frontiers in Digital Health

ORIGINAL RESEARCH
published: 06 December 2021
doi: 10.3389/fdgth.2021.779091

Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service

Zhaolu Liu[1], *Robert L. Peach*[1,2,3], *Emma L. Lawrance*[4,5], *Ariele Noble*[5], *Mark A. Ungless*[5] *and Mauricio Barahona*[1]*

[1] Department of Mathematics, Imperial College London, London, United Kingdom, [2] Department of Neurology, University Hospital Würzburg, Würzburg, Germany, [3] Department of Brain Sciences, Imperial College London, London, United Kingdom, [4] Institute of Global Health Innovation, Imperial College London, London, United Kingdom, [5] Mental Health Innovations, London, United Kingdom

OPEN ACCESS

Edited by:
Angus Roberts,
King's College London,
United Kingdom

Reviewed by:
Brett South,
IBM Watson Health, United States
Jianlin Shi,
The University of Utah, United States

***Correspondence:**
Mauricio Barahona
m.barahona@imperial.ac.uk

Specialty section:
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

Received: *17 September 2021*
Accepted: *12 November 2021*
Published: *06 December 2021*

Citation:
Liu Z, Peach RL, Lawrance EL, Noble A, Ungless MA and Barahona M (2021) Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service. Front. Digit. Health 3:779091. doi: 10.3389/fdgth.2021.779091

The current mental health crisis is a growing public health issue requiring a large-scale response that cannot be met with traditional services alone. Digital support tools are proliferating, yet most are not systematically evaluated, and we know little about their users and their needs. Shout is a free mental health text messaging service run by the charity Mental Health Innovations, which provides support for individuals in the UK experiencing mental or emotional distress and seeking help. Here we study a large data set of anonymised text message conversations and post-conversation surveys compiled through Shout. This data provides an opportunity to hear at scale from those experiencing distress; to better understand mental health needs for people not using traditional mental health services; and to evaluate the impact of a novel form of crisis support. We use natural language processing (NLP) to assess the adherence of volunteers to conversation techniques and formats, and to gain insight into demographic user groups and their behavioural expressions of distress. Our textual analyses achieve accurate classification of conversation stages (weighted accuracy = 88%), behaviours (1-hamming loss = 95%) and texter demographics (weighted accuracy = 96%), exemplifying how the application of NLP to frontline mental health data sets can aid with *post-hoc* analysis and evaluation of quality of service provision in digital mental health services.

Keywords: mental health, crisis, deep learning, transformers, digital mental health, natural language processing, machine learning

1. INTRODUCTION

Experiences of mental health difficulties and emotional distress are increasing globally and, according to the World Health Organisation, 700,000 people die by suicide world-wide each year. Even before the COVID-19 pandemic, the number of individuals meeting diagnostic criteria for emotional disorders or self-harming was growing in the UK, particularly among young women (1, 2). The COVID-19 pandemic has only worsened this troubling picture with increased

distress and mental health strain experienced by many individuals, and health systems struggling to meet demand (3). Yet there is still much we do not know about the "who," "what," and "how" of experiences of distress and mental health difficulties.

Evidence-based, good quality, timely and accessible support is needed to reach people in moments of distress and ensure they are heard, cared for, and equipped with appropriate resources. However, traditional mental health services, such as talking therapies provided through the UK National Health Service, cannot meet current demand, nor can they reach people in distress in a discreet and accessible format. To provide around-the-clock support, digital mental health services are increasingly being developed as an accessible alternative, or as adjunct resources to traditional face-to-face services. While such services hold great promise, and various digital mental health apps and tools continue to enter the market, few resources are properly evaluated, holding back progress in developing high quality support (4, 5).

The data from non-traditional digital mental health support services offers a unique window into mental health needs and experiences of distress, and provides an opportunity to evaluate personalised support approaches. In particular, big data analysis techniques, such as natural language processing (NLP) and machine learning, allow us to examine, at scale, expressions of distress and user interactions with mental health services (6). To date, the challenges involved in accessing relevant data sets have partly prevented a full exploration of the value of such techniques for mental health insights and service evaluation [see (7) for a notable exception]. The present work is an exception resulting from an in-depth collaboration between a research organisation (Imperial College London) and a frontline mental health service provider (Mental Health Innovations, MHI) with direct access to data, experts and volunteers.

In an effort to support individuals facing mental health difficulties, MHI launched Shout in 2019 as a 24/7 mental health crisis text line available to anyone in the United Kingdom (see mentalhealthinnovations.org/impact-report-2021). The service connects highly trained volunteers with texters in distress. The text conversations aim to guide the texter to a calmer state and identify appropriate next steps. The volunteers are trained to follow a six-stage conversation structure that includes, in order: conversation initialisation (initialise); building rapport with the texter (build rapport); identifying the challenges that brought them to Shout (explore); identifying helpful next steps (identify goal); creating action plans (problem-solve); and closing the conversation (end the conversation). All conversations are overseen by clinically-trained supervisors, and over 900,000 conversations have been completed to date. Shout provides a discreet, anonymised, accessible and confidential mechanism to support individuals in distress, including those who are not in circumstances to express their feelings verbally but prefer texting for help. The textual data from each conversation is collected and undergoes an anonymisation process. Additional metadata and satisfaction ratings are collected from post-conversation surveys of volunteers and texters.

As the size of the Shout data set continues to grow, its value for understanding the mental health condition of the UK population increases. Additionally, there is a need to monitor and continually improve the quality of the service, while simultaneously increasing its scale as demand rises. Big data analysis methods offer the potential to understand at scale the texter population, their behaviour and experiences of the Shout service, and, in the long term, such models could potentially help monitor conversation progress in real-time.

Recent advances in NLP offer a valuable opportunity to interrogate the textual data collected by Shout (8–11). NLP is already present in our mobile phones and computers for tasks such as predictive text (12) or for translating between languages (13). Recent NLP models learn an embedding (a numeric vector) from the textual data (words, sentences, paragraphs) which encodes contextual information. The quality of the output embedding depends on the language model, on the data used for its training, and the data to which it is applied (which should not be hugely dissimilar to the training data). The embedding can then be used for downstream tasks such as predicting expressed sentiment (14) or, in our case, predicting aspects about mental health conversations, such as texter demographics or behaviours.

NLP has been used in the context of medicine and healthcare, e.g., BioBERT (15) and Med-BERT (16). In the context of mental health, NLP has also been used for predicting suicidal ideation (17, 18), analysing post-traumatic stress disorder (19), predicting psychosis (20) and other disorders (21), generating artificial mental health records (22), and for motivational interviewing (23). For a more complete review of NLP in mental health, see (6). However, NLP models have been rarely applied to digital mental health resources such as crisis text line services (7), and to date no published studies analyse the Shout data set.

In this study, we use state-of-the-art NLP models to embed Shout conversation data followed by deep learning models to perform downstream tasks on these embeddings to better understand and evaluate the Shout service and its users. Specifically, we identified the following three focussed tasks:

1. Predicting the **conversation stages** of messages.
2. Predicting the **behaviours** present in messages, from both texters and volunteers.
3. Classifying full conversations to extrapolate **demographic information** in texter surveys.

The first task helps us evaluate whether the Shout volunteers follow the conversation structure which they have been trained to follow. Success at this task would provide us with a tool that could be used in future for the analysis of which conversation structures or sub-stages are important for predicting outcomes, and how the structure of conversations differ across different texter or topic subgroups. The second task allows for exploration of the expressions of distress by texters (helpful for the understanding of mental health crises), and behaviours used by volunteers (helpful for ascertaining what behaviours are relevant for conversation outcomes) and could also be used to guide further research in future. The third task provides insights into the

true user-base for Shout, controlling for the possible bias in survey completion across demographics. Understanding the true user demographic of Shout can help MHI to provide specific guidance to volunteers or, indeed, partner with other charities that can signpost appropriate resources more specifically target to particular subpopulations of users. It can also help identify which groups of texters Shout might not be reaching, allowing MHI to better target these underrepresented groups. In our study, we chose three demographic classes for prediction: "aged 13 and under," "autism diagnosis," and "non-binary gender." These groups were suggested by MHI experts because they are either known to be higher risk for mental health challenges and may require tailored support or resources (for autism and non-binary gender), or for safe guarding (in the case of children 13 and under). Moreover, survey participation bias has previously been observed for gender (24, 25), age (26, 27), and disability (28, 29).

