRF architecture as the initial step for detecting information sources. After the entity identification process, the proposed architecture recognizes identical entities that may appear different due to variations in how they are cited or mentioned within news articles. This process is carried out by analyzing how information sources are referenced, enabling the model to determine when two or more seemingly distinct entities refer to the same source.

Identifying identical information sources presents significant challenges due to citation simplification, typographical errors, and semantic and contextual similarities. The developed architecture integrates all four aspects to address these issues as features, considering that initial citations are typically more complete than subsequent references. Furthermore, unintentional spelling errors are handled using string similarity techniques. Mawardi et al. [12] have demonstrated the effectiveness of Levenshtein Distance (LD) in identifying and correcting typographical errors by calculating the minimum number of character changes required. In addition to spelling errors, semantic similarity features help recognize entities with similar meanings. Word2Vec vector representations can be utilized to obtain semantic similarity features because they can represent words as vectors that capture semantic context. Babić et al. [13] compared techniques such as Word2Vec, TF-IDF, and fastText and demonstrated that Word2Vec outperforms the others in identifying semantic similarities between words. However, semantic similarity alone is not sufficient. Context plays a crucial role in recognizing identical entities, as models that rely solely on semantic similarity may lead to misunderstandings. For instance, without considering context, a model may misinterpret polysemous terms (words with multiple meanings) or fail to distinguish between identical names used in different contexts. XLNet, when trained on Indonesian language data, outperforms BERT and is also trained in Indonesian to understand sentence context [14].

This proposed study integrates a NER architecture with XLNet vector representations, BiLSTM layers, CRF and a dedicated architecture for identifying identical entities in the Indonesian language. The architecture leverages features such as citation simplification, spelling errors detected by the LD, semantic similarity through Word2Vec, and contextual information using XLNet. This research aims to extract unique information sources that can serve as additional features for identifying native advertisements, as previously explored in [7], [8].

This solution responds to a challenge currently faced by the media sector. Accurate identification of information sources enables news organizations to better detect native

advertisements disguised as editorial content. This helps protect readers from being misled by hidden ads, supports media transparency, and strengthens audience trust in news reporting.

2. Research Method

This research method is designed to extract unique information sources from news articles through four main stages: data pre-processing (including tokenization and formatting), entity extraction using a NER model based on XLNet-BiLSTM-CRF, identification of entities that refer to the same information source through feature extraction and integration such as word truncation, string similarity, semantic similarity, and contextual similarity and model performance evaluation using a CNN architecture, as illustrated in Figure 1.

