Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Probabilistic emotion and sentiment modelling of patient-reported experiences

Curtis Murray [a,b,c,*], Lewis Mitchell [a], Jonathan Tuke [a], Mark Mackay [d]

[a] The University of Adelaide, School of Computer and Mathematical Sciences, Australia
[b] RMIT University, School of Computing Technologies, Australia
[c] The University of Melbourne, School of Computing and Information Systems, Australia
[d] James Cook University, Australia

ARTICLE INFO

Keywords:
Healthcare
Patient-reported experiences
Natural language processing
Topic modelling
Sentiment analysis
Emotion modelling

ABSTRACT

Patient feedback is necessary to assess the extent to which healthcare delivery aligns with public needs and expectations. Surveys provide structured feedback that is readily analysed; however, they are costly, infrequent, and constrained by predefined questions, limiting a comprehensive understanding of patient experience. In contrast, the unstructured nature of online reviews and social-media posts can reveal unique insights into patient perspectives, yet that very lack of structure presents a challenge for analysis. In this study, we present a methodology for interpretable probabilistic modelling of patient emotions from patient-reported experiences. We employ metadata-network topic modelling to uncover key themes in 13,380 patient-reported experiences from Care Opinion (2012-2022) and reveal insightful relationships between these themes and labelled emotions. Our results show positivity and negativity relate most strongly to aspects of patient experience, such as patient-caregiver interactions, rather than clinical outcomes. Patient educational engagement exhibits strong positivity, whereas dismissal and rejection are linked to suicidality and depression. We develop a context-specific probabilistic emotion recommender system that predicts both multi-label emotions and binary sentiments with a Naïve Bayes classifier using topics as predictors. We assess performance with nDCG and Q-measure and achieve an F1 of 0.921, significantly outperforming standard sentiment lexicons. This methodology offers a cost-effective, timely, and transparent approach to harness unconstrained patient-reported feedback, with the potential to augment traditional patient-reported experience collection. Our R package and interactive dashboard make the approach readily accessible for future research and clinical practice applications, enabling hospitals to integrate emotional insights into surveys and tailor care to patient needs. Overall, this study provides a new avenue for understanding and improving patient experience and the quality of healthcare delivery.

1. Introduction

1.1. Background on patient-reported experience

Patient-reported outcomes (PROs) and patient-reported experiences (PREs) are increasingly being used to develop a more holistic view of healthcare in a patient-centred care approach [1]. PROs and PREs can provide valuable insights into the quality and effectiveness of healthcare services and their alignment with patient expectations [2]. As such, patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) are being adopted as important indicators of healthcare performance [3,4].

1.2. Traditional acquisition methods and their limitations

Despite their increasingly recognised importance in medical literature, patient-reported experience acquisition practices, such as surveying and focus groups, may fail to capture a complete picture of patient-reported experience. The power of a survey lies in its ability to methodically gather answers to specific questions posed to its audience. This is achieved through the structured format that lends itself to rapid analysis of these specific questions. The specificity and structure of a survey, while appealing, inhibits the exploration and understanding of issues that are not explicitly addressed in the survey [5]. Questions not posed in a survey have no way to be answered, and survey-style multiple-choice questions may fail to capture the complexity

[*] Corresponding author at: The University of Adelaide, School of Computer and Mathematical Sciences, Australia.
E-mail address: curtis.murray@adelaide.edu.au (C. Murray).

https://doi.org/10.1016/j.artmed.2025.103178
Received 9 January 2024; Received in revised form 16 May 2025; Accepted 23 May 2025
Available online 8 June 2025

of the patient experience. In this way, the most common technique for collecting patient-reported experiences suppresses the patient from voicing their full experience. On the other end of the spectrum are focus groups. This qualitative-driven approach to harnessing opinions can succeed in providing scope to discuss issues the participants are interested in [6]. Where this approach suffers, however, is in its lack of scalability [7], and personal nature that may lead to self-censorship [6]. This leads to focus groups often being used as qualitative explorations used to drive survey questions [8]. There is a need to extend how we currently harness patient-reported experiences, to have enough relevance to capture issues that the patient wants to voice, in a cheap, scalable way.

1.3. Emergence of online patient feedback and challenges

Recently, increasing effort has been directed online, to *the cloud of patient experience*, as a means to provide real-time and low-cost methods to explore healthcare [9]. Online, uninhibited expositions document patient experiences without the constraints imposed by surveys, providing richer representations of experiences. However, the lack of structure in these expositions presents challenges in large-scale analysis.

1.4. Online patient feedback analytical methods

The lack of structure in free-text means that conventional surveying techniques do not apply. Instead, researchers turn to Natural Language Processing (NLP) to elicit latent structure from the reports [10–16]. This latent structure, once uncovered, can be exploited to capture patient-reported experiences for individuals, as well as varying strata. Sentiment analysis is a popular NLP technique that attributes sentiments to documents. Approaches in sentiment analysis typically use a sentiment lexicon, a vocabulary that associates sentiments with words to summarise a document's sentiment. Alternatively, machine learning approaches can be used to train a sentiment analysis model. Without labelled training data, pre-trained models known as language transformers are often used. While the simplicity of sentiment lexicons facilitates their easy adoption, their over-generalised nature inhibits their ability to accurately model sentiments in varied contexts. Large language models (LLMs), such as BERT [17], and OpenAI's GPT models [18–21] are often considered as *black-boxes* that lack interpretability, and may also suffer from a lack of contextual awareness if not fine-tuned. Moreover, these models suffer from intrinsic biases from the data they are trained on, and often *hallucinate* untrue responses [22,23]. These issues collectively contribute to a mistrust in the application of LLMs for healthcare applications [24].

Ensuring the interpretability of a model fosters trust in its results. This is crucial for the successful implementation and utilisation of models in healthcare, where the delivery of healthcare services has high potential to significantly impact health and well-being. As such, scalable patient-reported experience mining must be met with interpretability to foster trust that leads to its successful adoption in patient-reported experience measures.

1.5. Related work

Sentiment analysis and emotion detection in healthcare contexts have gained significant attention in recent years due to their potential to provide valuable insights into patient experiences and healthcare quality [25]. However, as [26] recognised, sentiment analysis in a medical context requires domain-specific models due to the varied meanings of words from a medical perspective. Several studies have approached this challenge in different ways. One such study trained a model to detect six broad classes of emotions from online health community discussions, creating training labels through manual annotation of discussions [27]. While this approach allows for the detection of simple emotions, it relies heavily on the subjective interpretations of annotators, which may not always accurately reflect patients' experiences. In addition, the six classes of emotions are likely to be insufficient to accurately capture the full complexities in patient emotions. In a different approach, [28] explored relationships between patient sentiments and topics in patient feedback using data from the NHS Choices website. By leveraging patient-reported ratings alongside reviews, they ensured their training data more accurately reflected patients' experiences. They applied sentiment analysis techniques, including Support Vector Machines and Naïve Bayes classifiers, to categorise feedback as positive or negative. Additionally, they used Latent Dirichlet Allocation [29] for topic modelling to identify common themes in patient comments and reported the mean sentiment scores for these topics. This method provides a more comprehensive view of patient feedback by leveraging patient-labelled review scores, as well as comparing sentiments across topics; however, the reduction of experience into positive or negative lacks granularity. Very recently, [30] examined data from the UK version of the patient review website Care Opinion [31], classifying patient-reported emotions into eight broad classes using SenticNet [32]. They then used these broad emotions as labels for the corresponding patient-reported experiences to train and compare multiple models for emotion classification. While this approach is effective at capturing simple emotions, it again lacks the granularity to uncover specific emotions that may more effectively convey the complexity of a patient's emotional state. These studies demonstrate the evolving approaches to sentiment analysis and emotion detection in healthcare contexts. However, there remains a need for methods that can capture the nuanced and complex emotions often present in patient feedback while maintaining the authenticity of patient-reported experiences.

1.6. Study objectives

The Australian website Care Opinion [33] is an independent, not-for-profit charitable institution that publishes user-submitted reports of patient experiences related to Australian hospitals. These reports contain both feedback on the healthcare system in Australia in free-text comments, each tagged with basic metadata such as patient-labelled emotions.

While there is structure in aspects of the Care Opinion reports, such as a catalogue of prompted answers to questions, the free text comments about healthcare contain detailed depictions of patient experience that may be otherwise overlooked. Relationships between free-text comments and labelled responses to prompted questions offer a semi-structured source for an understanding of patient-reported experiences in healthcare through natural language processing. Natural language processing tools such as topic modelling and sentiment analysis provide methods to uncover latent structure in free-text comments. This structure can provide an overview of what is being discussed and particular sentiments conveyed.

Our goal is to design an approach that allows healthcare researchers to analyse patient-reported experience narratives at scale, producing fine-grained emotion and sentiment information in a way that is low-cost, timely, and transparently interpretable.

Specifically, this study sets out to (1) develop a topic-based representation that reduces corpus dimensionality while preserving clinical meaning; (2) model the probabilistic association between topics, emotion labels, and derived sentiment categories; and (3) analyse key associations to reveal patient-reported experience insights, and (4) develop and evaluate a probabilistic emotion recommender system that provides clinicians and service managers with an interpretable tool for real-time monitoring and analysis of patient-reported experiences.

2. Methods

2.1. Care Opinion data

A corpus of 13,380 reports from February 2012 to October 2022 containing patient-reported experiences in hospitals was collected from the popular review website Care Opinion (formerly known as Patient Opinion). Each report in the Care Opinion corpus contains information on the title, the report text, the date reported, the location of the report, tags, and prompted answers to *What's good?*, *What could be improved?*, and *Feelings*.

2.2. Methods to analyse patient narratives on Care Opinion

In the analysis of Care Opinion reports, we employ the Design-Acquire-Process-Model-Analyse-Visualise (DAPMAV) framework, which we previously introduced, to systematically approach the natural language processing of patient narratives [34]. The framework guides us through designing the study around specific patient experience themes, processing the text to identify meaningful patterns, modelling to uncover latent structures, analysing the relationships between topics and sentiments, and visualising the results for actionable insights.

2.2.1. Preprocessing

In the preprocessing stage, we normalised the patient narratives to lowercase, contracted hyphenated words and conjunctions. All remaining non-alphabetic characters were replaced with spaces. Additionally, we removed words occurring fewer than five times to reduce noise and enhance data quality for our analysis. Further, we remove stop words (words with little-to-no contextual meaning such as *the*, *it*, and *and*), from the R *tidytext* package [35]. Narratives where no patient-reported emotions were present were removed from the corpus prior to topic modelling, resulting in 10,509 patient-reported experience narratives.

2.2.2. Topic modelling

Different themes of a patient's experience in healthcare may be reported online. In topic modelling, we attempt to model themes in text through collections of words that appear in statistically similar ways, called *topics*. The objective of a topic model is to express documents, or in this case, patient reports, as mixtures of meaningful topics that capture a specific theme of conversation, where the topics themselves are mixtures of words. This expression of documents as mixtures of topics acts as a dimension reduction, where documents, originally embedded in the high-dimensional ($\approx 10,000$) vocabulary space known as a bag-of-words, are mapped to a relatively low-dimensional (≈ 100) topic-space. By summarising the documents in this manner, we obtain an insightful representation of the data that would otherwise be unnecessarily specific and hence cumbersome for a general overview.

Recent advancements in topic modelling have introduced network-based approaches using a hierarchical stochastic blockmodels (hSBM) [36–38]. These methods cluster words in document-word networks to form topics, generalising the probabilistic Latent Semantic Indexing (pLSI) objective [39] of traditional latent Dirichlet allocation (LDA) models [29]. This approach offers several advantages over conventional techniques. First, it allows for hierarchical topic structures. Second, it avoids the unjustified unimodal Dirichlet distribution, which often poorly represents real text [36]. Additionally, it provides a non-parametric Bayesian method for determining the optimal number of topics, and is robust to overfitting [40]. From a minimum description length perspective, this approach offers an optimal compression of a corpus.

Network topic modelling using hSBMs can be extended to consider additional metadata in documents with the addition of new nodes corresponding to the metadata that links to documents [37]. For example, patient-reported emotion metadata can be linked to documents by using a metadata-network topic model, allowing this information to be pooled into the topic model. Fig. 1 conceptually summarises network topic modelling and metadata-network topic modelling, showing a clustering of the document-word network (with patient-reported emotion metadata).

2.2.3. Relationships between topics and patient-reported emotions

In our analysis, we quantitatively link patient-reported emotions to the thematic structures extracted from healthcare narratives using topic modelling. This is achieved by marginalising document-topic distributions, denoted as $p(T=t|d)$, across all documents tagged with a specific emotion. Here we use T to represent a random topic variable, k a specific topic, and d a specific document. The emotion-topic density for each topic k given an emotion e is computed as

$$p(T=k|E=e)=\frac{1}{|D_e|}\sum_{d\in D_e} p(T=k|d) \tag{1}$$

where $|D_e|$ is the count of documents associated with emotion e. This approach enables us to systematically associate each topic with corresponding emotional responses, providing a nuanced understanding of patient experiences in healthcare settings.

The likelihood of a topic given *any* positive (or negative) emotion is found by marginalising the joint emotion-topic density over the positive (or negative) emotions,

$$\begin{aligned} p(T=k|E\in E_p) &= \sum_{e_i\in E_p} p(T=k|E=e_i)p(E=e_i|E\in E_p) \\ &= \frac{\sum_{e_i\in E_p} p(T=k|E=e_i)p(E=e_i)}{\sum_{e_i\in E_p} P(E=e_i)}. \end{aligned} \tag{2}$$

We define the **positivity** of a topic as the ratio of positive topic likelihoods to negative topic likelihoods,

$$\text{Positivity}(k)=\frac{p(T=k|E\in E_p)}{p(T=k|E\in E_n)}, \tag{3}$$

which tells us how many times more likely topics are to be used with positive emotions than with negative ones. Topic **negativity** is defined similarly, being the inverse of topic positivity. Topics that have a high positive-to-negative topic ratio, or positivity, are associated more often with positive sentiments, and topics with a low positive-to-negative topic ratio, conversely, are associated more often with negative sentiments.

