Judul	:	5W1H Information Extraction Using Indobert Name Entity Recognition for Sports News
Penulis	:	1. George Wielianto 2. Novario Jaya Perdana 3. Jeanny Pragantha
Penerbit	:	IEEE
Konferensi	:	2025 International Conference on Applied Artificial Intelligence, Data Engineering and Sciences (ICAIDES)
Tanggal Konferensi	:	11-12 December 2025
Tanggal Terbit	:	26 Februari 2026
Tautan	:	https://ieeexplore.ieee.org/abstract/document/11404027

5W1H Information Extraction Using Indobert Name Entity Recognition for Sports News

Publisher: IEEE

George Wielianto ; Novario Jaya Perdana ; Jeanny Pragantha All Authors

Abstract
Document Sections
I. Introduction
II. Literature Review
III. Method
IV. Result
V. Conclusion
Authors
Figures
References
Keywords
Metrics

Abstract:
Football, basketball, and badminton sports news in today's online media often contain very rich information in the news, such as the names of athletes, teams, championships, tournaments, to important information such as scores, stadiums are often hidden in unstructured text. This study aims to help extract information using NER (Named Entity Recognition) with the IndoBERT model to obtain important information in sports news in the 5W1H format (What, When, Who, Where, Why, and How) using a rule-based method. The fine-tuned IndoBERT model will then be tested with Indonesian sports news test data and obtained the best F1-Score of 0.8191 in the 10th epoch, indicating that the performance and performance of this IndoBERT model works well in recognizing NER entities.

Published in: 2025 International Conference on Applied Artificial Intelligence, Data Engineering and Sciences (ICAIDES)

Date of Conference: 11-12 December 2025
DOI: 10.1109/ICAIDES67265.2025.11404027
Date Added to IEEE *Xplore*: 26 February 2026
Publisher: IEEE
ISBN Information:
Conference Location: Jakarta, Indonesia

5W1H Information Extraction Using Indobert Name Entity Recognition for Sports News

1st George Wielianto
Faculty of Information Technology
Universitas Tarumanagara
West Jakarta, Indonesia
george.535220090@stu.untar.ac.id

2nd Novario Jaya Perdana
Faculty of Information Technology
Universitas Tarumanagara
West Jakarta, Indonesia
novariojp@fti.untar.ac.id

3rd Jeanny Pragantha
Faculty of Information Technology
Universitas Tarumanagara
West Jakarta, Indonesia
jeannyp@fti.untar.ac.id

***Abstract*—Football, basketball, and badminton sports news in today's online media often contain very rich information in the news, such as the names of athletes, teams, championships, tournaments, to important information such as scores, stadiums are often hidden in unstructured text. This study aims to help extract information using NER (Named Entity Recognition) with the IndoBERT model to obtain important information in sports news in the 5W1H format (What, When, Who, Where, Why, and How) using a rule-based method. The fine-tuned IndoBERT model will then be tested with Indonesian sports news test data and obtained the best F1-Score of 0.8191 in the 10th epoch, indicating that the performance and performance of this IndoBERT model works well in recognizing NER entities.**

Keywords—Named Entity Recognition, 5W1H, IndoBERT, Sports, rule-based

I. INTRODUCTION

The internet is now an essential component of contemporary civilisation and a vital infrastructure for the distribution of information. In Indonesia, there are 221 million internet users, or almost 79.5% of the country's entire population, according to a 2024 study by the Association of Internet Service Providers (APJII) [1]. The emergence of internet media as a main source of information, including political, sports, and economic news, has increased due to this quick expansion [2]. Because digital platforms are convenient and quick, a large percentage of Indonesians spend a lot of time using news portals and streaming services, according to studies on the country's media consumption [3].

Despite the increasing accessibility of information, reading interest in Indonesia remains relatively low. According to UNESCO, only 0.001% of Indonesians have a strong reading habit [4], and research from the University of Canberra's News and Media Research Centre (N&MRC) shows that Indonesian audiences typically favour news summaries produced by artificial intelligence (AI) over reading complete articles [5]. This requirement emphasises the necessity of an information extraction system that can succinctly convey important information. Important information like athlete names, teams, venues, events, and match scores are frequently incorporated into unstructured language when discussing sports news. In order to overcome this difficulty, this study uses a Named Entity Recognition (NER) methodology that integrates rule-based techniques with the IndoBERT model, a BERT-based pre-trained language model trained on extensive Indonesian corpora, to generate structured and informative outputs [6]–[10].

