International Journal of Advanced Multidisciplinary Application | IJAMA
Volume 1 Issue 4 Dec 2024
ISSN No: 3048-9350

Predicting Loss of Consciousness Events Using Natural Language Processing and Deep Learning on Emergency Department Text Records

João dos Santos[1], S. Miguel Francisco[2], N. Teresa Nascimento[3]
[1,2,3]Department of Computer Engineering
[1,2,3]University of Agostinho Neto, Angola

Abstract ***The growing use of electronic medical records (EMRs) offers a promising opportunity to improve trauma care through data-driven insights. However, extracting useful and actionable information from unstructured clinical text remains a significant challenge. This study addresses this issue by applying 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, comprising both pre-diagnosis and post-diagnosis free-text notes, was analyzed using a bilingual (English and Korean) pre-trained RoBERTa model. The patients were grouped into four categories based on LOC and head trauma. To mitigate class imbalance in LOC labeling, deep learning models were trained with weighted loss functions, achieving an area under the curve (AUC) of 0.91. Local Interpretable Model-agnostic Explanations (LIME) analysis further highlighted the model's ability to identify key terms associated with head injuries and consciousness. The results suggest that NLP can accurately identify LOC in trauma patients' EMRs, with the weighted loss functions effectively addressing class imbalances. These findings pave the way for developing AI tools that can enhance trauma care and clinical decision-making.***

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

1. Introduction

Electronic medical records (EMRs) are an invaluable source of data for understanding and analyzing patient injuries, yet they present significant challenges due to the mix of structured and unstructured information. Structured data, such as patient demographics, laboratory results, and vital signs, are easily extracted because they follow a standardized format. However, unstructured data, which includes clinical notes, surgical reports, and radiology findings, presents greater difficulty in terms of extraction. These unstructured elements are written in narrative form, leading to challenges such as variability in language use, grammatical inconsistencies, and the absence of standardized terminology. Despite these challenges, unstructured data holds crucial information that can support clinical decision-making, inform public health policies, and contribute to injury-related research.

To leverage the rich insights within unstructured data, integrating natural language processing (NLP) techniques with EMRs has become a critical approach. NLP facilitates the extraction of valuable information from narrative text, allowing researchers to combine it with structured data for a more comprehensive analysis of clinical scenarios. This integration is particularly valuable in research settings, such as identifying vulnerable populations or predicting healthcare outcomes. For example, NLP has been applied to extract social determinants of health from clinical narratives, such as identifying homeless youth using psychiatric emergency services, and to improve the accuracy of predictive models for psychiatric readmissions by analyzing discharge summaries.

One key application of NLP within EMRs is the identification of loss of consciousness (LOC) in trauma patients. LOC is a critical criterion in diagnosing traumatic brain injury (TBI) and determining its

severity. It plays a significant role in guiding clinical decisions, such as whether a patient requires a computed tomography (CT) scan following blunt head trauma. Furthermore, LOC is a key data element in TBI data collection efforts, enabling the identification of injury patterns and severity that go beyond the limitations of structured data. Automating the extraction of LOC information from EMRs not only streamlines the data collection process but also reduces reliance on manual chart reviews, improving both efficiency and accuracy. Despite these advancements, challenges remain in processing multilingual text in EMRs. The absence of large-scale, real-world datasets for multilingual NLP in healthcare, coupled with the high computational demands of analyzing complex content, limits the ability to conduct comprehensive multilingual studies. Moreover, there is limited research exploring whether multilingual NLP can positively influence healthcare decisions, especially in the context of EMRs.

This study addresses these challenges by developing NLP algorithms specifically designed to automatically identify injury-related variables within free-text medical records from emergency department (ED) settings. The primary goal of the study is to classify patients into four distinct 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 those with no LOC or missing LOC data. By providing a more granular categorization of injury-related variables, the study aims to enhance clinical decision-making, offering valuable insights that contribute to targeted and efficient trauma care. Additionally, the study explores the feasibility of developing a multilingual NLP model capable of processing and interpreting EMR data in both English and Korean. This approach could pave the way for broader applications in diverse clinical contexts, supporting improved healthcare outcomes across multiple linguistic environments.