2.2.4. A landscape of patient-reported emotions

The emotion-topic profiles provide summaries of the thematic compositions of experience narratives associated with each emotion. These thematic compositions, when shared across different emotions, suggest a commonality in the underlying patient experiences. Uncovering relationships between emotions that are reflective of common experiences can help to reveal collective patient experience in more detail. This is especially true when considering that some affective states, such as *happy* and *sad*, describe one's intrinsic experience, whereas states such as *rejected* or *empowered* describe an extrinsic experience that shapes one's feelings. Recognising relationships between experiences that result in intrinsic and extrinsic emotions allows a deeper comprehension of intrinsic emotions in the context of patient experience. To investigate the thematic relationships, we examine the distances between emotion-topic densities. Applying Uniform Manifold Approximation and Projection (UMAP) [41] to these densities, we represent the emotion-topic space in a two-dimensional projection that preserves local distance. Consequently, emotions with thematic similarities are positioned in proximity within this projection.

2.3. Modelling patient emotions from patient experience narratives

2.3.1. Sentiment analysis

The natural language processing tool *sentiment analysis* allows for the detection and quantification of a text's *sentiment* by either inferring the text's positivity or identifying the emotions it presents. A simple methodology to do so involves comparing a body of text to a sentiment lexicon, such as AFINN, which expresses the sentiment of words as integer values ranging from −5 (most negative) to 5 (most positive) [42]. Other popular sentiment lexicons include VADER [43],

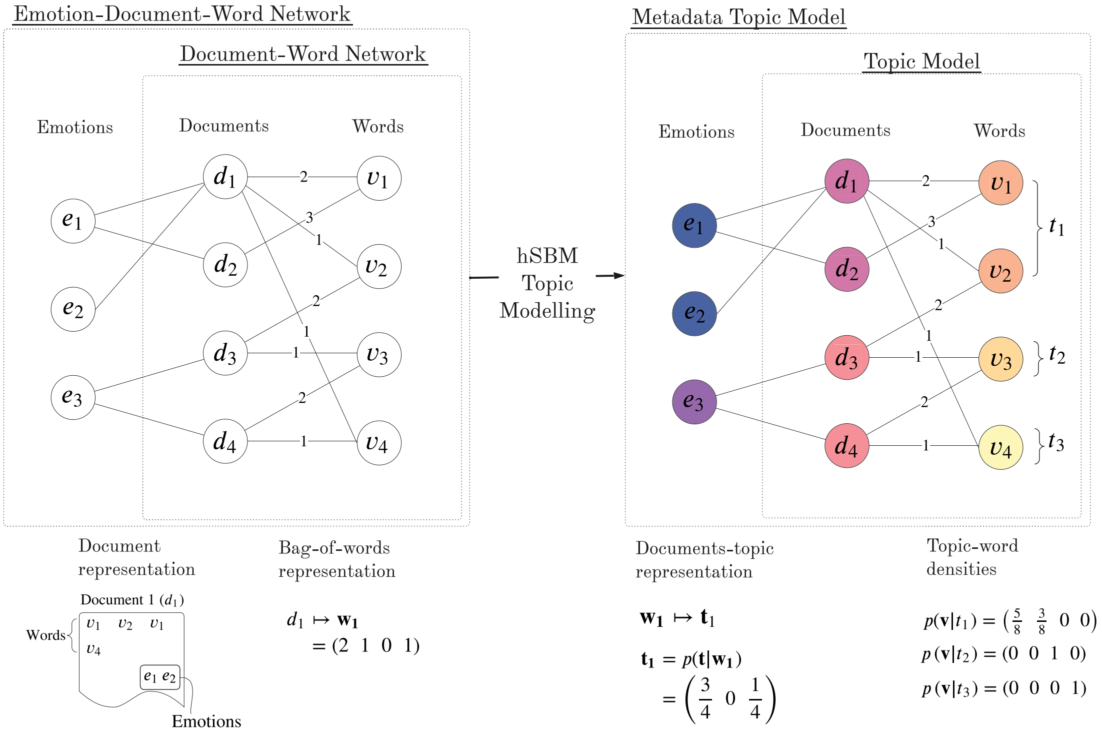

Emotion-Document-Word Network
Document-Word Network
Emotions
Documents
Words
e_1 e_2 e_3
d_1 d_2 d_3 d_4
v_1 v_2 v_3 v_4
2 3 1 2 1 1 2 1
hSBM Topic Modelling
Metadata Topic Model
Topic Model
t_1 t_2 t_3
Document representation
Document 1 (d_1)
Words
v_1 v_2 v_1 v_4
e_1 e_2
Emotions
Bag-of-words representation
$d_1 \mapsto \mathbf{w_1} = (2\ 1\ 0\ 1)$
Documents-topic representation
$\mathbf{w_1} \mapsto \mathbf{t_1}$
$\mathbf{t_1} = p(\mathbf{t}|\mathbf{w_1}) = \left(\frac{3}{4}\ 0\ \frac{1}{4}\right)$
Topic-word densities
$p(\mathbf{v}|t_1) = \left(\frac{5}{8}\ \frac{3}{8}\ 0\ 0\right)$
$p(\mathbf{v}|t_2) = (0\ 0\ 1\ 0)$
$p(\mathbf{v}|t_3) = (0\ 0\ 0\ 1)$

Fig. 1. Metadata Topic Modelling Overview: Documents are represented in a document-word network where edges between document nodes and word nodes count the number of occurrences of word v_i in document d_j (left). Emotion labels e_i are added to this network in the emotion-document-word network by connecting document nodes to emotion nodes (with an implicit edge count of one). Below this network Document 1 is depicted as consecutive words v_1, v_2, v_1, v_4, and tagged with emotions e_1, e_2. We also show the bag-of-words representation as the vector of word counts. Both networks undergo hSBM topic modelling. The hSBM performs community detection, which we visualise through the colouring of nodes to indicate group membership. Topics are the communities of word nodes. We show the mapping from d_1, or its bag of words representation $\mathbf{w_1}$, to the document-topic representation $\mathbf{t_1} = p(\mathbf{t}|\mathbf{w_1})$, which is taken as the empirical densities of topic use. To find this, we calculate the denominator as the number of edges out of the document (4), and numerators as the number of edges from the document to each respective topic $(3, 0, 1)$. The corresponding topic-word densities that indicate $p(\mathbf{v}|t_i)$ are again the empirical densities, taken as the number of edges to the topics $(8, 1, 1)$ as denominators, and the number of edges to each word as the numerators ($(5, 3, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$) for topics t_1, t_2, and t_3 consecutively. In Appendix B we show the calculation of the posterior distribution $p(E = e|d_1)$.

SentiWordNet [44], NRC [45], and Bing [46]. Dictionary approaches such as this are easily implemented, and can be effective; however, general dictionary approaches often fail to account for contextual information that may alter the sentiment of the words being used. In medical reports, context can significantly alter the implied sentiment of words and phrases. For instance, a *positive* diagnosis is paradoxically not positive, suggesting the presence of a condition or abnormality. Similarly, the word *intense* can convey contrasting sentiments based on context: in a review of a workout program, *intense* might be perceived positively, reflecting a high level of challenge and effectiveness, whereas, in patient feedback about pain experience, *intense* is indicative of severe discomfort, a decidedly negative sentiment. The contrast in sentiments conveyed by identical words across different contexts demonstrates the challenges and complexities inherent in sentiment analysis.

Recognising that sentiment varies through context is a challenge for sentiment analysis. The website Care Opinion provides the option for patients to express self-labelled emotions. Associating these labelled emotions with the respective free-text comments provides the possibility for the development of emotion and sentiment analysis models for detecting sentiments from patient-reported experiences. We develop one such model in the following section.

2.3.2. Probabilistic emotion recommender system

In this paper, we construct a context-specific probabilistic recommender system for emotions from the viewpoint of patients and their experiences in hospitals from the Care Opinion reports by exploiting the relationship between free-text reports and corresponding patient-labelled emotions. We find estimates for the posterior distribution of all 263 labelled emotions found in the Care Opinion corpus, given a new patient-reported experience narrative bag-of-words,

$$p(E = e|\mathbf{w}) \tag{4}$$

through a Naïve Bayes approach, assuming class-conditional independence of words. This non-parametric approach does not require hyperparameter tuning and is also reasonably robust to overfitting. We use topic modelling as a dimension reduction tool to (1) extract contextually meaningful predictors that (2) reduce the high dimensionality of the text data, associating reduced topic spaces with emotions. This feature engineering approach aims to capture more general relationships between patient experiences and emotions by pooling information from contextually similar words into themes. By doing so, overfitting from sparse interactions between emotions and words is mitigated, and meaningful interpretation of how model predictions are made is fostered. To ensure numerical stability in our calculations, techniques like the log-sum-exp trick are utilised. The model also incorporates practical adjustments to maintain stability and accuracy. For a detailed exposition of the model's mathematical framework, readers are referred to Appendix A.

This methodology provides a statistical foundation for a context-specific, probabilistic emotional recommender system, where recommendations are given as the top-ranking emotions under the posterior distribution. In addition to making the model we produce publicly available in both an R package and online dashboard, we have also made the code used to produce the model available in our supplementary GitHub repository [47]. This repository contains all code needed to reproduce our model, and does not include the Care Opinion data.

2.3.3. *Binary sentiment classification of patient reports*

We can further benefit from this model by using it for probabilistic sentiment analysis, where we only consider how positive (or negative) documents are by marginalising the posterior over positive (or negative) emotions.

$$p(E \in E_p|\mathbf{w}) = \sum_{e_i \in E_p} p(E = e_i|\mathbf{w}), \tag{5}$$

where E_p is the set of positively labelled emotions (and E_n the set of negatively labelled emotions).

Since E_p and E_n partition the set of emotions,

$$P(E \in E_n|\mathbf{w}) = 1 - P(E \in E_p|\mathbf{w}).$$

We perform binary classification of a document's sentiment as either positive or negative by hard classifying to the highest density sentiment class according to the marginalised posterior of Eq. (5),

$$\hat{S} = \underset{c \in \{p,n\}}{\operatorname{argmax}} P(E \in E_c|\mathbf{w}). \tag{6}$$

As there are patient narratives labelled with both positive and negative emotions, we hard-classify similarly to Eq. (6) using the empirical data. In both instances, we deal with ties by classifying a post as positive, reflecting the overall more prevalent class in the Care Opinion corpus.

2.3.4. *Model evaluation*

We employed a k-fold cross-validation method, partitioning the Care Opinion data into ten equal folds. Within each fold, we trained three iterations of the probabilistic recommender system, as described in Section 2.3.2. Two of these iterations utilise hierarchical Stochastic Block Models (hSBM) – a standard topic model and a metadata-enhanced variant – to cluster words into topics. The third iteration, in contrast, bypasses this dimension reduction step and uses the full vocabulary, as per the model detailed in Eq. (A.1).

To establish baseline comparisons, two additional models were incorporated. The first is a Maximum-Likelihood Estimate (MLE) model, which assigns emotion probabilities based on the empirical densities of emotion classes observed in the training set. This model serves as a baseline that takes into account the imbalances in emotion classes. The second baseline model adopts a uniform-random approach, assigning equal probability to all emotion classes, thereby not compensating for class imbalances.

We evaluate the performance of the recommender systems using precision and recall, interpolated precision at k, as well as Q-measure and normalised Discounted Cumulative Gain (nDCG). In all cases, we report the mean metric across all 10 folds. Both Q-measure and nDCG incorporate partial relevance of misclassifications into their model evaluations. This is particularly beneficial as there are over 200 patient-reported emotions present in the Care Opinion corpus. Considering partial relevance in model evaluations allows for lower penalisation when partially correct emotions are predicted, as opposed to when completely irrelevant emotions are predicted, through a **relevance** metric. We define the relevance, or equivalently gain, of a predicted emotion e_p given a true labelled emotion e as

$$\mathrm{rel}(e_p, e) = \begin{cases} \frac{\max_i(d(e_i,e)) - d(e_p,e)}{\max_i(d(e_i,e))} & \text{if } C(e_p) = C(e) \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $C(e)$ is an indicator of the positivity of e,

$$C(e) = \begin{cases} 1 \text{ if } e \text{ is positive} \\ 0 \text{ otherwise,} \end{cases}$$

and $d(\cdot, \cdot)$ is the Euclidean distance between topic-emotion densities,

$$\begin{aligned} d(e, e_p) &= \|p(\mathbf{t}|e) - p(\mathbf{t}|e_p)\| \\ &= \sqrt{\sum_{k=1:n_t} \left(p(T=k|e) - p(T=k|e_p)\right)^2}. \end{aligned}$$

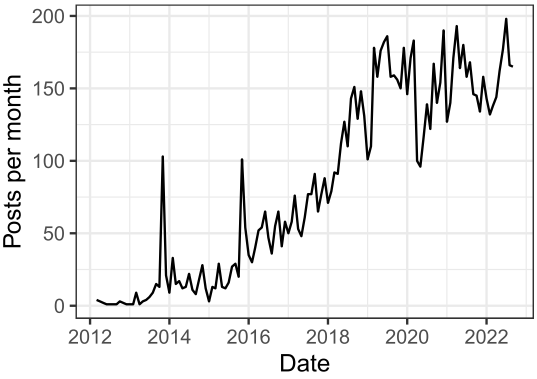


Fig. 2. Number of monthly reports to Care Opinion.

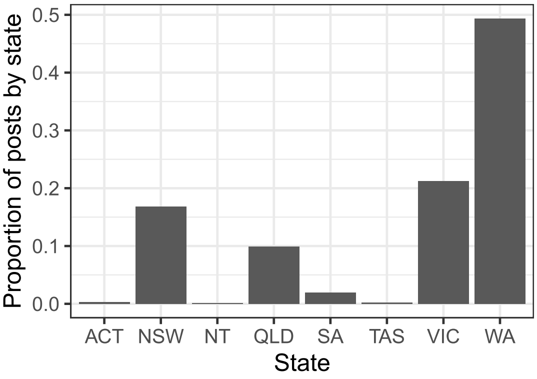


Fig. 3. Number of reports to Care Opinion by Australian states and territories.