II. LITERATURE REVIEW

Previous studies on Named Entity Recognition (NER) across multiple domains and languages have consistently shown that Transformer-based models, particularly BERT and its variants, outperform traditional approaches such as Conditional Random Fields (CRF). While Darji and Mitrovic (2023) reported notable performance gains using domain-specific German BERT for legal documents, highlighting the significance of language- and domain-adapted models, Tunsakul (2020) showed that BERT achieved superior contextual understanding in a tourism corpus [9], [10]. Although domain-specific elements are still under-represented, Koto et al. (2020) developed IndoBERT and IndoLEM as comprehensive benchmarks for a variety of NLP tasks in the Indonesian setting [11]. Wilie et al. (2020) further demonstrated that, despite the scarcity of specialised datasets, IndoBERT consistently performs better in Indonesian NER tasks than multilingual BERT (mBERT). NER is still difficult in linguistically diverse situations, according to studies in particular fields like esports news, where a BERT–CRF strategy only achieves mediocre results [12]. Overall, Transformer-based models have become the standard approach for entity extraction tasks [6].

III. METHOD

This section will focus on the research methods used, including BERT as the general approach and model used in this design. The aim is to develop and test an information extraction system for Indonesian sports news using the Named Entity Recognition (NER) method. This design involves several stages, starting with system design, dataset preparation, and system development, followed by implementation in a website application to display the extraction results in the 5W1H format (Who, What, Where, When, Why, and How).

System Design

The four primary components of the proposed sports news information extraction system flowchart, rule-based algorithm, and dataset details are intended to be summarised in the system design. Sports news items are first gathered from DetikSport, KompasSport, and BolaSport. LabelStudio is then used to manually label the articles, which are then separated into training, validation, and test sets. The AdamW optimiser and BCEWithLogitsLoss are used to refine the IndoBERT model for Named Entity Recognition using the annotated data. Performance is then assessed on the test set. As shown in Table 1, the trained model outputs are further processed using a rule-based approach that maps the retrieved entities into a structured 5W1H format and combines sequential tokens with identical labels.

TABLE I. ENTITY MAPPING INTO 5W1H

No	NER Label	5W1H Category
1.	ATHLETE	WHO

2.	TEAM	WHO
3.	ORGANIZATION	WHO
4.	CHAMPION	WHERE
5.	STADIUM	WHERE
6.	STATISTIC	WHAT
7.	AWARD	WHAT
8.	POSITION	WHO
9.	NATIONALITY	WHO
10.	AGE	WHO
11.	DATE	WHEN
12.	SCORE	WHAT
13.	ACTiON	HOW
14.	REASON FOR THE EVENT	WHY

The system is divided into four main parts: (1) an output section that displays extracted data in the 5W1H format; (2) a news input section with tab-based options for text or URL input; (3) example URLs from supported online media sources (DetikSport, KompasSport, and BolaSport); and (4) a header that displays the application name and supported media and categories. News text or URL input is the first step in the system's workflow; for URL inputs, web scraping and preparation are handled automatically. A refined IndoBERT model for Named Entity Recognition (NER) is then used to process the cleaned text. The extracted entities are then mapped into the 5W1H structure as the final output. Fig. 1 shows the total system workflow.

