

Novel application of deep learning to evaluate conversations from a mental health text support service

Daniel Cahn^a, Sarah Yeoh^a, Lakshya Soni^{b,c}, Ariele Noble^d, Mark A. Ungless^d,
Emma Lawrance^{c,d}, Ovidiu Șerban^{e,a,*}

^a Department of Computing, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

^b Department of Surgery & Cancer, Imperial College London, St. Mary's Campus, London W2 1NY, UK

^c Institute of Global Health Innovation, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

^d Mental Health Innovations, London, UK

^e Data Science Institute, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

ARTICLE INFO

MSC:

68T07

68T50

91E45

Keywords:

Applied mental health research

Understanding user conversations

Transformer-based architectures

ABSTRACT

The Shout text support service supports individuals experiencing mental health distress through anonymous text conversations. As one of the first research projects on the Shout dataset and one of the first significant attempts to apply advanced deep learning to a text messaging service, this project is a proof-of-concept demonstrating the potential of using deep learning to text messages. Several areas of interest to Shout are identifying texter characteristics, emphasising high suicide-risk participants, and understanding what can make conversations helpful to texters. Therefore, from a mental health perspective, we look at (1) characterising texter demographics strictly based on the vocabulary used throughout the conversation, (2) predicting an individual's risk of suicide or self-harm, and (3) assessing conversation success by developing robust outcome metrics. To fulfil these aims, a series of Machine Learning models were trained using data from post-conversation surveys to predict the different levels of suicide risk, whether a conversation was helpful, and texter characteristics, such as demographic information. The results show that language models based on Deep Learning significantly improve understanding of this highly subjective dataset. We compare traditional methods and basic meta-features with the latest developments in Transformer-based architectures and showcase the advantages of mental health research.

1. Introduction

Shout is a text line that supports individuals struggling with their mental health. It offers a free and confidential platform where people can text trained volunteers for emotional support, intervention, and referrals to local resources. Shout aims to provide immediate assistance to those in need, particularly during acute distress.

The service operates 24/7, allowing individuals to seek help anytime, anywhere. Trained volunteers, known as Shout Volunteers, undergo extensive training to handle various distress situations, including those involving suicidal ideation, self-harm, depression, anxiety, and other mental health concerns.

The large dataset of anonymised Shout conversations offers a valuable new lens to view how mental health issues affect individuals across the UK and provides an opportunity to learn about the nature of mental health distress by hearing directly from those affected at scale. This can help elucidate the factors that can lead to or contribute to mental

distress and what can best help. In parallel, as demand for the service grows, a need arises for computational solutions to assist the Shout Volunteers and ensure that conversations are consistently high quality.

1.1. The success of support textlines

Conversation success metrics provide a vital opportunity to evaluate and track service impact. More generally, developing appropriate success metrics is a key first step to shed light on the factors contributing to making a conversation with individuals in mental distress helpful or unhelpful. Before this study, the primary conversation success metric came from a question in the texter survey asking whether a texter found their conversation helpful and, if so, how helpful. However, a texter's opinion on conversation helpfulness may depend on factors external to Shout's control. For instance, whether the problem texters are dealing with can benefit from talking with a Shout Volunteer or

* Corresponding author at: Data Science Institute, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.
E-mail address: o.serban@imperial.ac.uk (O. Șerban).

their level of distress before entering the conversation. Gould et al. note that more distressed texters tended to find conversations less helpful. They also found that conversations are more effective with texters who discuss physical abuse, are younger than 25 years, or are of Hispanic origin [Gould et al. \(2022\)](#); all variables outside of the Shout Volunteer's control. It is, therefore, essential to develop a robust success metric that not only reflects the mental state of the texter but also validates the quality of the conversation or degree of skill shown by the Shout Volunteer.

Measuring volunteer skill. One potential success metric relates to the skill of the Volunteer supporting the texter, either in this particular conversation or in general. To study success for distress textlines, Althoff et al. focus analyses on counsellors rather than conversations and use factors such as distances between TF-IDF encoded conversations, conversation lengths and message lengths to determine factors influencing counsellors' success [\(Althoff et al., 2016\)](#) - an approach which is adapted for use in this study.

1.2. Suicide risk

Conversations with Shout often involve individuals at significant risk of suicide or self-harm. Shout Volunteers are required to assess suicide risk using the four rungs of the *suicide risk ladder*: thoughts (suicidal desire), plans (intent), means (capability), and time frame. This assessment is essential for identifying and escalating high-risk conversations to Clinical Supervisors and performing an *active rescue* (contacting emergency services where a texter is in imminent danger). Risk assessment results, collected separately from texters and Shout Volunteers, are collated in the post-conversation survey.

A significant concern is that risk is missed. Machine learning can help highlight conversations with high risk, either post-conversation - for review purposes - or in real time - for pre-emptive action. However, pre-existing literature highlights the complex nature of predicting suicide risk, with BERT-based methods outperforming simpler models like neural networks [\(Mohammadi et al., 2019; Ragheb et al., 2021\)](#), except where datasets are small [\(Morales et al., 2021; Bayram and Benhiba, 2021; Macavaney et al., 2021\)](#). Additionally, model performance for discriminating between categoric risk and no risk was better than model performance in discriminating level of risk, e.g. no or low risk from moderate or severe risk [\(Zirikly et al., 2019; Ragheb et al., 2021\)](#), suggesting that discriminating against no-risk cases is more manageable than identifying levels of risk.

1.3. Texter demographics

Demographic information enhances Shout's understanding of its texter base, aiding service improvement by providing Shout with information which allows them to tailor training for their Shout Volunteers and reach out to populations that text in less. However, this information relies on voluntary surveys with low response rates (~10%) and potential participation bias, wherein the characteristics of individuals who respond to surveys may not represent the entire texter population. This bias can skew the data and compromise the accuracy of any conclusions.

Deep learning helps mitigate bias, to a limited degree, by extrapolating survey answers onto conversations without responses. There has been some research into predicting demographics from text. Certain words, phrase selections, and topics have been found to correlate with specific age groups [\(Nguyen et al., 2013\)](#). However, with simpler models like SVMs, these differences seem insufficient to predict age from text. In a shared task on predicting age, the best-performing model (an SVM linear model) obtained an accuracy of only 58.9% [\(op Vollenbroek et al., 2016\)](#), highlighting the complexity of the task.