The relevance score $\mathrm{rel}(e_p, e)$ quantifies the closeness of a predicted emotion to the true emotion by normalising the emotion distance to a 0-to-1 scale, assigning a relevance of 1 for an exact match and 0 for the greatest emotion distance, while also truncating the score to 0 for any predicted emotions that differ in sentiment classification from the true emotion, reflecting zero relevance. Full definitions for Q-measure and nDCG are deferred to Appendix A.1.

The models we consider are adapted into binary sentiment classifiers as outlined previously, and their performances are evaluated using accuracy, balanced accuracy, and the macro-averages of the F1 score, precision, and recall. Additionally, we benchmark against other sentiment analysis models – namely, Bing, NRC, SentiWordNet, VADER, and AFINN – by calculating the average sentiment score per word and assigning the document to the most dominant sentiment class. In cases where positive and negative sentiments have equal weight, we default to classifying the document as positive, the most prevalent sentiment class in the Care Opinion corpus.

3. Results

3.1. *Analysis of patient narratives on Care Opinion*

Fig. 2 illustrates the rise in popularity of reports to Care Opinion, growing from approximately 2 reports per month in 2012 to approximately 160 reports per month in 2019. We see a sharp drop in monthly reports at the start of 2020 that corresponds with the COVID-19 pandemic. The COVID-19 pandemic had a significant impact on access to some healthcare services in Australia due to restrictions on travel, in combination with strains on other aspects of healthcare from increased demand for COVID-19 testing and treatment [48].

In Fig. 3, we observe a noticeable disparity in the distribution of reports across Australian states, with a notably higher proportion

Table 1
Summary of positive, negative, and mixed posts. Positive/negative posts have entirely positive/negative tagged emotions, whereas mixed have a combination. The mean positivity of a post is reported, with positivity defined here as the proportion of tagged emotions that are positive. The total number (count) of each type of post is given, as well as the proportion of each type of post.

Sentiment	Positivity	Count	Proportion	Tags per post
Positive	1.000	5807	0.434	2.355
Negative	0.000	3993	0.234	2.358
Mixed	0.539	719	0.054	4.267
None	–	2861	0.210	0

originating from Western Australia. This trend suggests a potentially greater awareness or adoption of Care Opinion in this region among healthcare practitioners and the public. Such regional variations in engagement with the platform can provide insights into differing levels of digital health literacy and public health communication strategies across states.

The prompted answers to the *Feelings* tag, for example, *happy*, *disappointed*, and *thankful*, associate emotions with patient experience narratives. We see the use of 263 distinct emotions in the Care Opinion corpus, which we manually classify as either positive or negative. In total, there were 26,163 tags made, 58.7% of which were positive, and the remaining 41.3% negative. Posts tagged with solely positive sentiments make up 43.4% of posts. Solely negatively tagged posts make up 23.4% of the corpus. Mixed posts make up 5.4% of the corpus, with average positivity, i.e. average proportion of tags that are positive, being 0.539. There are also an additional 21% of posts that do not contain tags. This is summarised in Table 1. These results correspond to those found in those found for a smaller study of 427 reports in the UK version of Care Opinion [31,49]. There, through manual sentiment analysis, they found 59.7% positively classified reports, 23.7% negatively classified reports, and 16.7% mixed reports.

While there is an overall tendency for posts to be positive, and the mean number of tags per post is comparable for both wholly positive and negative posts, we observe a more expansive vocabulary of tagged emotions for negative sentiment posts (96 unique positive emotions and 167 unique negative emotions); there are nearly twice the number of unique negative sentiments than positive sentiments. This phenomenon is not unique to this corpus, nor even the English language; a paper that analysed multiple sources in three languages noticed that words with negative emotions are used less than words with positive emotions, but because of their rarity they carry more information [50].

Fig. 4 shows a significantly heavier tail in the rank–size distribution for negative sentiments than in the positive counterpart. This indicates that, overall, a smaller proportion of people tend to make negative posts, although those who do tend to express themselves using a richer vocabulary through their tagged emotions.

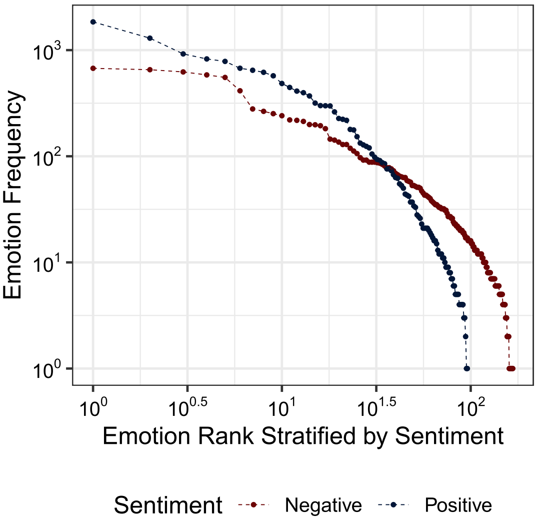


Fig. 4. Log-Log Distribution of Emotion Frequencies Stratified by Sentiment in Patient Feedback.

Table 2
Most prevalent topics and their frequencies in patient narratives from Care Opinion.

Topic	Density
information, system, late, required, concerned	0.041
found, im, lot, issues, hope	0.040
extremely, level, spent, difficult, understanding	0.037
wonderful, professionalism, kindness, appreciated, exceptional	0.035
amazing, grateful, comfortable, fantastic, safe	0.034
time, day, times, feeling, informed	0.033
opinion, dont, understand, leave, speak	0.030
didnt, couldnt, wasnt, finally, happened	0.028
caring, professional, friendly, excellent, impressed	0.027
night, morning, arrived, explained, busy	0.025
service, team, happy, helpful, provided	0.025
support, helped, recovery, knowledge, journey	0.020
upset, recall, disappointed, telling, requested	0.020
hospital, doctors, hospitals, regional, performed	0.020
due, offered, completely, allowed, previous	0.019
health, services, centre, community, provide	0.018
blood, symptoms, breathing, cold, bleeding	0.018
staff, nursing, pass, cleaning, polite	0.018
recently, received, attended, short, attention	0.017
private, lack, request, form, mentioned	0.016

3.1.1. *Topic themes in Care Opinion reports*

Our network topic modelling analysis using a hierarchical stochastic block model on the Care Opinion patient reports reveals 105 distinct topics at the deepest level in the topic hierarchy. In Table 2, we show the ten most prevalent topics with the five most common words in each, along with their respective topic densities across the corpus. We consulted a domain expert to assist in interpreting, naming, and categorising topics. Through this consultation, we observe that topics broadly fall under the following three themes, with some topics relating to more than one general theme at once.

Clinical Care, Procedures, Recovery, Rehabilitation, and Outcomes: This theme of topics encompasses interactions with healthcare professionals, medical interventions, and specific treatments or conditions, and covers many of the topics identified. It includes discussion surrounding healthcare staff such as nurses and doctors, as well as procedures such as surgeries and ultrasounds, and vast discussion around condition-specific care, including heart disease, cancer treatment, broken bones, pregnancy and birth, and hip replacement. This theme also covers patient health outcomes, through terms including *healed* and *improved*, as well as discussing post-treatment care, focussing on recovery processes, rehabilitation activities, and the role of exercise and therapy in patient recuperation through words such as *rehab*, *exercise*, *speech*, *physiotherapy*, *balance*, *medication*, *discharged*, *died*, *passed*, and *remission*. For example, a topic that relates to this general theme is a *cancer* topic, using words such as *cancer*, *chemotherapy*, *radiotherapy*, *oncologist*, *biopsy*, *mass*, *metastatic*, and *remission*.

Patient Experience, Emotion, Engagement, and Support: This topic theme captures apparent emotional responses and comfort levels of patients, including feelings of anxiety, appreciation, and the support received from healthcare staff. It also describes the quality of the patient-caregiver interaction through discussion on personalised care, using both positive and negative words such as *cared* and *abused*. Communication is discussed, for example, with terms such as *listened*, *told*, and *rude*, as well as the emotional impact of the healthcare journey, through words such as *safe*, *comfortable*, and *scared*. Patient support includes discussion on family members and personal relationships in the

Table 3
Topics and their conditional sentiment likelihoods for least-to-most positive topics from the Care Opinion corpus.

Topic	Likelihood		Positivity
	Positive	Negative	
upset, recall, disappointed, telling	0.005	0.045	0.117
told, stated, replied, script	0.001	0.009	0.156
front, rude, door, behaviour	0.004	0.014	0.253
private, lack, request, form	0.008	0.029	0.274
pay, cost, paid, afford	0.001	0.005	0.283
waiting, wait, waited, sitting	0.003	0.012	0.292
mri, chronic, spinal, spine	0.001	0.004	0.305
booked, letter, date, list	0.003	0.010	0.315
…			
support, helped, recovery, knowledge	0.030	0.009	3.301
drug, lived, alcohol, addiction	0.008	0.002	3.407
class, education, classes, online	0.004	0.001	4.140
program, sessions, learnt, tools	0.008	0.002	4.519
caring, professional, friendly, excellent	0.036	0.006	5.842
amazing, grateful, comfortable, fantastic	0.057	0.010	5.902
positive, supportive, recommend, supported	0.016	0.003	6.207
wonderful, professionalism, kindness, appreciated	0.052	0.008	6.388

healthcare experience, using words such as *family*, *loved*, and *dignity*, highlighting the presence of the support network in patient experience.

Healthcare Environment, Operations, and Administration: Topics within this theme discuss aspects of the healthcare environment, as seen with words such as *ward*, *bed*, and *clean*. Additionally, discussion surrounding healthcare operations and logistics is present, featuring discussion around clinic appointments, waiting times, service accessibility, and administrative processes. This is evidenced by topics making use of words such as *appointment*, *admitted*, *follow-up*, *waiting*, *sitting*, *late*, *system*, *park*, and *entrance*.

The identification of these themes through our topic modelling approach provides a complementary understanding of the patient-reported experience to traditional techniques, reflecting a comprehensive patient journey that spans clinical interactions, emotional responses, and the operational environment of healthcare delivery.

For each topic, we find the likelihood of the topic given an emotion as $p(T = k|E = e)$ (Eq. (1)) and the likelihood of a topic given *any* positive (or negative emotion $p(T = k|E \in E_p)$ (Eq. (2)) and define topic positivity in Eq. (3) as the ratio of positive topic likelihood to negative topic likelihood. We indicate the eight most positively-associated and eight most negatively-associated topics according to this ratio in Table 3, along with their likelihoods $p(T = k|E \in E_p)$ (Positive) and $p(T = k|E \in E_n)$ (Negative). We visualise topics at the polar extremes of positivity in Fig. 5, with the lowest positivity in Fig. 5(a), and the highest in Fig. 5(b).

Fig. 5 reveals a dichotomy in patient-reported experiences within the healthcare system. The topic characterised by the lowest positivity, as depicted in Fig. 5(a), uncovers a vast array of negative sentiments. The likelihood of this topic in the presence of negative emotions is quantified at 0.05, whereas it falls to 0.0045 for positive emotions, as detailed in Table 3. It encapsulates a spectrum of adverse emotions such as *upset*, *disappointed*, and *angry*, accompanied by descriptors such as *unacceptable* and *appalling*, and actions expressed through verbs such as *questioned*, *refused*, and *forced*. Significantly, the emphasis of this topic is on the subjective experiences of patients rather than on the clinical outcomes of their care.

Conversely, the topic with the highest positivity (likelihood of 0.052 given a positive emotion compared to 0.008 for a negative emotion), illustrated in Fig. 5(b), offers a contrasting portrayal. It accentuates the commendable attributes of healthcare workers, as evidenced by the prevalence of adjectives such as *wonderful*, *exceptional*, *compassionate*, *dedicated*, and *empathetic*. Additionally, it includes nouns that convey a sense of appreciation, such as *gratitude*, *professionalism*, and *kindness*, and is peppered with specific names, likely reflecting patient gratitude towards individual caregivers.

A complete collection of the topics identified in our analysis – along with their respective likelihoods under positive and negative sentiments – is provided in the online supplementary materials, available on the dedicated GitHub repository [47].

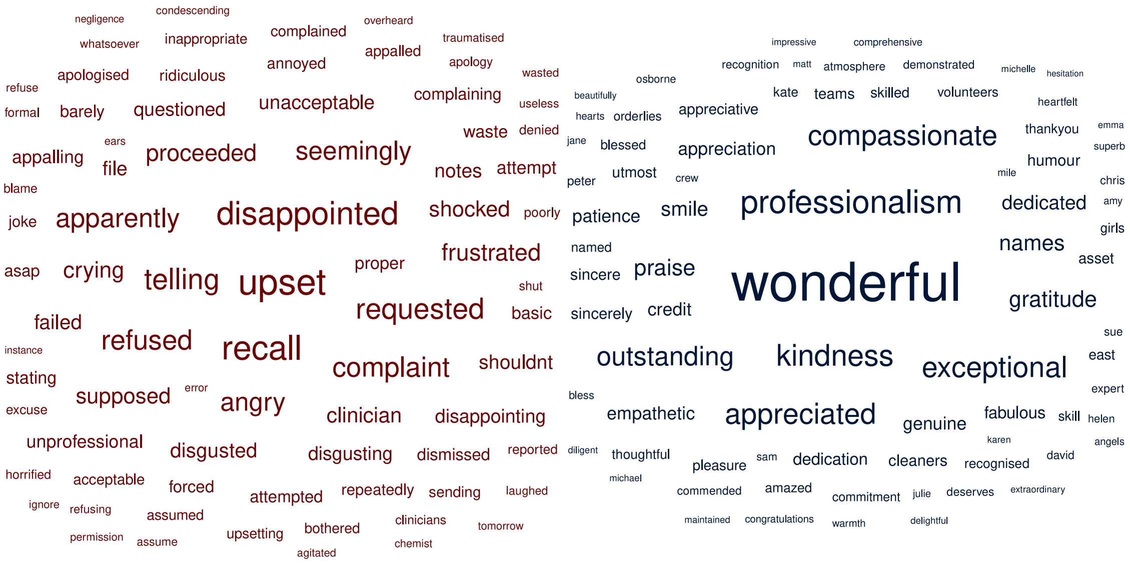
3.1.2. Relationships between topics and patient-reported emotions

Relationships between patient-reported emotions and healthcare topics discussed can reveal systematic emotional responses to experiences and capture the subjective quality of clinical encounters. These insights may inform providers and policy-makers about areas requiring compassionate engagement, contribute to patient-centred care models, and enable improvements in healthcare delivery.