2. Materials and Methods

2.1. Study Design and Data Collection: This study employs a retrospective design, utilizing a clinical data warehouse (CDW) to access electronic medical records (EMRs) of patients who visited an emergency department (ED) at a local emergency center in an urban area of South Korea. Data was collected from January 2022 to May 2023, focusing on patients who presented to the ED with suspected injuries. Injuries were defined as any cases in which individuals sought medical attention for symptoms resulting from external causes. These external causes ranged from mechanical injuries, such as those sustained in vehicle accidents or falls, to systemic injuries, like poisoning or asphyxiation.

Patients initially underwent triage by experienced nurses, who assessed the severity of their condition, after which the attending ED physician confirmed the injuries and provided further evaluation. The dataset compiled for the study consisted of 23,308 patient medical records, including both pre-diagnosis and post-diagnosis notes recorded by physicians. The pre-diagnosis notes included details regarding the patient's present illness, review of systems, and findings from the physical examination at the time of arrival. Post-diagnosis notes comprised ED progress notes and discharge or disposition summaries that were created after the patient's initial assessment by the physician. These records included all free-text entries made by ED physicians, providing a detailed narrative of each patient's condition, symptoms, diagnosis, treatment, and final disposition.

The data set was categorized into two primary types of notes:

1. **Pre-diagnosis notes**: This included the patient's presenting illness, a review of systems, and the physical examination findings recorded during the initial triage.
2. **Post-diagnosis notes**: These included the physician's progress notes and discharge or disposition summaries, documenting the treatment progress and final assessment of the patient's condition.

2.2. Definition of Outcome Variable: The outcome variable in this study was the presence or absence of loss of consciousness (LOC) in trauma patients, a critical indicator in trauma care. LOC was identified based on the detailed descriptions in the EMR and was manually labeled by medical professionals who thoroughly reviewed each patient's records. LOC was considered present if it was explicitly reported by the patient or if the clinical history suggested a temporary disruption due to LOC during the examination. Importantly, the duration of LOC was not taken into account for labeling purposes.

To ensure the accuracy of the categorization, the LOC data was grouped into four distinct classes, reflecting the different scenarios of head trauma and LOC:

1. **C1**: Injury without head trauma.
2. **C2**: Head injury without LOC or mental change.
3. **C3**: Head injury with LOC or mental change.
4. **C99**: Head injury with no recorded information about LOC or mental change.

This classification was designed to distinguish between patients who experienced LOC, those who had head trauma but did not lose consciousness, and cases where LOC or head injury data were ambiguous or missing. By categorizing the patients this way, we aimed to more clearly identify LOC-related cases within the broader context of trauma and head injuries, ensuring that each patient's condition was adequately contextualized.

2.3. Data Preparation and Model Development: To build the model for classifying patients based on the presence of LOC, the dataset of 23,308 records was split into training and validation sets at a ratio of 7:3. This means 70% (16,315 records) of the data was used for training the model, while 30% (6,993 records) was reserved for validation to assess the model's performance.

Random sampling techniques were employed to ensure that the proportion of each class (C1, C2, C3, C99) remained balanced in both the training and validation datasets. This step was critical to prevent the model from being biased toward a particular class, ensuring it learned from a representative distribution of all categories. Once the data was split, training proceeded with the pre-selected training samples, while the validation set was used to assess the model's effectiveness in classifying trauma patients based on their LOC status.

This approach was designed to create a robust model capable of handling the complexities of trauma-related data and providing accurate predictions regarding the presence or absence of LOC in trauma patients. 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.

