

Article

Natural Language Processing-Based Deep Learning to Predict the Loss of Consciousness Event Using Emergency Department Text Records

Hang A. Park ^{1,†}, Inyeop Jeon ^{2,†}, Seung-Ho Shin ², Soo Young Seo ², Jae Jun Lee ³, Chulho Kim ^{2,4,*}
and Ju Ok Park ^{1,*}

¹ Department of Emergency Medicine, Hallym University College of Medicine, Dongtan Sacred Heart Hospital, Hwaseong 18450, Republic of Korea; hangapark@hallym.or.kr

² Artificial Intelligence Research Center, Hallym University Sacred Heart Hospital, Chuncheon 24253, Republic of Korea; in_yeop@hallym.ac.kr (I.J.); sin12ho@hallym.or.kr (S.-H.S.); syseo96@gmail.com (S.Y.S.)

³ Department of Anesthesiology, Hallym University Sacred Heart Hospital, Chuncheon 24253, Republic of Korea; iloveu59@hallym.or.kr

⁴ Department of Neurology, Hallym University Sacred Heart Hospital, Chuncheon 24253, Republic of Korea

* Correspondence: gumdol52@hallym.or.kr (C.K.); juokpark@hallym.ac.kr (J.O.P.)

† These authors contributed equally to this work.

Abstract: The increasing adoption of electronic medical records (EMRs) presents a unique opportunity to enhance trauma care through data-driven insights. However, extracting meaningful and actionable information from unstructured clinical text remains a significant challenge. Addressing this gap, this study focuses on the application of natural language processing (NLP) techniques to extract injury-related variables and classify trauma patients based on the presence of loss of consciousness (LOC). A dataset of 23,308 trauma patient EMRs, including pre-diagnosis and post-diagnosis free-text notes, was analyzed using a bilingual (English and Korean) pre-trained RoBERTa model. The patients were categorized into four groups based on the presence of LOC and head trauma. To address class imbalance in LOC labeling, deep learning models were trained with weighted loss functions, achieving a high area under the curve (AUC) of 0.91. Local Interpretable Model-agnostic Explanations analysis further demonstrated the model's ability to identify critical terms related to head injuries and consciousness. NLP can effectively identify LOC in trauma patients' EMRs, with weighted loss functions addressing data imbalances. These findings can inform the development of AI tools to improve trauma care and decision-making.

Keywords: natural language processing; text mining; deep learning; emergency departments; clinical decision support

Citation: Park, H.A.; Jeon, I.; Shin, S.-H.; Seo, S.Y.; Lee, J.J.; Kim, C.; Park, J.O. Natural Language Processing-Based Deep Learning to Predict the Loss of Consciousness Event Using Emergency Department Text Records. *Appl. Sci.* **2024**, *14*, 11399. <https://doi.org/10.3390/app142311399>

Received: 18 October 2024

Revised: 2 December 2024

Accepted: 6 December 2024

Published: 6 December 2024



Copyright: © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Electronic medical records (EMRs) are a valuable source of data for injury analysis but present challenges due to their combination of structured and unstructured information [1–3]. Structured data, such as patient demographics and laboratory results, follow a standardized format and is easily extracted. However, unstructured data, including clinical notes, surgical records, and radiology reports, contain essential information in narrative form that is difficult to extract due to variability in language, grammatical errors, and inconsistent terminology [4]. Despite these challenges, unstructured data provides essential insights that support clinical decisions, inform public health policies, and support injury research [5].

Integrating natural language processing (NLP) with EMRs is crucial for extracting meaningful information from unstructured data [6,7]. NLP enables the extraction of

narrative data, allowing researchers to combine it with structured data for a more comprehensive analysis of specific clinical scenarios. For example, NLP has been used to identify vulnerable populations, such as homeless youth using psychiatric emergency services, by extracting social determinants from narrative data combined with demographic information [8]. Additionally, NLP can enhance the accuracy of predictive models by analyzing discharge summaries to detect risk factors for psychiatric readmission [8].

Detecting loss of consciousness (LOC) in EMRs is particularly important for assessing traumatic brain injury (TBI) severity, as LOC is a critical criterion in TBI diagnosis and classification. LOC information guides clinical decisions, such as the need for computed tomography in blunt head trauma cases [9,10], and is a key data element for TBI data collection [11]. Automating the extraction of LOC from EMRs streamlines the data collection process, reducing reliance on manual chart reviews. NLP-based EMR analytics can identify injury patterns and severity, offering insights beyond what structured data alone can provide. However, the lack of real-world datasets for multilingual text processing like EMRs [12], the computational power to handle complex content, and the lack of research on whether multilingual clinical NLP has helped with healthcare decisions has made it difficult to conduct multilingual NLP studies using EMRs [13].