2. MATERIALS AND METHODS

In this study, we aim to construct models that can learn the relationship between mental health conversation context (message text) and the labels of interest (such as survey outcomes or texter behaviour). In this section, we briefly introduce two essential components of our study: (1) the Shout data set and (2) the NLP models.

After providing some high-level statistics for the Shout data set, we describe the steps used to clean and filter the data set. We then describe the ground truth labels at the message and conversation levels that the NLP models will be trained to predict. Finally, we describe the necessary pre-processing steps for use of the conversation text within our NLP models.

For NLP and embedding of textual data, we use the Longformer model pre-trained on a large corpus of publicly available text. We detail the additional pre-training, where we updated the pre-trained weights using text from the Shout data set. We then describe the three experiments we performed in this study, where we fine-tuned the Longformer model to classify the chosen ground truth labels. We also describe hyper-parameter optimisation and our model evaluation measures.

2.1. Data

2.1.1. Ethics

The study received ethical approval (19IC5511) and the data set was anonymised by MHI prior to researcher access. More explicitly, anonymisation removed any personally identifiable information such as names, phone numbers and addresses, by comparison with a curated word list. Moreover, researchers were trained in security and ethical considerations and were only able to access the data on a secure server.

2.1.2. Shout Service

The data was provided by Shout, a mental health crisis text line launched publicly by the digital mental health charity Mental Health Innovations in May 2019. Individuals of all ages who live in the UK can text 85258 to begin a conversation 24/7 whenever they are experiencing any mental or emotional distress and seeking help.

Once the conversation is initiated, the texter will receive an automatic reply from a bot asking for further information about the issues the texter is experiencing and stating the service terms and conditions. A trained Shout volunteer, who is supervised by a Clinical Supervisor, will be assigned the conversation and communicate with the texter following the initial messages. The volunteer will then respond to the texter and provide relevant support across six "conversation stages," which they have been trained to follow in order (details of the stages can be found in section 2.1.4.1). The conversation stages have been developed to help de-escalate a texter from a place of distress or overwhelm to a calmer place from where they can move forward.

After the conversation, texters are sent a link to an optional post-conversation survey that contains various questions about the texter's experience of the conversation; what brought them to Shout; and their demographic information. In our data sample, 13.75% of texters completed the survey on at least one of their conversations with Shout. The volunteer also completes a short survey after every conversation providing information about the key topic categories raised in the conversation, and whether the texter experienced and expressed any level of suicide risk.

2.1.3. Conversation Textual Data

Our analysis was conducted on data recorded from 12 February 2018 to 3 April 2020, comprising 271,445 conversations and 10,809,178 messages. Conversations were of varying length (**Figure 1A**). In very short conversations, it was assumed that the texter did not engage with the conversation, as the conversation would be largely comprised of standard system messages from the automated bot. Very long conversations are generally rare and atypical: long conversations typically appear under very high risk scenarios, and, when such risk is not present, are discouraged due to the limited availability of volunteers and the importance of closing the conversation so the texter can move forward with any agreed next steps.

In addition to the raw textual information, recorded meta-data includes the time-stamp, the actors (texter, volunteer or Clinical Supervisor), and Volunteer ID. To provide the reader with an intuition for the Shout conversations, we provide a short constructed example *emulating* a Shout conversation in **Figure 1E**.

2.1.3.1. Filtering the Conversations

Conversations were filtered to remove extremely short (texter did not engage) or extremely long conversations (high-risk scenarios); hence the peaks to the left of the red box and the tail to the right of the red box in **Figure 1A** were treated as outliers and removed (85,803 conversations). The resulting data set used for analyses in this paper is a sub-set (68.4%) of the total conversations which fell within the range of 81–2,129 words (as shown in **Figure 1B**).

2.1.4. Ground Truth Labels

We use supervised learning NLP models in this study, for which it is necessary to train and test the model against ground truth data. To train a model to predict conversation stages and texter/Volunteer behaviours (tasks 1 and 2), we use annotations

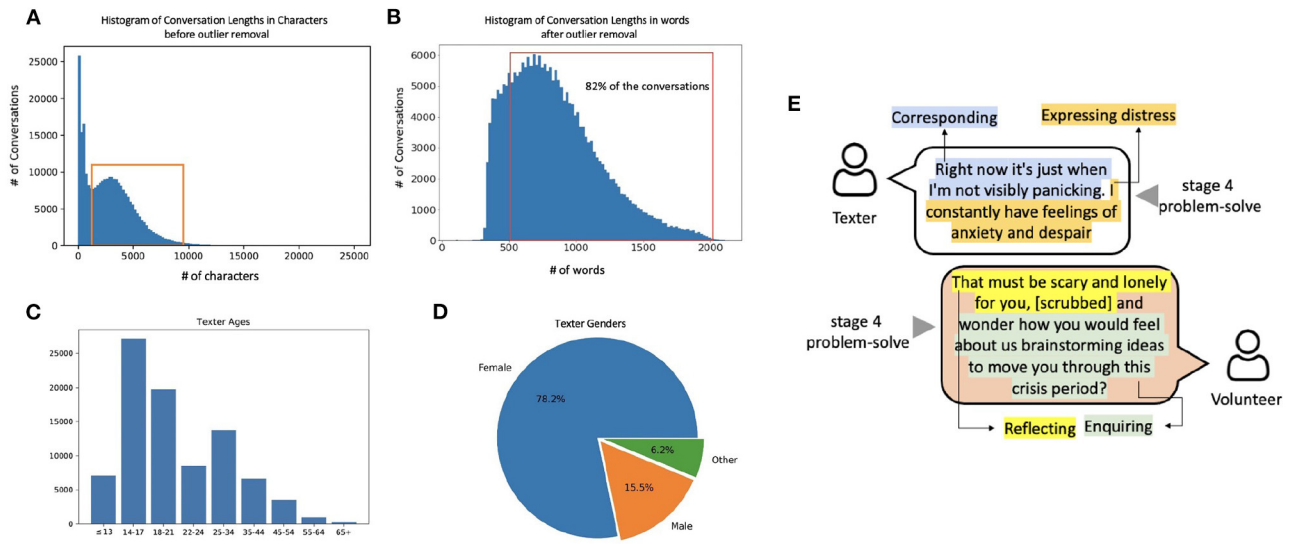


FIGURE 1 | High-level overview of the Shout data set. **(A)** The distribution of conversation lengths as measured by the number of characters. The red box indicates the conversations between 1,200 and 9,200 characters which passed the inclusion criteria for analysis. **(B)** The histogram of the number of words in a conversation. Since 82% of the conversations exceed the maximum word limit (512 words) of traditional BERT-style NLP models, we use the Longformer model. **(C)** Age distribution and **(D)** gender distribution of texters as obtained from the texter surveys (which we show in section 3.3 to be biased toward particular demographic groups). **(E)** A synthetic example of a portion of a conversation to illustrate annotations of behaviours and conversation stages.

at the individual message level, performed by expert Clinical Supervisor and research psychologist (AN). To train a model to predict texter demographics from the entire conversation content (task 3), we use as ground truth labels the texter survey responses available.

