# Why Do Self-Harm Prediction Models Struggle to Generalise? Lexical and Semantic Variations in Emergency Department Triage Notes

**Liuliu Chen[1], Mike Conway[1], Jo Robinson[2,3], Vlada Rozova[1,4]**

[1]School of Computing and Information Systems, The University of Melbourne, Australia
[2]Orygen, The National Centre of Excellence in Youth Mental Health, Australia
[3]Centre for Youth Mental Health, The University of Melbourne, Australia
[4]Centre for Digital Transformation of Health, The University of Melbourne, Australia
**Correspondence:** liuliuc@student.unimelb.edu.au

## Abstract

Self-harm presentations to emergency departments (EDs) are strongly associated with higher suicide risk. NLP models have shown robust performance in detecting self-harm from triage notes within single hospitals, yet performance often declines across institutions. To examine potential causes, we compare ED triage notes from two hospitals by analyzing lexical characteristics, highly associated predictive features, and salient topics. Our results reveal variation in lexical expression and feature importance related to self-harm across hospitals, despite consistent core themes such as self-poisoning and self-injury. These documentation differences are associated with reduced cross-site performance. Our findings provide insight into how institutional variation affects the identification of self-harm in clinical text and highlight potential methods to improve model generalisability.

## 1 Introduction

Generalisability of predictive models is important in the clinical domain, where data collection and annotation are often expensive and logistically challenging (Goetz et al., 2024). Yet achieving robust generalisation remains a key challenge in real-world clinical applications for various reasons. For example, limited representativeness in the training data can impact model transfer to new datasets (Goetz et al., 2024), while the predictive factors that support or hinder generalisability are poorly understood (Futoma et al., 2021). Prior work on clinical NLP portability has also shown that model performance can degrade substantially when systems are applied to new clinical domains, motivating domain adaptation strategies that account for differences in documentation style, institution, specialty, and data-sharing constraints (Laparra et al., 2020).

Self-harm – defined as "an intentional act of self-poisoning (e.g., drug overdose) or self-injury (e.g., self-cutting) regardless of suicidal intent" – has become increasingly common among young people and is a major global health concern (Witt et al., 2023). This has driven researchers to develop methods for self-harm detection in clinical documentation over the years (Obeid et al., 2020; Ayre et al., 2021; Iorfino et al., 2020; Rozova et al., 2022). Recently published studies using natural language processing have shown strong performance. For instance, Obeid et al. (2020) reported an AUROC of 0.88 and an F1-score of 0.77 on electronic health records (EHRs). Rozova et al. (2022) reported an AUPRC of 0.85 when applying a Gradient Boosting model to emergency department (ED) triage notes. However, these results were obtained on test sets with the same distributions as the training data. Whether these models can generalise to other contexts remains under-explored.

Building on Rozova et al.'s (2022) study, we aim to examine the model's external validity in a different hospital context. The model was trained on ED data from the Royal Melbourne Hospital (RMH), a major tertiary referral centre in Melbourne, Victoria, Australia. When applied to Latrobe Regional Health (LRH), a regional hospital serving rural communities in Victoria, model performance declined, with AUPRC decreasing from $0.85 \pm 0.01$ to $0.78 \pm 0.01$. To investigate the potential contributors to this cross-site generalisation gap, we conduct a corpora comparison analysis on ED triage notes between RMH and LRH.

Triage notes describe the reason for patient presentation to the ED and are typically short and hastily written, resulting in non-standard grammar, misspellings and extensive use of abbreviations, which can vary considerably between hospital systems. Through the analysis of lexical characteristics, features highly associated with the outcome variable, and topic modelling, we aim to identify potential contributors to the model's diminished performance when ported to a different context.



*Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pages 383–388
July 4, 2026 ©2026 Association for Computational Linguistics

Table 1: Comparisons between RMH and LRH.

| | RMH | LRH |
|---|---|---|
| Funding | Public | Public |
| Location | Metropolitan | Regional |
| ED records, 2012–2017 | 399,111 | 171,170 |
| Catchment area | 137 sq. km | 42,000 sq. km |
| Population served | 550,000 | 300,000 |
| Emergency mental health team | Yes | No |
| Age | 48±21 | 47±23 |
| Sex | | |
| Female | 48% | 52% |
| Male | 52% | 48% |
| Intersex | < 1% | < 1% |
| Unknown | < 1% | < 1% |

## 2 Methodology

### 2.1 Datasets

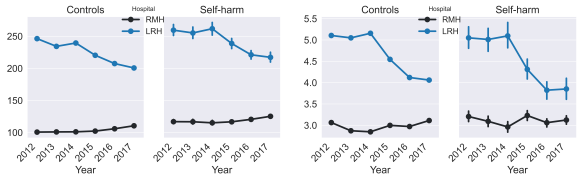
This study uses ED triage notes from RMH and LRH between 2012 and 2017 (Witt et al., 2023), which were manually annotated by psychology experts in suicide prevention as positive (**Self-harm**) or negative (**Control**). The RMH dataset consisted of N=399,111 notes with 1.4% notes annotated as positive. The LRH dataset contained N=171,170 notes, of which 1.7% were considered positive.

RMH is a large metropolitan public tertiary hospital, while LRH is the principal referral hospital for the Gippsland region, serving a predominantly rural and regional catchment. Table 1 summarises key contextual differences between the two hospitals.

### 2.2 Corpora Comparison

To compare the corpora, we examined lexical characteristics (e.g., number of characters and sentences), identified important features in each dataset, and applied topic modelling to explore the most frequent topics in the self-harm group.

**Important Features Selection** Following Rozova et al. (2022), we employed TF–IDF representations to ensure comparability and enable external validation. We computed the TF-IDF matrix and selected important features using both the Chi-Square and XGBoost algorithms. Chi-Square measures the correlation between a feature and the target variable (i.e., self-harm), and features with $p < .001$ were retained. XGBoost feature importance reflects each feature's predictive contribution, and we selected features above the 95th percentile (top 5%). We applied Chi-Square test and XGBoost independently on each dataset, both across all years and separately for each year. The final set of important features was the intersection of features selected by both methods. We further compare these final selected important features between the two hospitals.




(a) Character count (b) Sentence count

Figure 1: Lexical characteristics

**Topic Modelling** To explore common themes in the self-harm class, BERTopic was applied (Grootendorst, 2022) separately to each dataset, using only the positive class. We selected 'all-mpnet-based-v2' as our sentence embedding model, which currently performs the best in capturing semantic information (HuggingFace, 2025). The topic-wise embedding semantic similarity between hospitals was calculated by cosine similarity. To identify unique topics, we set a threshold of 0.75, selected based on a sensitivity analysis across the 0.60–0.90 range.

## 3 Results

### 3.1 Lexical Characteristics

Figure 1 provides descriptive statistics regarding the number of characters and sentences.

### 3.2 Comparison of Important Features

#### 3.2.1 Overall Comparison

Across the entire dataset, we identified 472 unigram, 701 bigram, and 637 trigram important features at RMH and 365 unigram, 460 bigram, and 426 trigram features at LRH. The top 10 selected features for each hospital are presented in Appendix, Table 5.

Of those, 213 unigram, 156 bigram, and 136 trigram features were shared between the two datasets, indicating their transferability. Table 2 presents the 10 most important shared features. As shown, these features mostly contain generic self-harm related keywords (e.g., *self harm*, *self inflicted*), self-harm methods (predominantly self-poisoning, e.g., *od*, *intentional od*), and self-injury (e.g., *superficial cuts to*). Suicide-related keywords are also common in shared features, such as *suicidal*, *suicide*, and *suicide intent*.

