A-DERC: Enhancing Clinical Text Preprocessing through Abbreviation Detection, Expansion, and Correction for Effective Trauma Care

Automated text cleaning can improve performance of downstream AI-ML NLP tasks in medical and operational settings

Background

Clinical text narratives, such as admission notes and injury cause descriptions, contain crucial information that drives medical decision-making for optimal patient outcomes. Artificial Intelligence-Machine Learning (AI-ML) powered Natural Language Processing (NLP) can extract actionable insights from these unstructured data sources. However, the reliability of NLP-derived embeddings depends on clean, interpretable input. Abbreviations, typos, and grammatical errors can distort semantic text representations, limiting the effectiveness of downstream tasks. Here, we present the Abbreviation Detection, Expansion, Replacement, and Correction (A-DERC) algorithm, designed to provide an end-to-end text cleaning solution for clinical language data, improving embedding quality and supporting downstream AI–ML applications.

Objectives

- Support the development of an AI-assisted transfusion decision-making system that is high-performing, interpretable, practical, and deployable in military-relevant trauma care settings.
- Improve the semantic representation of noisy clinical text through abbreviation expansion, contextual replacement, and typo correction using the A-DERC preprocessing approach.
- Apply AI-enabled topic modelling to extract, compare, and interpret injury themes associated with transfusion and no-transfusion cases.
- Generate interpretable text-derived insights that can strengthen Phase II AI transfusion research and inform future validation with operationally relevant military datasets.

Methods

This work used Ontario Trauma Registry injury-cause text as a proof-of-concept civilian dataset to support the development of an AI-assisted transfusion decision-support framework. Because military transfusion datasets remain limited, the civilian registry data provided a scalable testbed for evaluating whether unstructured trauma narratives could be transformed into clinically useful features for downstream modelling. The methodological workflow comprised two linked components: text cleaning and topic analysis

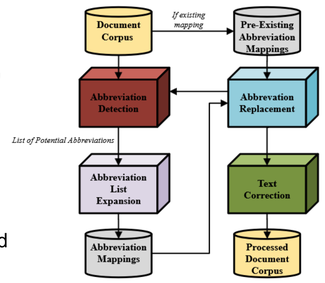

Figure 1. A-DERC preprocessing workflow for noisy clinical injury-cause text. Conceptual workflow showing how the Abbreviation Detection, Expansion and Replacement with Correction (A-DERC) algorithm improves noisy trauma-registry narratives.

Injury-cause documents were preprocessed using the Abbreviation Detection, Expansion and Replacement with Correction (A-DERC) algorithm (**Fig. 1**). A-DERC was designed to address common limitations of clinical free-text data, including acronyms, abbreviations, typographical errors, and spelling inconsistencies that can degrade the quality of text embeddings. The algorithm detects abbreviations, expands them according to contextual meaning using predefined mappings, replaces abbreviated forms with their expanded equivalents, and applies correction steps to spelling errors and typographical variants.

Methods (cont.)

To evaluate A-DERC, a manually annotated reference set of 500 injury-cause documents containing at least one abbreviation or typographical error was created. In this reference set, abbreviations were contextually expanded and typos were corrected to represent the desired textual form. Sentence or document embeddings generated from unprocessed text and A-DERC-processed text were compared against embeddings from the reference set, with embedding quality assessed using cosine similarity.

The preprocessed injury-cause documents were then analysed using a BERTopic-based topic modelling framework (**Fig. 2**). BERTopic uses transformer-based sentence embeddings, dimensionality reduction, clustering, and topic-word extraction to identify latent semantic themes within text corpora. Separate topic models were developed for transfusion and no-transfusion injury-cause groups to extract group-specific topics and compare their thematic structure. To improve model quality, a two-stage hyperparameter tuning process was implemented. Approximately 1,000 randomly selected hyperparameter combinations were evaluated from a larger search space of more than 10,000 possible settings, and the best-scoring configuration was selected for downstream analysis.

Topic modelling outputs were further characterized using hierarchical similarity analysis and network analysis to compare the structure, organization, hub nodes, communities, and inter-topic relationships of transfusion and no-transfusion injury causes.

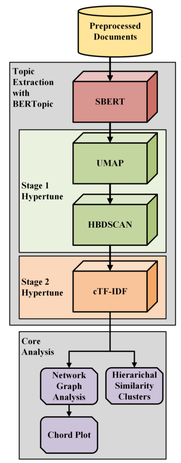

Figure 2. BERTopic modelling framework for transfusion and no-transfusion injury-cause narratives, with network analysis. Topic-modelling pipeline applied to A-DERC-processed injury-cause documents. Transfusion and no-transfusion narratives should enter parallel BERTopic workflows consisting of SBERT-based document embedding, dimensionality reduction, clustering, and topic-word extraction.

Results and Discussion

A-DERC substantially improved the quality of injury-cause text embeddings (**Fig. 3**). Compared with unprocessed text, A-DERC-processed documents showed a 13.3% increase in embedding quality relative to the manually curated reference set. The preprocessing pipeline was also computationally efficient, processing 73,117 injury-cause documents in approximately 50 seconds. These findings indicate that targeted abbreviation expansion and correction can materially improve semantic representation of clinical trauma text without requiring retraining of the underlying embedding model.

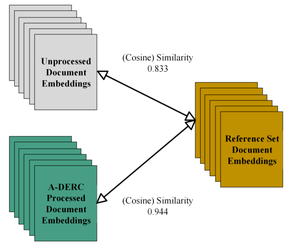

Figure 3. A-DERC improves embedding quality and preserves computational efficiency. Quantitative summary of A-DERC performance. This figure compares embedding quality for unprocessed and A-DERC-processed text relative to the manually annotated reference set, highlighting the 13.3% improvement after preprocessing.