2.1.4.1. Message-Level Labels

A subset of 169 conversations (8,844 messages) was randomly selected from the 271,445 total conversations and annotated at the message level to indicate the corresponding "conversation stage." The annotation was performed by the Shout Clinical Supervisor and research psychologist, according to the definitions of each conversation stage established by Shout to train Volunteers. Each message belongs to one of six conversation stages listed in **Table 1**, with stage 0 incorporating preliminary engagement of the texter with the Shout service and automated bot, and stages one to five comprising the conversation with the Volunteer.

Similarly, the Shout clinical supervisor and research psychologist (AN) annotated the behaviours present in each message. A process of codebook development (30, 31) was applied in an inductive thematic analysis (32, 33). Codebook development is a well-established methodology in qualitative research, which allows for iterative strengthening of the definitions through interactions with other members of the team, and leads to a structured trail of evidence and enhanced replication of findings. Specifically, once working definitions were developed for each theme, two other MHI Clinical staff applied the codebook to the data. These additional Shout staff independently coded a subset of the data as a validity check (34, 35), and in the case of any initial coding discrepancies the definitions associated with relevant codes were updated and refined. Therefore, the codebook was iterated until there was convergence in coding agreement between the Shout researchers. As a result of this process of iteration, a total of six distinct behaviours were identified and defined, namely: setting intention, enquiring, expressing distress, reflecting, corresponding, and discord.

Messages in the 169 conversations comprising our sample were annotated with up to a maximum of 3 behaviours for each message, in the order in which they appear. We provide further definitions of the behaviours in **Table 2**.

TABLE 1 | Pre-defined conversation stages which Shout Volunteers are trained to follow to best support the texters.

Stage code	Name	Brief definition
0	Initialise	Greetings and bot messages
1	Build rapport	Active listening and good contact techniques.
2	Explore	Understand what is the crisis and asses risk.
3	Identify goal	Clarify what support the texter needs.
4	Problem-solve	Identify current resources and create an action plan.
5	End the conversation	Review the action plan and close warmly

2.1.4.2. Conversation-Level Labels

After the conversation, texters have the option to complete a survey that contains questions about the texter's demographic information (including age, gender identify, ethnicity and

TABLE 2 | Definitions of behaviours.

Behaviour code	Name	Brief definition
1	Setting intention	Communicate an immediate (near future) aim
2	Enquiring	Explore one's understanding of another's experience
3	Expressing distress	Communicate an offloading of negative feelings
4	Reflecting	Mirror something the texter has said
5	Corresponding	Show comparability between both parties
6	Discord	Involves a lack of harmony between both parties

disability), and several questions related to the conversation such as "Did you find the conversation helpful?". In our data set, only 13.75% of the conversations had an associated completed texter survey, meaning that there is large uncertainty as to the demographic profile of the texters. In addition, previous studies suggest there may be completion bias in texter surveys (36, 37); therefore it is unreliable to extrapolate the distribution of demographics from the texter survey to the entire data set. **Figures 1C,D** show the distributions of age and gender in the texter survey. We used the demographic labels in the available texter surveys to train a model to map conversation content to demographic labels.

Here, we focussed on three classification categories, namely (1) self-declared autism diagnosis; (2) self-declared non-binary gender; (3) aged 13 and under. These were chosen to both understand mental health needs in the population, and help Shout identify use of their service by groups for whom more tailored support and resources may be warranted.

2.1.5. Pre-processing

2.1.5.1. Special Tokens

The Shout data set is a collection of conversations that involve three "actors": the texters, the volunteers and the Shout bot (automated messages generated for initiating and ending conversations). The tokens "[texter]," "[Volunteer]," and "[bot]" are added in front of the messages from the corresponding speaker to separate the text of the different actors.

2.1.5.2. Data Augmentation for Messages

Due to the small sample size of annotated messages (169 conversations comprised of 8,844 messages), data augmentation techniques (e.g., word deletion, word swapping, and synonym replacement) were employed to boost performance and ensure robustness of the trained model (38). The augmented messages retain the same label as the original messages, as it has been empirically shown that augmented data obtained from simple deletion, swapping, and replacement inherit the label from the original labelled text (38).

2.2. Computational Methods

2.2.1. NLP: Longformer

The Longformer NLP architecture was used to model the conversation content (39) since 82% of the pre-processed conversations exceeded the 512 word limit of standard NLP models (10, 11). The Longformer model uses specially designed attention patterns (local attention and global attention) to cope with long word sequences. The Longformer model was implemented via Python using the Huggingface package (40), which comes with pre-trained weights (39), i.e., the model had already been trained with a large corpus of text comprised of a wide selection of non-domain specific documents found across the internet.

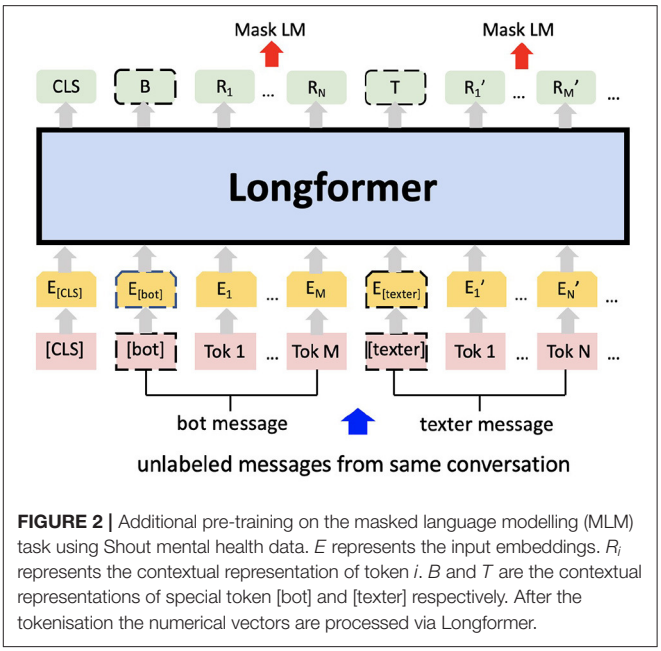


FIGURE 2 | Additional pre-training on the masked language modelling (MLM) task using Shout mental health data. *E* represents the input embeddings. R_i represents the contextual representation of token *i*. *B* and *T* are the contextual representations of special token [bot] and [texter] respectively. After the tokenisation the numerical vectors are processed via Longformer.

2.2.2. Additional Pre-training

In addition to the pre-training performed by Huggingface, the Longformer checkpoint (39) was further pre-trained on the masked language model (MLM) using the entire pre-processed Shout data set (see **Figure 2**). Words within a text sample are randomly masked; predicted from their surrounding context; and predictions evaluated against the actual masked word. This additional pre-training updates the weights to account for deviations of our data set from the original, non-specific text corpus used by Huggingface.

The maximum conversation length is set to 2,048, as 99.98% of the conversations have less than 2,048 words. Padding and truncation were done for conversations less than 512 words or greater than 2,048 words, respectively.

2.2.3. Fine-Tuning and Classification

After pre-training, the Longformer model was trained for classification of the chosen labels. In order to inherit the learned weights from the MLM after the additional pre-training phase described in section 2.2.2, the language modelling head (i.e., the top layers for generating the predictions of masked words) is removed while the other layers and their corresponding weights are retained. In other words, we keep the encoder that generates the numerical embeddings of the text, and remove the

decoder that converts the numerical embeddings back into text. An additional classification layer with a non-linear activation function (Sigmoid or Softmax) was then added to the encoder. This additional layer provides a set of additional parameters that can be learnt to map the numerical embeddings (encoded text) to the ground truth class labels. The choice of non-linear activation function for the additional classification layer depends on the task; for example, multi-class classification requires a Softmax activation function which normalises the output into a probability distribution.