In contrast, predicting gender from text has met with more success, with average accuracies of 83% across articles in a review of

91 articles [\(Cesare et al., 2017\)](#). Even when identifying information like names or user profile information is omitted, simpler models like SVMs or neural networks achieve accuracies of over 80% [\(Aljohani and Cristea, 2021; Basile et al., 2018\)](#). Models using bi-directional composition functions also achieve better results than models that do not [\(Aljohani and Cristea, 2021\)](#). These findings indicate that using BERT-based models for predicting demographics shows promise.

1.4. Deep learning for documents

As mentioned above, a significant limitation of the literature is that it mainly uses small chunks of text or features engineered from text. This approach is insufficient because it fails to exploit Shout's most extensive and informative database — its conversations.

However, the maximum token length in BERT is limited to 512 tokens because its attention mechanism comes with quadratic complexity ($O(N^2)$) in terms of sequence length, n . To address this limitation, sparse attention mechanisms are proposed, which confine attention to predefined patterns like local windows or block patterns. [Zaheer et al. \(2020\)](#), [Beltagy et al. \(2020\)](#), [Ainslie et al. \(2020\)](#), [Child et al. \(2019\)](#). Combining multiple patterns improves coverage. Big Bird, for example, combines local, global, and random attention, allowing it to handle sequences up to eight times longer (4096 tokens) with complexity $O(N)O(N)$. We adopt Big Bird due to its minimal complexity and superior performance compared to other sparse-attention models (e.g. Longformer, [Zaheer et al. \(2020\)](#)).

Another approach deals with length limitations by overcoming context loss when long documents are split. Recurrence over BERT (RoBERT) and Transformer over BERT (ToBERT) extract features from BERT-encoded sequences. RoBERT uses a small RNN, while ToBERT utilises a Transformer. Both methods perform similarly, with ToBERT slightly outperforming in some cases. Neither involves fine-tuning of BERT, and the authors did not provide exact implementation details [\(Pappagari et al., 2019\)](#); adjustments to the ToBERT model were therefore made for our work to incorporate fine-tuning.

1.5. Multi-task learning

In this work, a key to successful supervised learning without many labelled training samples was to use unlabelled training data and auxiliary labels with multi-task or multi-label classification.

For models not deep neural networks-based, performing multi-label classification is non-trivial since most models can only classify in a binary manner. However, with deep neural networks, a final layer can be trained to output multiple label predictions simultaneously, and the internal representation of features will be taken into account in the model's intermediate layers [\(Zhang and Zhou, 2006\)](#). Using multi-task training with deep neural networks can provide significantly improved results on NLP tasks [\(Liu et al., 2015\)](#), even in semi-supervised settings [\(Collobert and Weston, 2008\)](#), because deep neural networks can learn better feature representations of text through multi-task training. Multi-task training can exploit correlations between various targets [\(Nam et al., 2014\)](#), providing success even on noisy labels when auxiliary less-noisy labels are used [\(Duan and Wu, 2016\)](#). Additionally, it is suggested that multi-tasking provides a regularisation effect by reducing overfitting to specific tasks, allowing representations to be generalised across tasks [\(Liu et al., 2019a\)](#).

Text-to-Text Transfer Transformer, or T5 [\(Raffel et al., 2020\)](#), is a model developed for multi-task training. Various GLUE [\(Wang et al., 2018\)](#) NLP tasks, such as summarisation, sentiment analysis, and others, are converted into text-to-text problems, and the model is then trained simultaneously on multiple tasks. While pre-training on a multi-task objective and fine-tuning on a specific task is generally beneficial, the authors also discuss the problems they encounter when training a single network for multiple tasks, related to overfitting or underfitting to particular tasks due to having different ratios of data available

for training each task, the difficulty of regularising the network appropriately for each task, and the fear of task interference “where achieving good performance on one task can hinder performance on another” (Raffel et al., 2020).

1.6. Objectives

The main objectives relate to an improved understanding of mental health needs and the impact of the Shout service. The specific outcomes are as follows:

1. **conversation success metrics:** to develop robust success metrics for each conversation based on an estimate of the Shout Volunteer’s skill and from predictions of texter-reported conversation helpfulness ratings extrapolated from the post-conversation surveys to the whole dataset.
2. **texter suicide risk:** to predict texter suicide risk from conversation content to help screen conversations for risk of suicide or self-harm.
3. **texter demographics:** to extrapolate from texter survey responses to the entire conversation dataset to determine and partially reverse survey participation bias, especially for understanding trends related to demographic groups.

This project makes two primary contributions to machine learning methodology to achieve these aims. The first contributes to deep learning, including two novel methods of using the BERT model on sequences exceeding the length at which BERT can typically be used. These methods are Attention Over BERT, which involves combining multiple chunked forward passes through BERT using an attention mechanism, and Transformer Over BERT or ToBERT, which involves end-to-end training of hierarchical Transformers and is based on a model of the same name from Pappagari et al. (2019).

Secondly, this project studies multi-task training as a form of semi-supervised machine learning in the presence of sparse labels. This includes comparing single-task, multi-task, and multi-task pre-training with single-task fine-tuning.

2. Dataset exploration

This project uses the Shout dataset, comprised of a snapshot of heavily redacted, for anonymisation purposes, text message conversations between texters and Shout Volunteers on the Shout text support service from two years of conversation data, including 10,808,136 messages from 271,405 conversations (12 February 2018 through 3 April 2020). Throughout this paper, we refer to this as Dataset 1. The second round of experiments was carried out on an expanded dataset, which included 21,394,851 messages across 538,075 conversations from 12 February 2018 to 24 December 2020, referred to as Dataset 2 throughout the paper. The nature of these conversations is discussed in Section 2.1. The dataset also contains metadata related to messages, conversations, texters, and Shout Volunteers, including post-conversation surveys described in Section 2.3.

2.1. Conversations

Shout facilitates anonymous text message conversations between individuals experiencing significant distress and Shout Volunteers. Individuals applying to become Shout Volunteers undergo an extensive selection and training process of over 25 h and are supported by a coach at all stages. In addition to Shout Volunteers, Clinical Supervisors, clinically trained with appropriate psychological expertise, oversee all conversations on the platform.