This study aims to develop NLP algorithms for the automatic identification of injury-related variables within free-text medical records from emergency department (ED) settings. Specifically, the goal is to classify patients into four categories based on the presence or absence of head trauma and loss of consciousness (LOC). This classification not only identifies cases with LOC but also distinguishes them from other head trauma-related conditions, including cases with no LOC or missing LOC information. By providing a granular categorization of injury-related variables, this study supports clinical decision-making and contributes to more targeted trauma care. In addition, other studies explore the feasibility of developing a multilingual NLP model (English and Korean) capable of processing and interpreting EMR data across diverse clinical contexts.

2. Materials and Methods

2.1. Study Design and Data Collection

This study is a retrospective analysis using a clinical data warehouse (CDW) to access electronic medical records (EMRs) of patients who visited an ED serving as a local emergency center in an urban area of South Korea. The data, collected between January 2022 and May 2023, included patients who visited the ED for suspected injuries.

Injuries were defined as cases where individuals presented to the ED with symptoms resulting from external causes. These patients were initially assessed by experienced nurses during triage, and their conditions were subsequently confirmed by the attending ED physician. The injuries included a range of types, from mechanical causes such as vehicle accidents and falls to systemic causes like poisoning and asphyxiation.

A total of 23,308 medical records were collected for this study. The records included descriptions from pre-diagnosis notes obtained upon the patient's arrival at the ED and post-diagnosis notes recorded after initial evaluation. The dataset comprised all free-text entries written by ED physicians in the EMR, detailing the patient's present illness, review of systems, physical examination findings, ED progress notes, and discharge or disposition summaries. These notes were categorized into pre-diagnosis notes (present illness, review of systems, and physical examination) and post-diagnosis notes (progress notes and discharge or disposition summaries).

2.2. Definition of Outcome Variable

For the outcome variable, the history of each patient's loss of consciousness (LOC) was labeled based on information from the EMR. To ensure data accuracy, medical professionals meticulously reviewed all medical records and manually assigned labels to instances of LOC. LOC was identified as present if reported by the patient or if the clinical

history strongly suggested a temporary disruption due to LOC during the examination, with the duration of LOC not being considered. The refined LOC data were categorized into four groups based on the presence of head injury and LOC experience: (1) C1: injury without head injury, (2) C2: head injury without LOC or mental change, (3) C3: Head injury with LOC or mental change, and (4) C99: head injury without recorded information for LOC or mental change. The classification objective is to distinguish between cases with LOC, those with head trauma without LOC, and those with no head trauma or ambiguous LOC information. This approach ensures that LOC is explicitly identified and contextualized within the broader spectrum of trauma-related conditions.

2.3. Data Preparation and Model Development

To build the LOC classification model, we divided the total of 23,308 patients into a 7:3 ratio for training and validation datasets, respectively. Random sampling was conducted while maintaining the proportion of each class across the entire dataset. Subsequent analyses were performed using the pre-selected 16,315 training samples, and the model was tested using 6993 validation samples (Figure 1).

