Vietnam Journal of Computer Science

Comparative Analysis of SciBERT, BioBERT, and PubMedBERT for Named Entity Recognition in Seaweed Literature

Nabila Nurasyifa

Department of Informatics, Sultan Maulana Hasanuddin State Islamic University Banten, Serang, Banten, 42171, Indonesia
nabilanurasyifa020@gmail.com

Anne Parlina

Research Center for Data and Information Science, National Research and Innovation Agency, Bandung Jawa Barat, 40135, Indonesia
anne.parlina@brin.go.id

Received (Day Month Year)
Revised (Day Month Year)

The exponential growth of scientific literature on seaweed necessitates automated information extraction methods to efficiently identify relevant entities. This study presents a comparative analysis of three domain-specific transformer-based models SciBERT, BioBERT, and PubMedBERT for Named Entity Recognition (NER) tasks on seaweed literature. A dataset comprising 1,122 scientific abstracts from Scopus was manually annotated using Label Studio, identifying four entity categories: Seaweed, Location, Compound, and Benefit. The BIO labeling scheme was employed for token-level annotation, and models were fine-tuned using weighted cross-entropy loss to address class imbalance. Experimental results demonstrate that PubMedBERT achieved the highest performance with precision of 92.74%, recall of 85.93%, F1-score of 88.30%, and accuracy of 85.93%. SciBERT obtained an F1-score of 87.19%, while BioBERT achieved 86.97%. Per-class analysis revealed that PubMedBERT excelled in recognizing Seaweed and Location entities, whereas BioBERT performed best on Compound entities. These findings indicate that domain-specific pre-training significantly influences NER performance, with PubMedBERT being the most suitable model for biomedical entity extraction in seaweed research domains.

Keywords: Named Entity Recognition; Transformer Models; Seaweed Literature; BioBERT; PubMedBERT.

1. Introduction

The field of seaweed research has experienced substantial growth due to the increasing recognition of seaweed's potential in food security, pharmaceutical applications, and environmental sustainability.[1] Scientific publications in this domain contain valuable information about species identification, geographical distribution, bioactive compounds, and health benefits.[2] However, the manual extraction of such information from vast

literature collections is time-consuming and impractical, necessitating automated information extraction approaches.[3]

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that aims to identify and classify named entities in text into predefined categories.[3] Statistical methods such as Conditional Random Fields (CRF) have been widely used as foundational probabilistic models for segmenting and labeling sequence data.[4] The advent of deep learning has revolutionized NER, with recurrent neural networks and Long Short-Term Memory (LSTM) networks demonstrating superior performance.[5-6]

The introduction of the Transformer architecture, which relies entirely on self-attention mechanisms, enables the modeling of global dependencies and comprehensive contextual representations without regard to their distance in the input sequence .[7] Building upon this, transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers), have established new benchmarks in various NLP tasks including NER.[8] However, general-domain BERT models may not capture domain-specific terminology and linguistic patterns prevalent in scientific and biomedical literature.[9]

To address this limitation, domain-specific BERT variants have been developed. SciBERT was pre-trained on a large corpus of scientific publications from Semantic Scholar, encompassing computer science and biomedical domains.[10] BioBERT extended the original BERT model by continuing pre-training on PubMed abstracts and PubMed Central full-text articles.[11] PubMedBERT introduced a from-scratch pre-training approach exclusively on PubMed abstracts, demonstrating that domain-specific vocabulary and pre-training can enhance biomedical NLP performance.[12]

Despite the availability of these domain-specific models, their comparative effectiveness for NER in specialized scientific domains such as seaweed research remains underexplored. Previous studies have primarily evaluated these models on established biomedical NER benchmarks,[13-14] but limited research has investigated their applicability to marine biology and related fields.

This study addresses this gap by conducting a comprehensive comparative analysis of SciBERT, BioBERT, and PubMedBERT for NER on seaweed literature. The research objectives are: (1) to develop an annotated NER dataset from seaweed scientific abstracts with domain-specific entity categories; (2) to implement and fine-tune three transformer-based models for entity recognition; and (3) to evaluate and compare model performance using standard metrics to identify the most effective approach for this domain.

The remainder of this paper is organized as follows. Section 2 describes the methodology including dataset construction, annotation scheme, preprocessing, and model implementation. Section 3 presents the experimental results and analysis. Section 4 concludes the paper with findings and future research directions.