We carried out three experiments aimed to classify three different labels to match the identified tasks outlined above:

Experiment 1: Classification of Conversation Stages
This is a message-level multi-class classification task, where one of the six conversation stages (**Table 1**) is assigned to a given message. In this experiment, the non-linear activation function for the additional layer was a Softmax function.

Experiment 2: Classification of Behaviours
Up to three different behaviours were assigned to each message, reflecting the fact that texters and volunteers may display multiple behaviours within a text message. Hence this is a message-level multi-label classification task, where the six behaviours (**Table 2**) are represented by six independent binary labels. To generate independent probabilities for each label, the non-linear activation function used in the classification layer was a Sigmoid function. Note that in this task of classification of behaviours, data augmentation was not used, as perturbations to the messages can change the multi-label structure.

Experiment 3: Classification of Texter Demographics
Texter demographics were recorded in the texter survey at the conversation level. We chose categories recorded in the Shout data set in line with demographics of particular interest to MHI. The binary demographic categories (model labels) are: age (13 and under or over 13); autism (does the texter self-identify as having an autism diagnosis or not); and non-binary gender (does the texter self-identify as either agender, genderqueer, trans female, or trans male). In a similar manner to Experiment 1, the top layers of the Longformer model were substituted by a classification layer followed by Softmax activation function.

2.2.4. Hyperparameter Optimisation

In addition to learning the parameters of the Longformer model, there are various hyper-parameters that must be pre-defined. Here, we describe the choices of hyper-parameters and optimisation procedures for the pre-training and for the learnt classification models.

2.2.4.1. Hyper-Parameters of MLM Pre-training
A learning rate of 5×10^{-5} was chosen as it has been proved successful in previous relevant studies (10, 39, 40). The number of epochs (3) and batch sizes (2) were adopted as a reasonable choice given the computational power available.

2.2.4.2. Hyper-Parameters of Classification Models
The hyperparameters for the three classification models (Experiments 1-3) were optimised through a grid search of learning rate ($[8 \cdot 10^{-6}, 10^{-5}, 2 \cdot 10^{-5}, 5 \cdot 10^{-5}]$), number of epochs ($[3, 4, 5]$), and batch size ($[2, 4, 8, 16]$). The parameter combination yielding the best classification performance for all three classification models was: learning rate= $8 \cdot 10^{-6}$, number of epochs= 5, batch size= 8.

2.2.5. Evaluation and Performance

To evaluate the quality of the classification models, we use weighted accuracy, i.e., accuracy averaged for each class (41). For the behaviour classification model, we also report two additional measures. The (1 - Hamming loss), where values closer to 1 imply better results (42), provides a more meaningful output for multi-label classification as it considers the fraction of labels that are incorrectly predicted, whereas standard weighted accuracy would require all labels for a particular example to be correct to be considered accurate. The label ranking average precision score (LRAP) is a multi-label ranking metric that gives higher scores to predictions closer to the ground truth labels (43).

3. RESULTS

3.1. Classifying Messages Into Conversation Stages (Experiment 1)

We first evaluated our ability to classify messages into the six conversation stages (**Table 1**), using a multi-class classification model. Here, a message could only be assigned to a single conversation stage label.

3.1.1. Model Optimisation and Prediction Performance

The weights from the MLM were fine-tuned (see section 2.2.3) against the conversation stage class labels. For robustness, and to provide insights into the necessary components for learning mental health conversations, we examined the effects of changing: (i) the percentage of data used to pre-train the MLM; (ii) the percentage of training data used to fine-tune the model for conversation stage classification; (iii) the use of text augmentation; and (iv) the inclusion of text from both the preceding and subsequent messages to aid classification (see **Figure 3** for a visual description of including additional context).

Table 3 shows our results for the different models trained. The optimal model (model 3) achieved accuracy of 87.75%, and was pre-trained on 100% of textual data, used 100% of training data for fine-tuning, included text augmentation, and included context of the preceding and following messages. The different classification performance of the seven models in **Table 3** allows us to understand the importance each component. The hyper-parameters of all models were optimised using five-fold cross validation, see section 2.2.4.

From the trained models in **Table 3**, we draw four main observations. First, we find that without pre-training, the accuracy of the MLM (model 4, acc = 0.2010) is significantly lower than pre-training with just 33% (model 5, acc = 0.7512) or 100% (model 2, acc = 0.8620) of the total

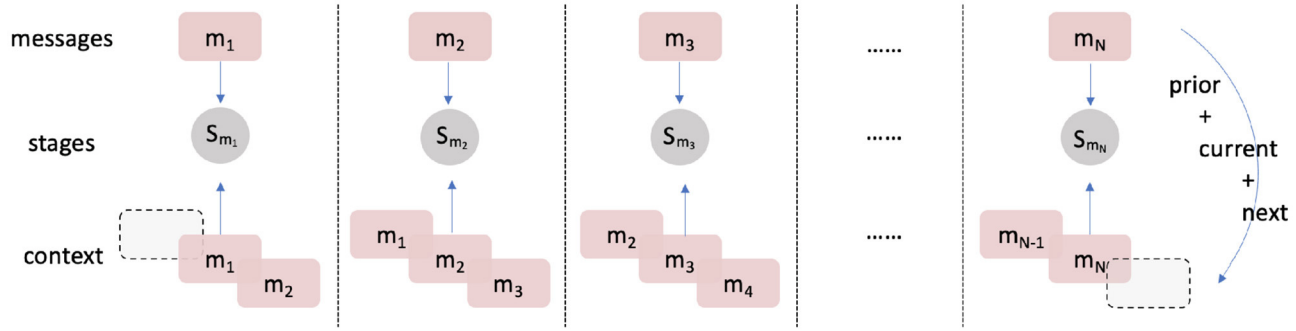


FIGURE 3 | Mechanism to include context in input data: m_i represents the message at time i and S_j represents the stages assigned for message j. Exactly one prior and one post message are concatenated with the message of interest. The label of the contextualised text inherits the original label of the current message.

TABLE 3 | Comparison of conversation stage classification performance for models with different performance improvement techniques.

Model ID	% Pre-training data	% Classification training data	Augmentation	Context	Accuracy
1	100	100	×	No context	0.6231
2	100	100	×	± 1 message	0.8620
3	100	100	✓	± 1 message	**0.8775**
4	0	100	×	± 1 message	0.2010
5	33	100	×	± 1 message	0.7512
6	100	25	×	± 1 message	0.6364
7	100	50	×	± 1 message	0.6966

*"No context" means that only the message of interest is input to the model, whereas "± 1 message" refers to the context concatenation scheme introduced in **Figure 3**, wherein we included the text from the messages preceding and subsequent to the message of interest. Bold values indicate the model with the best performance.*

available textual data. This result confirms the importance of conducting additional pre-training on the raw conversation text to improve the model's understanding of text specific to the Shout data set.

Second, increasing the proportion of training data for fine-tuning the model for classification improves accuracy (model 6, acc = 0.6364 < model 7, acc = 0.6966 < model 2, acc = 0.8620). This result is expected as training with more data typically results in higher accuracy. However, models trained with substantially smaller data sets (models 6 and 7) also produced reasonably accurate outcomes.

Third, using text augmentation increases the accuracy of message classification (model 2, acc = 0.8620 < model 3, acc = 0.8775). Data augmentation is used in most implementations of deep learning algorithms, most commonly as a regularisation strategy to prevent overfitting (44). In NLP, learning of high-frequency numeric patterns (e.g., token embeddings) or memorisation of particular forms of language prevent generalisation. Therefore, we would expect an increase in accuracy with augmentation. Since texters often use very informal language via their phones augmentation reduces the likelihood of the model overfitting to the unique writing styles present in a subset of conversations that may not generalise to a test set of conversations.

