Automated Triage Classification in Emergency Services Using Spanish Clinical Notes: A Comparative Analysis between ALBERT and Classical Machine Learning Approaches

Triage in emergency departments relies on nursing clinical notes, but traditional systems are variable and subjective, with the risk of undertriage or overtriage. Natural language processing (NLP) offers a promising alternative for improving accuracy and consistency of the analysis of clinical notes. Although most studies focus on English language, this work addresses this gap by evaluating NLP techniques on triage clinical notes in Spanish language, by comparing classical machine learning algorithms and deep learning approaches in a Chilean healthcare setting.

Clinical notes in Spanish from an emergency department were used. The dataset classifies patients into five triage categories: C1 (vital emergency), C2 (severe emergency), C3 (moderate emergency), C4 (minor emergency), and C5 (non-emergency). Preprocessing included text standardization and the construction of a domain-specific abbreviation dictionary. Tokenization was performed with SentencePiece, configured with a 31,000-token vocabulary. The study was structured around three experimental setups: (1) benchmarking of machine learning models (CatBoost, XGBoost, Random Forest, MLP, Extra Trees, and Logistic Regression) employing only structured variables (e.g., blood oxygen saturation, pulse, age, blood pressure etc.); (2) a natural language processing approach based on ALBERT employing clinical notes; and (3) a hybrid model combining structured data with clinical notes information, where embeddings from the fine-tuned ALBERT model in experiment 2 were integrated with structured features. The data were divided into 80/20 (training/testing) with stratified sampling. The performance evaluation included accuracy, F1-score and AUROC metrics.

The results across experiments were as follows: in the case of experiment 1 (only structured data) the accuracy was 59%, the F1 score 56%, and the AUROC was 0.88. For experiment 2 (clinical notes) the metrics obtained were accuracy of 75%, f1-score of 74%, and an AUROC of 0.89. Finally, for experiment 3 (hybrid model, structured data plus clinical notes) the metrics were accuracy of 77%, f1-score of 76%, and an AUROC of 0.96. The increase in AUROC from 0.88 to 0.96 confirms the conclusion of Matos' systematic review (2024), which indicates that combining structured and unstructured data improves performance. In fact, the AUROC of 0.96 exceeds the average of 0.91 reported in that review for models that integrate NPL with clinical notes in English language. Similar to the study by Chang et al. (2024), where adding free text improved both the AUROC (from 0.812 to 0.847) and the F1-score (from 0.534 to 0.580). In specific, regarding triage critical cases (C2 and C3), the best model corresponds to the hybrid model trained with XGBoost, where an accuracy of 83% and 77% was obtained for C2 and C3 cases, respectively. Besides, the F1 score obtained for both was 83% and 78%, respectively.

Our results show that integrating structured and unstructured data significantly improves triage classification in Spanish emergency notes. Future work will expand datasets, incorporate other contextual variables, and explore multimodal approaches to enhance clinical applicability and support real-world deployment.

References

Chang, Y.-H., Lin, Y.-C., Huang, F.-W., Chen, D.-M., Chung, Y.-T., Chen, W.-K., y Wang, C. (2024). Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department. BMC Emergency Medicine, 24(1).

Matos, B. (2024). Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. BMC Emergency Medicine, 24(1).