Anyone in the UK can initiate a conversation at any time, 24/7, by texting Shout’s five-digit number. After doing so, they receive an automatic machine reply enquiring what the texter is reaching out about and information about Shout’s privacy policy. Shout triages

Table 1

Size of both datasets and statistics for conversation length (measured in characters) after initial preprocessing and before unsuitable conversations are filtered out.

	Dataset 1 12 Feb 18 – 3 Apr 20	Dataset 2 12 Feb 18 – 24 Dec 20
Number of conversations	271,405	538,075
Number of messages	10,808,136	21,394,851
Conversations removed (%)		
Too long	1.14	1.00
Too short	30.47	31.48
Conversation length		
Minimum	1	1
Maximum	25,213	41,750
Mean	2852	2852
Median	2582	2563
Inter-quartile range	3351	3535
Standard deviation	2303	2365

conversations based on texters’ early messages to ensure that high-risk messages are responded to with the highest priority.

Once a conversation has been initiated, Shout Volunteers are trained to follow five stages, as seen in Fig. 1

Throughout a conversation, alerts may appear to the Shout Volunteer as pop-ups to remind them of goals for each stage of the conversation or when to move to the next stage. Shout Volunteers are also required to assess the risk of suicide and/or self-harm using the suicide risk ladder, which will be discussed in more detail in Section 4.2.

An Active Rescue may occur if the risk is sufficiently imminent, where authorities are contacted to intervene with a suicide attempt physically. Shout Volunteers can also receive feedback or comments from a Clinic Supervisor during a conversation. Conversations may occasionally be transferred to a Supervisor when the Shout Volunteer requires extra assistance, for example, in circumstances of high suicide risk.

2.2. Conversation length

Measuring conversation length by the number of characters reveals a bi-modal distribution (see Fig. 2). After initial preprocessing, conversation lengths in both experiments are similar and displayed in Table 1 for comparison. Shout staff explain that texters can have concise conversations for several reasons. Sometimes, they wish to test the Shout system before initiating a conversation. This testing allows them to understand precisely how Shout works — for example, the service is free, fully anonymous, private, and with real people. On other occasions, individuals merely want to vent something; they send one or a few messages to get something off their chest and then end the conversation without needing a response. Conversations can also end early when wait times are high, and it takes too long for the texter to be matched with a Shout Volunteer, especially during the night when the texter may have fallen asleep by the time they are matched. Based on these reasons and the bi-modal distribution, it is clear that concise conversations contain less meaningful interaction with a Volunteer and are not representative of the dataset as a whole.

Similarly, very long conversations will not represent the bulk of conversations. Due to the limited number of Shout Volunteers and clinical understanding of the value of the Volunteer setting clear boundaries during support giving, long conversations are discouraged, with warnings appearing to Shout Volunteers when conversations have gone on too long that encourage them to bring the conversation to a close. Nonetheless, in very heated situations or when there is an imminent high risk, the texter should be encouraged to stay engaged, and conversations can continue for several hours. Instead of removing long conversations from the dataset, another option would be to trim them, keeping only the first part of the conversation and simply not considering messages after a certain point. An upside of this approach is



Fig. 1. 5 stages of conversation Shout Volunteers are trained in.

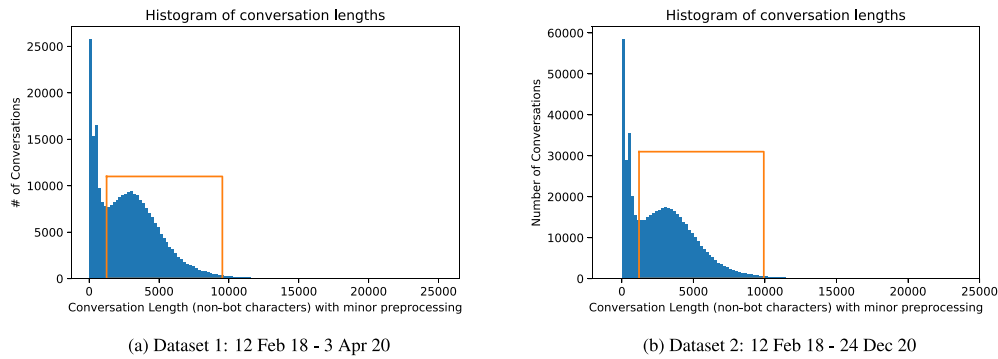


Fig. 2. Distribution of conversation lengths and a box indicating the conversations used in the analysis.

that long conversations are likely well-labelled and might contain an essential signal for a classifier. However, conversation stages are likely to be quite different in long conversations because the conversations are likely to go through a greater number of stages and because stages will inevitably last longer. These conversations are, therefore, also unlikely to be representative of the bulk of the dataset.

Based on a quick analysis of conversation lengths (Fig. 2) and topics, we removed conversations with lengths below 1,200 characters. A quick analysis reveals that rapport or issue statements are often not frequently established in discussions of such length (Section 2.1). On the high end, we remove conversations with a size greater than two times the interquartile range above the mean. In the first dataset, 30.47% of conversations were removed as they were too short, 1.14% were removed for being too long, and 14.14% individual messages were removed at this step. A similar procedure was used for eliminating conversations in the second dataset – with the same threshold of 1,200 characters – and the percentages of conversations removed are identical (Fig. 1).

2.3. Metadata

Shout volunteer survey. At the end of each conversation, Shout Volunteers are required to complete a short survey with six parts, and data is available for almost all conversations:

- (1) Select which were discussed in the conversation from a list of issues. Shout Volunteers have 15 options including “Abuse, Physical”, “Depression/Sadness”, “Self-harm” and “Suicide”
- (2) If “self-harm” is ticked as an issue discussed: Select whether self-harm is in progress (yes, no, or not sure)
- (3) If “suicide” is ticked as an issue discussed: Suicide Risk Ladder assessment (checkboxes: Thoughts, Plans, Means, Timeline)
- (4) Write if any “coping skills or safety plans” that were agreed to by the texter (free response, multiple separate answers are allowed)
- (5) Write any “signposts” that were agreed to by the texter (free response, multiple separate answers are allowed)
- (6) “How are you feeling?” (“I’m good!”, “Ehhh”, “Upset”)

Texter survey. Following conversation completion, texters receive an invitation to complete a post-conversation survey,¹ and in our data subset, the survey is completed by 10.24% of conversations. All questions

are optional, and texters are instructed to answer the questions they are comfortable responding to. The total number of questions is roughly 25, though the questions that appear vary depending on earlier response patterns. Questions have also varied slightly over time; for example, a Coronavirus-related question was added in early 2020.