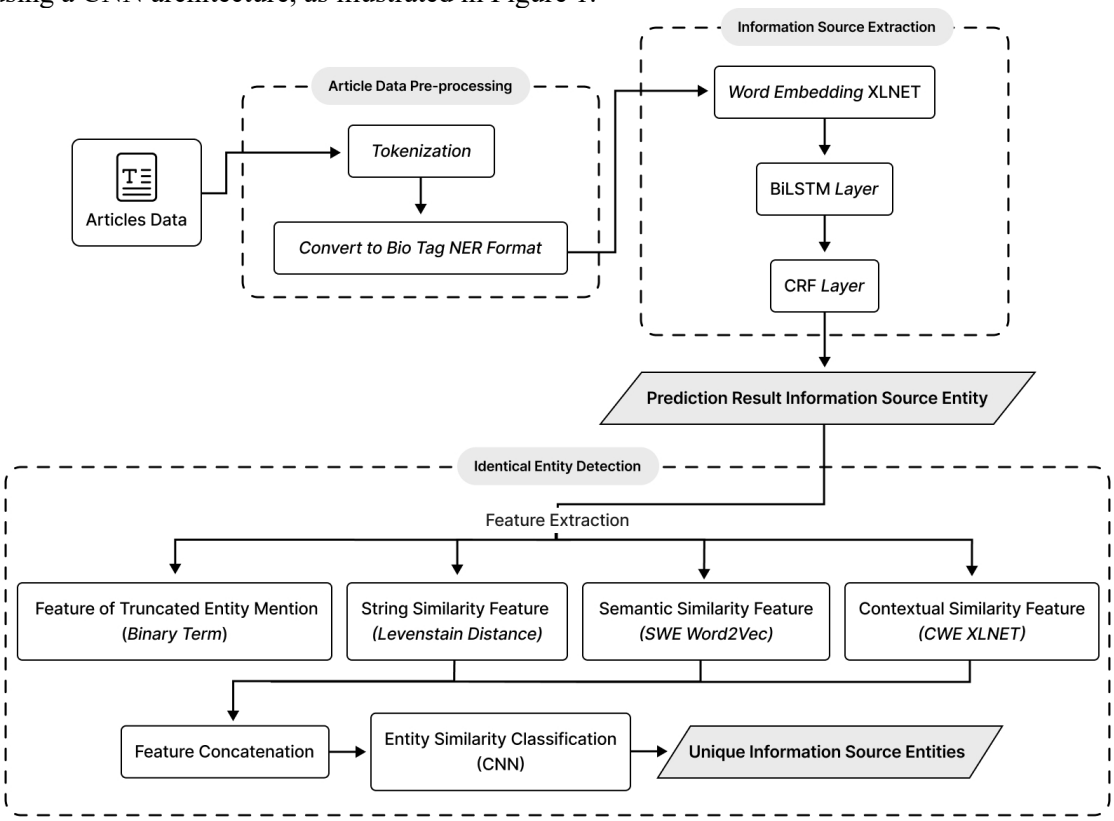

Figure 1. System Workflow for Unique Information Source Entity Extraction

As illustrated in Figure 1, after entities have been extracted by the XLNet–BiLSTM–CRF-based NER model, the process continues with merging similar or identical entities into unique information sources. This merging is performed through a systematic integration of various similarity measures, as previously stated, and explicitly depicted in the workflow diagram. The final stage involves assessing the overall capability of the integrated identification model through a CNN-based evaluation, focusing specifically on validating the accuracy of entity grouping results. This evaluation aims to confirm whether entities referring to the same real-world information source have been correctly consolidated, thus ensuring the robustness of the information source extraction pipeline.

2.1. Dataset and Pre-processing

The dataset was obtained from the results of web crawling and data processing based on the research by Darnoto et al. [15], initially consisting of 12,000 articles. These articles were selected to ensure a representative sample of online news in Indonesia. However, not all 12,000 articles were used in this study. The exact number of articles utilized is provided in Table 1. The articles in the analysis had already undergone preprocessing steps, such as noise removal. The detailed data distribution for each dataset is presented in Table 1.

https://doi.org/10.31849/digitalzone.v16i2. 27687

Table 1. Dataset Distribution for Unique Information Source Entity Extraction

Dataset	Number of Articles	Number of Sentences	Number of Information Source Entities
Training	1.960	8.000	8.120
Testing	477	3.000	3.029

During the data preprocessing stage, tokenization was performed using the spaCy library to segment each text into relevant tokens, such as words and punctuation marks. spaCy was chosen due to its ability to produce consistent tokenization results and its adaptability to the specific characteristics of news text. The pre-processed dataset was converted into a list of sentences and labeled using the BIO tagging format, where "B-IS" indicates the beginning of an entity, "I-IS" denotes the inside of an entity, and "O" refers to text outside of entities. Indonesian language experts annotated, and the dataset was then adjusted to the CoNLL-style BIO tag format, which is commonly used in NER model development [16]. The label distribution in the training data consists of 8,119 B-IS tokens, 16,611 I-IS tokens, and 318,412 O tokens. The testing data shows 3,029 B-IS tokens, 5,163 I-IS tokens, and 91,536 O tokens.

Subsequently, information source entities labeled in the BIO format and their respective article IDs were extracted. Each entity from the same article was then paired and labeled as "Identical" or "Not Identical," as illustrated in Figure 2, to compare whether the entity pairs refer to the same information source or different ones.