We plotted the ranking of shared unigram features in Figure 2. Although present in both hospi-

| | Shared features |
|---|---|
| **Unigram** | od, pain, self, tablets, overdose, sob, resolved, suicidal, suicide, attempt |
| **Bigram** | self inflicted, od of, self harm, intentional od, suicide attempt, to end, to kill, polypharmacy od, wanted to, suicide intent |
| **Trigram** | superficial lac(s) to, superficial cuts to, self harm to, self harm lacs, states has taken, with razor blade, with suicidal intent, to kill self, wants to die, pt has taken |

Table 2: Top 10 features identified in both RMH and LRH, sorted by average XGBoost importance ranking.

tals, these features differed in their statistical association with self-harm (Chi-Square) and predictive importance (XGBoost). Additionally, statistical association does not always indicate predictive power for classification. For example, *suicide* in RMH ranks highly based on the Chi-Square score but does not necessarily hold the highest predictive importance in model predictions.

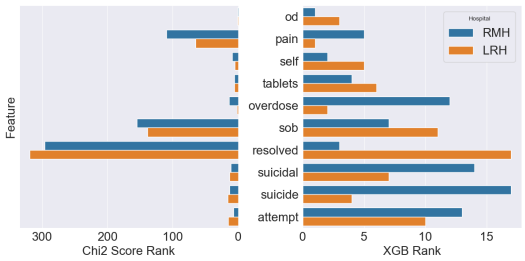
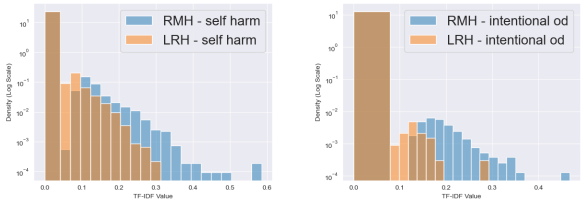



Figure 2: Ranking of shared unigram features, sorted by average XGBoost importance score ranking.

We compared the TF-IDF representations for each hospital. While feature selection was previously performed independently on each dataset, the TF-IDF representation here was computed using a shared vocabulary derived from both datasets to enable direct comparison.

Overall, the cosine similarity between site-level mean TF-IDF vectors for RMH and LRH was 0.822 for unigrams, 0.628 for bigrams, and 0.473 for trigrams. For the top 10 shared features listed in Table 2, we further compared the distributions of document-level TF-IDF values between hospitals. All unigram features differed significantly between the two hospitals ($p < .01$ for *resolved*, $p < .001$ for all others). Among bigram features, TF-IDF values for *suicide intent* and *suicide attempt* were similar, while the remaining bigram features showed significant differences ($p < .01$ for *self-inflicted*, $p < .001$ for all others). All trigram features differed significantly between hospitals ($p < .001$). Figure 3 shows examples of different distributions of TF-IDF values for terms *self harm* and *intentional od*




(a) *Self harm* (b) *Intentional od*

Figure 3: Distribution of TF-IDF values

| | RMH | LRH |
|---|---|---|
| **Unigram** | jump, dizzy, bpd, dementia, stab, life, changes, xanax | biba, dose, bandaged, tied, children, battery, community, droop |
| **Bigram** | sudden onset, social stressors, anxiety depression, attempt to, phx depression, flank pain, with etoh | on palpation, biba after, with police, analgesia taken, dose of, with nil, attempted od, alleged overdose |
| **Trigram** | recent relationship breakdown, previous self harm, recent social stressors, drowsy at triage, with stanley knife | to left forearm, under section 351, under section 10, with self inflicted, biba with police |

Table 3: Features selected only in RMH or LRH.

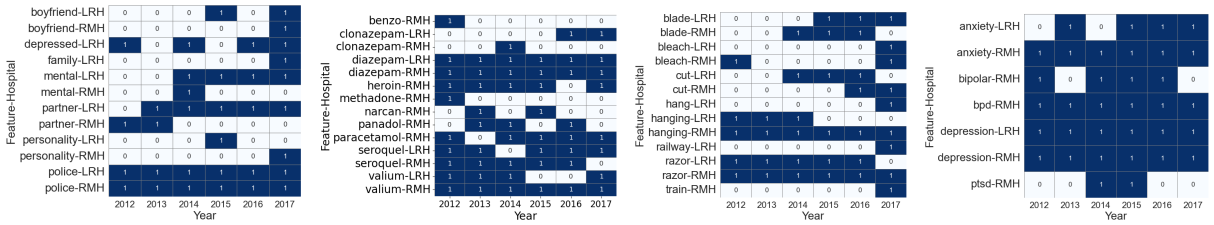
between hospitals.