Questions include some related to the texter’s experience of the conversation:

- “Did you find this conversation helpful?” and if so “How helpful was it?” from 1 (slightly helpful) to 5 (very helpful)
- Open text area to leave a note for the Volunteer
- Do they feel more, less, or equally “in control”, “hopeful”, “alone”, “depressed”, “overwhelmed”, “upset”, “suicidal” following the conversation
- Did they “mention an experience or feel that you have not shared with anyone else?”
- Do they “believe my Shout Volunteer was genuinely concerned for my well-being”?
- Did they agree to a plan; if so, how likely are they to follow it; and if not, why not.

Other questions relate to the texter’s demographic information and personal history, including age, gender, sexual orientation, race, disabilities, and past military experience. Texters are also asked about how often mental health issues were experienced in the prior two weeks, including feeling “nervous or anxious”, “worrying”, having “little interest in doing things”, “feeling depressed or hopeless”, “feeling left out”, “feeling isolated”. Finally, texters are asked “In addition to texting in today, how else do they get help when in distress?”, and “how did they hear about Shout?”.

Texters can complete the survey multiple times if they return to Shout for another conversation but are not asked demographic questions on future occasions. Demographic responses can be extrapolated across discussions from the same phone number.

2.4. Conversation helpfulness

Only 10.2% of conversations analysed respond to whether the conversation was helpful. As such, there is likely significant survey participation bias relating to conversation helpfulness — with research

¹ Texters usually get an invitation to fill in the post-conversation survey with a few exceptions: (1) they do not reply to a volunteer, (2) they text STOP, (3) if an active rescue is called.

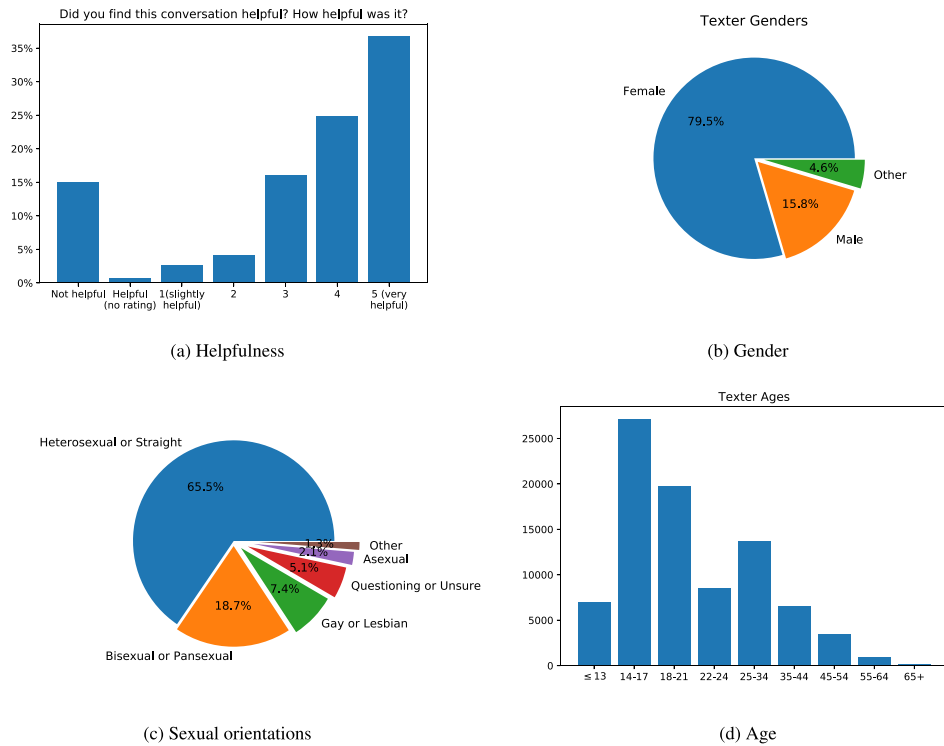


Fig. 3. The distribution of Shout texters' helpfulness responses and demographics, as per the texter survey. The helpfulness is shown for the "Did you find this conversation helpful?" and "How helpful was it?" survey questions. The charts provide a detailed description of Dataset 1. Dataset 2 exhibits a comparable distribution of responses and demographic characteristics.

indicating that those with very positive or negative experiences are more likely to provide a review (Chamberlain and Smart, 2017). Indeed, with 37.6% of texters who responded to the question rating it 5/5 on helpfulness (see Fig. 3(a)), the conversation success metrics are possibly skewed positively. There is also likely to be noise, with different people interpreting questions differently and quantifying how 'helpful' a conversation is in a highly subjective way.

Preliminary tests suggest some heuristics by which helpful conversations can be distinguished from less helpful ones. Based on the helpfulness responses from the texter survey, one can see that conversations found helpful are longer on average than those found unhelpful. Texters in helpful conversations write 81.56 more words on average,² in 2.32 extra messages on average² and Shout Volunteers in these conversations write 58.10 additional words² and 2.90 additional messages.² Shout Volunteers also, on average, write longer words in these conversations, by 0.054 letters per word²; texters write 2.08 more words per message²; and time between messages is 15.85 s less on average.²

2.5. Texter demographics

Due to the rarity of the texter survey being filled out, the survey results are likely to show significant participation bias; demographics from the texter survey can thus hardly be taken as representative of all texters. Nonetheless, based solely on survey data, it appears that females (79.5%) and heterosexual individuals (65.5%) make up the majority of conversations (see Fig. 3(b), 3(c)), and ages sharply peak among individuals age 14–21 (see Fig. 3(d)). Without accounting for participation bias, it is striking that respondent demographics closely mirror trends drawn from voluntary surveys by Crisis Text Line, a similar service operating in the United States. The majority of the respondents identified as being heterosexual/straight (52.2%) and female

(79%), while 66% of conversations were with users aged between 14 to 24 years of age (Pisani et al., 2022).

Looking at the correlations between word occurrences and specific labels, several words can be used heuristically to differentiate one group from another. For example, identifying as under eighteen years old in the texter survey correlates with the use of abbreviations including "ik", "idk", "ye", "tho" and "rlly", a heuristic that may be helpful for Shout Volunteers and modelling age.