2. Methods

2.1. *Dataset Description*

The dataset utilized in this study was obtained from the Research Center for Data and Information Science (PRSDI), National Research and Innovation Agency (BRIN), Indonesia. The corpus comprises 1,122 scientific abstracts retrieved from the Scopus database, focusing on seaweed-related research. All documents are written in English and stored in CSV format containing metadata and abstract text.

2.2. *Annotation Process*

Manual annotation was performed using Label Studio, an open-source data labeling platform.[15] Four entity categories were defined based on domain relevance: (1) Seaweed, which refers to scientific names, genus, species, or common names of seaweed; (2) Location, representing geographical locations related to seaweed occurrence, cultivation, or sampling sites; (3) Compound, encompassing nutritional content, bioactive compounds, chemical components, vitamins, minerals, pigments, antioxidants, and polysaccharides found in seaweed; and (4) Benefit, denoting health benefits, biological activities, physiological functions, nutritional value, and potential applications. The annotation followed established guidelines for biomedical NER,[16] ensuring consistency across annotators. Each abstract was carefully examined, and relevant text spans were labeled according to the defined categories.

2.3. *BIO Labeling Scheme*

The annotated data was converted to the BIO (Beginning-Inside-Outside) labeling scheme, a standard approach for sequence labeling tasks.[17] Under this scheme, the first token of an entity is labeled with B- prefix, subsequent tokens within the same entity receive I- prefix, and non-entity tokens are labeled O. This resulted in nine labels: O, B-SEAWEED, I-SEAWEED, B-LOCATION, I-LOCATION, B-COMPOUND, I-COMPOUND, B-BENEFIT, and I-BENEFIT.

2.4. *Data Preprocessing*

Preprocessing involved tokenization and label alignment. Tokenization was performed using the respective tokenizers of each model, converting text into subword tokens according to the WordPiece algorithm.[18] Since a single word may be split into multiple subword tokens, label alignment was necessary to ensure each token received the appropriate label. The offset mapping from tokenization was utilized to match tokens with entity positions in the original text.

Padding and truncation were applied to standardize input lengths across all samples. Maximum sequence length was set according to model specifications, with longer sequences truncated and shorter sequences padded.

2.5. *Dataset Splitting*

The dataset was divided into training, validation, and test sets following a 70-15-15 ratio. Specifically, 70% of the data was allocated for training, while the remaining 30% was equally split between validation and test sets. This partition ensures sufficient data for model training while maintaining independent sets for hyperparameter tuning and final evaluation.[19]

2.6. *Model Architecture*

Three transformer-based models were implemented for comparative analysis: SciBERT, BioBERT, and PubMedBERT. SciBERT was pre-trained on 1.14 million scientific papers from Semantic Scholar, covering computer science (18%) and biomedical (82%) domains.[10] It employed a custom vocabulary of 30,000 tokens derived from the scientific corpus.

BioBERT was developed by further pre-training the original BERT-Base model on PubMed abstracts (4.5 billion words) and PubMed Central full-text articles (13.5 billion words).[11] Unlike SciBERT, BioBERT retained the original BERT vocabulary while adapting its language representations to biomedical texts.

PubMedBERT was pre-trained from scratch exclusively on PubMed abstracts using a domain-specific vocabulary.[12] This training strategy enabled both the vocabulary and model parameters to be optimized specifically for biomedical text processing.

All three models were configured for token classification by adding a linear classification head on top of the transformer encoder. The classification layer mapped the contextualized hidden representations to the nine predefined Named Entity Recognition (NER) labels.

2.7. *Training Configuration*

Training was conducted using the Hugging Face Transformers library.[20] Table 1 presents the training hyperparameters applied consistently across all models.

Table 1. Training Configuration for NER ModeL

Parameter	Value
Learning Rate	2e-5
Batch Size	8
Maximum Epochs	7
Weight Decay	0.01
Early Stopping Patience	2
Loss Function	Weighted Cross-Entropy

To address class imbalance in the dataset, weighted cross-entropy loss was employed. The O label, being the majority class, was assigned a lower weight compared to entity labels. This approach prevents the model from being biased toward predicting non-entity tokens.[21] Early stopping with patience of 2 epochs was implemented to prevent overfitting. Training was terminated if no improvement in validation F1-score was observed for two consecutive epochs, and the best-performing checkpoint was retained.

2.8. *Evaluation Metrics*

Model performance was evaluated using precision, recall, F1-score, and accuracy, calculated as follows:

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

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

$$\boldsymbol{F1\text{-}score} = \boldsymbol{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

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

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.[22] While evaluation was performed at the token level following standard baseline protocols, confusion matrices were subsequently generated to provide a detailed error analysis.