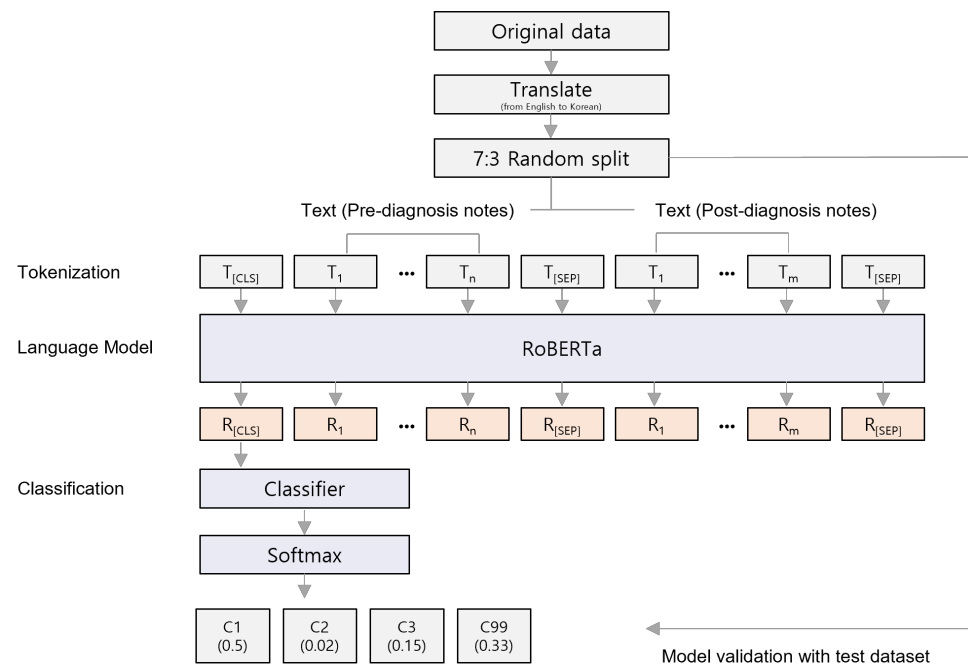


Figure 1. The architecture of the fine-tuned RoBERTa classification model. The model architecture is based on the basic RoBERTa framework. The tokenized input data consists of two types of clinical diagnostic notes: pre-diagnostic notes and post-diagnostic notes, with each note separated by the [SEP] special token. The output from the model is a set of probabilities and values that are calculated using the softmax function.

The RoBERTa model is developed based on BERT, which employs transformer architecture [14]. By optimizing the training process of the BERT model, RoBERTa achieves enhanced performance. The RoBERTa model leverages byte pair encoding and dynamic masking, demonstrating superior performance compared to BERT, thereby proving its suitability for handling complex sentence processing [15]. The pre-diagnosis notes comprising the medical records exhibit a high proportion of non-professional terms due to being derived from patient experiences or environments of injury. To achieve accurate classification performance, we employed byte pair encoding to decompose sentences into sub-word units. Therefore, we utilized the RoBERTa model to develop the classification model.

The notes were written in a mix of English and Korean, which is common practice in South Korea. In the case of English words, they were translated into Korean using a

custom word dictionary created by the author (C. Kim) before tokenization. We employed the transformer library (v4.31.0) to load a pre-trained RoBERTa model, which was adapted for transfer learning [14]. To tailor the RoBERTa model for Korean language analysis, we utilized modified word embeddings specific to our analysis [16]. In addition, we tokenized the Korean input texts using the tokenizer from the pre-trained model. The tokenizer applies morpheme-based byte pair encoding, which effectively separates words in the Korean language. A linear classification layer was added on top of the pre-trained model, with embedded weights from each layer being fine-tuned using the training dataset.

The EMRs were processed by concatenating different sections into a single input, with each section separated by the special token '[SEP]'. These combined inputs were then used to fine-tune the pre-trained model, while the test data, kept isolated, was used to evaluate model performance. During fine-tuning, we retained the default architecture of the pre-trained model and set the key training parameters as follows: (1) batch size of 16, (2) 50 training epochs, and (3) a learning rate of 2.4×10^{-6} . Training was stopped early if the model's performance did not improve after 5 consecutive epochs. Optimal hyperparameters were selected based on the best area under the receiver operating characteristic (AUROC) score. The model was trained using the AdamW optimizer to minimize the loss function [17], which was calculated using cross-entropy for multi-class classification. To address class imbalances, a weighted loss function was applied [18].

2.4. Model Performance Evaluation and Feature Interpretation

To evaluate the results of the multi-class classification task, we measured the model's predictive performance on both the training and test datasets. Recall and precision were calculated for each class to compare predictive performance. To quantitatively assess the prediction accuracy for each class, we performed AUROC analysis with the scikit-learn library (v1.3.0) [19–21]. For interpretability, we used the LimeTextExplainer function from the LIME (Local Interpretable Model-Agnostic Explanations) package (v0.2.0.1) to evaluate and visualize the contribution of individual words in a sentence to the model's class predictions [22]. The LimeTextExplainer function measures the change in classification probability when words are excluded one by one, enabling an indirect comparison of each feature's contribution to the model's predictions. Furthermore, to assess the potential biases within the model, ten repeated 5-fold cross-validations were performed.

2.5. Ethics Statement

The present study protocol was reviewed and approved by the Institutional Review Board of Hallym University Dongtan Sacred Heart Hospital (IRB No. 2023-08-013, 26 September 2023). Informed consent was waived for all subjects, as this study involved only the analysis of pre-existing text records and did not involve any direct interaction with human subjects.

3. Results

3.1. Analysis of Experimental Dataset

The distribution of the data was compared between the groups, and it was found that patients without head injuries (C1 group) represented the largest proportion at 67.7% (Table 1). The next largest group, accounting for 26.8%, consisted of patients with head injuries but without LOC (C2 group). Patients with both head injuries and LOC (C3 group) made up 3.4% of the total. Due to the low frequency of patients in the LOC group, we performed random sampling while maintaining the proportion of each group during data splitting for model training (Table S1).

Table 1. Clinical characteristics of the participants.