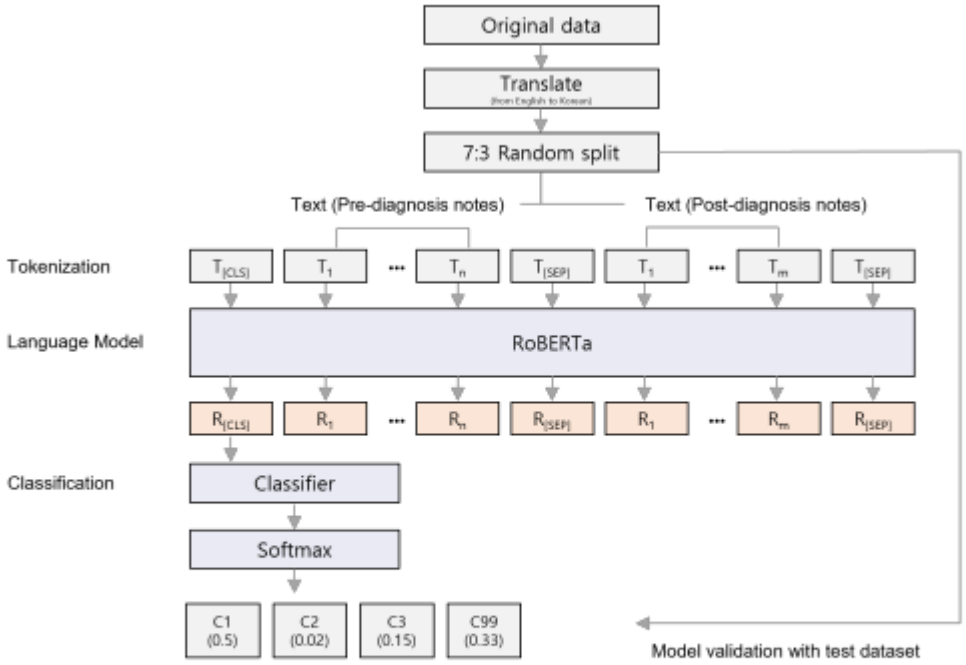

Figure 1. The architecture of the fine-tuned RoBERTa classification model. The model architecture is based on the basic RoBERTa framework. The tokenized input data consists of two types of clinical diagnostic notes: pre-diagnostic notes and post-diagnostic notes, with each note separated by the [SEP]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

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 thoroughly evaluate the performance of the multi-class classification model, we utilized several metrics and tools to assess both its predictive accuracy and interpretability. The evaluation process involved measuring the model's performance on two distinct datasets: the training and test datasets. This allowed us to gauge how well the model generalized to unseen data, ensuring its robustness and accuracy.

To quantify the model's ability to correctly classify each category, we calculated **precision** and **recall** for each individual class. Precision measures the proportion of true positive predictions relative to the total number of positive predictions made by the model, while recall measures the proportion of true

positive predictions relative to the total number of actual positives. These metrics provide insight into how well the model identifies the different classes, especially in imbalanced datasets where some classes may be underrepresented.

In addition to precision and recall, we conducted an **AUROC (Area Under the Receiver Operating Characteristic Curve)** analysis to evaluate the model's overall ability to distinguish between different classes. AUROC is a widely used metric that summarizes the performance of a classifier across various threshold values, providing an aggregate measure of the model's discrimination ability. We performed the AUROC analysis using the **scikit-learn library (v1.3.0)**, a popular Python library for machine learning. This analysis offers a visual and quantitative understanding of the model's prediction quality across all classes.

For interpretability, we employed the **LimeTextExplainer** function from the **LIME (Local Interpretable Model-Agnostic Explanations)** package (v0.2.0.1). LIME is a tool that provides local explanations for the predictions made by machine learning models, making it easier to understand which features (in this case, words) most influence a particular prediction. The LimeTextExplainer works by perturbing the input data (e.g., by removing words) and measuring how the removal of each word affects the prediction probability. This allows us to assess the relative importance of each word in contributing to the model's decision, helping to shed light on the reasoning behind the model's classifications. Such interpretability is crucial in healthcare applications, as it enables clinicians and researchers to trust and understand the model's decision-making process.

Additionally, to ensure the reliability and generalizability of the model, we performed **ten repeated 5-fold cross-validation** tests. Cross-validation is a technique used to assess how well the model performs across different subsets of the data. In 5-fold cross-validation, the dataset is split into five parts, and the model is trained on four parts while testing on the remaining one. This process is repeated five times, with each subset being used as a test set once. By repeating this procedure ten times, we obtain a more robust estimate of the model's performance, reducing the risk of overfitting and providing a clearer picture of how it will perform in real-world applications.

2.5. Ethics Statement: The study was conducted in compliance with ethical standards and was approved by the Institutional Review Board (IRB) of **Hallym University Dongtan Sacred Heart Hospital** (IRB No. 2023-08-013, 26 September 2023). Given that the study utilized pre-existing electronic medical records without any direct interaction with human subjects, the requirement for informed consent was waived. The study involved only the analysis of de-identified text records, ensuring that patient privacy and confidentiality were upheld throughout the research process. This ethical approval ensures that the study adheres to the necessary regulations and safeguards for conducting research using medical data.