Fig. 6 presents a parallel-coordinates plot that elucidates the relationships between themes in patient-reported experiences and their corresponding emotional responses. This visualisation captures the likelihood of specific topics conditioned on the spectrum of emotions expressed by patients, highlighting the prevalent sentiments tied to various facets of healthcare delivery. Along the x-axis, we list the eight most negatively-associated and eight most positively-associated topics from the patient reports. These are represented by their four most frequently used words, and arranged from left to right to correspond to an increase in topic *positivity*, defined in Eq. (3). The y-axis of Fig. 6 shows the topic-emotion likelihood $p(T = k|E = e)$ for each emotion e. The two emotions with the highest topic-emotion likelihood are labelled to reveal the emotions that are most strongly associated with each topic. The remainder of emotions are unlabelled, with a low opacity, and instead simply coloured to indicate positivity (blue) and negativity (red) to provide an overall view of the spectrum of positive and negative emotions for each topic. We add in bold the topic density marginalised over all positive emotions $p(T = k|E \in E_p)$, labelled *Positive (Collective)* and all negative emotions $p(T = k|E \in E_n)$, labelled *Negative (Collective)* to show the average topic-emotion likelihood for both sentiments.

Negatively-associated topics. At the far left on the x-axis in Fig. 6, the topic represented by the terms *upset*, *recall*, *disappointed*, and *telling* – previously identified in Fig. 5(a) – carries the lowest positivity among all Care Opinion topics. Predominantly, the emotions *abused* and *ridiculed* demonstrate the highest association with this topic, appearing to capture patients' extremely negative experiences and perceptions of care received from healthcare personnel. The overall negativity of this topic is evidenced by the distribution of negative emotions (red lines) sitting well above the distribution of positive emotions (blue lines). This pattern is evident for the following seven topics with low positivity.

The pattern of negativity being most strongly exhibited in topics related to patient-experience continues in the following two topics, with prominent emotions such as *embarassment*, and *aggressive* present. The fourth, fifth, sixth, and eighth topics (from left to right) appear to fall under the theme of *Healthcare Environment, Operations, and Administration*, and elicit emotions such as *uncertain*, *inconvenienced*, and *very distressed*. The emotion *very distressed* is most strongly associated with a *waiting* topic, indicating that high levels of distress are associated with patients having to spend time waiting for an aspect of care. The seventh topic, which mentions *mri*, *chronic*, and *spinal*, as well as *muscle*, *neurology*, *arthritis*, *paralysed*, and *chiropractor*, is more closely related to the *Clinical Care, Procedures, Recovery, Rehabilitation, and Outcomes* topic theme, and appears to discuss care around the musculoskeletal and nervous systems. Interestingly, the positive term *free* emerges as the most prevalent emotion within this context, a finding that may potentially reflect a sense of relief patients may experience upon recovery from enduring pain or mobility limitations. Conversely, the negative emotion *embarrassed* also features prominently, potentially signalling the psychological distress or stigma that patients often confront when dealing with chronic illnesses. These dichotomous emotional responses demonstrate the complex interplay between the physical alleviation of symptoms and the social-emotional challenges encountered during the patient journey.



(a) Topic with lowest positivity (b) Topic with highest positivity

Fig. 5. Extremes in sentiment-topic associations in the Care Opinion corpus, illustrating the dichotomy of patient-reported experiences: (a) encapsulates critical negative experiences, while (b) reflects commendations. Both extremes centre around patient experience rather than patient outcomes.

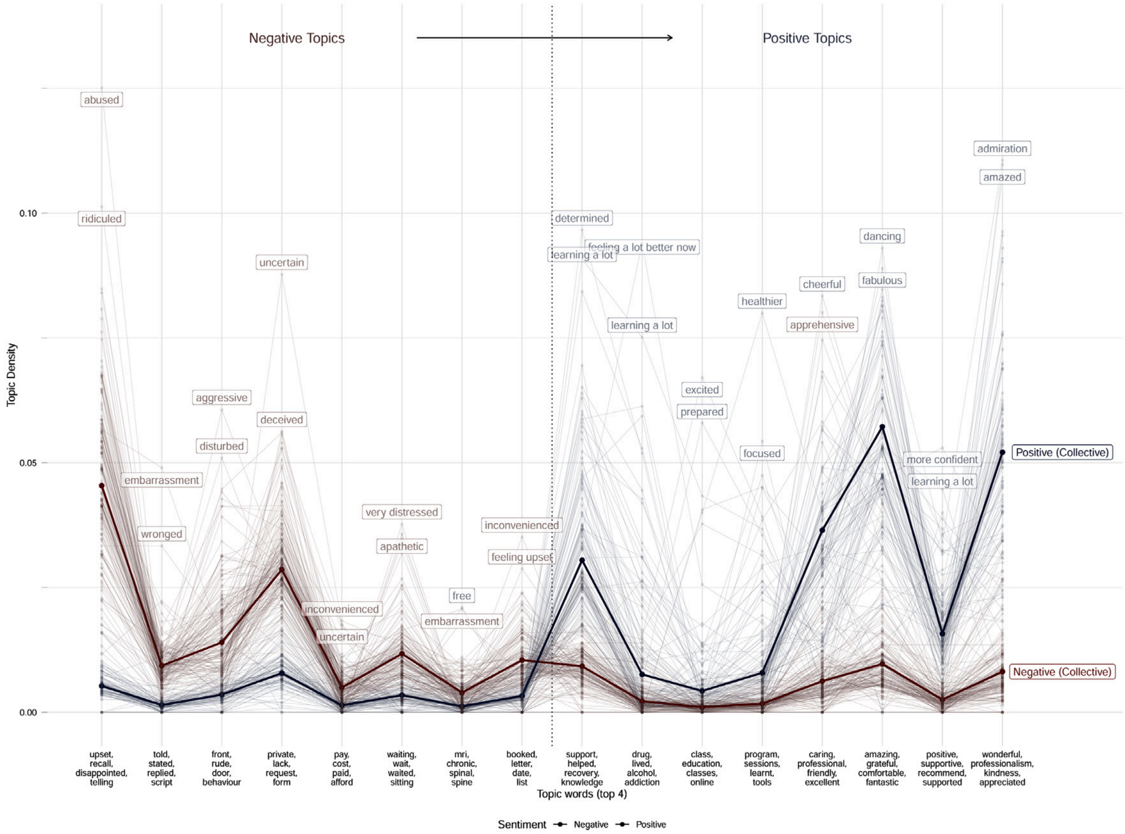


Fig. 6. Interactions between Emotions and Topics Discussed on Care Opinion: This parallel coordinates plot illustrates the interplay between patient-reported emotions (depicted as lines) and Care Opinion topics (arranged on the x-axis), with the likelihood of each topic given the presence of specific emotions quantified on the y-axis. Among 105 total topics analysed, the plot selectively emphasises the eight most negative and eight most positive topics from left to right, highlighting the extremities of themes in patient narratives. The red spectrum signifies a collection of negative emotions, while the blue spectrum represents positive emotions. The strongest two emotions associated with each topic are labelled, providing clear indicators of the significant emotional drivers in patient experiences. The density and spread of lines corresponding to unlabelled emotions across the topics reveal the position and variation in emotional responses.

Positively-associated topics. Most of the eight topics to the right of Fig. 6 that exhibit the highest positivity of all topics, appear to fall under the theme of *Patient Experience, Emotion, Engagement, and Support*. The rightmost, most positively associated topic (seen in Fig. 5(b)) is indicated by the words *wonderful*, *professionalism*, *kindness*, and *appreciated*. The likelihood of this topic is maximised for the emotions *admiration*, and *amazed*, showing signs of sincere patient gratitude. Several interesting observations can be made. The seventh most positive topic (from the right) appears to discuss addiction, which may fall more under the theme of *Clinical Care, Procedures, Recovery, Rehabilitation, and Outcomes*. This topic, while appearing negative, may be a result of a shift from addiction to recovery. This is supported by the prominent emotions this topic exhibits, which are *feeling a lot better now* and *learning a lot*. The fifth and sixth most positive topics both are indicative of patient education and involvement, using words such as *class*, *education*, *program*, *sessions*, *learnt*, and *tools*. That this topic is of such high positivity indicates that patients who are actively included and educated in classes and programs express positive sentiments, with emotions such as *excited*, *prepared*, *healthier*, and *focused* appearing most prominently. The fourth most positive topic once again shows signs that positive patient-caregiver interactions, where the patient feels cared for, can have a beneficial response to patient concerns, as the negative term *apprehensive* appears as one of the most prevalent emotions for this topic.

This analysis shows a clear delineation between topics associated with patient sentiment. Positive topics reflect quality interactions with healthcare workers and general experience, while those that are negatively connoted appear emblematic of adverse experiences within the healthcare system. In both cases, the extremes in positivity are strongly tied to patient-caregiver interactions rather than clinical outcomes.

This detailed analysis is supported by comprehensive online supplementary materials, which present the full spectrum of topic-emotion relationships derived from the study. These materials are accessible for further review and exploration at the corresponding GitHub repository [47].

3.1.3. A landscape of patient-reported emotions

Our analysis through UMAP reveals two distinct clusters of emotions that show strong spatial separation and correspond to positive and negative emotions. In addition, we manually classify each emotion as either positive or negative and find that the UMAP clustering agrees with our manual labelling with an accuracy of 0.985.

Figs. 7 and 8 show the division of the UMAP-projected regions containing each cluster. Each emotion label in this figure is scaled according to the frequency of occurrences in the corpus and coloured by the manual sentiment classification—blue for positive and red for negative. This bifurcation of the emotion-topic space in its two-dimensional representation effectively captures the dichotomy between positive and negative emotions. The closeness of emotions in the UMAP space suggests that there are thematic similarities in the narratives that elicited these emotions. Two emotions being close together in the UMAP visualisation implies that the patient-experiences leading to these emotions share common themes or topics.

Negative emotions. This visualisation reveals several insights into patient experience. In the negative emotion cluster of Fig. 7, the co-occurrence of the intrinsic emotions *frustrated* (upper right) and *stressed* with the extrinsic emotions *uninformed*, *forgotten*, and *inconvenienced* highlights the similarities and shared experience between these emotions that are characterised by neglect or lack of information. The proximity between *angry* (upper middle) and the extrinsic emotions *dismissed*, *patronised*, and *being lied to* reveals negative experiences that may cause patients to feel anger. Similarly, *in pain*, *scared*, *frightened*, and *distressed* appearing together show that experiences resulting in traumatic emotions share similar themes.

The clustering (lower right) of *suicidal*, *depressed*, and *hopeless* in close proximity to *invalidated*, *rejected*, and *unsupported* demonstrates a significant correlation between profound negative emotional states and experiences of neglect or dismissal in patient narratives. This alignment reinforces existing understanding of the impact of emotional validation (or lack thereof) on patient mental health and the necessity of empathetic and supportive communication in healthcare settings.

Positive emotions. In the positive emotion cluster of Fig. 8, *thankful* (upper middle), *grateful*, *looked after*, *cared for*, *respected*, and *involved* are clustered together. This grouping suggests that experiences, where patients are cared for, included, and respected, are thematically similar to those where patients feel appreciation and contentment. This reinforces that fostering an environment of respect, inclusion, and quality care in a patient-centred care approach has a strong association with patients' satisfaction with the care they receive.

The positive clustering of *informed* (lower middle), *listened to*, *heard*, *trust*, and *confident* illustrates the relationship between experiences where patients are actively engaged and those that they feel trust and confidence in. This finding emphasises the need for strong communication skills in addition to clinical competency to bolster patient confidence and trust.

Similarly, the proximity of *supported* (middle right), *helped*, *encouraged*, and *empowered* with *loved*, *prepared*, and *hopeful* reflects a narrative where patients feel actively and emotionally supported, as well as involved in their healthcare journey. This clustering suggests a relationship between these experiences and those where patients feel empowered and have a positive outlook towards their future.

The group comprising *safe*, *calm*, *at ease*, and *reassured* highlights the significant emotional impact that a secure and supportive healthcare environment can have on patients. This clustering suggests that when patients feel safe and reassured, it may cultivate a positive and calming environment. Notably, the emotion *nervous* is also used within this positive cluster, an apparent paradox given its conventional connotation; however, in this context, it is plausible for patients to be nervous and yet report positive experiences if subsequently reassured. Similarly, *apprehensive* (upper middle), a decidedly negative emotion, appears in the positive cluster situated next to *extremely relieved*. This may be indicative of a potential shift from apprehension to relief, in a positive emotional transition in response to compassionate care.

3.2. Probabilistic emotion recommender system

We have developed a probabilistic emotion recommender system, as detailed in Section 2.3.2, which employs a network-based topic modelling approach on the Care Opinion corpus. By identifying distinct topics, the system projects new text inputs into a multidimensional topic space, with the resulting topic densities serving as predictors within a Naïve Bayes classification framework. An interactive version of this model is available online [51], complemented by a package in the R statistical computing environment [52], which is freely available [53].

3.2.1. Model evaluation

The evaluation of our probabilistic emotion recommender system is divided into two distinct categories. In this section, we assess the system's ability to predict specific emotions from the comprehensive range of patient-reported emotions, as derived in Eq. (1). Later, we evaluate its performance in the binary classification task of discerning between positive or negative emotions, outlined in Eq. (5).

Results from the 10-fold cross-validation of the three probabilistic recommender systems, using metadata topic modelling, topic modelling, and the full-vocabulary model, are presented, with comparisons to the baseline models using maximum likelihood estimates and uniform random guessing.