Variable	Number	%
Gender		
Male	13,371	57.4
Female	9937	42.6
Age, years		
0–19	8114	34.8
20–39	6497	27.9
40–59	5470	23.5
60–79	2469	10.6
≥80	758	3.3
LOC ¹ status		
C1 ²	15,775	67.7
C2 ³	6240	26.8
C3 ⁴	780	3.4
C99 ⁵	513	2.2
Disposition at emergency department		
Discharge (to home)	21,038	90.3
Admission	1993	8.6
Transfer	101	0.4
Death	55	0.2
NA ⁶	121	0.5
Death		
Yes	119	0.5
No	23,189	99.5

¹ Loss of consciousness. ² Injury without head injury. ³ Head injury without no LOC or mental change. ⁴ Head injury with LOC or mental change. ⁵ Head injury with no recorded information for LOC or mental change. ⁶ Missing value.

3.2. Classification Results and Model Evaluation

An early stopping method was applied to prevent overfitting, particularly for the C1 group, which comprised 68.4% of the dataset, as the model was trained on unbalanced data. Training was halted when the AUROC score did not improve for five consecutive epochs. The training parameters for our model are detailed in Table S2. After fine-tuning the pre-trained model, the evaluation showed strong performance, with an AUROC of 0.91, an accuracy of 0.91, and an F1 score of 0.91 (Table 2). For the groups without LOC (C1 and C2), the recall rates were 94.0% and the precision rates were 90.7%, respectively. Although the group with LOC (C3) only represented 3.4% of the total data, the model predicted 148 patients with LOC out of 234 observed cases, achieving a recall and precision rate of 63.3% (Figure 2). The results of AUROC's 5-fold cross validation on the test dataset are shown in Figure S1.

Table 2. Performance metrics of loss of consciousness classification according to input dataset.

Input Data	AUROC	Accuracy	Precision	Recall	F1	Weighted F1
Pre-/Post-diagnosis note	0.916	0.910	0.735	0.737	0.735	0.911
Pre-diagnosis note	0.906	0.904	0.724	0.723	0.718	0.905
Post-diagnosis note	0.746	0.818	0.695	0.428	0.447	0.797

Precision, recall, and F1 score were calculated on macro-averaging and weighted F1 score was calculated by taking into account the size of the group. AUROC, area under the receiver operating characteristic.

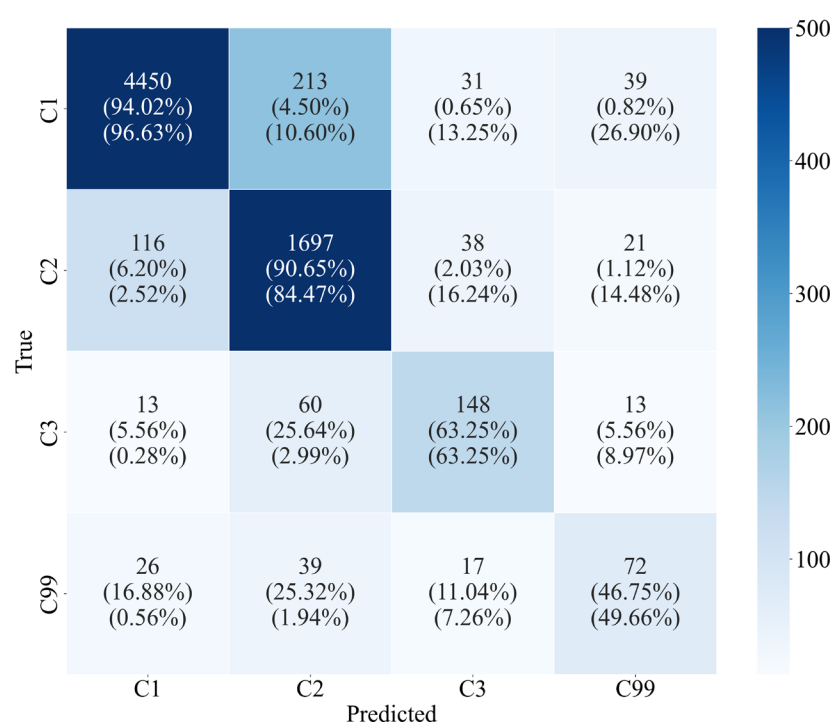
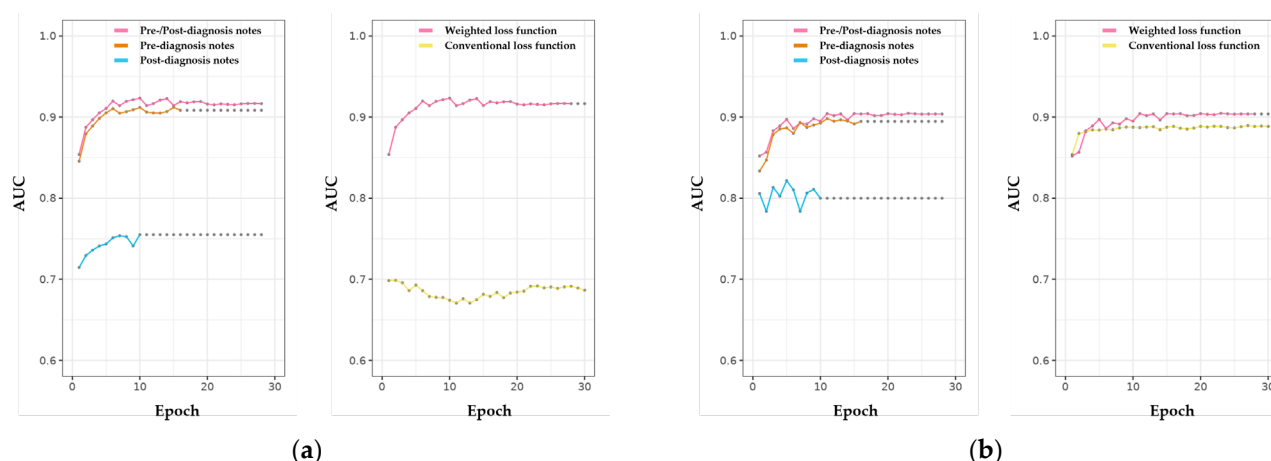