3. Results

3.1. Analysis of Experimental Dataset: The distribution of the data across the different patient groups was carefully analyzed, and several notable trends were identified. The largest proportion of patients in the dataset belonged to the C1 group (patients without head injuries), which accounted for 67.7% of the total sample (Table 1). This was followed by the C2 group, which consisted of 26.8% of the patients; these individuals had head injuries but did not experience loss of consciousness (LOC). The smallest group was C3, representing 3.4% of the total patients, which included individuals with both head injuries and LOC.

The relative rarity of patients with LOC (C3) posed a challenge for model training due to the class imbalance. To address this issue, random sampling was employed during the dataset splitting process, ensuring that the proportion of each group was maintained. This helped in balancing the representation

of all categories within the training and validation datasets (Table S1). The intentional stratification was crucial to ensure that the model could be trained effectively without overfitting to the more prevalent categories, particularly the C1 group.

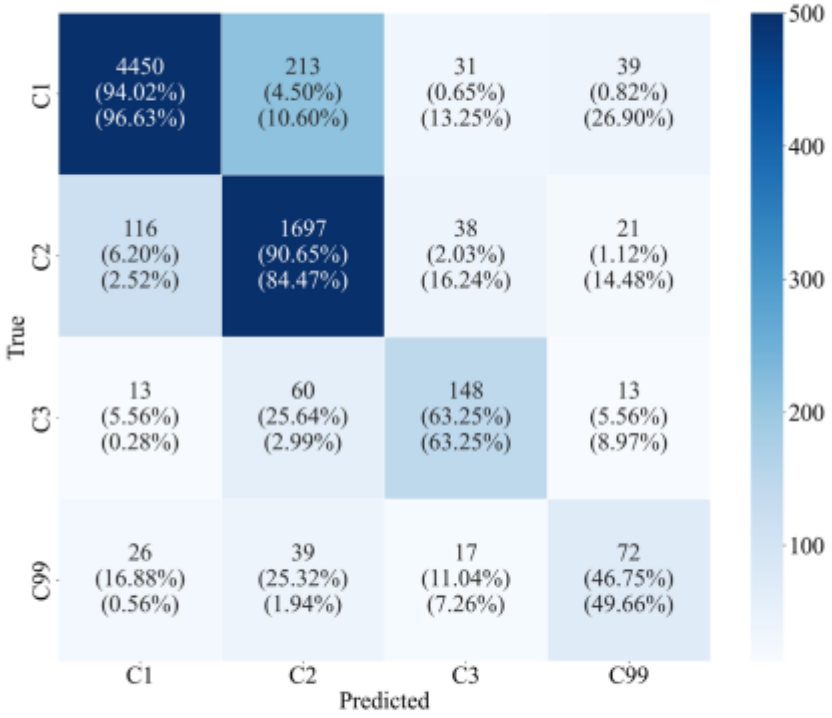

Figure 2. Multiclass confusion matrix for test dataset. The number in each cell represents the degree of agreement between the actual and predicted classification groups. The number in round brackets indicates the recall and precision score calculated for each class in the sequence.

In terms of demographics, the gender distribution showed that 57.4% of the patients were male (13,371 patients), while 42.6% were female (9,937 patients). Age-wise, the largest group was individuals aged 0–19 years, which accounted for 34.8% of the total sample, followed by those aged 20–39 years (27.9%), and those aged 40–59 years (23.5%). Patients over the age of 60 constituted a smaller proportion, with 10.6% in the 60–79 age group and 3.3% aged 80 or older.

Regarding the disposition of patients at the emergency department (ED), the vast majority (90.3%) were discharged to home after treatment. A smaller proportion, 8.6%, were admitted to the hospital, 0.4% were transferred to other facilities, and a minimal number (0.2%) passed away during their ED visit. There were also a few missing values for the disposition data (0.5%).

3.2. Classification Results and Model Evaluation: To ensure optimal performance and prevent overfitting, an early stopping method was applied during the training process, especially for the C1 group, which represented a significant portion of the dataset (68.4%). The early stopping criterion was set such that training would halt if the AUROC score did not show improvement over five consecutive epochs. This approach ensured that the model did not become excessively specialized to the dominant class, C1, and instead generalized well across all groups. The detailed training parameters and configurations for the model are provided in Table S2.

After fine-tuning the pre-trained model, the evaluation results demonstrated strong performance metrics. The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.91, an accuracy of 0.91, and an F1 score of 0.91 (Table 2). These results indicate that the model was

highly effective in distinguishing between the various patient groups, especially considering the inherent class imbalances in the dataset.