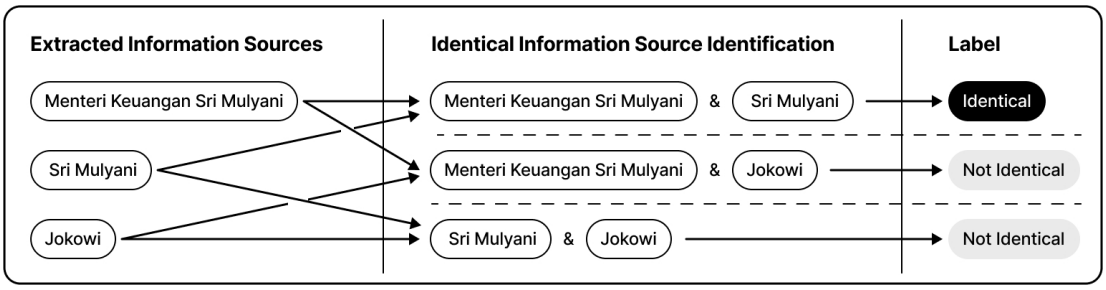

Figure 2. Example of Identical Information Source Entity Labeling

The training dataset for the information source entity identification model was obtained by removing duplicate entity pairs from the initial 22,866 extracted pairs. It resulted in 7,519 unique pairs (Identical: 4,110; Not Identical: 3,409) for training. For the testing dataset, there are 7,460 entity pairs (Identical: 4,320; Not Identical: 3,140) without further filtering, as this set is used to evaluate the comparison between the number of information sources identified by the model and those identified without the model.

2.2. Information Source Extraction

In the information source extraction stage, an NER architecture based on XLNet+BiLSTM+CRF is used, as described in the introduction. The XLNet model employed is IndoXLNet, developed by Pratama et al. [14], which was pre-trained on the Indo4B Indonesian corpus consisting of approximately 3.5 billion words compiled by Wilie et al. [17]. This model generates contextual word representations, which are further processed by BiLSTM and CRF layers to determine entity labels. All parameters are optimized using the AdamW algorithm to improve accuracy [18]. The model is trained with a learning rate of 5e-5, a batch size of 16, and two BiLSTM layers with a hidden size of 256.

2.3. Identical Entity Detection

The identical entity detection model was developed by extracting and integrating four main features: word truncation, string similarity, semantic similarity, and contextual similarity. In news articles, word truncation of information sources frequently occurs, where a partial name or pronoun follows an initially complete mention in subsequent references. Word truncation is an important feature for identifying identical entities; entities exhibiting truncation are labeled "0," while those without are labeled "1."

String similarity between entities is measured using the LeD, which calculates the minimum number of insertions, deletions, or substitutions required to transform one string into another. This method was chosen because of its ability to handle typos and spelling variations common in Indonesian. For example, "Sri Mulyani" versus "Sri Mulyono" results in an edit distance of one. Calculations are performed using the python-Levenshtein library, with the output being an integer value representing the degree of string difference; the smaller the value, the more similar the entities.

Semantic similarity features are extracted using a pre-trained Word2Vec model based on the Indo4B corpus for pre-training XLNet. The Word2Vec model has a vector dimension 300 and is implemented using the Gensim library. Each information source entity is pre-processed through tokenization and lowercasing and then directly mapped to a semantic vector from the Word2Vec embedding matrix.

This study utilizes the IndoXLNet PLM employed in the entity extraction stage to capture the contextual similarity between entities. The model is fine-tuned to function as a feature extractor. Each sentence containing an entity is mapped to a contextual representation vector using mean-pooling on the output of the final layer. This context-based representation enables the system to recognize entities that are spelled differently. Integrating contextual features from XLNet with other features makes the model more robust in the entity identity classification stage.

After obtaining the four main features (word truncation, string similarity, semantic similarity, and contextual similarity), all features are concatenated into a single vector for each entity pair and normalized before input to a CNN architecture. CNN is chosen because its Conv1D layer effectively extracts local patterns among features, enabling the model to detect micro-level differences that distinguish identical and non-identical entities. The model is trained using the AdamW optimizer and categorical cross-entropy loss on labeled data and evaluated on the test set to measure classification performance.

2.4. Performance Evaluation

a) Evaluation of Information Source Extraction Model

The first evaluation assesses the NER model's performance in extracting Information Source entities from news articles using test data manually labeled with the BIO scheme and not previously seen by the model during training. Evaluation is conducted at two levels: token level, which measures the consistency of B-, I-, or O tagging for each token, and entity level, which ensures that entity spans are correctly detected. Accuracy, Precision, Recall, and F1-Score are calculated at both evaluation levels, with basic definitions presented in Equation 1.

$$Precission = \frac{TP}{TP+FP}\ , Recall = \frac{TP}{TP+FN}\ , F1 = \frac{2\ x\ Precission\ x\ Recall}{Precission+Recall} \qquad (1)$$

TP (True Positive) refers to correctly predicted entities or tokens, FP (False Positive) denotes incorrect predictions, and FN (False Negative) indicates entities or tokens that were missed. Token-level results highlight tagging accuracy for individual tokens, whereas entity-level results reflect the model's ability to extract complete information from source entities.