Figure 2. Multiclass confusion matrix for test dataset. The number in each cell represents the degree of agreement between the actual and predicted classification groups. The number in round brackets indicates the recall and precision score calculated for each class in the sequence.

The weighted-loss method was applied to prevent the model's performance from disproportionately favoring the non-LOC group, which makes up a large portion of the dataset. With a conventional loss function, the AUROC result did not align with the model's actual loss under the same conditions (Figure 3a). By contrast, the weighted loss model showed higher evaluation metrics than the conventional loss model (Figure 3). Specifically, training with the weighted loss function led to a 31.9% improvement in AUROC (from 0.69 to 0.91), as well as increases of 2.3% in accuracy and 4.6% in F1 score relative to the unweighted model. Notably, there was no significant difference in performance between the two loss functions on the training dataset (Figure 3b).

Next, we analyzed the contribution of different types of text used in the model. Models trained with only pre-diagnosis notes showed a slightly lower AUROC score than models incorporating both pre- and post-diagnosis notes, though the difference was minimal (Figure 3a). Furthermore, the early stopping points for models using only pre-diagnosis notes occurred sooner than for models using both pre- and post-diagnosis notes, with AUROC levels remaining comparable in subsequent epochs. Although evaluation performance on the test dataset improved with each epoch, the model trained solely on post-diagnosis notes did not show a sustained performance improvement over training iterations on the training dataset (Figure 3). In other words, post-diagnosis notes did not provide additional information in classifying LOC or head trauma compared to the pre-diagnosis notes.



(a)

(b)

Figure 3. Evaluation of the performance of LOC classification among NLP-based models using different clinical diagnosis notes. (a) Model performance evaluated in the test dataset. The left panel showed a difference in classification performance according to types of clinical notes. The performance improvement for unbalanced data using the weighted loss function is shown in the right panel. The results of the analysis performed on the training dataset are displayed in (b). Points in the scatter plot indicate last updated model performance according to the early stop method.

3.3. Quantitative Interpretation of LOC Classification Model

To investigate how word-level information in the text contributes to the classification model, we used the LIME package, which provides insights into the keywords that are most influential for the prediction model. Figure 4 illustrates the impact of the vocabulary used in the data on the predicted group classifications. In the C2 (head injuries but no LOC) group, vocabulary related to head injury is particularly significant (Figure 4A). In contrast, the C3 (head injuries with LOC) group primarily focuses on words directly associated with LOC rather than head injury, resulting in relatively straightforward decisions compared to the C2 group (Figure 4B). The results of LIME on the original version of Korean are shown in Figure S2.

Interestingly, while both the C2 and C3 groups share words related to LOC, they are classified as C2 when the context includes a denial of LOC (Figure 4). These findings suggest that the model is capable of making contextual judgments regarding head injuries and levels of consciousness.



Figure 4. Interpretation of LOC classification and the importance of each word feature (English version). The figures indicate the importance of each word in the LOC classification for patients

classified as C2 (A) and C3 (B). The right panel displays descriptive input consisting of pre- and post-diagnosis notes. The left panel of each figure shows the weight of features for visualization.