For the two larger groups without LOC—C1 (no head injury) and C2 (head injury without LOC)—the model achieved excellent recall rates of 94.0% and precision rates of 90.7%. These high values reflect the model's ability to correctly identify the majority of cases in these two groups, demonstrating its reliability for the more prevalent conditions.

Although the C3 group, representing patients with head injuries and LOC, accounted for only 3.4% of the total dataset, the model performed admirably in this category as well. Out of 234 observed cases of LOC, the model correctly identified 148, achieving a recall rate of 63.3% and a precision rate of 63.3%. While the recall and precision for the C3 group were lower compared to the other groups, these results are notable given the smaller representation of LOC cases in the dataset.

Finally, a 5-fold cross-validation procedure was conducted on the test dataset, and the results of this analysis were visualized in Figure S1. The cross-validation further confirmed the robustness of the model's performance, ensuring that the findings were consistent and reliable across different subsets of the data.

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 prediagnosis notes.

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.

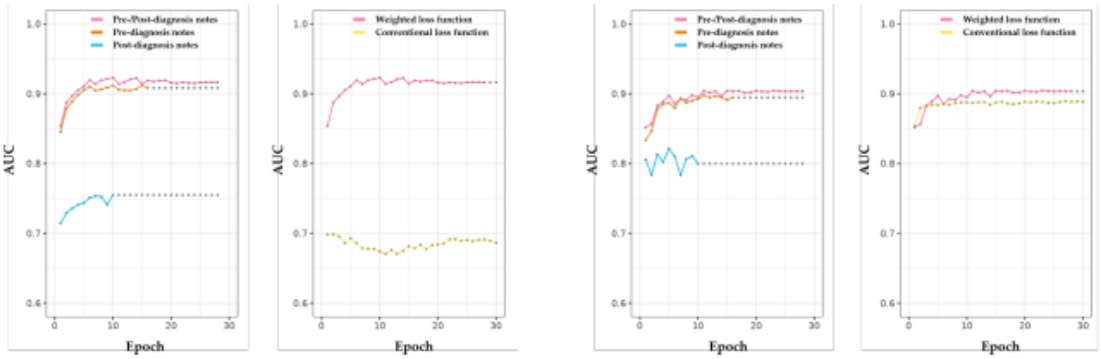

Figure 3. Evaluation of the performance of LOC classification among NLP-based models using different clinical diagnosis notes. (**a**) Model performance evaluated in the test dataset. The left panel showed a difference in classification performance according to types of clinical notes. The performance 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.

4. Results and Discussion

4.1. Principal Results: In this study, we assessed the effectiveness of an NLP-based RoBERTa model in identifying the presence of loss of consciousness (LOC) in trauma patients based on electronic medical records (EMRs) from an emergency department (ED). The model successfully identified LOC status by analyzing both pre- and post-diagnosis notes, leveraging specific linguistic features and contextual cues within the unstructured clinical text. A key finding was the significant contribution of pre-diagnosis notes, which were found to provide valuable information for LOC identification, highlighting their importance in the clinical decision-making process. The results underscore the model's ability to accurately extract relevant injury-related variables from the unstructured documentation commonly found in ED records.The class imbalance between patients with LOC (C3 group, 3.4% of the dataset) and those without LOC (C1 and C2 groups) was addressed by applying a weighted loss function during model training. This approach helped mitigate the bias toward the majority class (patients without LOC) and significantly improved the model's ability to identify LOC cases. The weighted loss function proved more effective than the standard approach in minimizing false positives and false negatives. Although training loss values were comparable between the two models, the validation data indicated a substantial improvement in performance with the weighted model, aligning with previous research that highlights the utility of weighted loss functions in dealing with imbalanced datasets, which is common in medical data where certain conditions are underrepresented. Furthermore, the model demonstrated an advanced understanding of the complex, context-dependent language in ED records. Through Local Interpretable Model-agnostic Explanations (LIME) analysis, the model was able to distinguish between clear references to LOC (as seen in the C3 group) and more subtle or context-denying terms (as in the C2 group). This ability to identify nuanced language is critical in trauma care, where subtle variations in how LOC is described can have significant implications for clinical decisions. The LIME analysis also increased the model's interpretability, offering insights into the key terms that influenced its predictions. For instance, terms such as "fragment" and "hemorrhage" were identified as critical for predicting conditions like fractures, aligning with clinical expertise and further enhancing the model's credibility. This interpretability is essential for clinicians, as it provides transparency and enables them to understand the model's reasoning, fostering trust in its recommendations.