3. Results and Discussion

3.1. *Dataset Statistics*

Table 2 presents the distribution of entity labels in the annotated dataset. The O label constitutes the majority class, which is typical for NER datasets where most tokens are non-entities.

Table 2. Distribution of NER Labels in Dataset

Label	Count
O	184180
B-SEAWEED	3293
I-SEAWEED	3555
B-LOCATION	1118
I-LOCATION	241
B-COMPOUND	6583

I-COMPOUND	4289
B-BENEFIT	5571
I-BENEFIT	17355

The class imbalance justifies the use of weighted loss during training to ensure adequate learning of minority entity classes.

3.2. *SciBERT Results*

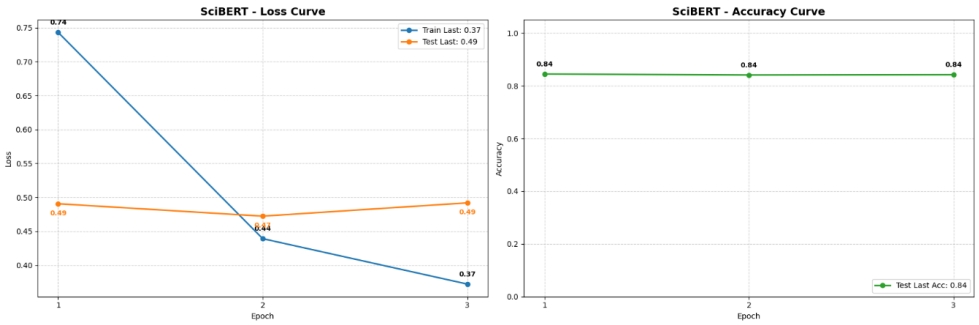

Fig 1. Visualisasi loss curve & accuracy curve SciBERT

SciBERT achieved optimal performance at epoch 1 with early stopping triggered thereafter. Figure 1 illustrates the training and validation loss curves during the fine-tuning process. The training loss decreased from 0.74 to 0.37 by epoch 3, indicating effective pattern learning. Validation loss reached its minimum of 0.47 at epoch 2 before increasing, suggesting potential overfitting with continued training.

Fig 2. Confusion Matrix SciBERT

Figure 2 presents the normalized confusion matrix of the SciBERT model, providing a detailed view of the classification performance for each entity class. Analysis of the confusion matrix revealed strong performance on Seaweed entities, with B-SEAWEED and I-SEAWEED achieving 0.94 and 0.95 normalized accuracy, respectively. Location entities also demonstrated good recognition with B-LOCATION at 0.89. However, Benefit entities showed relatively lower performance, with B-BENEFIT at 0.66 and I-BENEFIT at 0.75. Table 3 presents the final evaluation metrics for SciBERT.

Table 3. SciBERT Evaluation Results

Metric	**Value (%)**
Precision	92.38
Recall	84.46
F1-Score	87.19
Accuracy	84.46

3.3. *BioBERT Results*

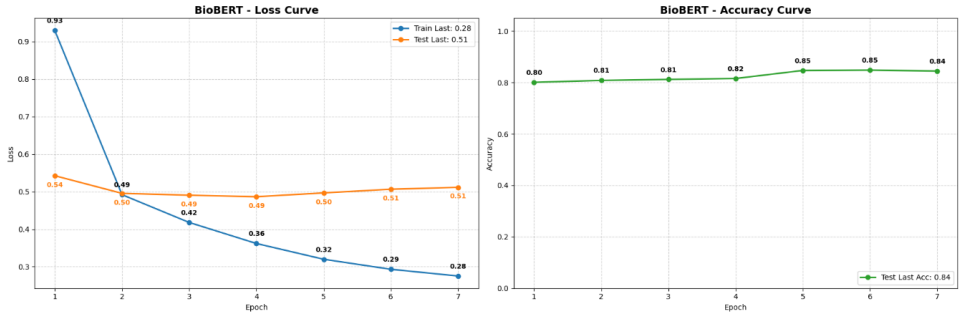

Fig 3. Visualisasi loss curve & accuracy curve BioBERT

BioBERT demonstrated stable training dynamics with training loss decreasing from 0.93 to 0.28 across epochs. Figure 3 illustrates the training and validation loss curves, as well as the validation accuracy during the fine-tuning process. Validation loss remained relatively stable between 0.49 and 0.51, indicating consistent generalization. Validation accuracy improved from 0.80 to 0.84, with peak performance of 0.85 at epochs 5 and 6.