4. Discussion

4.1. Principal Results

In this study, we evaluated the performance of an NLP based RoBERTa model to automatically identify a patient's LOC within the EMR of the ED. The model effectively determined LOC status by analyzing both pre- and post-diagnosis notes, leveraging specific words and contextual cues in these free-text medical records. Our comparative analysis highlighted the importance of pre-diagnosis notes, which provide valuable information for assessing LOC. These results demonstrate the model's capability to accurately identify injury-related variables, specifically LOC, within unstructured ED documentation.

We employed a weighted loss function to address the imbalance in our data, as only 3.4% of patients were classified as experiencing LOC, resulting in a significant imbalance between this minority class (patients with LOC) and the majority class (patients without LOC). This underrepresentation can bias the model toward the majority class, potentially compromising its accuracy in identifying LOC cases. Our results indicated that the weighted loss model was more effective in correctly identifying positive LOC cases, while maintaining low rates of false positives and false negatives compared to the conventional (non-weighted) model. Interestingly, while training loss values were similar across both approaches, validation data revealed a marked improvement with the weighted model. These findings align with previous research suggesting that weighted loss functions can effectively address challenges associated with imbalanced medical data, which is often due to the prevalence of certain conditions or patient characteristics [23].

The model also demonstrated a nuanced understanding of contextual language, as revealed through LIME analysis, by effectively distinguishing between straightforward LOC terms (as in group C3) and those with context-denying nuances (as in group C2). This sensitivity to the complex language often found in ED records is critical for accurately identifying injury-related conditions, capturing subtle variations in LOC expressions, and making the model particularly well-suited for ED documentation. Previous study has shown that LIME-based models enhance interpretability by highlighting word-level contributions, helping clinicians understand how specific terms influence predictions [24]. For example, terms like "fragment" and "hemorrhage" have been identified in fracture prediction, aligning with physicians' expertise and reinforcing the model's credibility. This transparency fosters trust and supports clinical decision-making by clearly illustrating the model's reasoning process.

The choice of data source—pre-diagnosis, post-diagnosis, or combined notes—plays a critical role in the development of efficient and accurate NLP-based classification models. Pre-diagnosis notes, which are recorded upon a patient's arrival at the emergency department, contain a high proportion of patient-reported symptoms and triage assessments. These notes are available earlier in the clinical workflow, making them particularly valuable for rapid classification. Our findings demonstrate that models trained solely on pre-diagnosis notes achieved comparable performance to those trained on combined notes. Notably, significant differences in model performance were observed when using only pre-diagnosis notes or post-diagnosis notes, as evaluated through five-fold cross-validation (Student's *t*-test, $p < 1.16 \times 10^{-119}$) (Table S3). This suggests that pre-diagnosis data alone can effectively support the rapid identification of LOC information, especially in time-sensitive scenarios where immediate clinical decisions are critical. In addition, utilizing pre-diagnosis data exclusively can offer a practical advantage by reducing data processing requirements and computational complexity.

NLP applications provide significant advantages in healthcare by automating the processing of vast amounts of clinical text. While rule-based algorithms were previously the most common, recent trends reveal a marked increase in studies using NLP to retrieve and extract medical concepts and terminology [25]. This shift reflects the growing reliance

on NLP in healthcare as EMRs and other health information systems expand, resulting in increased volumes of textual data that necessitate advanced methods for efficient data retrieval. NLP addresses this need by reducing or eliminating the need for manual narrative reviews, enabling the quick and accurate assessment of large datasets.

Additionally, NLP is instrumental in extracting critical factors from EMRs. By applying these techniques to LOC identification, this model fills critical gaps in EMR documentation that ED physicians may overlook. A systematic review has also demonstrated NLP's effectiveness in identifying social determinants of health from EMRs, factors such as smoking status, substance use, and homelessness, which are essential for comprehensive patient assessments and health risk evaluations [26].

NLP models offer substantial benefits for healthcare providers, patients, and caregivers. In healthcare AI applications, NLP is primarily used for information extraction, either as part of a larger input set or to generate training target values for downstream AI models [27]. By integrating LOC information from patients' or caregivers' narrative accounts and other key characteristics related to traumatic brain injury, this model enhances clinical decision support systems. Such systems can provide recommendations for ED visits after a TBI, enabling timely intervention for patients at risk of post-concussion complications. Additionally, the model can track patient recovery by analyzing follow-up records, aiding in the monitoring of symptom resolution or progression over time, which supports continuous care and informed clinical management.