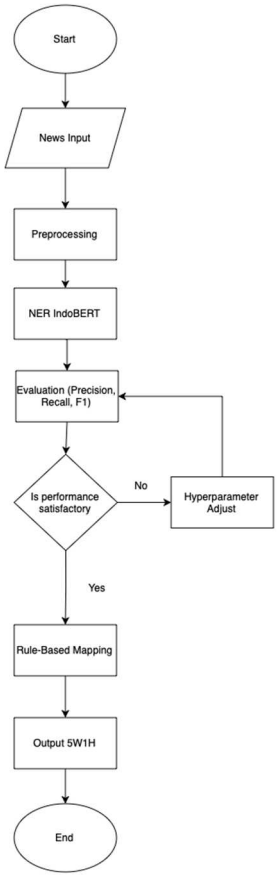

Fig. 1. System Flowchart

Dataset Preparation

Prior to system development, an Indonesian sports news dataset covering football, basketball, and badminton was prepared for training, validation, and testing the IndoBERT model. The data were collected from online media using the Scrapy framework with BeautifulSoup over a one-year period (January 1, 2024, to August 30, 2025), where separate scraping scripts were implemented to handle different website structures. The scraping process involved retrieving article URLs, extracting relevant metadata and news content, removing irrelevant elements such as advertisements, and storing the cleaned text in JSON format. The collected data were then imported into LabelStudio for manual NER annotation, exported in a span-based CSV format to preserve entity boundaries and handle overlapping entities, and validated by journalism experts to ensure label consistency. Finally, the annotated dataset was split into training (80%), validation (10%), and testing (10%) sets, with dataset statistics summarized in Table 2.

TABLE II. DATASET NUMBER DETAILS

Media	Sports	Amount of Data
DetikSport	Basketball	233
	Football	234
	Badminton	233
KompasSport.com	Basketball	233
	Football	234
	Badminton	233
BolaSport.com	Basketball	233
	Football	234
	Badminton	233
Total		2100

The system will then be built using the Python programming language and the Flask framework, which will serve as the backend. The trained and tested IndoBERT model will then be integrated with a website application. The system also integrates automatic scraping using the BeautifulSoup library to extract news content from user-entered URLs.

Model Training

Bidirectional Encoder from Transformers (BERT) is a pre-trained language model introduced by Devlin et al. The BERT model utilizes the basic Transformer architecture, particularly its encoder, which generates context-rich text representations bidirectionally. Unlike previous models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs), Transformers do not rely on sequential data processing [11]. This architecture utilizes a self-attention mechanism to understand the relationships between words in a sentence in parallel and is more efficient at understanding long contexts [12].

The Transformer architecture consists of two main components: an encoder and a decoder. However, in the BERT LLM model, only the encoder component is used because the model's primary focus is text understanding. Each encoder layer has core components such as Token Embedding and Positional Encoding, Multi-Head Self-Attention, and a Feedforward Neural Network [13].

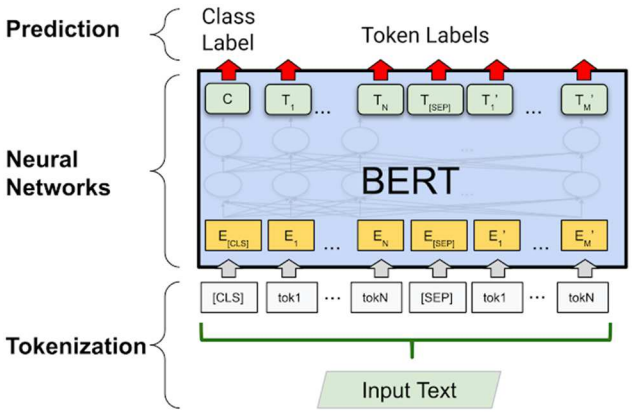

Fig. 2. Transformer Architecture

The model used in the design and development of this system is IndoBERT base-p1 from the IndoBERT Benchmark, a BERT model specifically trained for Indonesian. IndoBERT-base is the base model of IndoBERT, trained on a corpus of 5.5 billion words, encompassing several forms of Indonesian text. This model can be used for NLP tasks. The model itself consists of 12 transformer layers with 12 heads per layer and 110 million parameters [14].

The model is fine-tuned using a sports news dataset labeled with entities in the sports domain. This training will be conducted using PyTorch and the Transformers Library with the following hyperparameters:

TABLE III. HYPERPARAMETER USED

Hyperparameter	Value
Base Model	indobenchmark/indobert-base-p1
Batch size	2
Learning rate	2e-5
Weight decay	0.01
Epoch	15
Max_length	512
Optimizer	AdamW

During this training process, the model will generate checkpoints at each epoch to select the best model based on the highest F1-Score value in the model training process.