2.6. Ethical considerations

Given the sensitive nature of conversations and the imperative to protect texters' anonymity, the Shout platform and dataset demand careful ethical handling. This project required training in GDPR and safe computing practices to ensure legal and ethical responsibilities could be respected, with all data accessed through secure machines and connections. Additionally, while anonymising data could compromise the effectiveness of machine learning models (e.g. where the scrubber³ confuses everyday words with personally identifiable information), this is a necessary sacrifice to protect texters' privacy — which remains paramount. Lastly, even after the data is anonymised, it is essential that the dataset is carefully protected and that no part of it is publicly available.

3. Model implementation

This section will discuss implementing this project's five increasingly complex Machine Learning models. We start with two baselines, namely a Random Forest using hand-crafted conversation features, followed by TF-IDF word occurrence vectors. Next, we use a relatively

³ The scrubber is an automated process applied to the message data to remove names, location details and anything else with a significant impact on the texter's anonymity.

² $p \ll 0.001$.

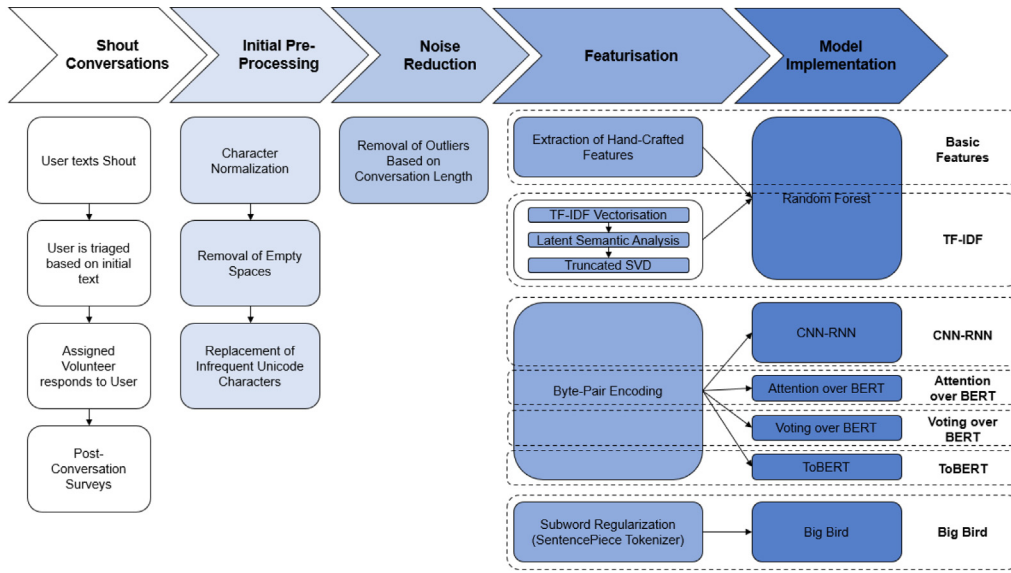


Fig. 4. A pipeline of data collection and processing for each experiment. The exact architecture of each model is described and depicted in the subsequent sections.

Table 2

A summary of the dataset, train-test protocol and tasks for each implemented model.

	Basic Features	TF-IDF	CNN-RNN	BERT Voting	BERT Attn	ToBERT	Big Bird
Dataset ^a	1	1	1	1	1	1	2
Train-Test split (%)	95–5	95–5	95–5	95–5	95–5	95–5	90–10
Multi-Task?	×	×	✓	✓	✓	✓	✓
Semi-Supervised Learning?	×	×	×	✓	✓	✓	×
Labels per Model ^b	1	1	489				9
Predicted Labels ^c							
Topic:							
Self harm	✓	✓	✓	✓	✓	✓	✓
Depression	✓	✓	✓	✓	✓	✓	✓
Substance Use	✓	✓	✓	✓	✓	✓	✓
Suicide risk:							
Desire	✓	✓	✓	✓	✓	✓	✓
Capability	✓	✓	✓	✓	✓	✓	✓
Texter survey:							
Helpful?	✓	✓	✓	✓	✓	✓	✓
Age ≤ 21	✓	✓	✓	✓	✓	✓	×
Heterosexual	✓	✓	✓	✓	✓	✓	×
Race: White	✓	✓	✓	✓	✓	✓	×
Gender: Male	✓	✓	✓	✓	✓	✓	×

^a Dataset 1 and 2 refer to the time period when conversations were extracted and are discussed in Section 2.1.

^b The number of labels per model is based on the multi-task model, where a multi-task model was used. The results presented in Section 4.1 are based on these models.

^c These are the labels used for evaluation in Section 4.1, but are not the only labels predicted by the models.

small Residual Convolutional and Recurrent Neural Network (or CNN-RNN) with pre-trained word embeddings. We then turn to using BERT models, fine-tuning a pre-trained model on a Masked Language Modelling task. Due to Shout's conversations being generally longer than BERT's maximum sequence length of 512 tokens, we experiment with three potential methods of applying BERT on longer sequences, namely Voting over BERT, which involves independently forming predictions on overlapping sub-sequences and combining the forecasts from each sub-sequence; Attention over BERT, which involves five methods of combining multiple BERT forward passes using an attention mechanism; and ToBERT, a method slightly modified from Pappagari et al. (2019) for end-to-end training. We also explore the benefit of multi-task training by training on auxiliary labels, including conversation metadata and non-target survey data (see Table 2). Results from these models are evaluated in the next section (see Fig. 4).

3.1. Pre-processing

Conversation texts were preprocessed before use in training models. This process includes the normalisation of characters, namely converting emojis into their byte form; removing Unicode characters such as those for underlining and italicising; converting both forms of dashes (“-” and “_”) into one form, and converting multiple forms of single and double quotation marks (e.g. “ and ”) into single forms for each (e.g. "). Additionally, leading and trailing spaces are removed from messages and empty messages are removed. Finally, Unicode characters that occur very infrequently, such as Chinese characters, are replaced with a special “unknown” token.