In this study, we used a bilingual (English and Korean) dataset in which all text was translated into English prior to the classification task. This approach was chosen to leverage pre-trained language models like RoBERTa, which are optimized for English text processing. Our model effectively captures contextual information, achieving a high AUROC score of 0.91. However, it is important to consider the potential impact of using monolingual datasets directly, particularly in scenarios where translation may result in a loss of nuanced clinical details specific to the source language. Therefore, the integration of multilingual and monolingual approaches presents a promising avenue for future research. Comparing performance metrics across translation-based and native language models will help determine the optimal approach for specific clinical contexts.

4.2. Limitations

This study has some limitations. First, as the results show, the low contribution of post-diagnosis notes to model performance may be due to data bias from the pre-diagnosis note collection system, which varies by hospital or location. In order to be effective, pre-diagnostic notes must contain words and contexts that are highly associated with head injury. Therefore, based on the results displayed in Figure 3b, the large fluctuations in the performance metrics of the model with post-diagnosis notes may be due to the presence of more diverse information compared to the pre-diagnosis note model. Second, in this study, a word-by-word exclusion technique was used to determine the importance of words. This is due to the lack of explanatory power of the model considering the context, so the evaluation results of the trained model could not be interpreted in context. The issue can be resolved by utilizing the attention map of the transformer model. Further study based on multi-center data will help develop a generalized and explainable AI-based model for the classification of LOC patients, which will provide insights into the development of rapid and reliable analytical models for clinical narratives and directly contribute to improving patient outcomes. Lastly, differences in writing style and terminology among ED clinicians may influence the model's performance. However, we did not capture this stylistic and terminological information in our study. Further study should be conducted incorporating such variables to evaluate and mitigate potential biases stemming from inter-clinician variability in EMR documentation.

5. Conclusions

This study utilized an NLP-based RoBERTa model to classify patients with LOC from ED text records, achieving high accuracy. By applying a weighted loss function, we effectively mitigated the impact of data imbalance, a common issue in medical data. LIME analysis demonstrated the model's ability to capture nuanced language. Our findings suggest that NLP is a robust approach for classifying LOC and other imbalanced medical data. Future work will focus on enhancing contextual understanding through explainable deep learning strategies, further improving model accuracy and transparency in clinical settings.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app142311399/s1>, Table S1: Distribution of patients across the entire dataset based on LOC status; Table S2: Training parameters of the model; Figure S1: Evaluation of model performance with cross-validation; Figure S2: Interpretation of LOC classification and the importance of each word feature (original Korean version); Table S3. Classification performance according to input data evaluated using 5-fold cross-validation.

Author Contributions: Conceptualization, J.O.P., C.K., H.A.P. and S.-H.S.; Data curation, J.O.P., H.A.P., I.J. and S.-H.S.; Formal analysis, I.J., S.-H.S. and S.Y.S.; Funding acquisition, J.J.L. and C.K.; Investigation, H.A.P. and J.O.P.; Methodology, H.A.P., I.J., S.-H.S., C.K. and J.O.P.; Software, I.J.; Validation, I.J.; Visualization, S.-H.S.; Writing—original draft, H.A.P., S.-H.S., J.O.P. and C.K.; Writing—review and editing: H.A.P., S.-H.S., I.J., J.O.P., C.K. and J.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR21C0198, HI22C1498); by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A5A8019303); and supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-005), and Hallym University Research Fund (HURF).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Hallym University Dongtan Sacred Heart Hospital Institutional Review Board/Ethics Committee (IRB No. 2023-08-013, 26 September 2023).

Informed Consent Statement: Informed consent was waived by the institutional review boards because this study used only de-identified image data which were retrospectively collected.

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Polnaszek, B.; Gilmore-Bykovskiy, A.; Hovanes, M.; Roiland, R.; Ferguson, P.; Brown, R.; Kind, A.J. Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. *Med. Care* **2016**, *54*, e65–e72.
2. Chan, K.S.; Fowles, J.B.; Weiner, J.P. Electronic health records and the reliability and validity of quality measures: A review of the literature. *Med. Care Res. Rev.* **2010**, *67*, 503–527.
3. Bian, J.; Lyu, T.; Loiacono, A.; Viramontes, T.M.; Lipori, G.; Guo, Y.; Wu, Y.; Prosperi, M.; George Jr, T.J.; Harle, C.A. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1999–2010.
4. Cho, H.; Yoo, S.; Kim, B.; Jang, S.; Sunwoo, L.; Kim, S.; Lee, D.; Kim, S.; Nam, S.; Chung, J.H. Extracting lung cancer staging descriptors from pathology reports: A generative language model approach. *J. Biomed. Inform.* **2024**, *157*, 104720.
5. Hossain, E.; Rana, R.; Higgins, N.; Soar, J.; Barua, P.D.; Pisani, A.R.; Turner, K. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Comput. Biol. Med.* **2023**, *155*, 106649.
6. Tignanelli, C.J.; Silverman, G.M.; Lindemann, E.A.; Trembley, A.L.; Gipson, J.C.; Beilman, G.; Lyng, J.W.; Finzel, R.; McEwan, R.; Knoll, B.C. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. *J. Trauma Acute Care Surg.* **2020**, *88*, 607–614.
7. Kulshrestha, S.; Dligach, D.; Joyce, C.; Baker, M.S.; Gonzalez, R.; O'Rourke, A.P.; Glazer, J.M.; Stey, A.; Kruser, J.M.; Churpek, M.M. Prediction of severe chest injury using natural language processing from the electronic health record. *Injury* **2021**, *52*, 205–212.