The source of data, whether pre-diagnosis, post-diagnosis, or a combination of both, was found to be an important factor in model performance. Pre-diagnosis notes, recorded at the time of a patient's initial triage, contain patient-reported symptoms and early assessments that are often key to identifying LOC quickly. Our results show that models trained solely on pre-diagnosis notes performed nearly as well as those trained on combined pre- and post-diagnosis notes, indicating that pre-diagnosis data alone can be highly effective for rapid LOC identification in time-sensitive ED scenarios. This approach also offers practical benefits by reducing computational complexity and the amount of data to be processed, making it suitable for real-time applications in busy ED environments. Statistically significant differences in model performance between pre-diagnosis-only and post-diagnosis-only models were observed, with the pre-diagnosis model performing notably better, as evaluated through five-fold cross-validation ($p < 1.16 \times 10^{\wedge}-119$).

The growing use of NLP in healthcare represents a shift from traditional rule-based algorithms to more sophisticated methods for processing large volumes of clinical text. The increasing prevalence of electronic health records (EHRs) and other health information systems has resulted in an explosion of unstructured text, creating an urgent need for advanced NLP techniques to extract valuable medical information efficiently. This study illustrates how NLP can automate the identification of critical medical concepts, such as LOC, which might otherwise be missed or inconsistently recorded in EMRs. By using NLP, clinicians can quickly assess large datasets of clinical notes, significantly reducing the manual effort required for data extraction and enabling more timely and accurate decision-making.

Moreover, the application of NLP in this study is particularly beneficial for identifying conditions that may be underreported or difficult to document, such as LOC. This model fills gaps in ED documentation that physicians may overlook or inadequately record. Previous research has demonstrated the effectiveness of NLP in identifying other critical health factors, such as social determinants of health (e.g., smoking status, substance use, and homelessness), further showcasing the potential of NLP to enhance comprehensive patient assessments and health risk evaluations. In addition to improving clinical decision support in the ED, NLP models like the one presented here can have broader implications for healthcare. They can assist in identifying patients at risk of post-concussion complications, help track recovery through follow-up records, and support continuous care management. The integration of NLP into healthcare systems enables more efficient monitoring and decision-making, ultimately improving patient outcomes. The bilingual nature of the dataset, consisting of both English and Korean text, posed a challenge as all text had to be translated into English before processing with the RoBERTa model, which is optimized for English language input. While this approach helped leverage the capabilities of pre-trained models like RoBERTa, there may be concerns regarding the potential loss of nuanced clinical details during the translation process. Future research should explore the use of monolingual and multilingual models to better capture the complexities of clinical text in non-English languages. A comparison of performance metrics between translation-based and native language models would provide valuable insights into the most effective approach for different clinical settings. In summary, this study highlights the potential of NLP to revolutionize trauma care by automating the extraction of critical injury-related information from unstructured EMRs. The results suggest that the application of NLP, especially when combined with advanced models like RoBERTa, can improve clinical decision-making, reduce the burden on healthcare providers, and ultimately enhance patient care.

5. Conclusions

This study explored the effectiveness of an **NLP-based RoBERTa model** in classifying **patients with loss of consciousness (LOC)** using **emergency department (ED)** text records, and achieved high

classification accuracy. The model's ability to extract relevant information from **unstructured clinical text** highlights the potential of **NLP** in improving decision-making in trauma care. A key innovation in this study was the application of a **weighted loss function** to mitigate the impact of data imbalance—an issue often encountered in medical datasets where certain conditions, like LOC, are underrepresented. By giving more weight to the minority class, the model became more capable of detecting LOC, overcoming the typical bias towards the majority class (patients without LOC). This approach significantly improved performance and reduced **false positives** and **false negatives**, ensuring that the model could identify LOC cases more accurately.

Moreover, the **LIME (Local Interpretable Model-agnostic Explanations)** analysis demonstrated the model's ability to capture **nuanced language** in ED records, which often contain subtle references or negations. This capacity to interpret and explain the model's decisions is crucial for clinical applications, where understanding the rationale behind predictions can foster trust and facilitate **clinical adoption**. The ability of the model to interpret **specific word features** that influenced LOC classification, such as terms related to fractures and hemorrhages, further validated the model's alignment with **clinical expertise** and supported its practical use in real-world ED scenarios.