3.2. Basic features and TF-IDF

Initial baseline experiments were used to assess the problem difficulty, understand the importance of different types of features in predicting different labels, and justify the need for more complex BERT

models. The first document representation method is extracting hand-created features based on conversations with Shout staff on features that may be most predictive of labels, such as demographics and conversation helpfulness. Demographic variables like age were thought to relate to the speed of responses, length of messages, or length of words. In contrast, conversation length could vary by conversation topic and relate to the texter's experience of conversation helpfulness. A relative number of messages could explain the relative involvement of each speaker; for example, the average length of words used by a speaker or the speed at which they type could represent their mental state, and the total size of a conversation could indicate the level of distress. Hence, the selected features were conversation duration, number of messages, number of words, and number of characters sent by the Shout Volunteer and the texter separately, as well as the ratios of these values. While feature engineering could be used to identify features from a more extensive set (e.g. time of day, the entire distribution of time between messages) or task-specific features (e.g. the presence or frequency of specific keywords), it was expected that later deep learning models perform automatic feature engineering on message data. As such, these basic features suffice as a baseline.

A second baseline uses TF-IDF vectors with the implementation in Scikit-learn (Pedregosa et al., 2011). TF-IDF vectors were generated separately for the texter, Shout Volunteer, and bot (i.e. the automated messages) and concatenated to form one vector per conversation. Bot messages are included in case they provide valuable metadata, for example, about wait times, which may pertain, for instance, to texter-reported conversation helpfulness. Tokenisation and stop-word removal are performed using library defaults and words that occur in more than 40% of conversations (domain-specific stop-words) or fewer than 1% for Shout Volunteer and texter messages and 5% for bot messages. Thresholds were chosen by manually scanning through term frequencies to identify a reasonable boundary between stop-words and meaningful words. Similarly, for low frequencies, the boundary aims to reduce rare terms that increase the likelihood of overfitting. We then perform Latent Semantic Analysis (Deerwester et al., 1990) to reduce the dimensionality of the TF-IDF vectors, capturing redundancy in the vectors (e.g. if “thank” and “you” very frequently co-occur, this can be taken into account). LSA also makes these vectors more computationally manageable and reduces the ease of overfitting for labels with fewer examples. The Truncated SVD algorithm for dimensionality reduction (Halko et al., 2011) was used for efficiency at the slight cost of being sub-optimal due to the size and sparse nature of the TF-IDF vectors. The vector size after dimensionality reduction is 1000.

A Random Forest classifier is used for its ability to reduce model variance and bias. Using a Random Forest also allows for complex relationships between words to be considered, rather than the mere presence or frequency of a single word. Early stopping is introduced to the Random Forest to further balance model bias and variance in training. For hyper-parameters, a balanced weighting of classes is used; a maximum tree depth of 4 for basic features and 5 for TF-IDF; and an increasing number of trees with a step size of five trees per epoch and patience of 3 epochs before stopping training for the best validation metric. The Matthews Correlation Coefficient between the predicted and true label is used as the validation metric due to the imbalanced nature of many classes.

3.3. CNN-RNN

Next, we train a deep CNN-RNN with pre-trained word embeddings. The use of CNN layers allows for the detection of co-occurring words and phrases, and the use of RNN layers allows for distant dependencies to be accounted for. Using strided convolutions also helps reduce the input size to the RNN, thus improving efficiency. The RNN cell type used is the Gated Recurrent Unit (GRU), which allows for nearly identical performance compared to LSTMs at a fraction of the computational cost (Kaiser and Sutskever, 2016).

The CNN-RNN is designed as follows. First, conversation characters are tokenised with special tokens to indicate a conversation's start and end and separate messages. The same tokenisation strategy and pre-trained embedding are used for the CNN-RNN as in the BERT model experiments to allow for comparison with BERT models. Specifically, the pre-trained Byte-Pair Encoding tokeniser from RoBERTa is used, merging adjacent bytes to form increasingly large tokens (Sennrich et al., 2016). This methodology ensures that the dataset will have no unknown tokens and can handle emojis correctly. Tokens are also added to the tokeniser for common words that are not otherwise recognised, like “NHS”; a “[scrubbed]” token for redacted words; for all bot messages that occur a significant enough number of times; and several complete messages that occur with sufficient frequency, such as initial and final messages sent by Shout Volunteers.

In addition to word embeddings, for each token, a separate embedding is added corresponding with whether the texter or Shout Volunteer wrote the token, similar to the learnt segment embedding in BERT (Devlin et al., 2019). This embedding mechanism is demonstrated in Table 3.

Embeddings are then passed through a fully connected layer, which converts them into a form more useful for the CNN layers without the need for fine-tuning of individual word embeddings, as some words might appear in the test set and not the training set, keeping word embeddings as close as possible to those that are pre-trained can reduce model overfitting. The model then applies three convolutional blocks with residual ‘skip’ connections, based on ResNet (He et al., 2016). The first block uses a stride of one, the following two use a stride of two, and all convolutional layers use a kernel size of three.⁴ Outputs from the CNN are then passed through three layers of a bi-directional GRU to model the long-term semantic relationships within a conversation. The six vector outputs from the GRU (i.e. one vector per layer per direction) are concatenated and passed through two fully connected layers, separated by a LeakyReLU activation. A diagram of the CNN-RNN model is shown in Fig. 5.

Multi-task training. While Random Forests can only be trained to predict a single label simultaneously, neural networks can simultaneously predict any number of labels. We, therefore, experimented with multi-task training using the CNN-RNN to benefit from correlations between different labels. The model is trained to output 489 labels simultaneously in the multi-task setting. These include all questions from the Shout Volunteer and texters post-conversation surveys; the Shout Volunteer skill labels, discussed in Section 4.3; and data on whether a conversation was escalated or led to an active rescue. Additionally, regular expression-based labels are added to extract ages from relevant messages, e.g. when a texter writes “I’m 25”, which pushes the model to pay attention to explicit statements of a texter's age. Merging such labels with texter survey data is likely unwise as the regular expression data may be less reliably accurate than texter survey data; for example, texters may be more likely to be asked their age if they are young, or maybe less honest if they worry it will affect how they are perceived.

Ordinal labels were converted into binary rank variables, such that an ordinal variable with labels 1, 2, and 3 would be mapped to two binary variables with values (0,0), (0,1), (1,1), as in Niu et al. (2016). Free text response labels were converted into word indicator vectors for the most commonly occurring words, with each vector position representing only whether a given word is present in the text. All words were stemmed, stop-words were removed, and only the most common words after that were kept.