Fourth, using additional context to classify the message of interest improves accuracy (model 1, acc = 0.6231 < model 2, acc = 0.8620). Clearly, the model stands to benefit from the extra context surrounding the target message because changes in conversation stage are relatively slow (i.e., multiple consecutive messages in a short time period are likely to have the same conversation stage label). The input data "± 1 message" scheme therefore includes three times the information of the original target message for the conversation stage, resulting in better performance.

TABLE 4 | Classification accuracy for each conversation stage with the optimal model (Model ID 3 in **Table 3**).

Stage name	Initialise	Build rapport	Explore	Identify goal	Problem solve	End
Accuracy	0.9001	0.8890	0.8936	0.8995	0.8709	0.8929

3.1.2. Classification Performance of Individual Conversation Stages

Given that we achieved high classification accuracy for conversation stages, we now ask whether the classification accuracy of individual stages differed. In **Table 4**, we show the classification accuracy for each conversation stage. We find that the classification accuracy for each stage is comparable to that of the entire model, and that the variability among conversation stages is minimal (std = 0.009), suggesting that our model has generalised to all conversation stages.

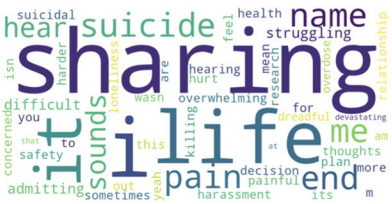
3.1.3. Feature Analysis and Interpretation

Despite boasting high accuracy predictions, deep learning methods often do so at the expense of interpretability. To get insight into mental health needs and factors affecting Shout service quality, it would be highly valuable not only to predict labels of interest, but also to understand what features of the conversations and language used are linked to these labels. To generate such insights, we used LIME (Local Interpretable Model-Agnostic Explanations) to provide "explanations" of the decisions of the trained model (45). LIME learns local information required to make a prediction and ranks individual words according to their contribution. We used LIME to extract the two most important words contributing to the final prediction of 300 random samples for each conversation stage (total of 600 words per conversation stage). Plotting word clouds of these word collections allows for a visual interrogation of which words the model finds important for predicting the conversation stage label. Below we provide summaries of the distinct features of each conversation stage according to the word clouds (**Figures 4A–F** for conversation stages 0–5):

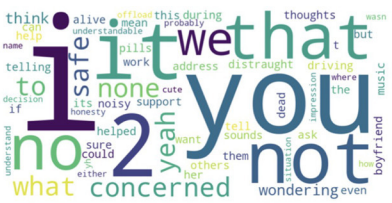
- Stage 0 - *Initialisation*: Dominated by introductory words ("welcome" and "hello") and includes messages from the "bot" and the "texter," showing that the model has learnt that Stage 0 is associated with the beginning/initialisation of the conversation.
- Stage 1 - *Build rapport*: Includes introductory words such as "hi" and "evening," but this time from the volunteer rather than the texter. Stage 1 also includes words that begin to explore how the texter is feeling as can be seen from "struggling" and "crisis."
- Stage 2 - *Explore*: Volunteers are trying to explore the problems in detail; we can see words such as "sharing" and "life," alongside descriptive terms such as "pain," "suicide," "overwhelming," "difficult," and "killing."
- Stage 3 - *Identify goal*: After establishing the problem, Volunteers try to clarify what support the texter is seeking and these messages usually start with first-person and second-person pronouns such as "I," "we," and "you." The goals identified at this stage appear to be texter-specific and no recurrent adjectives and nouns are present in the word cloud.
- Stage 4 - *Problem-solve*: Volunteers explore with the texter what they can identify as helpful steps to take after the conversation, from small things like listening to music, to bigger things like talking to a trusted person about their experiences and difficulties. Volunteers may also signpost texters to relevant further resources or organisations to provide ongoing support or information. Suggestions

A stage 0: initialise the conversation

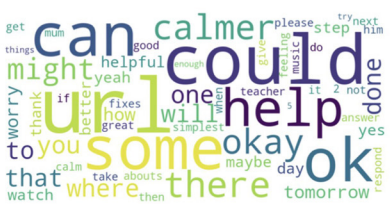
B stage 1: build rapport



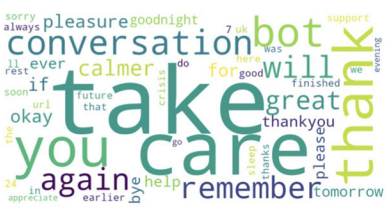
C stage 2: explore



D stage 3: identify goal



E stage 4: problem-solve



F stage 5: end the conversation

FIGURE 4 | Word clouds for LIME-derived words that best explain the model prediction for each conversation stage. **(A)** Bot messages are detected from the *initialisation*. **(B)** Greeting words are frequent when *building rapport*. **(C)** Volunteer begins to *explore* the texter's problem with key words from the discussion featuring "sharing" and "life." **(D)** Volunteer works to *identify goals* with the texter and these messages usually start with first-person and second-person pronouns. **(E)** Texter-specific *solutions* are suggested such as a "helpful" "URL." **(F)** The conversations are brought to a close using phrases like "take care" and "bye".

generally include words like "can," "could," and "might." "Help(ful)" resources can include signposts with a "URL."
- Stage 5 - *End the conversation*: Volunteers will bring the conversation to a close with phrases like "take care," "pleasure," and "bye." Texters express gratitude by saying "thank you" and "thanks."

Hence the word clouds are closely related to each conversation stage and the consistency of LIME-derived features for each class label means that the accurate predictions made by the model are based on meaningful and interpretable features of the messages, in line with similar features used by the human annotator when assigning class labels.

3.2. Predicting Texter Behaviour Expressed in Messages (Experiment 2)

In the previous section, we showed that the conversation stage for a message could be accurately classified using a multi-class model architecture. In this section, we explore the classification of messages into different psychologically-relevant "behaviours," as discussed in section 2.1.4.1 and defined in **Table 2**. In contrast to Experiment 1, where messages could only be assigned to a single conversation stage, multiple behaviours can be present in a single message and therefore can be annotated with multiple (up to a maximum of three) behaviours. Therefore, we used a multi-label classification architecture, where the six behaviours (**Table 2**) are represented by six independent binary labels.

3.2.1. Model Optimisation and Prediction Performance

Comparing models with and without message context (**Table 5**), we find that the optimal model does not use the text from the preceding and subsequent messages: all three metrics for the "no context" model (acc = 0.7701, "1-hamming" = 0.9502, LRAP = 0.9541) are higher than for the contextualised model (acc = 0.7407, "1-hamming" = 0.9410, LRAP = 0.9485). This means that the behaviours displayed by a texter or volunteer in one message do not predict the behaviour of the other actor in the subsequent message.

This result is in contrast with Experiment 1, where we found that context improved the prediction of conversation stages, since conversation stages are likely to remain unchanged for several messages. On the other hand, behaviours are unique to each message and rapidly varying across time and between actors, hence our metrics suggest that identifying behaviours based only on the current message is sufficient and appropriate. While there may be relationships between behaviours from message to message not currently picked up by our model, it is understandable that textual information from the volunteer (in a preceding or subsequent message) may not inform classification of texter behaviour, or vice-versa, in the current message. The behaviours of two consecutive messages are unlikely to be the same (and even less likely when including three messages), due to the interchange of actors across text messages and the rapid variation of behaviours.

The high accuracy demonstrated for classification of texter and Volunteer behaviours suggests that the codebook of defined categories of behaviour displayed by texters and volunteers are meaningful and distinct. The ability to classify psychologically-relevant behaviours at scale is the first step for future work to determine the relationship between such behaviours and important features such as conversation outcomes or different texter profiles.

TABLE 5 | Results of behaviour classification for individual messages.

Model	Accuracy (weighted)	1-hamming loss	LRAP
No context	**0.7701**	**0.9502**	**0.9541**
± 1 message	0.7407	0.9410	0.9485

Comparison of two models: without context and with message context, i.e., including the text of the preceding and following messages. Bold values indicate the model with the best performance.