b) Evaluation of Identical Information Source Identification Model

The second evaluation measures the accuracy of the CNN model in classifying whether two extracted entities refer to the same subject using pairs of entities annotated as identical or not identical. Evaluation is conducted using Accuracy, Precision, Recall, and F1-Score metrics, with results compared across different feature combinations to demonstrate the advantages of the feature-rich CNN-based approach.

c) Evaluation of Integrated Model

At this stage, the performance of the NER pipeline combined with identical entity identification is evaluated based on how completely the system extracts all Information Sources (IS) present in each news article. The process begins by utilizing test data manually annotated

https://doi.org/10.31849/digitalzone.v16i2. 27687

with the BIO format, which serves as the ground truth for determining the number of unique IS entities.

The proposed Information Source Extraction model is then applied to the test data to generate predictions. A predicted entity is considered valid only if its BIO labels and token spans (start and end indices, including the token sequence) exactly match the ground truth. In other words, the prediction must not miss or misalign a single token to be recognized as an accurate IS detection.

Once all valid entities are counted, the completeness of the system is evaluated using two measures. First, the coverage ratio is calculated by dividing the number of unique IS entities correctly detected by the model by the number of unique IS entities in the ground truth for each article. A value of 1 indicates full coverage, while values below 1 indicate that some entities remain undetected. Second, the absolute difference is calculated as the absolute value of the difference between the number of unique entities in the ground truth and the number of unique entities successfully predicted to determine how many entities were missed or redundantly predicted. Finally, these coverage and difference values are averaged across all articles, providing an overall picture of the system's ability to capture information sources regarding total entities and unique entities comprehensively. This evaluation provides an additional perspective beyond standard token or entity-level evaluation.

3. Results and Discussion

3.1. Evaluation of Information Source Extraction Model

This section presents the evaluation results of the NER models developed to extract Information Sources from Indonesian news articles. The evaluation used 15% of the training data as the test set. Token-level evaluation results are shown in Table 2, while entity-level results are presented in Table 3. Across both evaluation levels, the proposed model architecture consistently achieved the highest performance compared to other baseline models.

Table 2. NER Evaluation at the BIO Tagging Level

Model No.	Model	Metric	B- IS	I- IS	O
1	Word2Vec+BiLSTM+CNN [9]	Precision	95.86	93.70	98.91
		Recall	93.70	84.99	99.54
		F1-score	**98.90**	88.54	99.22
		Accuracy		98.50	
2	fastText+BiLSTM+CNN [9]	Precision	96.57	92.22	99.10
		Recall	91.09	88.15	99.53
		F1-score	93.75	90.14	99.32
		Accuracy		98.69	
3	Bert Linear [10]	Precision	96.42	95.08	99.37
		Recall	93.46	91.42	99.69
		F1-score	94.92	93.22	99.53
		Accuracy		99.07	
4	**XLNet+BiLSTM+CRF (proposed)**	Precision	**96.72**	**95.18**	**99.51**
		Recall	**96.24**	**92.27**	**99.70**
		F1-score	96.48	**93.71**	**99.61**
		Accuracy		**99.21**	

Model No. 4 achieved the highest accuracy (99.21%) and the best F1-scores for Information Source classes, with B-IS = 96.48% and I-IS = 93.71%. This performance gain is most pronounced on the I-IS label, where the F1-score difference between Model No. 4 and the second-best model, Model No. 3, reaches +0.49 percentage points. These results demonstrate that using a CRF layer combined with XLNet contextual embeddings is highly effective for accurately and consistently recognizing multi-token entity spans.

Both transformer-based models (Model No. 3 and Model No. 4) consistently outperform the CNN pipelines using word embeddings (Model No. 1 and Model No. 2) on almost all metrics, except for the F1-score of the B-IS label on Model No. 1, which is slightly higher than one of the transformer models. This finding highlights the importance of deep contextual representations that can capture relationships among tokens within a single entity, unlike static word vectors, which tend to lose context when entities consist of more than one word.

In addition, all models achieved F1-scores above 99% for the non-entity (O) class, reflecting the label imbalance in the dataset, where non-entity tokens dominate news article content. Consequently, Model No. 4's increase in overall accuracy (+0.14 percentage points compared to Model No. 3) is primarily driven by its superior ability to recognize and label the relatively rare Information Source tokens (B-IS and I-IS).