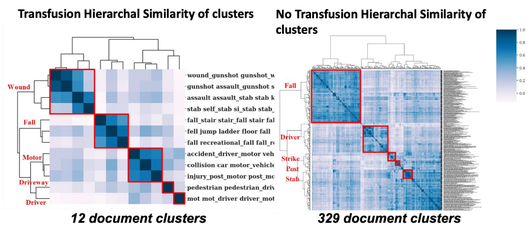

Figure 4. BERTopic output comparison between transfusion and no-transfusion groups. Comparative visualization of BERTopic outputs for transfusion and no-transfusion injury-cause narratives.

The BERTopic analysis identified clear qualitative and quantitative differences between transfusion and no-transfusion injury-cause narratives (**Fig. 4**). The transfusion group produced 329 document clusters, whereas the no-transfusion group produced 12 document clusters.

Results and Discussion (cont.)

Network analysis further showed that the no-transfusion topic network contained 17 distinct communities and 437 total nodes, while the transfusion topic network contained 5 distinct communities and 24 total nodes (**Fig. 5**). Graph comparison metrics confirmed that transfusion and no-transfusion topic structures were distinct. The Laplacian spectral distance was 1.674, indicating differences in overall graph structure. Portrait divergence was 0.790, indicating substantial differences in node-distance organization. The normalized Weisfeiler-Lehman kernel score was 0.825, suggesting that although some local structures were similar, the overall network organization differed meaningfully between groups.

Topic-word and hub-node comparisons suggested that transfusion-associated narratives were enriched for terms and themes related to mechanisms or injury patterns such as wound, motor, driveway, stab, self-inflicted injury, gunshot, shot, pedestrian, crash, collision, and assault. In contrast, no-transfusion narratives included more broadly distributed or lower-severity injury contexts, including fall, ladder, stairs, bike, dog, lawn, bath, pool, and recreational activities (**Fig. 6**). Overall, no-transfusion graphs appeared more interconnected, whereas transfusion graphs contained more isolated communities, suggesting more specialized or fragmented injury themes associated with transfusion-relevant cases.

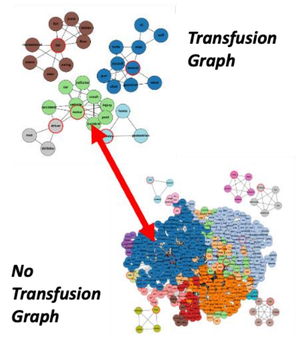

Figure 5. Network-analysis comparison of transfusion and no-transfusion topic structures. Visual and quantitative comparison of transfusion and no-transfusion topic networks. The figure shows paired networks with topic words or clusters as nodes, relationships as edges, and hub nodes highlighted as central community themes.

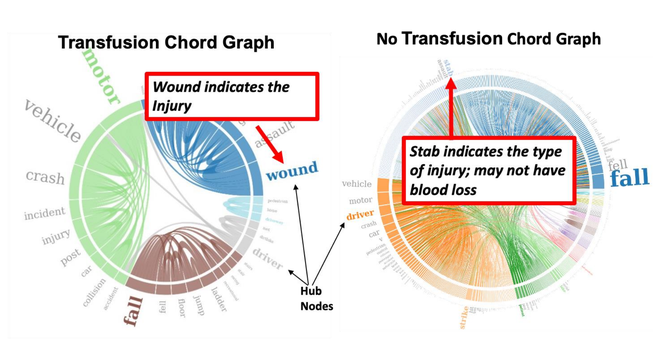

Figure 6. Topic-word and hub-node differences associated with transfusion relevance. Summary of topic-word and hub-node patterns distinguishing transfusion-associated and no-transfusion-associated injury narratives. Transfusion-associated terms should include wound, motor, driveway, stab, self-inflicted, gunshot, pedestrian, crash, collision, assault, driver, vehicle, and car, reflecting mechanisms such as penetrating injury, motor-vehicle trauma, and wound-related presentations. No-transfusion terms should include fall, ladder, stairs, bike, dog, lawn, bath, pool, recreational, floor, and home. The figure should contrast more isolated transfusion communities with more interconnected no-transfusion communities and note the need for military-relevant validation.

Conclusions

- Preprocessing is essential for clinical NLP. A-DERC provides a practical approach for improving embedding quality without retraining the vectorizer or embedding model.
- Topic modelling identified mechanisms and themes that may help characterize cases more likely to require transfusion.
- Transfusion and no-transfusion injury narratives show distinct semantic organization. Differences in clusters, communities, graph structure, and hub nodes suggest that the two groups are not simply variations of the same narrative space but have distinguishable thematic architectures.
- Network analysis improves interpretability. Hierarchical similarity, hub-node identification, and graph comparison metrics provide interpretable outputs that can help clinicians and researchers understand how injury-cause topics relate to one another.
- Future validation is needed in military-relevant datasets. Although civilian registry data provide a strong proof of concept, future work should assess whether the same text-derived themes and preprocessing benefits generalize to CAF or operational trauma contexts.

AUTHORS:Maxime Bouthillier[1, 2], Adrienne Sy[1], Tristan Bonnici[1], Henry T. Peng[1], Jing Zhang[1*], Luis da Luz3, Andrew Beckett[4,5]

[1]Defence Research and Development Canada, Toronto Research Centre, Canada; [2]University of Waterloo, Waterloo, Ontario, Canada; [3]Sunnybrook Health Sciences Center, Toronto, Ontario, Canada; [4]St. Michael's Hospital, Toronto, Ontario, Canada; [5]Royal Canadian Medical Services, Ottawa, Canada; *Corresponding Author