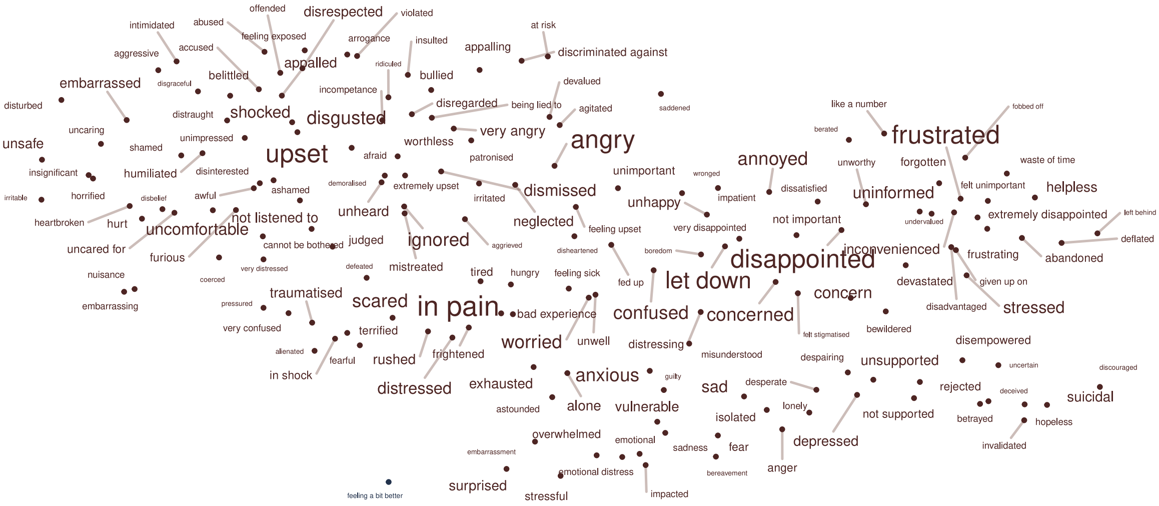


Fig. 7. UMAP dimension reduction of emotion-topic representation (negative cluster).

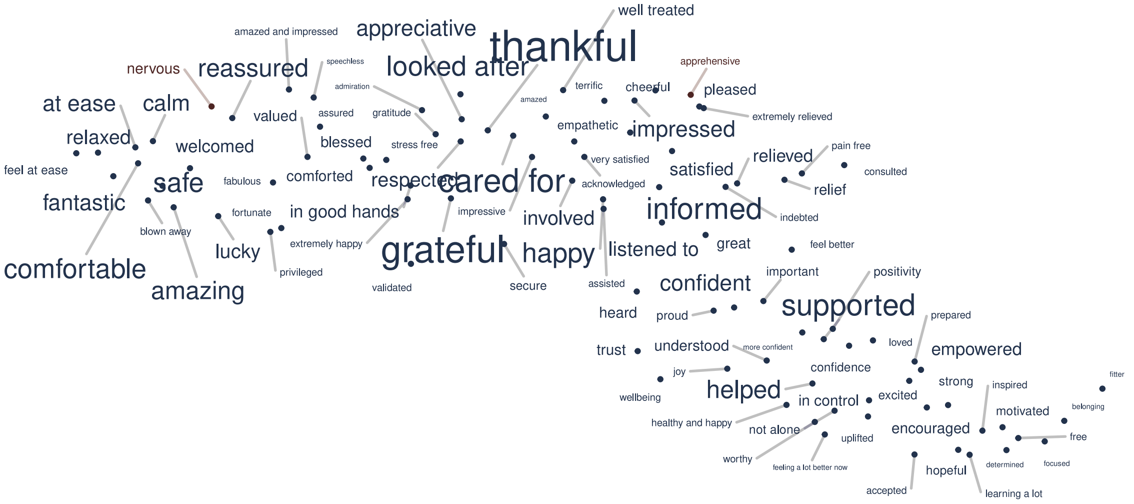


Fig. 8. UMAP dimension reduction of emotion-topic representation (positive cluster).

Precision and recall. We find the interpolated precision [54,55] $P(r)$ at recall r as the highest precision for any recall r' greater than r,

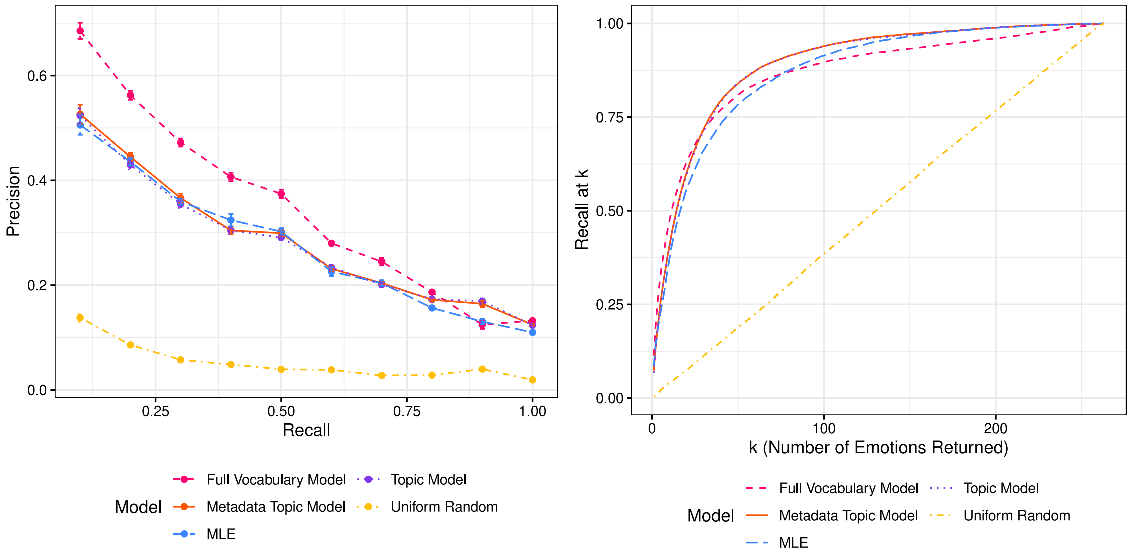
$$P_i(r_i) = \max_{r_i \le r'_i} p(r'_i),$$

and find the macro average across the validation sets. This is seen in Fig. 9(a), where we show standard errors of the macro averages across the folds with error bars. In Fig. 9(b) we show the recall at k (macro average across folds), where k denotes the k highest ranking emotions returned by the recommender system.

For small values of k the full vocabulary model has the greatest recall, however, is outperformed by the topic model and metadata topic model when over 30 emotions are returned. The full vocabulary model performs worse than simply selecting the highest density sentiment classes (MLE model) after 80 emotions are returned. We argue that the higher performance in the full vocabulary model for small values of k is due to the high sensitivity to detect specific emotions from specific mentions of words (e.g. if the word *angry* appears it may conclude with near certainty that the text is of emotion *angry*).

The highly subjective nature of emotions in individuals suggests that model predictions with near certainty are likely a sign of overfitting. Predictions involving synonymous emotions, such as *scared* and *frightened*, should account for this subjectivity with some density in prediction going to both emotions. However, potential overfitting in the full vocabulary model due to its high complexity may lead to predictions of single emotions with near certainty as a result of noise in the training data, amplified by the already sparse interactions between individual words and emotions. Poor model understanding of relationships between like-emotions may be missing when compared to more parsimonious approaches, which meaningfully reduce the feature space, such as topic modelling.

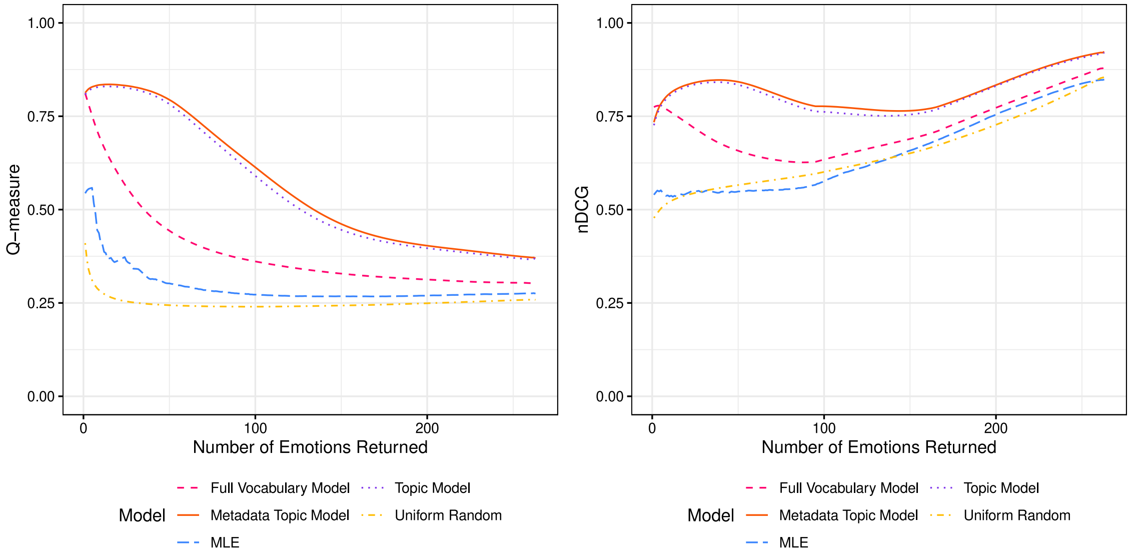
Accounting for partial relevance and penalising late arrival of predicted emotions. We more appropriately validate these models by capturing partial relevance, and demonstrate greater performance in the topic modelling approaches by considering the degree of closeness between the predicted emotions and the labelled emotions in evaluation metrics. Precision and recall metrics fail here, as they will only recognise an emotion prediction as relevant if it is an exact match for the emotion label. This is a concern, as a misclassification to *frightened* when the labelled emotion is *scared* is falsely considered just as bad as a misclassification to *happy*. By considering partial relevance in model evaluation, we can attribute value to predictions that have good, yet imperfect predictions. Graded relevance metrics, such as Q-measure and Normalised Discounted Cumulative Gain (nDCG), are more appropriate in evaluating information retrieval systems as they account for this partial relevance, and penalise the late arrival of relevant documents [56]. By incorporating partial relevance in our evaluation metrics, as defined



(a) Interpolated precision $P(r)$ at recall r.

(b) Recall at k for the k highest posterior density emotions.

Fig. 9. Evaluation of emotion recommender systems against baseline models for the Care Opinion corpus using precision and recall: The precision–recall plot in (a) shows strong performance of the full-vocabulary model, however, the recall at k plot in (b) shows that when more than 30 emotions are returned, the models using dimension reductions are superior under this metric, potentially signalling overfitting in the full vocabulary model.



(a) Q-measure for varying numbers of returned emotions.

(b) nDCG over the range of emotions returned.

Fig. 10. Comparative performance of emotion recommender systems with respect to the Q-measure (a) and nDCG (b). Each evaluation metric assesses the ranking order of emotions, with a score of 1 achieved under an ideal ranking of the preceding emotions. The metadata topic modelling shows the greatest performance under both metrics, whereas the high-complexity full-vocabulary model performs comparatively poorly.

in Eq. (7), we reduce the penalisation of predictions that are partially aligned with the true labels. Full definitions of Q-measure and nDCG may be found in Appendix A.1. Q-measure results for the probabilistic emotion recommender systems are shown in Fig. 10(a), and nDCG results in Fig. 10(b).

The data presented in Fig. 10 predominantly demonstrate the enhanced effectiveness of the metadata topic modelling approach, as evidenced by its performance in both Q-measure and nDCG metrics across the majority of the range for returned emotions. Notable exceptions are present: for instance, when only a single emotion is considered, both the topic modelling and full vocabulary approaches yield comparable Q-measure results; similarly, the full vocabulary model shows a slight advantage in nDCG when fewer than four results are returned. Nonetheless, these instances do not diminish the overall trend, which suggests that the dimension reduction inherent in topic modelling significantly bolsters the emotion recommender system's ability to discern associations among similar emotions. In contrast, the full vocabulary model exhibits a tendency towards overfitting, with its performance waning more sharply as the number of emotions increases, demonstrating the robustness of dimension-reduced models in handling a broader spectrum of emotions.

Additionally, both the topic models and the full vocabulary model demonstrate significant advancements over more rudimentary approaches, such as random guessing or defaulting to the most frequent emotions. This illustrates that, despite the complexity inherent in high-dimensional emotional data, the application of comparative metrics like Q-measure and nDCG can effectively showcase the relative performance enhancements afforded by these sophisticated models. Consequently,

Table 4
Evaluation metrics of the probabilistic emotion recommender system in the context of document-sentiment classification for three models compared to standard sentiment analysis lexicons. We show accuracy, balanced accuracy, F1, Precision, and Recall.

Model	Accuracy	Balanced accuracy	F1	Precision	Recall
Metadata Topic Model	0.910	0.911	0.921	0.939	0.903
Topic Model	0.908	0.911	0.918	0.945	0.893
Full Vocabulary Model	0.877	0.869	0.896	0.878	0.914
Bing [46]	0.846	0.844	0.812	0.794	0.830
VADER [43]	0.842	0.822	0.785	0.845	0.727
NRC [45]	0.705	0.647	0.491	0.788	0.357
SentiWordNet [44]	0.353	0.364	0.341	0.287	0.419
AFINN [42]	0.149	0.163	0.177	0.144	0.230

these models are validated as performing well beyond baseline levels, suggesting their practical utility in real-world applications.

3.2.2. Binary sentiment classification of patient reports

Model evaluations of predictions made about binary sentiment using $P(E \in E_p|\mathbf{w})$ (and $P(E \in E_n|\mathbf{w})$) further capture the superior performance of the topic modelling approaches, and allow for evaluation using standard binary classification metrics. Evaluation metrics for sentiment classification of each model are shown in Table 4, including metrics such as accuracy and balanced accuracy, as well as the macro averages across cross-validation folds of the F1 score, precision, and recall. Notably, the metadata-based topic modelling approach achieves the highest F1 score of 0.921. Among standard sentiment analysis lexicons, the Bing Sentiment Lexicon shows best performance, with an F1 score of 0.812, closely followed by VADER with 0.785. In contrast, models like SentiWordNet and AFINN exhibit significantly lower performance. This comparison illustrates the advantage of employing context-specific models over standard sentiment lexicons.