Rule-Based 5W1H

In this design, after the IndoBERT model generates NER entities, the results of these entities will be mapped into the 5W1H format using a rule-based method. This mapping is done by combining 5W1H with entity types, the entity mapping can be seen in Table 1. This rule-based approach is useful for producing extraction results that are neater, more structured, and easier to read.

System Testing

This stage will be carried out to verify and ensure that the extraction system is functioning properly and according to its intended function. Testing will cover three aspects:

1. IndoBERT performance testing using Precision, Recall, and F1-Score metrics.
2. Human evaluation testing, where the extraction system will be tested by the community.

Model Evaluation

In this information extraction project for Indonesian sports news, there are 14 entities to be extracted. Therefore, evaluation metrics are used to assess the extraction results from the trained model. The metrics used are Precision, Recall, and F1-Score to measure the performance of the model.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

Precision is used to measure the accuracy and precision of entity predictions performed by the IndoBERT model. Precision measures the number of correctly predicted entities compared to those predicted as positive. For example, if the model predicts 100 "ATHLETE" entities and 80 of them are correctly labeled, the precision value is 0.8 [15].

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

Recall measures how well a model successfully finds all the entities it should. This metric calculates the number of correctly predicted entities relative to the total number of entities actually present in the data. For example, if a news story contains 100 "ATHLETE" entities and the model only successfully predicts 85, the recall value is 0.85 [15].

$$F-1\,Score = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

The F1-Score is the weighted average of precision and recall. This metric serves to provide a balance between the two metrics, precision, and recall [15].

IV. RESULT

How to Apply 5W1H

The implementation of the 5W1H framework in this study is conducted by mapping the extracted named entities into structured information components. After the IndoBERT-based Named Entity Recognition model identifies relevant entities from Indonesian sports news articles, each entity is assigned to a corresponding 5W1H category. Subsequently, a post-processing stage is applied to refine the extraction results by removing duplicated entities that may arise due to repeated mentions within the same article. This post-processing step ensures that each 5W1H component contains unique and representative information, thereby improving the clarity and usability of the structured output.

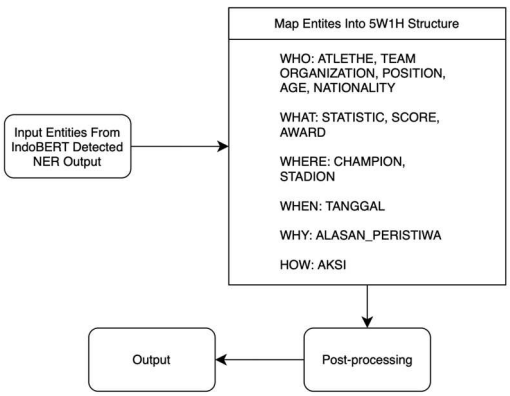

Fig. 3. Mapping IndoBERT NER Outputs into the 5W1H Structure with Rule-Based

IndoBERT Model Testing Result

This model was tested using 100 Indonesian sports news test data sets that had never been seen in previous training data. Based on the results of this test, the best F1-Score value was obtained with a value of 0.8191 at the 10th epoch. Meanwhile, in subsequent epochs, the F1-Score value continued to decline, this occurred because after the 10th epoch the model began to experience overfitting. The graph of the F1-Score and loss values can be seen in Figure 3. This evaluation was carried out on all 14 NER entities. Therefore, the results of the F1-Score and loss will produce the overall performance of this model itself in recognizing all categories of entities in Indonesian sports news.