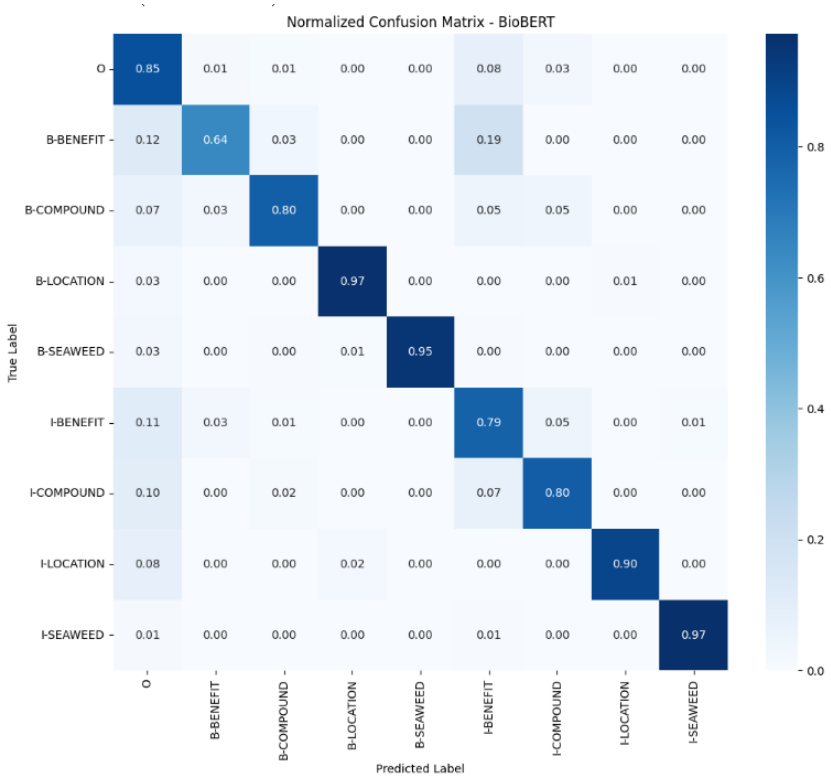

Fig. 4. Confusion Matrix BioBERT

Figure 4 presents the normalized confusion matrix of the BioBERT model. The confusion matrix analysis showed BioBERT's strength in Location and Seaweed recognition. B-LOCATION achieved 0.97 normalized accuracy, followed by I-SEAWEED at 0.97 and B-SEAWEED at 0.95. Similar to SciBERT, Benefit entities proved challenging, with B-BENEFIT at 0.64. Table 4 summarizes BioBERT's evaluation metrics.

Table 4. BioBERT Evaluation Results

Metric	Value (%)
Precision	91.29
Recall	84.82
F1-Score	86.97
Accuracy	84.82

3.4. *PubMedBERT Results*

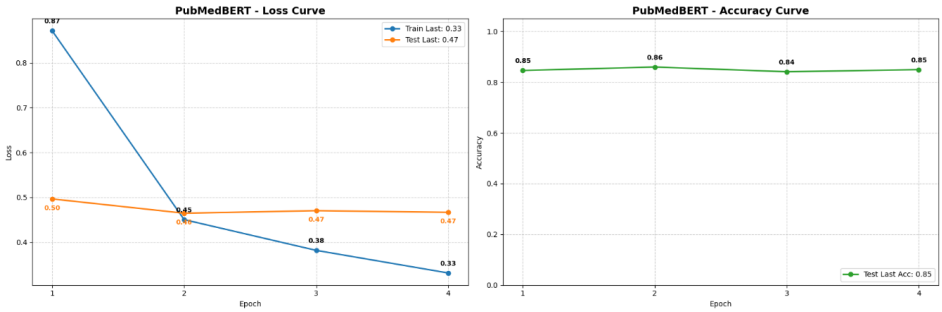

Fig. 5. Visualisasi loss curve & accuracy curve PubMedBERT

PubMedBERT achieved optimal performance at epoch 2, demonstrating rapid convergence. Figure 5 illustrates the training and validation loss curves, as well as the validation accuracy during the fine-tuning process. Training loss decreased from 0.87 to 0.33, while validation loss stabilized around 0.46–0.47 after the initial decrease. Validation accuracy ranged from 0.84 to 0.86, with consistent performance throughout training.