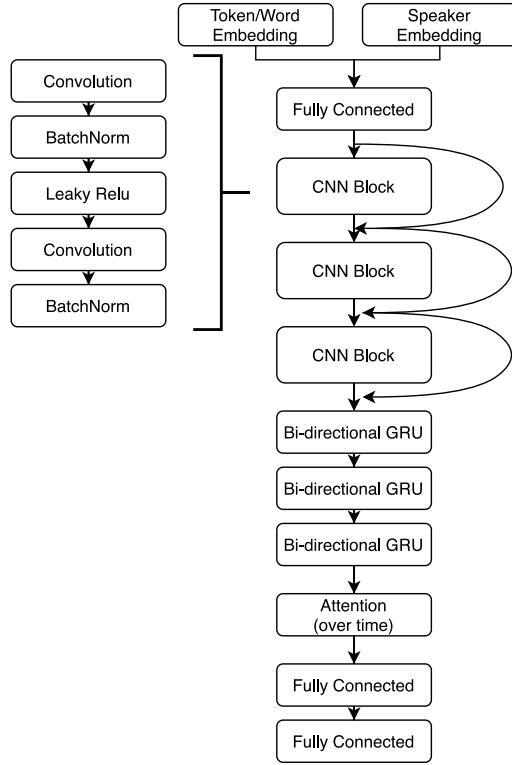
Due to the sparsity of various labels (e.g. texter survey answer labels), only labels present for a given example can be included in the loss function. Additionally, the losses for each label need to be

⁴ For skip connections for an underlying strided convolutional block, a single convolutional layer is applied with a kernel size of one and the relevant stride.

Table 3

Example of a conversation embedding, with the sequence of tokens in the conversations encoded simultaneously for both speakers. In this example, we abbreviated the Shout Volunteer as SV.

Input:	[START]	Help	[SEP]	Hi	,	please	tell	me	what's	wrong	[END]
Token embedding:	$E_{[\text{START}]}$	E_{Help}	$E_{[\text{SEP}]}$	E_{Hi}	$E_{,}$	E_{please}	E_{tell}	E_{me}	$E_{\text{what's}}$	E_{wrong}	$E_{[\text{END}]}$
	+	+	+	+	+	+	+	+	+	+	+
Speaker embedding:	E_0	E_{texter}	E_0	E_{SV}	E_{SV}	E_{SV}	E_{SV}	E_{SV}	E_{SV}	E_{SV}	E_0

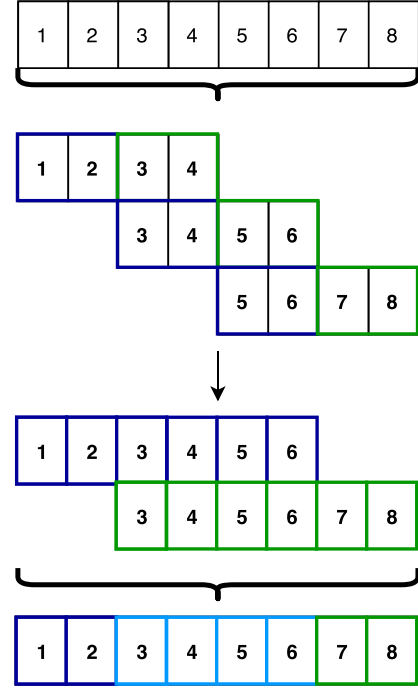
**Fig. 5.** Architecture of the CNN-RNN. Arrows denote skip connections.

combined in a weighted manner to minimise the variance in losses between examples. To achieve this, a mean loss score is computed per question by averaging the loss scores for all possible answers corresponding to a single question; then, all questions containing labels for a given example are combined by taking the mean. Binary Cross Entropy and Mean Squared Error are the loss functions for binary and continuous labels, respectively.

3.4. BERT

For all but the final experiment, a pre-trained “DistilRoBERTa” is used for its smaller size and performance. The somewhat bigger RoBERTa-base model was trained in the final experiments to understand the impact of increasing model size and predictive ability. The tokeniser used is described in Section 3.3.

The model is fine-tuned on an MLM task for five epochs with hyperparameters taken directly from BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b). Sequences for the MLM task are 512 token sub-sequences extracted from conversations with 256 token overlaps between sequences. A validation set ensured the model did not overfit the MLM task training set. Next, BERT is fine-tuned in a supervised learning fashion, predicting the relevant set of labels. As with RoBERTa (Liu et al., 2019b), there is no pre-train task associating the [CLS] tag at the beginning of sequences with the entire sequence, and therefore an attention layer was instead used to combine all BERT outputs.

**Fig. 6.** A visualisation of how inputs are passed through BERT with overlap and then recombined. In this example, the maximum sequence length is 4, and an overlap of 2 is used between segments. The blue segments are the left embeddings, and the green segments are the right embeddings.

To use BERT with sequences longer than 512 tokens, we combined outputs from multiple BERT forward passes over overlapping sequences with the maximum sequence length. Fig. 6 shows how long sequences can be segmented with overlap, passed through BERT, and reshaped to form a full sequence embedding. We then experimented with three methods of increasing complexity for converting these BERT output embeddings into label predictions: voting, attention, and a Transformer.

First, voting involves simply taking the mean of the model outputs from each sequence. Voting can be unweighted, where all sequences are considered equally, or weighted, whereby sequence outputs are weighted proportionally to the number of tokens they contain compared to another forward pass, reducing the impact of tokens appearing in two forward passes being counted twice. Both methods are used separately and evaluated in Section 4.1.3.

Combining forward passes with the attention mechanism is a slightly more sophisticated approach. At training time, BERT token-wise embeddings can be combined using an attention layer that takes a feature-wise weighted mean based on the importance of each feature of each token. This attention mechanism can map any number of inputs to a single output and combine token-wise embeddings from multiple forward passes. Again, to reduce the effect of overlapping tokens being counted twice, these token embeddings can be down-weighted by decreasing the magnitude of the tokens’ attention vectors, either before or after the softmax function is applied. Applying the down-weighting before softmax allows softmax to normalise the attention scores to

Table 4

Comparison of results from each main model across a select subset of labels. The metric shown is the Matthews Correlation Coefficient, chosen for its capacity to deal with the imbalanced nature of the dataset appropriately. For Voting over BERT, the results for the weighted method are shown here, and for Attention over BERT, the weight before softmax method results are displayed. The models marked with a * symbol reuse the same pre-trained distilRoBERTa model, and ToBERT-base uses a pre-trained RoBERTa-base model.