TABLE 6 | Classification performance for three demographic variables.

Demographic category	Accuracy (weighted)	Survey proportion	Predicted proportion	Δ proportion
Age 13 or under	0.9575	0.0663	0.0435	−0.0228
Autism	0.9533	0.0618	0.0128	−0.0490
Non-binary gender	0.9592	0.0494	0.0160	−0.0334

We report the proportion of texters that self-identified in each demographic category in the survey. Using the trained models, we then predicted the class of the remaining (unlabelled) conversations and report the percentage of total texters predicted in each demographic category. We also report the difference between the survey-reported and predicted proportions of each category.

3.3. Revealing Texter Demographics Using Full Conversation Classification (Experiment 3)

Finally, we examine the classification of entire conversations using the texter survey results as ground truths (see section 2.2.3). In particular, we focus on particular texter demographics of special interest to better understand the users of the Shout service.

3.3.1. Model Optimisation and Prediction Performance

To build and test a capable model for conversation level classification, we chose three texter demographic variables of interest: (1) Age: 13 or under/over 13; (2) Autism: self-identification of an autism diagnosis or not, (3) Non-binary gender: self-identification of gender as non-binary, vs. all other answers to the question on gender. These choices of demographic subgroups was motivated by their interest to MHI experts.

The conversation-level classification models for the different demographic survey questions were fine-tuned at the checkpoint of the MLM. The classification accuracy of the three trained models for each binary label are reported on validation data (**Table 6**) with excellent performance (above 95% accuracy). Revealingly, while the results of texter surveys showed that 6.63, 6.18, and 4.94%, of texters were, respectively, of age 13 or under, autistic, and non-binary gender, the model-predicted

proportions for the whole data set were 4.35, 1.28, and 1.60%, respectively. This suggests that younger texters, and/or those who have been diagnosed with autism and/or identify as having non-binary gender may be more likely than others to complete the survey.

3.3.2. Feature Analysis and Interpretation

We used LIME to provide insight into the classification models for each demographic class. For the age label, LIME indicates that, as could be expected, numbers less than or equal to 13 are predictive of being aged 13 or less, and likely occur in response to the volunteer asking the texter's age, which can occur when the volunteer suspects they are talking to a child, for safeguarding purposes. Additionally, words that refer to school life, such as "school work," "bullying," or "lesson" also appear as important, and may indicate that bullying and school pressures are key issues faced distinctly by young texters. Moreover, we find that words corresponding to friends and parents are predictive of the texter age (**Figure 5A**).

Application of LIME to the autism classification model identified obvious words (**Figure 5B**), such as "autistic," "autism," and "aspergers," but words representing other disabilities that are not directly related to autism were also present in the word cloud, such as "adhd" or "dysphoria." This may suggest that texters with these disabilities tend to express themselves in similar ways, or that texters with autism are more likely to have other diagnoses, which is a well-known phenomenon (46).

For the non-binary gender class label (**Figure 5C**), we again find obvious words as predictive, including "transgender" and "trans"; however, less obvious words such as "gonna," "yeah," and "like" appear as important. We also see that "autism" is quite a significant word in the word cloud of non-binary gender. This implies that individuals with non-binary genders are more likely to discuss autism in their conversations with Shout compared to cisgender individuals, which is consistent with the co-occurrence of autism and gender dysphoria previously documented (47–49).

4. DISCUSSION

We have demonstrated the use of NLP to gain insights into users and service provision of a digital mental health crisis service. We trained and tested our NLP models using a unique large-scale data set from the Shout crisis text messaging service. To our knowledge, this is the first demonstration of NLP to generate predictions of attributes of mental health crisis conversations at both a message-level and conversation-level, including conversation stages, psychologically-relevant behaviours, and texter demographics.

While increasing numbers of people are turning to non-traditional forms of mental health support, few services are evaluated for impact or quality. In our first experiment, we developed a model to predict the "conversation stage" of each text message. Shout Volunteers are trained to follow six distinct conversation stages to guide a texter from a place of distress to a calmer place, and assessing adherence to this structure is a potential indicator of general conversation quality and volunteer skill. Our model is able to accurately predict (weighted accuracy = 87.75%) the conversation stage of messages on an unseen test set, which, given the randomised sampling of annotated conversations, should extrapolate to unannotated conversations. Analysing the structure of conversations has previously been shown as predictive of conversation outcome. Althoff et al. (7) derived linguistic features related to conversation outcomes that allowed the identification of optimal conversation strategies for volunteers. Although our analysis of conversation stages has some similarities to this work, we do not rely on feature engineering and instead derive a latent space embedding of conversations using a deep learning NLP model. Further, our analysis of conversation structure examines both Volunteer and texter, whilst Althoff et al. focussed only on the volunteer (7).

As well as a potential marker of conversation quality or efficacy, the importance of accurate prediction of conversation stages opens several other opportunities. For example, coaches could use the predicted stages of conversations as a tool to review conversations with volunteers and improve their training. Moreover, using the predicted conversation stage for each message, one could potentially track conversation stages in real time and direct prompts to volunteers or their Clinical Supervisors if the conversation appears "stuck," or stages appear to be missed. Additionally, the conversation structures and timings can be compared for different issues or subgroups (e.g., in conversations with extremely high levels of distress or suicidal ideation), or for conversations rated as helpful or unhelpful by texters. Such comparisons could generate insights into conversation formats that are most appropriate for texter needs, allowing for appropriate guidance for Volunteers to tailor

FIGURE 5 | Word cloud for words that best explains the model predictions for **(A)** age group 13 or under, **(B)** autism, and **(C)** non-binary gender.

conversations, and for supporting those in mental health crisis more broadly.

The Shout data set also provides potential for insight into the psychologically-relevant behaviours of people in mental distress as expressed during a conversation, and behaviours that are most helpful from a supporter. Language-based deficits are common symptoms of mental health crises and associated behaviours (50), and NLP techniques can generate accurate numeric embeddings of human-developed behavioural codes (i.e., annotated behavioural categories) (23). The ability to label at scale the psychologically-relevant behaviours of Shout texters and volunteers would enable deeper exploration of behaviours demonstrated by different texter groups (e.g., those experiencing different levels of distress) and the volunteer behaviours most predictive of helpful outcomes. Here we demonstrate the successful development of an NLP model to predict the presence of up to three of six psychologically-relevant behaviours (setting intention, enquiring, expressing distress, reflecting, corresponding, discord) in each message from a volunteer or texter. The model performs with 95.02% prediction accuracy. Such a result supports the validity of the behavioural classifications developed using a qualitative inductive thematic analysis (32, 33) and their distinctiveness, and shows they can be predicted from linguistic features. Our results complement a previous study examining behaviours of texters using data from the USA Crisis Text Line (51). Using machine-learning techniques the researchers queried message-level data to identify conversations that included phrases related to suicidality and revealed three distinct behaviours among texters expressing suicide risk. Here, we present a more general study with a wider range of behaviours and included all texters and volunteers within the analysis, not just those at high risk. The ability to predict behaviours of both texters and Volunteers will allow in future work for examination of texter-volunteer interactions and their relative efficacy (relationship to conversation outcomes), and whether volunteers do and/or should adapt their behaviours to different texters.

It is worth noting that while message-level predictions could potentially be made in real time, the current technique of context concatenation sets a limit on predicting messages, i.e., a lag of specifically one time step (message). A possible solution for achieving real-time monitoring is to fine-tune a unidirectional NLP model such as GPT-3 (12), which is pre-trained on an auto-regressive language modelling task so that only previous context is considered to predict the current word.