8. Edgcomb, J.B.; Zima, B. Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. *Psychiatr. Serv.* **2019**, *70*, 346–349.
9. Roy, D.; Peters, M.E.; Everett, A.D.; Leoutsakos, J.-M.S.; Yan, H.; Rao, V.; Bechtold, K.T.; Sair, H.I.; Van Meter, T.; Falk, H. Loss of consciousness and altered mental state as predictors of functional recovery within 6 months following mild traumatic brain injury. *J. Neuropsychiatry Clin. Neurosci.* **2020**, *32*, 132–138.
10. Waseem, M.; Iyahan, P., Jr.; Anderson, H.B.; Kapoor, R.; Kapoor, R.; Leber, M. Isolated LOC in head trauma associated with significant injury on brain CT scan. *Int. J. Emerg. Med.* **2017**, *10*, 30.
11. Maas, A.I.; Harrison-Felix, C.L.; Menon, D.; Adelson, P.D.; Balkin, T.; Bullock, R.; Engel, D.C.; Gordon, W.; Langlois-Orman, J.; Lew, H.L.; et al. Standardizing data collection in traumatic brain injury. *J. Neurotrauma* **2011**, *28*, 177–187.
12. Torres-Silva, E.A.; Rúam, S.; Giraldo-Forero, A.F.; Durango, M.C.; Flórez-Arango, J.F.; Orozco-Duque, A. Classification of Severe Maternal Morbidity from Electronic Health Records Written in Spanish Using Natural Language Processing. *Appl. Sci.* **2023**, *13*, 10725.
13. Qiu, P.; Wu, C.; Zhang, X.; Lin, W.; Wang, H.; Zhang, Y.; Wang, Y.; Xie, W. Towards building multilingual language model for medicine. *Nat. Commun.* **2024**, *1*, 8384.
14. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
15. Delobelle, P.; Winters, T.; Berendt, B. RobBERT: A Dutch RoBERTa-based language model. *arXiv* **2020**, *1*, 3255–3265.
16. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T. Klue: Korean language understanding evaluation. *arXiv* **2021**, arXiv:2105.09680.
17. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
18. Ho, Y.; Wookey, S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* **2019**, *8*, 4806–4813.
19. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *1*, 12.
20. Hand, D.J.; Till, R.J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **2001**, *45*, 171–186.
21. Obuchowski, N.A.; Bullen, J.A. Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* **2018**, *63*, 07TR01.
22. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
23. Chamseddine, E.; Mansouri, N.; Soui, M.; Abed, M. Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss. *Appl. Soft Comput.* **2022**, *129*, 109588.
24. Ling, T.; Jake, L.; Adams, J.; Osinski, K.; Liu, X.; Friedland, D. Interpretable machine learning text classification for clinical computed tomography reports—a case study of temporal bone fracture. *Comput. Meth. Programs Biomed. Update* **2023**, *3*, 100104.
25. Gholipour, M.; Khajouei, R.; Amiri, P.; Hajesmaeel Gohari, S.; Ahmadian, L. Extracting cancer concepts from clinical notes using natural language processing: A systematic review. *BMC Bioinform.* **2023**, *24*, 405.
26. Patra, B.G.; Sharma, M.M.; Vekaria, V.; Adekanattu, P.; Patterson, O.V.; Glicksberg, B.; Lepow, L.A.; Ryu, E.; Biernacka, J.M.; Furmanchuk, A.o. Extracting social determinants of health from electronic health records using natural language processing: A systematic review. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2716–2727.
27. Wen, A.; Fu, S.; Moon, S.; El Wazir, M.; Rosenbaum, A.; Kaggal, V.C.; Liu, S.; Sohn, S.; Liu, H.; Fan, J. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digit. Med.* **2019**, *2*, 130.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.