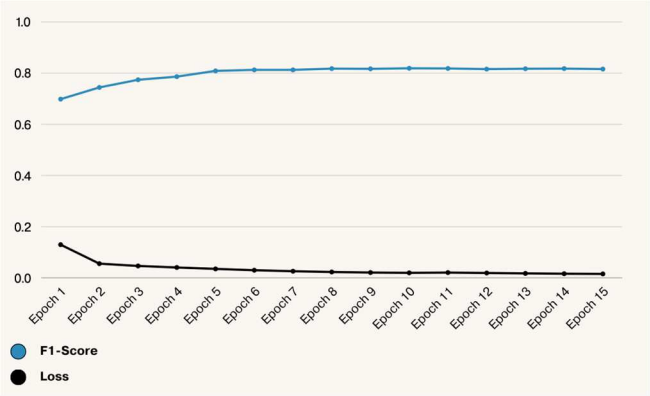

Fig. 4. F1-Score and Loss Graph

In addition to performing an F1-Score evaluation on all labels, this design also evaluates each label or entity using precision, recall, and F1-Score metrics. This evaluation aims to assess the IndoBERT model's ability to recognize each entity label individually. The results of each label evaluation can be seen in Figure 5, which show the performance of the IndoBERT model. From these evaluations, we can also identify which entity labels have the highest scores and which ones still require improvement.

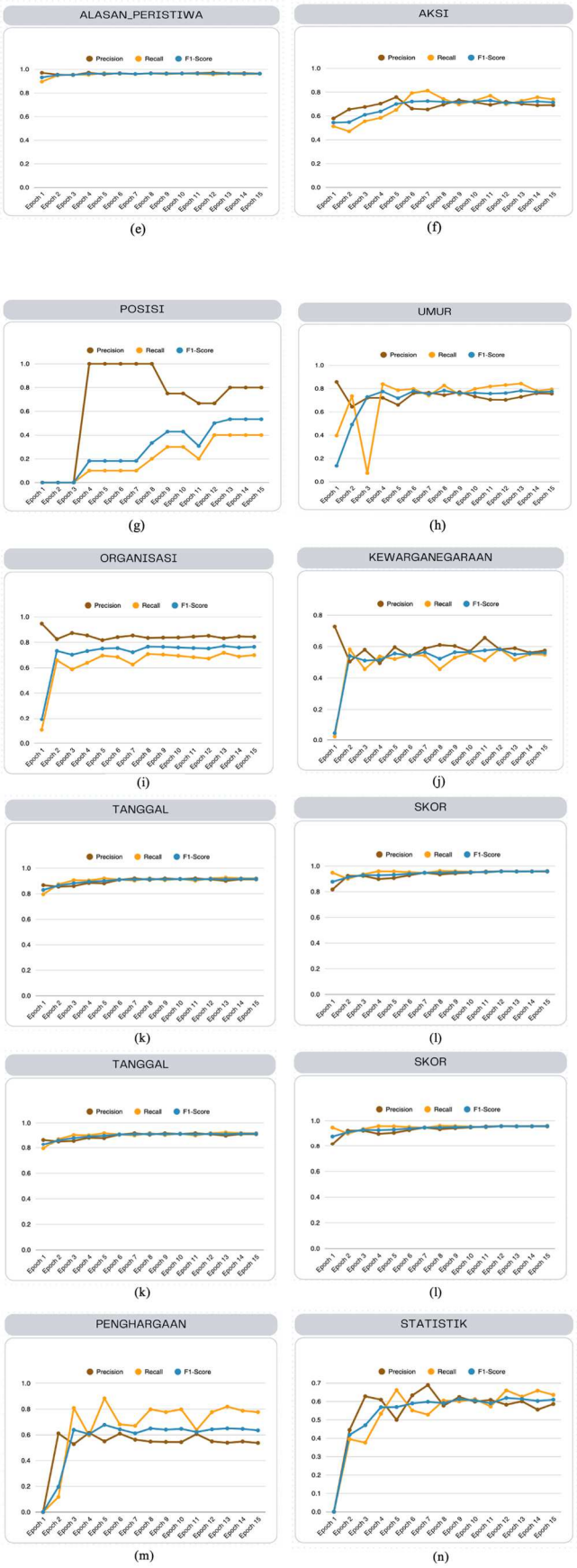

Fig. 5. Evaluation Results of Each Label

The per-label evaluation results indicate varying model performance across entity types, primarily influenced by entity frequency and linguistic complexity. High and stable