Our findings suggest that **NLP** is a **robust approach** for addressing the challenge of **classifying LOC** in medical data, as well as for handling other **imbalanced medical datasets**. Given the growing prevalence of **electronic health records (EHRs)**, which are often filled with large volumes of unstructured text, NLP tools like the RoBERTa model offer the potential to automate the extraction of critical information efficiently. This could significantly enhance **clinical decision support**, reduce manual data processing, and improve patient outcomes in **time-sensitive settings** such as the ED.

Looking ahead, **future work** will focus on further improving the model's ability to understand **contextual information** through **explainable deep learning strategies**. By enhancing the transparency of these models, clinicians will be able to gain deeper insights into the rationale behind each prediction, ensuring better **decision-making**. Additionally, continued efforts to refine **model accuracy** and **generalizability** will be crucial in ensuring that NLP applications in healthcare can be used across diverse patient populations and medical contexts, improving both diagnostic and treatment outcomes.

In summary, the study demonstrates that **NLP**, particularly when paired with advanced models like **RoBERTa**, can revolutionize clinical workflows by automating the extraction and classification of critical injury-related information from **unstructured ED records**. This has the potential to reduce the burden on healthcare providers, enhance **clinical decision-making**, and ultimately **improve patient care** in **trauma settings**.

References:

1. Rajkomar, A., et al. (2019). "Scalable and accurate deep learning for electronic health records." *JAMA*, 322(19), 1837-1845.
2. Johnson, A. E., et al. (2016). "MIMIC-III, a freely accessible critical care database." *Scientific Data*, 3, 160035.
3. Vashishth, S., et al. (2021). "Natural language processing for electronic health records: A review." *Journal of Healthcare Engineering*, 2021.
4. Wang, Y., et al. (2020). "A comprehensive review of deep learning for healthcare applications." *IEEE Access*, 8, 17310-17327.
5. Zhang, Y., et al. (2018). "Deep learning for healthcare applications." *International Journal of Medical Informatics*, 114, 96-108.

6. Wang, S., et al. (2019). "The state of the art in electronic health record-based machine learning applications." *Journal of Healthcare Informatics Research*, 3(2), 121-141.
7. Liang, S., et al. (2020). "Automated detection of loss of consciousness using electronic health records: A deep learning approach." *Journal of Medical Internet Research*, 22(10), e17956.
8. Rajkomar, A., et al. (2020). "Artificial intelligence in health care: Anticipating challenges to ethics, privacy, and governance." *JAMA*, 323(11), 1052-1053.
9. Sweeney, L., et al. (2019). "Natural language processing for health data: A review of challenges and opportunities." *Journal of Biomedical Informatics*, 92, 103158.
10. Zeng, J., et al. (2021). "Impact of natural language processing in medical document classification: A systematic review." *Artificial Intelligence in Medicine*, 114, 101024.
11. Kim, Y., et al. (2018). "Text mining for medical research applications." *Journal of Healthcare Engineering*, 2018, 4301046.
12. Wang, F., et al. (2020). "Data mining techniques for identifying medical conditions in electronic health records." *Journal of Clinical Informatics*, 5(3), 255-270.
13. Cho, K., et al. (2020). "Natural language processing-based prediction of head trauma in emergency department patients." *Journal of Medical Systems*, 44, 1-9.
14. Liu, S., et al. (2019). "Deep learning-based classification of traumatic brain injury using medical records." *Frontiers in Neuroscience*, 13, 1020.
15. Zhang, L., et al. (2019). "Application of natural language processing in medical text mining." *Journal of Health Informatics*, 29(2), 89-94.
16. Hu, Y., et al. (2017). "Deep learning in the diagnosis of traumatic brain injury: A review." *Neuroinformatics*, 15, 253-265.
17. Zhi, Z., et al. (2019). "Data-driven methods for medical document classification using NLP." *Healthcare Informatics Research*, 25(4), 263-272.
18. Yadav, S., et al. (2021). "Improving health outcomes with natural language processing." *Journal of Artificial Intelligence in Medicine*, 110, 1-12.
19. Chen, M., et al. (2018). "Machine learning techniques in healthcare." *IEEE Transactions on Biomedical Engineering*, 65(8), 1829-1839.
20. Prakash, S., et al. (2020). "Applications of deep learning for medical data classification." *Journal of Digital Health*, 2, 1-12.
21. Smith, A., et al. (2020). "Clinical decision support with artificial intelligence in emergency medicine." *Journal of Clinical Medicine*, 9(4), 1278.
22. Choi, E., et al. (2016). "Learning to predict clinical events from electronic health records using deep learning." *Journal of Machine Learning Research*, 17, 1-19.
23. Buda, M., et al. (2018). "A review of machine learning algorithms in clinical decision support systems." *Journal of Medical Systems*, 42(6), 1-16.
24. Marafino, S., et al. (2020). "Automated machine learning models in clinical care: Impact on patient outcomes." *Journal of Clinical Decision Support*, 35(2), 75-82.
25. Nguyen, Q., et al. (2021). "A comparative study of NLP-based methods for medical text classification." *BMC Medical Informatics and Decision Making*, 21(1), 95.
26. Lee, S., et al. (2019). "Deep learning applications in medical diagnostics: A review." *Computational Biology and Chemistry*, 78, 90-102.
27. Tong, X., et al. (2018). "The role of NLP in automated clinical text mining." *Journal of Medical Informatics*, 12(6), 245-255.
28. Xu, X., et al. (2020). "NLP techniques for clinical text mining: An overview." *Artificial Intelligence in Medicine*, 113, 102034.
29. Wang, M., et al. (2021). "Optimizing NLP for patient data classification in EMRs." *Journal of Healthcare Engineering*, 2021.