Table 3. NER Evaluation at the Entity Level

Model No.	Precision	Recall	F1-score
1	87.02	82.14	84.51
2	89.36	84.29	86.75
3	93.19	90.33	91.74
4	**92.81**	**92.87**	**92.84**

Transformer-based models (Model No. 3 and Model No. 4) demonstrated significant advantages over the two CNN pipelines employing word embeddings (Model No. 1 and Model No. 2). The average F1-score of both transformer models exceeded 92.29%. In contrast, the best CNN-based model only achieved 86.75%. Therefore, it underscores the importance of deep contextual representations in accurately capturing the full-span boundaries of entities.

Model No. 4 exhibited an excellent balance between precision and recall (92.81% vs 92.87%), resulting in the highest F1 score of 92.84%. This balance indicates that the CRF layer effectively reduces false positive errors (yielding high precision) while minimizing missed entities (yielding high recall). The F1-score difference between Model No. 3 and Model No. 4 reached +1.10 percentage points, primarily attributable to the reduced boundary errors made possible by the CRF's ability to leverage global label dependencies, thus producing more consistent span predictions—particularly for multi-token entities.

Compared to token-level evaluation, all models experienced a decrease in entity-level scores. This decline is expected, as an error on a single token at the entity-level evaluation results in the entire entity is considered incorrect. The proposed model exhibited the most minor reduction in score, highlighting the robustness of the global decoding mechanism inherent in its architecture.

The entity-level evaluation further reinforces the findings from the token-level results, where the combination of XLNet contextual embeddings, sequential modeling via BiLSTM, and CRF-based decoding delivers the most reliable Information Source detection overall. This success is especially critical in practical applications, which require entities to be recognized entirely rather than partially tagged. To further illustrate these differences, Table 4 presents selected prediction samples in which baseline models failed to identify entity spans correctly. In contrast, the proposed model demonstrated superior accuracy and robustness in handling multi-token entities and ambiguous cases.

Table 4. Sample Prediction Errors of Information Source Extraction

Sentence	**Model No.**	**Prediction**	**Error**
Wakil Perdana Menteri China Sun Chunlan mengungkapkan varian Omicron melemah dan tingkat vaksinasi meningkat.	1	Chunlan (O)	Span Error
	4	Chunlan (I- IS)	-
Bukti menunjukkan serangan udara kembar Rusia sengaja menargetkan teater yang	2	Amnesty (O), International (O)	Missed Entity

Sentence	Model No.	Prediction	Error
digunakan sebagai tempat perlindungan di Mariupol, kata Amnesty International	4	Amnesty (B-IS), International (I-IS)	-
Token Dalam konferensi pers bersama Presiden Jokowi di Kyiv, Zelensky mengatakan mereka telah membahas krisis pangan global	3	Kyiv(I-IS)	Tagging Error
	4	Kyiv(O)	

In the qualitative evaluation, each architecture demonstrates distinct error patterns. For the sentence, "Wakil Perdana Menteri China Sun Chunlan …", Model No. 1 (Word2Vec + BiLSTM + CNN) labeled "Sun" as B-SI but assigned O to "Chunlan," resulting in a truncated entity span (span error). In contrast, Model No. 4 (XLNet + BiLSTM + CRF) consistently labeled "Sun" as B-SI and "Chunlan" as I-SI, thus correctly capturing the full entity span. In the case of "… kata Amnesty International", the limitation of Model No. 2 (fastText + BiLSTM + CNN) is evident, as both "Amnesty" and "International" were not recognized at all (missed entity). In contrast, Model No. 4 successfully identified both tokens as B-SI/I-SI. The third example, "… konferensi pers … di Kyiv …", highlights a labeling inconsistency in Model No. 3 (BERT Linear), where "Kyiv" was tagged as I-SI without a preceding B-SI (tagging error). In contrast, Model No. 4 appropriately labeled "Kyiv" as O, as it is not an information source entity.

These three cases illustrate a consistent trend: CNN pipelines with static word embeddings often fail to capture multi-token entities or out-of-vocabulary tokens. At the same time, BERT without CRF is prone to label sequence inconsistencies. The integration of XLNet (contextual representations), BiLSTM (bidirectional context), and CRF (global decoding) in the proposed model effectively maintains BIO scheme validity, handles multi-token entity spans, and minimizes missed entities. Thus, these qualitative findings align with the quantitative results presented in Tables 3 and 4, further confirming the reliability of the proposed model for information source extraction in Indonesian news articles.