By framing the problem within a binary classification setting, we provide a complementary perspective to high-dimensional emotion profiling. This binary viewpoint allows for the utilisation of well-established metrics such as accuracy, precision, recall, and the F1 score, which offer a more intuitive and explainable measure of performance. In doing so, we enhance the interpretability of the results, facilitating a clearer understanding of the model's efficacy in distinguishing between positive and negative sentiments. Moreover, this approach demonstrates the versatility of the topic modelling techniques, which outperform the other models at high-dimensional emotion prediction and provide tangible, interpretable outcomes when reduced to a binary sentiment classification framework.

3.2.3. Model application to simulated patient-reports

In conjunction with model evaluation, we provide a supplementary demonstration of the probabilistic emotion recommender system's capabilities (using the metadata topic model) by applying it to simulated patient reports. The results in Table 5, show input text and the top three emotions as determined by the model's posterior distribution, alongside empirical priors that align with the maximum likelihood estimates derived from the observed frequencies of emotions within the simulated reports. We invite readers to further explore the model's utility via its interactive online version [51], and R package [53].

The results from Table 5 demonstrate the model's ability to identify and measure emotions within patient narratives. The analysis reveals positive sentiment in contexts typically associated with negative emotions, such as bereavement, where expressions of thankfulness for support coexist with grief. In scenarios where negative responses might be expected, such as inadequate care or extended wait times, the model identifies low positive sentiment scores and distinguishes between specific emotional responses ranging from sadness to frustration. The results identify low positive sentiment in scenarios involving inadequate care or extended wait times, while distinguishing between specific emotional responses. In particular, the analysis suggests that surgical anxiety appears attenuated in narratives that mention supportive staff interactions. Such findings indicate the model can detect contextual emotional responses to healthcare experiences, potentially informing providers about the complexity of patient reactions.

4. Discussion

This study employs a two-phase approach for analysing and utilising patient-reported healthcare experiences. First, we conduct a detailed analysis of patient narratives from Care Opinion. Second, we address the research problem by developing a scalable and transparent model that enhances structured mining of patient-reported experiences. Our approach prioritises interpretability to foster trust and facilitate adoption, particularly for analysing free-text comments in patient experience surveys.

4.1. Analysis of patient narratives on Care Opinion

Our findings indicate that while patient posts are predominantly positive, expressions of negative emotions are more intense and frequent. This suggests a complex emotional landscape where negative experiences in healthcare evoke more expressive and emotionally charged responses.

4.1.1. Topic themes in Care Opinion reports

Through topic modelling, we reveal relationships between themes in patient experiences and a spectrum of patient-reported emotions. The aggregation of topic densities for each emotion label provides a foundation for our subsequent analysis.

We observe that topics fall under the three general themes broadly relating to (1) *Clinical Care, Procedures, Recovery, Rehabilitation, and Outcomes*, (2) *Patient Experience, Emotion, Engagement, and Support*, and (3) *Healthcare Environment, Operations, and Administration*.

Collectively, these topics and their themes provide a patient-centred perspective on healthcare that encompasses clinical outcomes, emotional responses, and operational factors. This multi-dimensional understanding of patient experience may be valuable to healthcare providers and policymakers working to enhance care quality and patient satisfaction. Our study offers an empirical analysis of patient narratives that could contribute to healthcare improvement initiatives and future research.

4.1.2. Relationships between topics and patient-reported emotions

Additionally, our analysis reveals thematic differences in patient narratives associated with patient sentiment. Topics linked to positive sentiments reflect positive interactions with healthcare workers and general experience. Conversely, topics associated with negative sentiment typically reflect adverse experiences within the healthcare system. Notably, the most pronounced positive sentiments are connected to the dynamics of patient-caregiver relationships, rather than clinical outcomes.

Negative aspects of *Healthcare Environment, Operations, and Administration*, such as *cost*, and *waiting*, demonstrate a strong association with adverse emotional responses. In contrast, positive aspects within the same theme, notably initiatives aimed at enhancing patient engagement and education elicit strong positive emotional reactions. Clinical care elements, particularly chronic and mobility-limiting conditions, are often associated with negative sentiments. Interestingly, discussions related to addiction recovery exhibit a trend towards positivity. This observation warrants cautious interpretation due to potential selection biases, as individuals in recovery might be more inclined to share their experiences on platforms such as Care Opinion.

These findings demonstrate the importance of interpersonal elements in shaping patient perceptions of care quality. They suggest that, alongside clinical outcomes, the humanistic dimensions of healthcare delivery significantly influence patient satisfaction. Healthcare

Table 5
Probabilistic modelling of emotions in simulated patient reports. This table displays the simulated report, the posterior of a positive sentiment, as well as the prior and posterior probabilities of the top three emotions identified by the model for each report.

Report	Positive Sentiment Posterior	Emotion	Emotion Prior	Emotion Posterior
Grandfather sadly passed away after a fall. Staff care was brilliant and family felt supported through a difficult time.	0.995	thankful grateful supported	0.071 0.050 0.032	0.194 0.155 0.139
Resident was left soiled and untreated for long periods of time while understaffed.	0.025	sad upset angry	0.005 0.021 0.022	0.071 0.057 0.046
The food was bland, cold, and made me feel sick.	0.117	hungry disappointed angry	0.001 0.024 0.022	0.086 0.065 0.043
The patient experienced a long wait at a cancer clinic that resulted in doctors identifying a problem. The appointment was rescheduled without the patient knowing, and they felt as if they were being treated as the problem.	0.021	inconvenienced frustrated annoyed	0.002 0.025 0.009	0.215 0.180 0.110
The patient was scared by the prospect of surgery, however, felt that the staff's help put them at ease. They express their gratitude towards a particular practitioner for providing a high level of information.	0.986	thankful grateful comfortable	0.071 0.050 0.026	0.128 0.107 0.077

providers and researchers can leverage this data to identify specific aspects of care that evoke strong emotional responses. This insight enables the development of targeted interventions to improve patient satisfaction and emotional well-being.

4.1.3. A landscape of patient-reported emotions

By conducting a dimension reduction of aggregate topic profiles for each emotion, we revealed the dichotomy between positive and negative emotions in an emotional landscape. Our results suggest that profound negative emotional states like *suicidal*, *depressed*, and *hopeless* are closely associated with experiences of neglect and dismissal, reinforcing the impact that emotional validation and supportive communication have on patient mental health.

The convergence of such severe emotional states with feelings of invalidation and rejection suggests that these external experiences can be significant contributors to, or exacerbations of, deep-seated emotional distress. This finding emphasises the need for healthcare providers to be acutely aware of the power of their interactions with patients. The data indicates that negative experiences, such as feeling unsupported or invalidated, can have far-reaching consequences, potentially culminating in extreme emotional responses like feelings of hopelessness or suicidal ideation.

This insight has practical implications for patient care strategies, suggesting that interventions aimed at fostering a sense of validation, support, and acceptance could be crucial in mitigating negative emotional outcomes. It also reinforces the importance of training healthcare professionals in emotional intelligence and communication skills, equipping them to recognise and respond to signs of such distress effectively.

On the positive side, the clustering of emotions such as *thankful*, *grateful*, and *respected* with *looked after* and *involved* reinforces the beneficial impact of respect, inclusion, and comprehensive care on patient satisfaction. These insights suggest that enhancing communication and emotional support in healthcare settings can substantially influence patient well-being and perceptions of care quality.

Additionally, the presence of typically negative emotion terms such as *nervous* or *apprehensive* within positive clusters may indicate a transition to positive states like relief, suggesting that the initial negative sentiment can be effectively addressed within a supportive healthcare environment. These findings support arguments for healthcare policies that prioritise emotional support and active patient engagement to foster positive patient experiences.

4.2. Probabilistic emotion recommender system

This paper contributes to the existing literature by developing and implementing a probabilistic emotion recommender system that functions in the context of patient-reported experiences. Through this paper, we have striven for a model that meets the following criteria:

1. High accuracy in both high-dimensional emotion recommendation and binary sentiment classification, with robustness to overfitting.
2. Interpretability and transparency in how predictions are made.
3. Accessibility, ease of use, and ready adoptability for healthcare researchers and practitioners.

The context-specific nature of this model has a stronger justification for use in health care than non-contextual models that may be too general to offer specific health-related insights, contributing to our improved accuracy. This is evidenced by the comparison made with existing sentiment analysis models. A Naïve Bayes model is a simple probabilistic model that is capable of high accuracy, particular robustness to overfitting, whose simplicity allows transparency and easy interpretation of the results. This interpretability is further enhanced when topics are used as features in the Naïve Bayes' model, as the feature space is significantly reduced to interpretable themes, and hence there are less moving parts to be interpreted. The dimension reduction that network topic modelling affords is inherently robust to overfitting, and also translates to both improved robustness to overfitting and improved accuracy in the final model. We quantitatively show this superior performance in both emotion recommendation and binary sentiment classification when compared to a full-vocabulary model, through evaluation metrics under 10-fold cross-validation. These evaluation metrics, including Q-measure and normalised discounted cumulative gain, have strong information-theoretic justifications that allow for partial relevance and penalise late arrival of relevant results.

In healthcare, the systematic evaluation of patient-reported experiences is critical for advancing quality improvement and fostering patient-centred care. The probabilistic emotion recommender system we have introduced allows for interpreting patient-reported experiences at both a granular emotional level, as well as at a binary sentiment level, offering a significantly more detailed analysis than what is typically achieved with binary sentiment analysis alone. This model has the potential to augment traditional patient-reported experience collection in healthcare by

- Offering healthcare providers and researchers with a more refined understanding of patient feedback by revealing emotional insights from patient narratives
- Facilitating a swift and accurate assessment of patient feedback that may aid in the timely and targeted response to patient needs.
- Integrating into existing or new public patient-feedback collection platforms to summarise patient feedback at scale, enhancing transparency of healthcare service performance at a public level for stronger accountability of healthcare services.
- Offering a new pathway to systematically distil the collective voice of patients into insights that may shape care-evaluation and decision-making processes.

While the full integration of this system into healthcare practice would require further research and development, we have identified potential pathways for implementation. These could include

- Integration in the analysis of existing patient-experience surveys that incorporate open-ended responses
- Real-time analysis of patient feedback from various sources, such as social media and online review platforms.
- Application in policy evaluation to assess the emotional impact of new healthcare interventions or policies based on catalogued interviews or open-ended survey responses undertaken during evaluation, or based on unsolicited views shared on social media platforms.
- Longitudinal tracking of patient-reported emotions to monitor changes in patient experience over time, either through the introduction of specific open-ended questions in patient experience surveys or analysis of unsolicited commentary on patient review platforms.

To enable such integrations, we have made our model available as an R package (pers) [53] and online dashboard [51], aiming to facilitate adoption by healthcare researchers and practitioners. This accessibility enables continued exploration and refinement of these techniques in real-world healthcare settings.

4.3. Limitations

This study, while offering valuable insights into patient-reported experiences and emotions in healthcare, is subject to several limitations that warrant careful consideration.

4.3.1. Data limitations

Single source constraints. Our reliance on patient narratives from the Australian online platform Care Opinion may limit the generalisability of our findings. While this data source is valuable for understanding patient experiences in the Australian healthcare context, it may not fully represent the diverse range of patient experiences across different healthcare systems or cultures.

Selection and demographic biases. The use of an online platform introduces potential biases towards more technologically inclined individuals, potentially underrepresenting older adults, rural populations, or those with limited internet access. This selection bias may skew the Care Opinion corpus towards certain age groups, socioeconomic statuses, or geographic locations, potentially making it non-representative of the broader patient population.

Self-selection biases. The likelihood of individuals with more extreme sentiments—either highly positive or negative—being more motivated or able to share their experiences introduces survivorship bias. This bias may skew our corpus towards success stories or particularly challenging experiences, potentially distorting the overall landscape of patient-reported emotions.

Survivorship bias. Narratives on Care Opinion are influenced by survivorship bias. Topics like palliative care, and events that lead to severe impairment and death, are likely to be underrepresented from a patient's perspective in Care Opinion, as patient's may not have an opportunity to share their experience. For example, experiences relating to addiction often show a strong association with positive sentiments. However, individuals who are successfully managing their addiction may find it more feasible to share their stories online, potentially overshadowing the experiences of those still struggling, leading to overly optimistic conclusions about the recovery process.

Language and cultural biases. By focusing on English-language posts, our study may exclude the experiences of non-English speaking patients or those with limited English proficiency. Cultural differences in expressing emotions or discussing healthcare experiences may not be fully captured by our current approach, potentially limiting the cross-cultural applicability of our findings.

Temporal bias. Our study captures patient experiences between 2012 and 2022, which may be influenced by contemporary healthcare policies, societal events (such as the COVID-19 pandemic), or evolving trends in patient expectations. These temporal factors could impact the nature of reported experiences and emotions, potentially limiting the long-term applicability of our findings.

4.3.2. Methodological limitations

Quantitative approach for qualitative experiences. Our study's quantitative approach, while robust in many respects, may not capture the full depth and nuance of qualitative patient experiences. The transformation of rich, narrative data into quantifiable metrics risks losing the subtlety and individuality of patient experiences. We argue that while this is indeed a limitation, accurate structured summaries of individual experience is a significant improvement to not considering individual experiences at all, which may often be the case for large-scale analyses.

Topic modelling and emotion classification challenges. While powerful, our approach of using topic modelling as a dimension reduction of documents may oversimplify complex patient narratives. The nuanced context of healthcare experiences may not always be fully captured by these methods. The simplification topic modelling offers helps to improve generalisability, but is at the cost of increasing epistemic uncertainty in our emotion modelling.