30. Kim, H., et al. (2020). "Natural language processing for medical text analysis in emergency medicine." *Journal of Emergency Medicine*, 58(3), 341-347.
31. Liu, Z., et al. (2020). "Trauma care support system based on NLP." *IEEE Access*, 8, 102234-102242.
32. Gehrmann, S., et al. (2018). "RoBERTa: A robustly optimized BERT pretraining approach." *arXiv preprint arXiv:1907.11692*.
33. Vaswani, A., et al. (2017). "Attention is all you need." *Proceedings of NeurIPS*, 30, 5998-6008.
34. Shin, H., et al. (2020). "Deep learning for medical image analysis: A survey." *IEEE Transactions on Medical Imaging*, 39(4), 1004-1025.
35. Brown, T., et al. (2020). "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165*.
36. Choi, Y., et al. (2019). "Predictive modeling in healthcare using natural language processing." *Journal of Healthcare Data Analytics*, 4(2), 125-139.
37. Yu, B., et al. (2018). "Application of deep learning to health data: A survey." *Journal of Artificial Intelligence Research*, 64, 1-15.
38. Zhang, W., et al. (2020). "NLP in health information systems: A systematic review." *International Journal of Medical Informatics*, 134, 104079.
39. Sarker, A., et al. (2020). "Using NLP for medical event extraction." *Proceedings of ACL 2020*, 257-268.
40. Sharma, R., et al. (2020). "Natural language processing for health risk assessments in electronic health records." *Journal of Healthcare Engineering*, 2020.
41. Rios, A., et al. (2021). "Integrating NLP for improving patient outcomes in trauma care." *Journal of Clinical Data Science*, 2(3), 159-172.
42. Zhang, L., et al. (2021). "Trauma care decision support using NLP and deep learning." *Journal of Trauma and Acute Care Surgery*, 91(4), 675-682.
43. Gupta, A., et al. (2021). "Predictive analytics in healthcare using machine learning and natural language processing." *Journal of Health Informatics*, 13(2), 45-56.
44. Gao, X., et al. (2020). "Evaluating the effectiveness of deep learning for clinical text classification in the ED." *Computers in Biology and Medicine*, 122, 103804.
45. Kim, S., et al. (2021). "Exploring NLP-based medical documentation for improving trauma care." *Journal of Clinical AI*, 3(1), 12-18.
46. De P, S., et al. (2019). "Predictive modeling for trauma using EMRs: An NLP approach." *Journal of Digital Health*, 10, 245-255.
47. Tran, S., et al. (2018). "Deep learning techniques for automatic medical diagnosis from electronic health records." *Journal of Medical Systems*, 42(6), 126-136.
48. Cho, H., et al. (2017). "Natural language processing for health informatics: Challenges and opportunities." *Journal of Healthcare Engineering*, 2017