3.2. Evaluation of Identical Information Source Identification Model

This section evaluates the capabilities of various classification architectures in determining whether two entities are identical. The validation data used consists of 15% of the training set. Of all entity pairs, 3,812 pairs contained the truncated mention feature. The evaluation results indicate that the proposed architecture performs better than the benchmark models. Table 5 presents the performance of four variants of entity identity classification models.

Table 5. Performance of Information Source Identity Models

Model No.	Model	Precision	Recall	F1 score	Accuracy
1	Levenstain Distance [12]	71	67	67	70
2	SWE -Word2vec [13]	88	89	88	88
3	CWE - XLNet [14]	89	89	89	89
4	Hybrid (proposed)	**94**	**95**	**94**	**95**

The character-based Levenshtein Distance model (Model No. 1) is the lowest benchmark, with an F1 score of 67% and an accuracy of 70%. Performance increases significantly when semantic representation is introduced: the static Word2Vec model (Model No. 2) achieves an F1-score and accuracy of 88%. Replacing static embeddings with contextual representations using XLNet (Model No. 3) offers only a marginal improvement, yielding an F1 score and an accuracy of 89%. This result suggests that context alone cannot resolve all ambiguities between entity pairs.

Model No. 4, which combines truncated mention features, Levenshtein distance, Word2Vec, and XLNet within a CNN architecture, achieves the highest performance, with a precision of 94%, recall of 95%, F1-score of 94%, and accuracy of 95%. This improvement is

particularly significant for the 3,812 pairs containing the truncated mention feature; combining word truncation indicators with string, semantic, and contextual similarity makes the system more sensitive to name spelling and pronoun usage variations. Thus, the quantitative results demonstrate that combining rich features is more effective than relying on a single type of representation for identifying identical information sources. Table 5 also presents three sample entity pairs illustrating the error patterns of each model as well as the success of the hybrid model (Model No. 4). Furthermore, Table 6 provides additional sample prediction results from the proposed model, highlighting its ability to identify complex cases that the baseline models misclassified correctly.

Table 6. Sample Prediction Errors of Identical Information Source Identification Model

Entity Pairs	Model No.	Prediction
Camat Setiabudi , Iswahyudi vs Komisari Besar Azis Adriansyah	1	Identical
	4	Not Identical
mantan Menteri Koordinator Pembangunan Manusia dan Kebudayaan (Menko PMK) vs Puan	2	Not Identical
	4	Identical
SYL vs Mentan	3	Not Identical
	4	Identical

For the pair "Camat Setiabudi, Iswahyudi" vs "Komisaris Besar Azis Adriansyah," Model No. 1, which relies solely on character distance, incorrectly predicts them as identical entities due to the similarity in the position-name pattern, despite differing contexts. Similarly, in the case of "mantan Menteri Koordinator Pembangunan Manusia dan Kebudayaan (Menko PMK)" vs "Puan," Model No. 2 (Word2Vec + CNN) fails to map the lengthy position phrase to the short nickname, resulting in an incorrect prediction. In contrast, the hybrid Model No. 4 leverages word truncation features and contextual embeddings to recognize that "Puan" is a short form referring to the same individual, as well as to distinguish between the titles "Camat" and "Komisaris Besar."

Meanwhile, the example "SYL" vs "Mentan" highlights the limitation of Model No. 3 (XLNet alone), which still struggles to infer that two different abbreviations may refer to the same person. The integration of string similarity, word truncation, semantic, and contextual embeddings in Model No. 4 effectively compensates for the shortcomings of each single-approach model, allowing the hybrid model to consistently yield correct predictions in all three cases handling abbreviation variations, truncated names, and complex position phrases with higher accuracy.

3.3 Evaluation of Integrated

This stage evaluates the effectiveness of integrating the NER model with the identical entity identification model for extracting and consolidating unique information sources from news articles. Based on the results of the proposed integrated model, performance was assessed using the coverage metric, where the average entity coverage reached 90.21% and the average unique entity coverage was 97.40%. These high coverage values indicate that the integrated model can accurately recognize and merge information sources regarding total entity coverage and the uniqueness of entities identified within the news articles.