Fig. 6. Confusion Matrix PubMedBERT

Figure 6 presents the normalized confusion matrix of the PubMedBERT model. Confusion matrix analysis revealed PubMedBERT's superior entity recognition capabilities. B-SEAWEED and I-SEAWEED both achieved 0.96 normalized accuracy, while B-LOCATION reached 0.93. Benefit entities remained the most challenging category, with B-BENEFIT at 0.63 and I-BENEFIT at 0.75. Table 5 presents PubMedBERT's evaluation metrics.

Table 5. PubMedBERT Evaluation Results

Metric	Value (%)
Precision	92.74
Recall	85.93
F1-Score	88.30
Accuracy	85.93

3.5. *Comparative Analysis*

Table 6 provides a comprehensive comparison of all three models.

Table 6. Comparative Performance of NER Models

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
SciBERT	92.38	84.46	87.19	84.46
BioBERT	91.29	84.82	86.97	84.82
PubMedBERT	**92.74**	**85.93**	**88.30**	**85.93**

PubMedBERT outperformed both SciBERT and BioBERT across all metrics. The 1.11 percentage point improvement in F1-score over SciBERT and 1.33 percentage point improvement over BioBERT demonstrates the advantage of domain-specific pre-training from scratch.Table 7 presents per-class F1-scores for detailed comparison.

Table 7. Per-Class F1-Score Comparison

Entity Class	SciBERT (%)	BioBERT (%)	PubMedBERT (%)
Seaweed	94.17	95.00	**95.74**
Location	85.37	90.80	**92.22**
Compound	80.40	**81.87**	79.30
Benefit	**83.84**	82.58	83.14

PubMedBERT achieved the highest F1-scores for Seaweed (95.74%) and Location (92.22%) entities. BioBERT demonstrated superior performance on Compound entities (81.87%), while SciBERT performed best on Benefit entities (83.84%).

3.6. *Discussion*

The experimental results support several important findings regarding transformer-based NER for scientific literature.

First, domain-specific pre-training significantly impacts NER performance. PubMedBERT's from-scratch pre-training on biomedical text resulted in vocabulary and representations better suited for scientific entity recognition.[23] This aligns with previous findings that domain-adaptive pre-training enhances downstream task performance .[24]

Second, the Benefit entity category proved most challenging across all models, with F1-scores ranging from 82.58% to 83.84%. This difficulty likely stems from the semantic complexity and diversity of benefit-related expressions, which often overlap with general language patterns. Future work could explore expanded annotation guidelines or hierarchical entity classification to improve Benefit recognition.

Third, the relatively rapid convergence of PubMedBERT (optimal at epoch 2) compared to BioBERT (optimal at epoch 6) suggests that models pre-trained from scratch on domain-specific data require less fine-tuning. This finding has practical implications for computational resource allocation in NER system development.[25]

The consistent strong performance on Seaweed entities across all models (F1-scores above 94%) indicates that species names and taxonomic terminology are well-represented in scientific pre-training corpora. Conversely, Location entities showed more variation,

potentially due to the diversity of geographical references in seaweed literature spanning multiple continents and marine environments.[26]

These results extend previous comparative studies on biomedical NER[27] by demonstrating model effectiveness on a novel domain-specific dataset. The findings suggest that researchers working with marine biology literature should prioritize PubMedBERT for information extraction tasks, while considering ensemble approaches to leverage the complementary strengths of different models.

4. Conclusion

This study conducted a comparative analysis of SciBERT, BioBERT, and PubMedBERT for Named Entity Recognition on seaweed scientific literature. An annotated dataset of 1,122 abstracts was developed with four entity categories: Seaweed, Location, Compound, and Benefit.

Experimental results demonstrated that PubMedBERT achieved the best overall performance with precision of 92.74%, recall of 85.93%, F1-score of 88.30%, and accuracy of 85.93%. SciBERT obtained an F1-score of 87.19%, while BioBERT achieved 86.97%. Per-class analysis revealed that PubMedBERT excelled in Seaweed and Location recognition, BioBERT performed best on Compound entities, and SciBERT showed advantages for Benefit entities.

These findings confirm that domain-specific pre-training significantly influences NER performance, with from-scratch pre-training on biomedical corpora providing advantages for scientific entity extraction. The developed dataset and comparative findings contribute to advancing information extraction methods for marine biology research.