Subjectivity of emotions and sentiment analysis. Healthcare is inherently personal, and what constitutes a positive experience for one patient may be neutral or negative for another. The inherently subjective nature of emotions and the potential for different emotional responses to the same experiences contribute to aleatoric uncertainty.

Naïve Bayes, assumptions and underfitting. The Naïve Bayes model makes a class conditional independence assumption between features, which is unlikely to be true in general. Additionally, whilst Naïve Bayes is unlikely to overfit, it may be prone to underfitting. Despite these limitations, we show that our model performs well, offers probabilistic predictions, is robust to overfitting, and fosters interpretability that is often desired in clinical settings.

Validation challenges. The lack of a standardised benchmark dataset for healthcare-specific emotion classification makes it difficult to compare our model's performance directly with other existing approaches in the medical informatics literature at an emotion level.

4.3.3. Practical and implementation limitations

Integration with existing healthcare systems, user adoption, and the real-world impact on healthcare outcomes are aspects that require careful consideration and remain to be thoroughly evaluated. Additionally, ethical and privacy considerations in handling confidential patient data within a healthcare system are of critical importance and require diligent attention to ensure responsible use of technology.

These limitations suggest areas for future research and refinement in the analysis of patient-reported experiences and emotions in healthcare settings. In addition, they also reveal the complexity of translating computational approaches to real-world healthcare applications and the need for continued interdisciplinary collaboration in this field.

4.4. Future research

Future research could expand upon these findings, employing longitudinal data to track the evolution of patient experiences over time and through various healthcare reforms. For instance, tracking topic densities associated with known positive and negative emotions, as well as the prevalence of specific patient-reported emotions, or overall sentiment, can supplement traditional healthcare performance monitoring. This could be especially effective in response to specific healthcare interventions or policy changes. Additionally, comparative studies between different healthcare systems or regions could offer insights into systemic influences on patient experiences. Further, implementing and evaluating the probabilistic emotion recommender system in real healthcare settings would provide practical insights into its effectiveness, user adoption, and impact on healthcare outcomes. Investigating how the probabilistic emotion recommender system can be used to tailor healthcare services to individual patients, such as through personalised interventions based on patient-reported emotions. Currently, our emotion recommender system utilises all 263 labelled emotions that appear on Care Opinion. In the future, if patients express a wider variety of more nuanced emotions on Care Opinion, our model may benefit from being retrained to include these as well.

While our study reveals certain aspects of patient care, we acknowledge its limitations, including potential biases inherent in self-reported data and the need for broader demographic representation. Further studies should aim to corroborate these insights with a more diverse patient cohort.

5. Conclusion

This study makes contributions to the evolving landscape of healthcare, where understanding and integrating patient experiences and emotions are becoming increasingly crucial for quality care. Our main contributions are (1) an analysis of the free-text narratives on the popular healthcare review website Care Opinion, and (2) the development and implementation of a probabilistic emotion recommender system.

In our analysis of patient experiences on Care Opinion, we conduct topic modelling to summarise patient experience into topics, broadly relating to themes in clinical care, patient experience, and healthcare logistics. By capturing relationships between patient-reported emotions and their aggregate thematic composition, we reveal numerous insights into patient-reported experiences that can help inform healthcare practitioners and researchers. We show that topics linked with both extremely high and low degrees of positive sentiment are most closely related to areas in subjective patient experience, such as interactions with healthcare staff, rather than outcomes of clinical care. Additionally, we reveal a landscape of patient-reported emotions that helps contextualise relationships between specific patient-reported emotions. For instance, we show that extremely negative self-reported emotion terms that capture intrinsic states, such as suicidality, depression, and hopelessness, exhibit a strong relationship with negative emotions relating to experiences such as dismissal, invalidation, and rejection. Our findings indicate that patient experience is greatly improved through positive patient-caregiver interactions and initiatives that increase patient engagement, such as educational programs. Additionally, we show that positive patient-caregiver interactions are associated with mitigating negative emotions such as apprehension. Our study also shows that strong states of negativity are present in experiences with chronic and mobility-limiting conditions, however, experiences in healthcare relating to addiction tend to be framed positively.

In addition to the patient narrative analysis, we develop a probabilistic emotion recommender system from the Care Opinion reports and their corresponding patient-reported emotion labels. This context-specific tool, leveraging topic modelling in a probabilistic Naïve Bayes approach, demonstrates a high degree of interpretability and transparency, as well as superior performance in emotion recommendation and binary sentiment classification compared to comparator and baseline models, including standard sentiment lexicons. This is seen through our quantitative evaluations, including 10-fold cross-validation under appropriate metrics such as Q-measure and normalised discounted cumulative gain. The systematic distillation of complex narratives into thematic composition both (a) reduces the risk of overfitting, enhancing its applicability and generalisability in real-world healthcare settings and (b) improves interpretability through the use of meaningful features.

The model effectively identifies and classifies emotions in patient narratives with high interpretability, making it useful for healthcare providers seeking to understand patient feedback. This facilitates more nuanced patient care and service improvement by providing a deeper understanding of how patient experiences link with emotions in a patient-centred care framework. By offering healthcare professionals a transparent, probabilistic tool to comprehend patient feedback comprehensively, available as an R package and online dashboard, it opens avenues for more tailored and empathetic patient care strategies and allows for the augmentation of free-text comments on patient-reported experience surveys. This approach aligns with the current healthcare emphasis on patient-centeredness and helps to enhance patient-provider communication. Future research should focus on expanding the model's application to more diverse healthcare contexts and exploring its utility in longitudinal patient experience studies.

CRediT authorship contribution statement

Curtis Murray: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Lewis Mitchell:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Jonathan Tuke:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Mark Mackay:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Ethical approval

Not applicable. As this study involves the analysis of publicly available data from Care Opinion (Australia), which is an open platform where patients share their healthcare experiences, which are classified as a public resource intended for broader public benefit by Care Opinion, this study is not considered human subjects research. Therefore, no consent for this research was required. In line with Care Opinion's terms of re-use under the Creative Commons licence Attribution-NonCommercial-ShareAlike 4.0, our research strictly adheres to non-commercial scholarly purposes, ensuring that individual privacy is respected, identifying information is not disclosed, and any patient narratives that we mention are entirely simulated.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by The University of Adelaide, Australia, LM is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210103700).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced this study.

Acknowledgements

We would like to acknowledge the authors who shared their experiences on Care Opinion. In addition, we would like to thank the staff at Care Opinion for creating an open platform for patients to voice their experiences.

Appendix A. Probabilistic emotion recommender system

For the purposes of this appendix, when we refer to probabilities in shorthand, where the random variable is assumed. For example, $p(A = a|B = b)$ is written succinctly as $p(a|b)$ and the random variables are inferred by context.

Probabilistic modelling of the emotional distribution of text can be conducted from the posterior distribution $E|d \sim p(e|d)$, where e is an emotion and d is a document in bag-of-words representation. This is equivalent to $E|d \sim p(e|\mathbf{w})$, where $\mathbf{w}$ is the bag-of-words vector, or $E|d \sim p(e|w_1, \dots, w_{|V|})$, where w_i is the word count of the word at index i in vocabulary V. Using Bayes' Theorem, we rewrite the posterior distribution as

$$p(e|\mathbf{w}) = \frac{p(\mathbf{w}|e)p(e)}{\sum_{j=1}^{n} p(\mathbf{w}|e_j)p(e_j)},$$

where the priors $p(e_j)$ are taken as the maximum likelihood estimates in an empirical Bayesian approach [57].

Since $|V|$ is typically in the order of thousands or tens of thousands, the feature space is large, making density estimates have poor generalisability [58]. A typical simplification that is often made in Natural Language Processing is an assumption of class-conditional independence of words in a Naïve Bayes approach [59].

Under this assumption, the conditional density is

$$\begin{aligned} p(\mathbf{w}|e) &= p(w_1|w_2, \dots, w_n, e) \cdots p(w_n|e) \\ &= p(w_1|e) \cdots p(w_n|e) \\ &= \prod_{\{i: w_i \neq 0\}} p(w_i|e) \end{aligned}$$

So that

$$p(e|\mathbf{w}) = \frac{\left(\prod_{\{i: w_i \neq 0\}} p(w_i|e)\right) p(e)}{\sum_{j=1}^{n} \left(\prod_{\{i: w_i \neq 0\}} p(w_i|e_j)\right) p(e_j)} \tag{A.1}$$

However, even with the reduction in the dimensionality of the feature space, the model is not particularly parsimonious, having $|V| \times |E|$ parameters, and may cause overfitting. Topic modelling can act as a dimension reduction tool, taking high dimensional documents, to relatively lower dimensional topic mixtures $TM : \mathbb{N}^{|V|} \to [0,1]^{n_t}$, $\mathbf{w} \mapsto \mathbf{t}$, where $\mathbf{t}$ is a n_t-dimensional vector of topic densities $p(t_k|\mathbf{w})$, and topics are $|V|$-dimensional word mixtures with elements $p(v_i|k)$ for topics T, with $v_i \in V$.

Under the dimension reduction, we can estimate

$$p(w_i|e) = \sum_{k=1}^{n_t} p(v_i|k, e)^{w_i} p(k|e)^{w_i}, \tag{A.2}$$

where the topic-emotion distributions from Eq. (1). Note the use of the exponent w_i in Eq. (A.2), the word count of word v_i, to account for words with word counts exceeding one.

Topic models such as those found through network-based topic modelling allow topics to partition the vocabulary, i.e. words belong to exactly one topic [36]. This allows the summation in (A.2) to collapse to the term with non-zero probability. This term corresponds to the topic that word w_i belongs to, say, k_{w_i};

$$p(w_i|e) = p(v_i|k_{w_i}, e)^{w_i} p(k_{w_i}|e)^{w_i}.$$

Further, we make an assumption that words are conditionally independent to emotions given their topic, that is to say, that

$$p(v_i|k_{w_i}, e) = p(v_i|k_{w_i}).$$

This gives,

$$p(e|\mathbf{w}) = \frac{\left(\prod_{\{i: w_i \neq 0\}} p(v_i|k_{w_i})^{w_i} p(k_{w_i}|e)^{w}_{i}\right) p(e)}{\sum_{j=1}^{n} \left(\prod_{\{i: w_i \neq 0\}} p(v_i|k_{w_i})^{w_i} p(k_{w_i}|e_j)^{w_i}\right) p(e_j)},$$

which can be simplified by noticing that the term

$$\prod_{\{i: w_i \neq 0\}} p(v_i|k_{w_i})^{w_i}$$

can be pulled out from the summation in the denominator as it does not depend on j, and hence cancels with the corresponding term in the numerator (it is always non-zero as v_i is of topic k_{w_i} by construction),

$$p(e|\mathbf{w}) = \frac{\left(\prod_{\{i: w_i \neq 0\}} p(k_{w_i}|e)^{w_i}\right) p(e)}{\sum_{j=1}^{n} \left(\prod_{\{i: w_i \neq 0\}} p(k_{w_i}|e_j)^{w_i}\right) p(e_j)}. \tag{A.3}$$

Numerical underflow as a result of repeated multiplication of near-zero numbers has the potential to influence these results. In order to combat this, we employ the log-sum-exp trick to avoid numerical underflow [60]. This is illustrated below.

By taking the log of the numerator in Eq. (A.3), the log of the products can be exchanged for the sum of the logs;

$$\begin{aligned} &\log\left(\left(\prod_{\{i: w_i \neq 0\}} p(k_{w_i}|e)^{w_i}\right) p(e)\right) \\ &= \sum_{\{i: w_i \neq 0\}} \log\left(w_i \; p(k_{w_i}|e)\right) + \log\left(p(e)\right) \end{aligned}$$

This exchange prevents the numerator from underflowing as the repeated product of near-zero probabilities is avoided. If we seek to find $\log(p(e|\mathbf{w}))$ by employing a similar strategy in the denominator,

$$\begin{aligned} \log(p(e|\mathbf{w})) = &\sum_{\{i: w_i \neq 0\}} \log\left(w_i \; p(k_{w_i}|e) + \log(p(e))\right) - \\ &\log \sum_{j=1}^{n} \exp\left(\sum_{\{i: w_i \neq 0\}} \log\left(w_i \; p(k_{w_i}|e_j)\right) + \log(p(e_j))\right), \end{aligned} \tag{A.4}$$

we see that the latter term, say U,

$$\begin{aligned} U = \\ \log \sum_{j=1}^{n} \exp\left(\sum_{\{i: w_i \neq 0\}} \log\left(w_i \; p(k_{w_i}|e_j)\right) + \log(p(e_j))\right) \end{aligned} \tag{A.5}$$

has the potential to underflow due to the exponentiation of large negative numbers that result from the summation of the log of many near-zero numbers. Fortunately, we can avoid this numerical underflow as follows. Firstly, let S_j denote the terms inside the exponentials of U,

$$S_j = \sum_{\{i: w_i \neq 0\}} \log\left(w_i \; p(k_{w_i}|e_j)\right) + \log(p(e_j))$$

so that

$$U = \log \sum_{j=1}^{n} \exp(S_j), \tag{A.6}$$

where each of the S_j is negative by definition. By subtracting the largest of these negative sums,

$$\hat{S} = \max_j S_j,$$

from within each of the exponentials of U in Eq. (A.6), and accounting for this with a tactful multiplication of the corresponding exponential

$$\begin{aligned} U &= \log \sum_{j=1}^{n} \exp(S_j - \hat{S}) \exp\left(\hat{S}\right) \\ &= \hat{S} + \log \sum_{j=1}^{n} \exp(S_j - \hat{S}), \end{aligned} \tag{A.7}$$

numerical underflow is avoided, as the most dominant term is captured directly without needing to exponentiate the log of a small number. In full, this gives the log posterior as

$$\begin{aligned}\log\left(p(e|\mathbf{w})\right) = &\\ \sum_{\{i:w_i\neq 0\}} \log\left(w_i\ p(k_{w_i}|e)\right) + \log\left(p(e)\right) - \hat{S} - &\\ \log\sum_{j=1}^{n}\exp\left(\sum_{\{i:w_i\neq 0\}}\log\left(w_i\ p(k_{w_i}|e_j)\right) + \log(p(e_j)) - \hat{S}\right).\end{aligned} \tag{A.8}$$

Additionally, as there is the potential for only few documents being associated with a particular emotion e, it may arise that $p(k_{w_i}|e) = 0$ for some topics k, resulting in posterior $p(e|\mathbf{w})$ collapsing to zero whenever the word v_i appears. We avoid this by adding a small, non-zero number to each $p(k_{w_i}|e) = 0$.