However, inspection of the prediction outputs reveals inconsistencies in the identical-label assignments for specific entity pairs. For example, in the evaluation results, the pairs “Nadiem” vs. “Mas Menteri” and “Nadiem” vs. “Nadiem Anwar Makarim” are both predicted as Identical, whereas “Mas Menteri” vs. “Nadiem Anwar Makarim” is predicted as Not Identical. Such inconsistencies can introduce consolidation errors, whereby two entities that should be merged are instead treated as distinct.

3.1 Discussion

The proposed two-model integration meets the study’s objective of extracting and identifying identical information sources. Mechanistically, the superiority of Model 4 (XLNet+BiLSTM+CRF) arises from the synergy between rich contextual representations

https://doi.org/10.31849/digitalzone.v16i2. 27687

(XLNet), sequential dependency modeling (BiLSTM), and global decoding (CRF), which enforces consistent BIO transitions and reduces span errors on multi-token entities. In comparison to CNN-based pipelines with static embeddings, transformer-based contextual representations more effectively capture inter-token relations and variability in name mentions, which explains the observed gains at the entity level.

For identical-entity resolution, the hybrid classifier outperforms single-feature models by combining edit string score (Levenshtein), static semantics (Word2Vec), contextual embeddings (XLNet), and a truncated-mention feature. This fusion increases sensitivity to aliases, acronyms, and the mapping between nicknames and full names, particularly across thousands of test pairs involving truncated mentions. Integrating the two modules yields high entity coverage (90.21%) and very high unique-entity coverage (97.40%), indicating that nearly all unique sources are captured and duplicate mentions are consolidated appropriately.

Nonetheless, several limitations warrant attention. First, error propagation can occur across stages: inconsistencies in identical-label assignments for a subset of pairs suggest the need for post-hoc verification based on cluster- or graph-level consistency rules, so that a single misclassification does not cascade into additional merges. Second, the corpus limited to Indonesian online news and dominated by the O label may constrain generalizability to other domains, genres, or languages.

Looking forward, promising directions include cross-domain (e.g., social media, transcripts, video captions) and cross-lingual evaluation to test robustness, as well as data diversification/augmentation to mitigate label imbalance. Methodologically, our findings validate a stable two-stage design for media analytics: a strong sequence tagger, when coupled with a hybrid-feature identity resolver, produces end-to-end outputs that are accurate, de-duplicated, and reliable. Beyond this, we envision a one-shot (end-to-end) architecture that jointly performs NER and identity resolution, e.g., multi-task models that output entity spans and cluster assignments simultaneously, reducing error propagation while leveraging global evidence for both extraction and consolidation.

4. Conclusions

This study successfully developed and evaluated an integrated system for extracting and identifying unique information sources in Indonesian news articles. The XLNet-BiLSTM-CRF-based NER model, which served as the foundation of this research, consistently achieved superior performance compared to other architectures, both at the token-level and entity-level evaluations. Additionally, the identical entity identification model based on CNN, which utilizes a combination of word truncation, string similarity, semantic, and contextual features, also achieved the highest accuracy and F1 score compared to other baseline models.

Overall, the proposed model addresses the limitations of conventional NER systems, which typically detect entities without consolidating unique entities within an article. With an average unique entity coverage of 97.40%, the system effectively recognizes and consolidates various references to information sources, including cases involving word truncation and abbreviations.

From a practical and industrial perspective, the results of this research directly support news organizations and media analytics companies by providing reliable tools for accurately identifying and consolidating information sources. Specifically, the developed system can significantly improve the detection of native advertisements disguised as editorial content, thus safeguarding readers from misleading information. By facilitating greater transparency and clarity in distinguishing genuine editorial content from hidden ads, media companies can uphold journalistic standards, maintain audience trust, and comply with evolving regulatory frameworks. Furthermore, the developed methodology is scalable and adaptable, offering significant potential for integration into existing media monitoring and analytic platforms, enhancing their capability to manage large-scale content effectively.

These results respond effectively to the research objectives and hypotheses, achieving improved accuracy and completeness in information source extraction and consolidation within news articles. The success of this approach also opens up further research opportunities, such as

applying similar methods to other entity types, developing broader media analytics systems, and optimizing the developed pipeline for efficient implementation in large-scale industrial applications.