Finally, we also considered prediction at a conversation level. The Shout post-conversation survey collects demographic data on perceived conversation helpfulness scores from texters, but only a minority of texters complete the survey, and there is considerable opportunity for survey bias (52–54). Our results show that NLP models can also predict with high accuracy three different conversation survey results (age 13 or under: 95.75%; autism: 95.33%; non-binary gender: 95.92%), based on the conversation content as a whole. This allows us to determine demographic data for the entire texter cohort. The high accuracy for each demographic class, all with imbalanced classes, suggests the model should achieve similarly high accuracy on other survey results, and could support development of novel triaging strategies for crisis services, or help with the stratification of different population subgroups for further in-depth analysis of their experiences. Whilst we focussed on predicting conversation demographics here, our analysis could be extended to other features such as suicide risk or conversation helpfulness in future work (7, 17, 18).

Using the predicted population demographics, we observe that the true demographic breakdown of Shout users differed from the survey results—the predicted proportions were lower across all tested demographic categories, suggesting survey bias for some demographic categories. Texters identified as autistic or with non-binary gender were over-represented $5\times$ in the survey relative to the predicted results. As it is important to understand who is using digital mental health services and experiencing mental health crisis, the ability to generate predicted texter demographics without survey bias provides vital information to inform service development. Further, the age at which younger individuals develop mental health problems and expressions of mental health crisis in children is still not well-understood (55, 56). Identifying the complete set of conversations made by young texters could offer an invaluable data set to evaluate mental health crises in younger texters, complementing machine learning studies aimed at predicting mental health issues in children and adolescents (57, 58).

While we have provided initial results in the use of NLP to analyse and monitor mental health conversations, there are potential limitations to our study. First, we had limited annotations for message-level data (both conversation stage and behavioural keys, but not for conversation-level demographic data) due to the manual labelling of messages being resource intensive for the clinically trained research psychologist. Hence, increasing the number of training samples would improve model performance. Additionally, although several Shout clinical professionals were involved in the iterative development of a codebook for annotation of psychologically-relevant behaviours that produced consistent annotations on a conversation subset (validation), our final message-level annotations were performed by a single clinical research psychologist which could potentially lead to bias in the train and test data (59). Despite this, the feature importance analysis returned words that sensibly reflected what would be expected for each class, suggesting that the model learned class categories did not appear to contain bias. Having provided a proof of concept for NLP, a key future aim will be to annotate more conversations and to do so with multiple raters.

Looking forward, we aim to train models to classify labels that could be considered directly actionable or provide a direct measure of conversation efficacy. For example, a key future aim is prediction of a texter's risk of suicide, Furthermore, we would like to predict the extent to which a conversation de-escalated a texter or mitigated their crisis, which would help us measure the efficacy of the Shout service. Moreover, using the predicted conversation stages, we aim to use Bayesian modelling (60) to understand how conversation structure may confer conversation outcomes, and use unsupervised learning (61) to identify natural clusters of conversations that might not relate to ground truth labels.

Overall, our study highlights the potential for NLP to help gain insights into mental health service provision and the experiences of people in mental and emotional crisis. The ability to accurately predict message level attributes of crisis text conversations provides novel insights into psychologically-relevant behaviours displayed by individuals in mental distress and those supporting them, and the structures of those conversations. The demonstrated prediction of post-conversation survey results from conversation content has wide implications for monitoring, triaging and mental health service improvement, as well as for understanding the complex interactions of texter demographics. Taken together, this application of NLP to a unique large-scale charity mental health data set opens routes for monitoring and improvement of digital mental health services, and to gain novel insights, at scale, for those who most need our help.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because access requires completion of Information Security Awareness test and Data Protection Awareness on General Data Protection Regulation test. Access only through a secure server. Requests to access the datasets should be directed to Emma L. Lawrance, emma.lawrance@mhiuk.org.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Imperial College Research Ethics Committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ZL, RP, and EL conceived the study. AN produced the message-level annotations. ZL conducted all machine learning and natural language processing analyses, supervised by RP, EL, AN, MU, and MB. ZL, EL, RP, and MB drafted the manuscript. All authors reviewed the manuscript before submission.

FUNDING

We acknowledge funding through EPSRC award EP/N014529/1 supporting the EPSRC Centre for Mathematics of Precision Healthcare at Imperial and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 424778381-TRR 295. Mental Health Innovations also supported the research through a charitable donation to Imperial College London. EL was partly supported by a grant to Mental Health Innovations from the Rayne Foundation.

ACKNOWLEDGMENTS

We thank Ovidiu Serban, Daniel Cahn, David Newman, Joy Li, Jianxiong Sun, Jingru Yang, and Sophia Yaliraki for valuable discussions. We thank Mental Health Innovations for providing training and access to their invaluable data set on their highly secure server environment. We thank Lia Maragou and Lauren Duncan for their help in codebook development for data annotation.

REFERENCES

1. Abrams LS, Gordon AL. Self-harm narratives of urban and suburban young women. *Affilia*. (2003) 18:429–44. doi: 10.1177/0886109903257668
2. Jenney A, Exner-Cortens D. Toxic masculinity and mental health in young women: an analysis of 13 reasons why. *Affilia*. (2018) 33:410–7. doi: 10.1177/0886109918762492
3. Ornell F, Schuch JB, Sordi AO, Kessler FHP. "Pandemic fear" and COVID-19: mental health burden and strategies. *Braz J Psychiatry*. (2020) 42:232–5. doi: 10.1590/1516-4446-2020-0008
4. Roland J, Lawrance E, Insel T, Christensen H. *The digital mental health revolution: transforming care through innovation and scale-up*. Doha: World Innovation Summit for Health (2020).
5. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions–a call for innovative evidence generation approaches. *NPJ Digit Med*. (2020) 3:1–14. doi: 10.1038/s41746-020-00314-2
6. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. *J Med Internet Res*. (2021) 23:e15708. doi: 10.2196/15708
7. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Trans Assoc Comput Linguist*. (2016) 4:463–476. doi: 10.1162/tacl_a_00111
8. Stewart R, Velupillai S. Applied natural language processing in mental health big data. *Neuropsychopharmacology*. (2021) 46:252. doi: 10.1038/s41386-020-00842-1
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R, editor. *Advances in Neural Information Processing Systems*. Long Beach, CA (2017). p. 5998–6008. Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN: Association for Computational Linguistics (2019). p. 4171–86. doi: 10.18653/v1/n19-1423
11. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv* (2019) abs/1907.11692. Available online at: http://arxiv.org/abs/1907.11692
12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020.* (2020). Available online at: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
13. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach*