F1-scores were achieved for frequently occurring and well-structured entities such as ATLETHE, TEAM, DATE, and SCORE, with the DATE label reaching the highest F1-score due to consistent date patterns and the SCORE label benefiting from highly structured numerical formats. Entities such as CHAMPION, STADIUM, ORGANIZATION, and POSITION showed gradual performance improvements as the model learned diverse contextual patterns, while more complex entities like ACTION and STATISTIC exhibited moderate performance due to wide phrase variability and complex numerical structures. Labels with limited data representation, including NATIONALITY and AWARD, achieved lower but relatively stable F1-scores. Overall, these results demonstrate that the IndoBERT model performs effectively in extracting sports-related entities, achieving strong overall performance, while highlighting the impact of data distribution and entity complexity on NER accuracy.

V. CONCLUSION

The experimental results demonstrate that the fine-tuned IndoBERT model achieves strong performance for Named Entity Recognition in Indonesian sports news, attaining a best F1-score of 0.8191 at the 10th epoch, while the subsequent decline indicates potential overfitting. This performance confirms that domain-specific fine-tuning improves the model's ability to capture contextual patterns and recognize diverse sports-related entities. Furthermore, the integration of a rule-based approach for mapping extracted entities into the 5W1H structure produces more organized and interpretable outputs, enhancing information usability. Despite these positive results, the system faces limitations related to entity imbalance, restricted sports coverage, and complex entity structures such as statistics and actions. Overall, the proposed IndoBERT and rule-based information extraction system effectively meets its objectives, and future work may extend the dataset to additional sports and media sources or explore hybrid architectures, such as combining Transformer models with sequence labeling methods, to further improve extraction accuracy.

ACKNOWLEDGMENT

This design is supported by the supervisors who have provided guidance, direction, and very useful and valuable input during the design process until the preparation of this publication. My gratitude is also conveyed to my closest family and friends whom I love. This gratitude is conveyed for the support given to support the publication of this design. All support that has been given is an important part in completing this design and this publication.

REFERENCES

[1] APJII, "Jumlah Pengguna Internet Indonesia 2024," 7 2 2024. [Online]. Available: https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang.

[2] C. M. Annur, "Pengguna Internet di Indonesia Tembus 213 Juta Orang hingga Awal 2023," 20 09 2023. [Online]. Available: https://databoks.katadata.co.id/teknologi-telekomunikasi/statistik/d109a45f4409c34/pengguna-internet-di-indonesia-tembus-213-juta-orang-hingga-awal-2023.

[3] Media Indonesia, "55 Persen Waktu Masyarakat Indonesia Dihabiskan di Open Internet," 15 02 2023. [Online]. Available: https://mediaindonesia.com/humaniora/558604/55-persen-waktu-masyarakat-indonesia-dihabiskan-di-open-internet.

[4] D. Mardiah, "Minat Baca di Indonesia: Systematic Literature Review," *Jurnal Pena Ilmiah,* pp. 33-44, 2023.

[5] R. Tuasikal, "Concentrated, Corporate, and Camouflaged: The Nature of AI News Coverage in Indonesia," *Asian Journal of Media and Communication,* vol. 8, no. 2, 23 12 2024.

[6] J. Li, A. Sun J. Han, and C. Li , "A Survey on Deep Learning for Named Entity Recognition," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* 2020.

[7] M. V. Koroteev, "BERT: A Review of Applications in Natural Language Processing and Understanding," *arXiv,* 2021.

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *arXiv,* 2020.

[9] C. Chantrapornchai, and A. Tunsakul, "Information Extraction based on Named Entity for Tourism Corpus," *arXiv,* 2020.

[10] H. Darji, J. Mitrović, and M. Granitzer, "German BERT Model for Legal Named Entity Recognition," *arXiv,* 2023.

[11] A. Vasnawi *et al.*, "Attention Is All You Need," *arXiv,* 2023.

[12] G. Tucudean, M. Bucos, B. Draguscu, and C. D. Caleanu, "Natural language processing with transformers: a review," *IEEE Access,* 2024.

[13] R. K. Kaliyar, "A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of BERT," *IEEE,* 2020.

[14] B. Willie *et al.,* "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *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*, 2020.

[15] J. Terven *et al.*, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review,* 2025.