References

[1] Nic Newman with Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen, "Reuters Institute Digital News Report 2022," Oxford, 2022. Accessed: Mar. 20, 2023. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

[2] W. Yao, J. W. B. Mohd Zawawi, A. M. @ Z. Ahmad, and T. J. Sern, "Recognizing Native Advertising and Its Challenge to Traditional Advertising," *International Journal of Academic Research in Business and Social Sciences*, vol. 11, no. 19, Dec. 2021. https://doi.org/10.6007/IJARBSS/v11-i19/11727

[3] M. A. Amazeen and B. W. Wojdynski, "The effects of disclosure format on native advertising recognition and audience perceptions of legacy and online news publishers," *Journalism*, vol. 21, no. 12, pp. 1965–1984, Dec. 2020. https://doi.org/10.1177/1464884918754829

[4] A. Kutlu, "Native Advertising the Effect of Native Advertising on Ad Credibility," *International Journal of Economics, Business and Management Research*, vol. 06, no. 11, pp. 152–165, 2022. https://doi.org/10.51505/IJEBMR.2022.61111

[5] C. C. Pasandaran, "Political Advertising Camouflage As News," *Jurnal Komunikasi Ikatan Sarjana Komunikasi Indonesia*, vol. 3, no. 2, Dec. 2018. https://doi.org/10.25008/jkiski.v3i2.239

[6] Y. Li, "The Role Performance of Native Advertising in Legacy and Digital-Only News Media," *Digital Journalism*, vol. 7, no. 5, pp. 592–613, May 2019. https://doi.org/10.1080/21670811.2019.1571931

[7] B. R. P. Darnoto, D. Siahaan, and D. Purwitasari, "Deep Learning for Native Advertisement Detection in Electronic News: A Comparative Study," in *2022 11th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, IEEE, Aug. 2022, pp. 304–309. https://doi.org/10.1109/EECCIS54468.2022.9902953

[8] B. R. P. Darnoto, D. Siahaan, and D. Purwitasari, "Automated Detection of Persuasive Content in Electronic News," *Informatics*, vol. 10, no. 4, p. 86, Nov. 2023, https://doi.org/10.3390/informatics10040086

[9] B. S. Jati, S. Widyawan, and S. T. Muhammad Nur Rizal, "Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, IEEE, Nov. 2020, pp. 180–184. https://doi.org/10.1109/ICOIACT50329.2020.9332027

[10] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - The 28th International Conference on Computational Linguistics*, Nov. 2020. https://doi.org/10.48550/arXiv.2011.00677

[11] R. Yan, X. Jiang, and D. Dang, "Named Entity Recognition by Using XLNet-BiLSTM-CRF," *Neural Process Lett*, vol. 53, no. 5, pp. 3339–3356, Oct. 2021. https://doi.org/10.1007/s11063-021-10547-1

[12] V. Christanti Mawardi, F. Augusfian, J. Pragantha, and S. Bressan, "Spelling Correction Application with Damerau-Levenshtein Distance to Help Teachers Examine Typographical Error in Exam Test Scripts," *E3S Web of Conferences*, vol. 188, p. 00027, Sep. 2020. https://doi.org/10.1051/e3sconf/202018800027

[13] K. Babić, F. Guerra, S. Martinčić-Ipšić, and A. Meštrović, "A Comparison of Approaches for Measuring the Semantic Similarity of Short Texts Based on Word Embeddings," *Journal of information and organizational sciences*, vol. 44, no. 2, pp. 231–246, Dec. 2020. https://doi.org/10.31341/jios.44.2.2

[14] T. Pratama and S. Rjito, "IndoXLNet: Pre-Trained Language Model for Bahasa Indonesia," *International Journal of Engineering Trends and Technology*, vol. 70, no. 5, pp. 367–381, Jun. 2021. https://doi.org/10.14445/22315381/IJETT-V70I5P240

[15] B. R. P. Darnoto, D. Siahaan, and D. Purwitasari, "Electronic News Dataset for Native Advertisement Detection," *Sci Data*, vol. 12, no. 1, p. 1045, Jun. 2025. https://doi.org/10.1038/s41597-024-04341-6

[16] C. Palen-Michel, M. Pickering, M. Kruse, J. Sälevä, and C. Lignos, "OpenNER 1.0: Standardized Open-Access Named Entity Recognition Datasets in 50+ Languages," Dec. 2024. https://doi.org/10.48550/arXiv.2412.09587

[17] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. https://doi.org/10.18653/v1/2020.aacl-main.85

[18] P. Chen, M. Zhang, X. Yu, and S. Li, "Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 315, Dec. 2022. https://doi.org/10.1186/s12911-022-02059-2