Uniquely important features were also identified in each dataset. Some were misspellings (e.g., *paracetOmol*) or dosage values (e.g., *55mg*), which were excluded to focus on features with potentially meaningful insights, as shown in Table 3. Compared to RMH, LRH appears to have more police involvement in triage notes, with terms such as *with police* and *section 351* (Victorian Mental Health Act - apprehension of a person by police). In contrast, RMH contains more features related to social and psychological stressors, including *recent relationship breakdown*, *social stressors*.

### 3.2.2 Longitudinal Analysis

We further applied feature selection independently to each year and each hospital (Appendix, Figure 5). Some features remained consistent across all years in both datasets (unigram: 20, bigram: 9, trigram: 2). These features are primarily related to self-poisoning (e.g., *OD, overdose, tablets*), suicide (e.g., *suicidal, suicide attempt*), and self-inflicted harm (e.g., *inflicted, forearm*).

Focusing on social and mental factors, self-harm methods, mental disorders, and medication, we examined the consistency of related unigram features, as shown in Figure 4. Certain word features, such as *police*, *diazepam*, and *depression*, appear consistently across years in both hospitals. Social factors like *partner* are more consistent in RMH, as well




(a) Social factors (b) Medication (c) Self-harm methods (d) Mental health disorders

Figure 4: Heatmap of feature presence for feature-hospital across years (1: present, 0: absent). Note: some features appear in only one hospital because features absent in all years for a given hospital were removed.

as suicide methods (e.g., *hanging/hang*). Medication terms also show some variations between hospitals. For example, *paracetamol* was present in most years in RMH but was not selected in LRH. Regarding mental health disorders, most related terms appear consistently in RMH, whereas some features, such as *bipolar*, were not selected in any year for LRH.

### 3.3 Topic Modelling

BERTopic models identified 30 topics in LRH and 74 in RMH. After manually reviewing, we found that most were related to self-harm methods, primarily overdosing on specific medications and dosages. Other self-harm methods also emerged, such as self-inflicted injuries and hanging.

The average semantic similarity across all documents was 0.46, while the overall topic representation (top-10 keywords) had a similarity score of 0.68. To further analyse topic overlap, we calculated pairwise topic semantic similarity and identified unique topics in each hospital. Based on sensitivity analysis across the 0.60–0.90 cosine similarity range, LRH consistently showed no unique topics across thresholds of 0.60–0.85, while the number of unique RMH topics varied. We therefore used 0.75 as a heuristic threshold, as it provided an intermediate level of strictness within the tested range.

We found 10 unique topics in RMH, while no unique topics were identified in LRH. Table 4 presents the top 5 unique topics in RMH. Topics 42 and 56 are associated with overdose and poisoning presentations. Topics 51 and 14 primarily reflect psychiatric complexity, such as *psychosis* and *BPD*. Topic 27 captures hanging-related presentations.

Interestingly, while semantic embeddings of topics showed high similarity, their Jaccard similarity was much lower. This suggests that while both hospitals discuss similar topics, the specific lexicons used may differ.

| No. | Representative words |
|---|---|
| 14 | etoh, valium, valiumx4, bpd, suicidal, anxiety, 22mg, 5mg, abuse, harm |
| 27 | rope, hung, hanging, hang, distress, neck, haematoma, fell, seizure, spine |
| 42 | olanzipine, 80mg, 10mg, benztropine, 4mg, fluoxetine, 5mg, 30mg, polypharmacy, haloperidolol |
| 51 | psychosis, meds, anxiety, pmhx, stressors, midazolam, discharged, abusive, bandaged, 351 |
| 56 | suicidal, suicide, poisoning, cpr, monoxide, unconscious, hypoxic, triaged, lethargy, catatonic |

Table 4: BERTopic identified unique topics in RMH

## 4 Discussion

The above analysis highlights key differences between two hospitals, offering insights into why the TF-IDF Chi-Square/XGBoost model from Rozova et al. (2022) performed less effectively at LRH.