Future work should explore ensemble methods combining multiple transformer models, investigate semi-supervised and weakly supervised approaches to expand training data, and extend entity categories to capture additional domain-relevant information such as extraction methods and industrial applications.

References

1. S. Lomartire, J. C. Marques and A. M. M. Gonçalves, An Overview to the Health Benefits of Seaweeds Consumption, Marine Drugs 19 (2021) 341.
2. M. D. Torres, S. Kraan and H. Domínguez, Seaweed biorefinery, Rev. Environ. Sci. Biotechnol. 18 (2019) 335–388.
3. J. Li, A. Sun, J. Han and C. Li, A Survey on Deep Learning for Named Entity Recognition, IEEE Trans. Knowl. Data Eng. 34 (2022) 50–70.
4. J. Lafferty, A. McCallum and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in Proc. 18th Int. Conf. Machine Learning (ICML), San Francisco, USA (Morgan Kaufmann, 2001), pp. 282–289.
5. W. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv:1508.01991 (2015).

6. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural Architectures for Named Entity Recognition, in Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics (NAACL), 2016, pp. 260–270.
7. J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proc. NAACL-HLT 2019 (Association for Computational Linguistics, Minneapolis, 2019), pp. 4171–4186.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention Is All You Need, in Advances in Neural Information Processing Systems 30 (2017), pp. 5998–6008.
9. S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey and N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online (2020), pp. 8342–8360.
10. I. Beltagy, K. Lo and A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text, in Proc. EMNLP-IJCNLP 2019, Hong Kong (Association for Computational Linguistics, 2019), pp. 3615–3620.
11. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining, Bioinformatics 36 (2020) 1234–1240.
12. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, ACM Trans. Comput. Healthc. 3 (2021) 1–23.
13. L. Weber, M. Sänger, J. M. Münchmeyer, M. Habibi, U. Leser and A. Akbik, HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition, Bioinformatics 37 (2021) 2792–2794.
14. S. Gao, O. Kotevska, A. Sorokine and J. B. Christian, A Pre-training and Self-training Approach for Biomedical Named Entity Recognition, PLoS ONE 16 (2021) e0246310.
15. Heartex, Label Studio: Data Labeling Software (2020), https://labelstud.io/.
16. K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. L. Johnson, C. Roeder, J. D. Choi, C. Funk, Y. Malenkiy, M. Eckert, N. Xue, W. A. Baumgartner Jr., M. Bada, M. Palmer and L. E. Hunter, A Corpus of Full-Text Journal Articles is a Robust Evaluation Tool for Revealing Differences in Performance of Biomedical Natural Language Processing Tools, BMC Bioinformatics 13 (2012) 207.
17. E. F. Tjong Kim Sang and J. Veenstra, Representing Text Chunks, in Proc. EACL'99 (1999), pp. 173–179.
18. Y. Wu et al., Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, arXiv:1609.08144 (2016).
19. K. Khotimah, S. Surono and A. Thobirin, Optimizing EfficientNet for Imbalanced Medical Image Classification Using Grey Wolf Optimization, Comput. Sci. Inf. Technol. 6 (2025) 112–121.
20. T. Wolf et al., Transformers: State-of-the-Art Natural Language Processing, in Proc. EMNLP 2020: System Demonstrations (2020), pp. 38–45.
21. X. Li, X. Sun, Y. Meng, J. Liang, F. Wu and J. Li, Dice Loss for Data-imbalanced NLP Tasks, arXiv:1911.02855 (2020).
22. D. M. W. Powers, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.
23. H. Poon and P. Domingos, Unsupervised Semantic Parsing, in Proc. EMNLP 2009, Singapore (Association for Computational Linguistics, 2009), pp. 1–10.
24. C. Sun, X. Qiu, Y. Xu and X. Huang, How to Fine-Tune BERT for Text Classification, in China National Conf. Chinese Computational Linguistics (2019), pp. 194–206.

25. E. Strubell, A. Ganesh and A. McCallum, Energy and Policy Considerations for Deep Learning in NLP, in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy (2019), pp. 3645–3650.
26. O. G. Mouritsen, C. Dawczynski, L. Duelund, G. Jahreis, W. Vetter and M. Schröder, On the Human Consumption of the Red Seaweed Dulse (Palmaria palmata (L.) Weber & Mohr), J. Appl. Phycol. 25 (2013) 1777–1791.
27. R. I. Doğan, R. Leaman and Z. Lu, NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization, J. Biomed. Inform. 47 (2014) 1–10.