A.1. Evaluation metrics

When considering documents labelled with multiple emotions, the gain for the rth ranked emotion, $g(r)$, is computed as follows:

$$g(r) = \max_{e_l\in L} \mathrm{rel}(e_r, e_l),$$

where e_r is the ranked emotion, and the calculation seeks the highest relevance score between e_r and the most closely related label emotion, e_l, in the given set of labels L.

Q-measure. Q-measure is a generalisation of average precision in a binary setting to account for partial relevance, defined as

$$Q(k) = \frac{1}{k}\sum_{r=1}^{k} \mathbb{1}_{\mathrm{rel}(r)>0}\mathrm{Br}(r), \tag{A.9}$$

where $Br(r)$ is the *blended ratio*,

$$Br(r) = \frac{\mathrm{cg}(r) + \sum_{i=1}^{r}\mathbb{1}_{g(i)>0}}{\mathrm{cg}_I(r) + r}, \tag{A.10}$$

$cg(r)$ is the *cumulative gain* of the top r ranked emotions,

$$\mathrm{cg}(r) = \sum_{i=1}^{r} g(i), \tag{A.11}$$

and $\mathrm{cg}_I(r)$ is the cumulative gain of the top r ranked emotions in an ideal ranking, i.e. that which ranks results with non-increasing relevance for increasing rank.

Normalised discounted cumulative gain (nDCG). Another metric that considers partial relevance, as well as penalises the late arrival of relevant documents is nDCG. We first introduce discounted cumulative gain (DCG), which also allows for the penalisation of the late arrival of relevant documents by *discounting* the rth ranked result's gain according to the rank. Traditionally, the discounting factor is logarithmic, and relevant retrieval is more strongly emphasised by exponentiating the gain,

$$\mathrm{DGC}_r = \sum_{i=1}^{r}\frac{2^{g(i)} - 1}{\log_2(i+1)}.$$

DCG should not be used to grade information retrieval systems across differing queries, as some queries may have more (or less) potential relevant results to select from, resulting in a greater (or lesser) DCG. To account for this, DCG is normalised by the *ideal* discounted cumulative gain (IDCG), the maximum DCG achievable, resulting from an ideal document ranking with non-increasing relevance as rank increases,

$$\mathrm{nDCG}_r = \frac{\mathrm{DCG}_r}{\mathrm{IDCG}_r}.$$

Appendix B. Calculating the posterior distribution of emotions given a document

Here we illustrate how to compute the vector of emotion-densities

$$\left[p(E = e_j|\mathbf{w})\right]_j$$

using the network topic model in Fig. 1 where we use words from document d_1; $\mathbf{w_1} = (2, 1, 0, 1)$.

First, we find the empirical densities of each emotion class. For each emotion, this is the number of edges out of the emotion, divided by the total number of edges out of all emotions. For example, there are two edges out of emotion e_1, out of the combined five edges out of all emotions, so $p(E = e_1) = \frac{2}{5}$. In full, the column vector $[p(E = e_j)]_j$ is

$$\left[p(E = e_j)\right] = \begin{bmatrix}\frac{2}{5}\\ \frac{1}{5}\\ \frac{2}{5}\end{bmatrix}. \tag{B.1}$$

The matrix $\left[p(T = t_i|d_j)\right]_{i,j}$ shows the empirical topic-class use for each document. For example, Document 1 uses topic t_1 three times (two occurrences of word v_1 and one of v_2), and topic t_3 once (v_4 is used once). This gives $p(T = t_1|d_1) = \frac{3}{4}$. In full, we find

$$\left[p(T = t_i|d_j)\right]_{i,j} = \begin{bmatrix}\frac{3}{4} & 1 & \frac{2}{3} & 0\\ 0 & 0 & \frac{1}{3} & \frac{2}{3}\\ \frac{1}{4} & 0 & 0 & \frac{1}{3}\end{bmatrix}.$$

The matrix $\left[p(d_i|e_j)\right]_{i,j}$ tells us the probabilities of selecting a random document d_i belonging to an emotion e_j. Since each document is assumed equally likely, each document has a probability equal to the inverse of the emotion's use:

$$\left[p(d_i|E = e_j)\right]_{i,j} = \begin{bmatrix}\frac{1}{2} & 1 & 0\\ \frac{1}{2} & 0 & 0\\ 0 & 0 & \frac{1}{2}\\ 0 & 0 & \frac{1}{2}\end{bmatrix}.$$

This allows us to calculate $\left[p(t_i|e_j)\right]_{i,j}$:

$$\begin{aligned}\left[p(t_i|e_j)\right] &= \left[p(T = t_i|d_j)\right]_{i,j}\left[p(d_i|E = e_j)\right]_{i,j}\\ &= \begin{bmatrix}\frac{7}{8} & \frac{3}{4} & \frac{1}{3}\\ 0 & 0 & \frac{1}{2}\\ \frac{1}{8} & \frac{1}{4} & \frac{1}{6}\end{bmatrix}.\end{aligned}$$

Note that the implicit summation in the matrix multiplication here marginalises over documents. This likelihood, and the emotion-class densities in Eq. (B.1) can be used with the model in Eq. (A.3) to find the posterior distribution:

$$\left[p(E = e_j|d_1)\right]_j \approx \begin{bmatrix}0.59\\ 0.37\\ 0.04\end{bmatrix}.$$

References

[1] Commonwealth of Australia. Royal Commission into Aged Care Quality and Safety. Commonwealth of Australia; 2019, [Accessed 18 May 2022] URL.

[2] Németh G. Health related quality of life outcome instruments. Eur Spine J 2006;15(1):S44–51.

[3] Weldring T, Smith SM. Article commentary: Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Heal Serv Insights 2013;6:HSI.S11093. http://dx.doi.org/10.4137/HSI.S11093, PMID: 25114561. arXiv:https://doi.org/10.4137/HSI.S11093.

[4] Australian Institute of Health and Welfare. Patient experience of health care. 2017, [Accessed 18 May 2022]. URL https://web.archive.org/web/20220322052449/https://www.aihw.gov.au/reports/australias-health/patient-experience-of-health-care.

[5] Simon MK, Goes J. Scope, limitations, and delimitations. 2013.

[6] Fitzpatrick R, Boulton M. Qualitative research in health care: I. The scope and validity of methods. J Eval Clin Pract 1996;2(2):123–30.

[7] Vicsek L. Issues in the analysis of focus groups: Generalisability, quantifiability, treatment of context and quotations.. Qual Rep 2010;15(1):122–41.

[8] Morgan DL. In: Focus groups as qualitative research, vol. 16, Sage publications; 1996.

[9] Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare: Table 1. BMJ Qual Saf 2013;22(3):251–5. http://dx.doi.org/10.1136/bmjqs-2012-001527.

[10] Antoniak M, Mimno D, Levy K. Narrative paths and negotiation of power in birth stories. 2019.

[11] Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB. Natural language processing of reddit data to evaluate dermatology patient experiences and therapeutics. J Am Acad Dermatol 2020;83(3):803–8.

[12] Du C, Lee W, Moskowitz D, Lucioni A, Kobashi KC, Lee UJ. I leaked, then I reddit: experiences and insight shared on urinary incontinence by reddit users. Int Urogynecol J 2020;31(2):243–8.

[13] Murray C, Mitchell L, Tuke S, Mackay M. Symptom extraction from the narratives of personal experiences with COVID-19 on reddit. In: Workshop proceedings of the 15th international AAAI conference on web and social media, special edition on healthcare social analytics. 2021, http://dx.doi.org/10.36190/2021.71, The version of this paper was also published as a pre-print on arXiv with DOI: 10.48550/arXiv.2005.10454, URL: https://arxiv.org/abs/2005.10454.

[14] Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016;25(6):404–13.

[15] Bovonratwet P, Shen TS, Islam W, Ast MP, Haas SB, Su EP. Natural language processing of patient-experience comments after primary total knee arthroplasty. J Arthroplast 2021;36(3):927–34.

[16] Clark EM, James T, Jones CA, Alapati A, Ukandu P, Danforth CM, et al. A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. 2018, arXiv preprint arXiv:1805.09959.

[17] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint arXiv:1810.04805.

[18] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. OpenAI; 2018.

[19] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9.

[20] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

[21] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730–44.

[22] Wake N, Kanehira A, Sasabuchi K, Takamatsu J, Ikeuchi K. Bias in emotion recognition with ChatGPT. 2023, arXiv:2310.11753.

[23] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2).

[24] Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009.

[25] Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: A scoping review. J Biomed Inform 2023;140:104336. http://dx.doi.org/10.1016/j.jbi.2023.104336, URL https://www.sciencedirect.com/science/article/pii/S1532046423000576.

[26] Denecke K, Deng Y. Sentiment analysis in medical settings: New opportunities and challenges. Artif Intell Med 2015;64(1):17–27.

[27] Khanpour H, Caragea C. Fine-grained emotion detection in health-related online posts. In: Proceedings of the 2018 conference on empirical methods in natural language processing. 2018, p. 1160–6.

[28] Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. In: Proceedings of the 3rd IEEE/ACM international conference on big data computing, applications and technologies. 2016, p. 94–9.

[29] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3(Jan):993–1022.

[30] Serrano-Guerrero J, Bani-Doumi M, Romero FP, Olivas JA. Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions. Artif Intell Med 2022;128:102298. http://dx.doi.org/10.1016/j.artmed.2022.102298, URL https://www.sciencedirect.com/science/article/pii/S093336572200063X.

[31] Care Opinion. [Accessed 22 November 2022], https://www.careopinion.co.uk/.

[32] Cambria E, Hussain A, Cambria E, Hussain A. Senticnet. Sentic Comput: A Common- Sense- Based Fram Concept- Lev Sentim Anal 2015;23–71.

[33] Care Opinion Australia. [Accessed 22 November 2022] https://www.careopinion.org.au/.

[34] Murray C, Mitchell L, Tuke J, Mackay M. Revealing patient-reported experiences in healthcare from social media using the DAPMAV framework. 2022, arXiv preprint arXiv:2210.04232.

[35] Silge J, Robinson D. Text mining with R: A tidy approach. " O'Reilly Media, Inc."; 2017.

[36] Gerlach M, Peixoto TP, Altmann EG. A network approach to topic models. Sci Adv 2018;4(7):eaaq1360.

[37] Hyland CC, Tao Y, Azizi L, Gerlach M, Peixoto TP, Altmann EG. Multilayer networks for text analysis with multiple data types. EPJ Data Sci 2021;10(1):33.

[38] Peixoto TP. Nonparametric Bayesian inference of the microcanonical stochastic block model. Phys Rev E 2017;95(1):012317.

[39] Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. 1999, p. 50–7.

[40] Peixoto TP. Bayesian stochastic blockmodeling. Adv Netw Clust Blockmodeling 2019;289–332.

[41] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. 2018, arXiv preprint arXiv:1802.03426.

[42] Nielsen FÅ. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Rowe M, Stankovic M, Dadzie A-S, Hardey M, editors. Proceedings of the ESWC2011 workshop on 'making sense of microposts': big things come in small packages. CEUR workshop proceedings, vol. 718, 2011, p. 93–8, URL http://ceur-ws.org/Vol-718/paper_16.pdf.

[43] Hutto C, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, (1):2014, p. 216–25, URL DOI: 10.1609/icwsm.v8i1.14550.

[44] Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh international conference on language resources and evaluation. LREC'10, Valletta, Malta: European Language Resources Association (ELRA); 2010, URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[45] Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Comput Intell 2013;29(3):436–65.

[46] Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. 2004, p. 168–77.

[47] Murray C. Supplementary materials for "probabilistic emotion and sentiment modelling of patient-reported experiences". 2024, [Accessed 04 January 2024]. https://github.com/curtis-murray/persSupplementaryMaterials.

[48] COVID-19: Impacts on health and the Australian health system. 2022, [Accessed 05 January 2023]. URL https://www.aph.gov.au/About_Parliament/Parliamentary_departments/Parliamentary_Library/pubs/BriefingBook47p/PandemicHealthSystem.

[49] Ramsey LP, Sheard L, Lawton R, O'Hara J. How do healthcare staff respond to patient experience feedback online? a typology of responses published on care opinion. Patient Exp J 2019;6(2):42–50.

[50] Garcia D, Garas A, Schweitzer F. Positive words carry less information than negative words. EPJ Data Sci 2012;1(1):1–12.

[51] Murray C. Pers: Probabilistic emotion recommender system interactive shiny dashboard. 2022, https://curtismurray.shinyapps.io/pers/.

[52] Team RC, et al. R: A language and environment for statistical computing. 2013.

[53] Murray C. Pers: Probabilistic emotion recommender system (r package). 2022, https://github.com/curtis-murray/pers.

[54] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press; 2008.

[55] Zuva K, Zuva T. Evaluation of information retrieval systems. Int J Comput Sci Inf Technol 2012;4(3):35.

[56] Sakai T. On the reliability of information retrieval metrics based on graded relevance. Inf Process Manage 2007-03;43(2):531–48. http://dx.doi.org/10.1016/j.ipm.2006.07.020, URL https://linkinghub.elsevier.com/retrieve/pii/S0306457306001129.

[57] Collins M. The naive bayes model, maximum-likelihood estimation, and the em algorithm. In: Lecture Notes, 2013, URL http://Www.cs.columbia.edu/mcollins/em.Pdf.

[58] Hastie T, Tibshirani R, Friedman JH, Friedman JH. In: The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer; 2009.

[59] Mitchell TM, Mitchell TM. In: Machine learning, vol. 1, (9). McGraw-hill New York; 1997.

[60] Murphy KP, et al. Naive bayes classifiers. Univ Br Columbia 2006;18(60):1–8.