**Variations in TF-IDF lexical features** TF-IDF analysis revealed that while certain features are transferable between hospitals, their associations and relative importance differ, and each hospital displayed unique features. Mental health–related and social factor terms were more characteristic of RMH than LRH, and the selected medication types also differed. This suggests that term frequency distributions across hospitals vary, potentially due to differences in documentation styles or the geodemographics of patient populations. For instance, RMH (but not LRH) has an ED mental health team specifically trained to assess mental health status. These findings suggest that models trained within a single site may partially rely on site-specific documentation patterns rather than consistently capturing stable indicators of self-harm risk.

**Semantic similarity with lexical variation in BERTopic** BERTopic analysis showed high semantic similarity across most detected topics, indicating the thematic consistency across RMH and LRH. The topics primarily focus on self-harm methods, especially on overdosing with specific

medications and dosages, aligning with previous findings that self-poisoning is present in over 50% of self-harm presentations (Witt et al., 2023). However, the low Jaccard similarity indicates differences in word or phrase choices between hospitals, rather than fundamental differences in patient conditions. Notably, 10 unique topics were identified in RMH, mainly related to psychiatric conditions and overdoses. The absence of these topics in LRH may reflect differences in case mix, transfer pathways, documentation templates, or local triage practices.

**Future direction** Our results suggest several potential directions to enhance model generalisability. First, as self-harm themes remain largely consistent across hospitals, future work could incorporate lexical semantics as predictive features. Second, considering that many unique features related to drugs and their dosages, standardising such alphanumeric combinations into higher-level features might enhance cross-hospital model performance (Sikora et al., 2024). Additionally, normalising triage notes into a standardised format could further help reduce lexical and grammatical variability, ensuring more consistent documentation and better model transferability across hospitals.

## 5 Conclusion

This study compared ED triage notes between RMH and LRH, using selected TF-IDF features and topic modelling. Our analysis shows differences in the importance of selected TF-IDF features, while BERTopic showed high semantic similarity between hospitals but notable lexical differences in topic representation. These differences may contribute to the performance drop in model generalisation, and point to future directions in enhancing cross-hospital model transferability by normalising text to reduce lexical variability.

## Limitations

This study is limited to two Australian hospitals, which constrains the generalisability of our findings. Differences in dataset size and possible discrepancies in annotation practices, even with expert involvement, may have influenced feature representations. We also note that TF-IDF and BERTopic were selected for comparability and interpretability, but alternative representations may yield different insights. In particular, the implications for contextual embedding models and large language models (LLMs) remain an open question. If the primary source of cross-site variation is lexical rather than thematic, as our BERTopic results suggest, models with greater capacity to generalise across surface-level lexical differences may partially mitigate the portability gap. However, empirical validation on such models is needed before drawing conclusions.

## Ethical Considerations

Ethical approval was granted by the Melbourne Health Human Research Ethics Committee (HREC; 2017.342). All data were de-identified and are analysed within a secure institutional infrastructure. Due to the sensitivity of the data, our corpus cannot be publicly shared.

Self-harm detection should be approached with particular caution in clinical contexts. In this study, the primary purpose of these models is public health surveillance rather than clinical decision-making. Our models are not intended to replace clinician judgement. We emphasize that predictive models require careful validation prior to deployment to avoid unintended harm, particularly when applied across institutions.

## Acknowledgements

We thank Jonathan Knott and Owen Connolly for facilitating data acquisition at the Royal Melbourne Hospital and the Latrobe Regional Health. We also thank Hannah Richards and Lu Zhang for their assistance with manual data coding.

## Funding

JR is supported by an NHMRC Investigator Grant (2008460) and the University of Melbourne Dame Kate Campbell Fellowship.