Learn Res. (2019) 21:140:1–140:67. Available online at: http://jmlr.org/papers/v21/20-074.html
14. Naseem U, Razzak I, Musial K, Imran M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generat Comput Syst*. (2020) 113:58–69. doi: 10.1016/j.future.2020.06.050
15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2020) 36:1234–40. doi: 10.1093/bioinformatics/btz682
16. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. (2021) 4:1–13. doi: 10.1038/s41746-021-00455-y
17. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput Math Methods Med*. (2016) 2016:8708434. doi: 10.1155/2016/8708434
18. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, Meyers G, Richey LA, Matykiewicz P, et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life Threat Behav*. (2016) 46:154–9. doi: 10.1111/sltb.12180
19. Shiner B, D'Avolio LW, Nguyen TM, Zayed MH, Young-Xu Y, Desai RA, et al. Measuring use of evidence based psychotherapy for posttraumatic stress disorder. *Admin Policy Mental Health Mental Health Serv Res*. (2013) 40:311–8. doi: 10.1007/s10488-012-0421-0
20. Viani N, Botelle R, Kerwin J, Yin L, Patel R, Stewart R, et al. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Sci Rep*. (2021) 11:1–12. doi: 10.1038/s41598-020-80457-0
21. Karystianis G, Adily A, Schofield P, Knight L, Galdon C, Greenberg D, et al. Automatic extraction of mental health disorders from domestic violence police narratives: text mining study. *J Med Internet Res*. (2018) 20:e11548. doi: 10.2196/11548
22. Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, et al. Generation and evaluation of artificial mental health records for Natural Language Processing. *NPJ Digit Med*. (2020) 3:1–9. doi: 10.1038/s41746-020-0267-x
23. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V. A comparison of natural language processing methods for automated coding of motivational interviewing. *J Substance Abuse Treat*. (2016) 65:43–50. doi: 10.1016/j.jsat.2016.01.006
24. Dunne MP, Martin NG, Bailey JM, Heath AC, Bucholz KK, Madden P, et al. Participation bias in a sexuality survey: psychological and behavioural characteristics of responders and non-responders. *Int J Epidemiol*. (1997) 26:844–54. doi: 10.1093/ije/26.4.844
25. Sigmon ST, Pells JJ, Boulard NE, Whitcomb-Smith S, Edenfield TM, Hermann BA, et al. Gender differences in self-reports of depression: The response bias hypothesis revisited. *Sex Roles*. (2005) 53:401–11. doi: 10.1007/s11199-005-6762-3
26. de Winter AF, Oldehinkel AJ, Veenstra R, Brunnekreef JA, Verhulst FC, Ormel J. Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *Eur J Epidemiol*. (2005) 20:173–81. doi: 10.1007/s10654-004-4948-6
27. Stormark KM, Heiervang E, Heimann M, Lundervold A, Gillberg C. Predicting nonresponse bias from teacher ratings of mental health problems in primary school children. *J Abnormal Child Psychol*. (2008) 36:411–9. doi: 10.1007/s10802-007-9187-3
28. Martelli MF, Zasler ND, Bush SS, Pickett TC. Assessment of response bias in impairment and disability examinations. (2006).
29. Locker D, Wiggins R, Sittampalam Y, Patrick DL. Estimating the prevalence of disability in the community: the influence of sample design and response bias. *J Epidemiol Commun Health*. (1981) 35:208–12. doi: 10.1136/jech.35.3.208
30. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol*. (2013) 13:117. doi: 10.1186/1471-2288-13-117
31. Roberts K, Dowell A, Nie JB. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Med Res Methodol*. (2019) 19:66. doi: 10.1186/s12874-019-0707-y
32. Braun V, Clarke V. What can "thematic analysis" offer health and wellbeing researchers? *Int J Qual Stud Health Well Being*. (2014) 29:1284–92. doi: 10.3402/qhw.v9.26152
33. Boyatzis RE. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage; Case Western Reserve University (1998).
34. Yardley L. Dilemmas in qualitative health research. *Psychol Health*. (2000) 15:215–28. doi: 10.1080/08870440008400302
35. Yardley L. Demonstrating validity in qualitative psychology. *Qual Psychol*. (2008) 2:235–51. Available online at: https://www.tandfonline.com/doi/abs/10.1080/08870440008400302
36. Gove WR, Geerken MR. Response bias in surveys of mental health: an empirical investigation. *Am J Sociol*. (1977) 82:1289–317. doi: 10.1086/226466
37. Olson K. Survey participation, nonresponse bias, measurement error bias, and total bias. *Int J Public Opin Q*. (2006) 70:737–58. doi: 10.1093/poq/nfl038
38. Wei J, Zou K. Eda: easy data augmentation techniques for boosting performance on text classification tasks (2019). *arXiv preprint arXiv:190111196*. doi: 10.18653/v1/D19-1670
39. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer (2020). *arXiv preprint arXiv:200405150*.
40. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: state-of-the-art natural language processing (2019). *arXiv preprint arXiv:191003771*. doi: 10.18653/v1/2020.emnlp-demos.6
41. Mosley L. *A balanced approach to the multi-class imbalance problem*. Graduate Theses and Dissertations. Paper 13537 (2013). Available online at: https://lib.dr.iastate.edu/etd/13537/
42. Tsoumakas G, Katakis I. Multi-label classification: an overview. *Int J Data Warehous Mining*. (2007) 3:1–13. doi: 10.4018/jdwm.2007070101
43. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recogn*. (2012) 45:3084–104. doi: 10.1016/j.patcog.2012.03.004
44. Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning. *J. Big Data*. (2021) 8:1–34. doi: 10.1186/s40537-021-00492-0
45. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM (2016). p. 1135–44. doi: 10.1145/2939672.2939778
46. Chien YL, Chou MC, Chiu YN, Chou WJ, Wu YY, Tsai WC, et al. ADHD-related symptoms and attention profiles in the unaffected siblings of probands with autism spectrum disorder: focus on the subtypes of autism and Asperger's disorder. *Mol Autism*. (2017) 8:1–12. doi: 10.1186/s13229-017-0153-9
47. Van Der Miesen AI, Hurley H, De Vries AL. Gender dysphoria and autism spectrum disorder: a narrative review. *Int Rev Psychiatry*. (2016) 28:70–80. doi: 10.3109/09540261.2015.1111199
48. van der Miesen AI, Cohen-Kettenis PT, de Vries AL. Is there a link between gender dysphoria and autism spectrum disorder? *J Am Acad Child Adolesc Psychiatry*. (2018) 57:884–5. doi: 10.1016/j.jaac.2018.04.022
49. Warrier V, Greenberg DM, Weir E, Buckingham C, Smith P, Lai MC, et al. Elevated rates of autism, other neurodevelopmental and psychiatric diagnoses, and autistic traits in transgender and gender-diverse individuals. *Nat Commun*. (2020) 11:1–12. doi: 10.1038/s41467-020-17794-1
50. Cohen AS, Mitchell KR, Elvevåg B. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. *Schizophrenia Res*. (2014) 159:533–8. doi: 10.1016/j.schres.2014.09.013
51. Szlyk HS, Roth KB, Garcia-Perdomo V. Engagement with crisis text line among subgroups of users who reported suicidality. *Psychiatr Serv*. (2020) 71:319–27. doi: 10.1176/appi.ps.201900149
52. Fowler FJ Jr. *Survey Research Methods*. 5th ed. Center for Survey Research, University of Massachusetts Boston (2013).
53. Hill BM, Shaw A. The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. *PLoS ONE*. (2013) 8:e65782. doi: 10.1371/journal.pone.0065782
54. Chamberlain A, Smart M. *Give to Get: A Mechanism to Reduce Bias in Online Reviews*. Technical Report. Glassdoor Research Report (2017).

55. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB. Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry*. (2007) 20:359. doi: 10.1097/YCO.0b013e32816ebc8c
56. Wiens K, Bhattarai A, Pedram P, Dores A, Williams J, Bulloch A, et al. A growing need for youth mental health services in Canada: examining trends in youth mental health from 2011 to 2018. *Epidemiol Psychiatr Sci*. (2020) 29:e115. doi: 10.1017/S2045796020000281
57. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. *PLoS ONE*. (2020) 15:e0230389. doi: 10.1371/journal.pone.0230389
58. Arya V, Mishra AK. Machine learning approaches to mental stress detection: a review. *Ann Optimizat Theory Pract*. (2021) 4:55–67. doi: 10.22121/aotp.2021.292083.1074
59. Geva M, Goldberg Y, Berant J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics (2019). p. 1161–6. doi: 10.18653/v1/D19-1107
60. Peach RL, Greenbury SF, Johnston IG, Yaliraki SN, Lefevre DJ, Barahona M. Understanding learner behaviour in online courses with Bayesian modelling and time series characterisation. *Sci Rep*. (2021) 11:1–15. doi: 10.1038/s41598-021-81709-3
61. Peach RL, Yaliraki SN, Lefevre D, Barahona M. Data-driven unsupervised clustering of online learner behaviour. *NPJ Sci Learn*. (2019) 4:1–11. doi: 10.1038/s41539-019-0054-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.