## References

Karyn Ayre, André Bittar, Joyce Kam, Somain Verma, Louise M Howard, and Rina Dutta. 2021. Developing a natural language processing tool to identify perinatal self-harm in electronic healthcare records. *PLoS one*, 16(8):e0253809.

Joseph Futoma, Morgan Simons, Finale Doshi-Velez, and Rishikesan Kamaleswaran. 2021. Generalization in clinical prediction models: the blessing and curse of measurement indicator variables. *Critical Care Explorations*, 3(7):e0453.

Lea Goetz, Nabeel Seedat, Robert Vandersluis, and Mihaela van der Schaar. 2024. Generalization—a key

challenge for responsible ai in patient-facing clinical applications. *npj Digital Medicine*, 7(1):126.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

HuggingFace. 2025. Pretrained Models &x2014; Sentence Transformers documentation — sbert.net. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. [Accessed 21-03-2025].

Frank Iorfino, Nicholas Ho, Joanne S Carpenter, Shane P Cross, Tracey A Davenport, Daniel F Hermens, Hannah Yee, Alissa Nichles, Natalia Zmicerevska, Adam Guastella, et al. 2020. Predicting self-harm within six months after initial presentation to youth mental health services: A machine learning study. *PLoS one*, 15(12):e0243467.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2):146–150.

Jihad S Obeid, Jennifer Dahne, Sean Christensen, Samuel Howard, Tami Crawford, Lewis J Frey, Tracy Stecker, and Brian E Bunnell. 2020. Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. *JMIR medical informatics*, 8(7):e17784.

Vlada Rozova, Katrina Witt, Jo Robinson, Yan Li, and Karin Verspoor. 2022. Detection of self-harm and suicidal ideation in emergency department triage notes. *Journal of the American Medical Informatics Association*, 29(3):472–480.

Andrea Sikora, Kelli Keats, David J Murphy, John W Devlin, Susan E Smith, Brian Murray, Mitchell S Buckley, Sandra Rowe, Lindsey Coppiano, and Rishikesan Kamaleswaran. 2024. A common data model for the standardization of intensive care unit medication features. *JAMIA open*, 7(2):ooae033.

Katrina Witt, Gowri Rajaram, Michelle Lamblin, Jonathan Knott, Angela Dean, Matthew J Spittal, Greg Carter, Andrew Page, Jane Pirkis, and Jo Robinson. 2023. Characteristics of self-harm presentations to the emergency department of the royal melbourne hospital, 2012–2019: data from the self-harm monitoring system for victoria. *Australasian emergency care*, 26(3):230–238.

# A Appendix

| | RMH | LRH |
|---|---|---|
| **Unigram** | od, self, resolved, tablets, pain, hanging, sob, depression, superficial, diazepam | pain, overdose, od, suicide, self, tablets, suicidal, police, razor, attempt |
| **Bigram** | self inflicted, intentional od, to end, suicidal intent, polypharmacy od, sudden onset, self harm, suicide attempt, od of, hx depression | selling to, od of, self harmed, overdose of, self inflicted, self harm, by gp, to kill, pain on, on palpation |
| **Trigram** | self inflicted stab, self inflicted lac, self inflicted lacs, with intent to, in attempt to, pmhx depression anxiety, superficial lacs to, superficial cuts to, section 351 self, self harm to | superficial lacs to, superficial cuts to, superficial lacerations to, to left forearm, under section 351, unknown quantity of, self harm lacs, self harm to, pt states took, wants to die |

Table 5: Top 10 selected features in RMH and LRH.

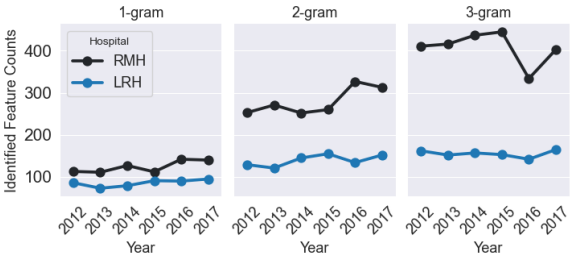



Figure 5: Numbers of selected features each year