	Basic features	TF-IDF	CNN-RNN	Big bird	BERT* voting	BERT* Attention	ToBERT*	ToBERT-base
Topic:								
Self harm	0.116	0.581	0.738	0.665	0.747	0.755	0.769	0.777
Depression	0.115	0.319	0.264	0.449	0.454	0.454	0.482	0.480
Substance use	0.014	0.348	0.068	0.362	0.542	0.555	0.555	0.559
Suicide risk:								
Desire	0.255	0.521	0.733	0.712	0.068	0.743	0.752	0.758
Capability	0.349	0.531	0.739	0.734	0.741	0.753	0.783	0.793
Texter survey:								
Helpful?	0.241	0.286	0.428	0.500	0.480	0.505	0.537	0.531
Age $\leq$ 21	0.180	0.455	0.518	NA	0.723	0.737	0.751	0.761
Heterosexual	0.079	0.158	0.168	NA	0.211	0.199	0.206	0.204
Race: White	0.158	0.080	0.079	NA	0.166	0.182	0.219	0.172
Gender: Male	0.05387	0.2491	0.142	NA	0.307	0.304	0.384	0.384

sum to one, as is done in training; however, doing so can change the distribution of attention scores in an undesirable way. Down-weighting after softmax should have less effect on the attention score distribution and can be followed by simple linear normalisation for attention scores to sum to one. Both methods are used separately and evaluated in Section 4.1.3.

Lastly, for the ToBERT model, left- and right-embeddings of 256 token sub-sequences are concatenated feature-wise, with 0 padding added for tokens with either no left or no right-embedding, these embeddings are first combined using a two-layer neural network; then passed through a two-layer Transformer; then combined to form a single embedding vector using an attention mechanism; and finally converted to a prediction using a linear layer, the two-layer Transformer inherits the same dropout, feature dimension, feed-forward dimension, and activation function as the underlying BERT model, the Transformer model uses sinusoidal position embeddings, unlike RoBERTa's learnt embeddings, to allow for unseen or rare input lengths at test time compared to training time.

3.5. Big bird

This project used Big Bird with a sequence classification head in a multi-task setting (for predicting suicide risk, depression, self-harm, substance use and conversation helpfulness). For tokenisation, conversations were truncated to 2048 tokens. This choice of sequence length was motivated by GPU limitations, which did not allow for the use of Big Bird's maximum capacity of 4096 tokens. The model was implemented using the Trainer API from the Transformers library (Wolf et al., 2020) and trained for five epochs with a batch size of two. The batch size was determined by GPU limitations, which arose when Big Bird was trained at longer sequence lengths. An Adam optimiser with weight decay regularisation (Loshchilov and Hutter, 2017) and a learning rate of 10^{-4} was used for training. The loss was calculated by combining the binary cross-entropy function with a sigmoid layer.

Before training the model on the classification task, the model was fine-tuned on a Masked Language Modelling (MLM) task for five epochs using the hyperparameters described above. A validation set was used to prevent overfitting the training set during the MLM task.

It is important to note that Big Bird was trained on a smaller subset of data than the other BERT models. While BERT exploits semi-supervised learning, Big Bird does not. As such, the dataset for Big Bird comprises fully labelled conversations (i.e. conversations with the nine labels described above), which total 54,482 conversations. Additionally, Big Bird models were not developed for predicting texter demographics or conversation success metrics (other than conversation helpfulness), for reasons which will be discussed in Section 4.1.4.

4. Results

In this section, we first introduce some of the evaluation results for the main model, tested across all mental health tasks introduced in previous sections. Second, we perform in-depth experiments on texter demographics, suicide risk and conversation success metrics.

4.1. Model evaluation

This section will evaluate model performances for each model described in Section 3 and will conclude with a discussion of future directions for model improvements. Results from each model across a small subset of labels can be found in Table 4. We used the Matthews Correlation Coefficient (MCC) as an evaluation metric for model comparison due to the imbalanced nature of these labels. Unsurprisingly, increasingly complex models show superior performance, with deep learning significantly outperforming baselines and transformer-based models, considerably improving the CNN-RNN.

4.1.1. Baseline models

The baseline models (Basic Features, TF-IDF and CNN RNN) show performance that varies significantly between labels, which may be for several reasons. The basic features model succeeds when a label highly correlates with length- or speed-related features. For example, the essential features model performs relatively well at predicting texter helpfulness reviews, perhaps because more extended conversations or conversations with faster responses are more helpful. Suicidal capability also appears to be predictable from these features, though the relationship may not be causal; it is perhaps more likely that conversations with high suicide risk are escalated, leading to more extended discussions or ones with faster response time.

TF-IDF performs surprisingly well at predicting age, as compared with the CNN-RNN, and performs even better at predicting the conversation topic of depression. TF-IDF can be expected to perform well on topic detection, as the presence and frequency of specific words and phrases can typically identify topics. For example, TF-IDF predicts the topic "self-harm" relatively well, likely because conversations regarding self-harm often use particular words like "cutting". Word usage may also play a significant role in differentiating between individuals of different ages, as discussed in Section 2. By contrast, the CNN-RNN has the benefit of accessing long-term relationships between tokens, which may be more helpful for some labels than others.

Table 5

Comparing single-task, multi-task, and multi-task fine tuning using the CNN-RNN model on two specific labels.

	Texter survey: Helpful	Risk ladder: Time frame
All tasks	0.4282	0.6989
Single task	0.4475	0.6865
Fine-tuned	0.4667	0.7113

4.1.2. Multi-task learning

The TF-IDF model's superior performance over the CNN-RNN suggests that multi-task learning may impair the ability to model specific labels. Though the strong TF-IDF performance for the conversation topic 'depression' label may be explained by the types of features relevant for detecting topics, the CNN-RNN model has access to precisely the same features. The likely reason for TF-IDF's superior performance on the depression label is that a new Random Forest is trained independently on each label. In contrast, only a small subset of model weights predict each label in the multi-task CNN-RNN model. Moreover, multi-task learning may also lead to compromises during optimisation — paying more attention to some labels, especially those that are easier to predict. The CNN-RNN model is likely too small to maintain relevant features for predicting all labels, resulting in poorer performance for some labels.

While a larger model could be used — as with the T5 model, which uses 11 billion parameters — the authors describe how task interference impairs performance on multi-task-only training and propose fine-tuning a multi-task model on a single or small set of tasks to improve performance (Raffel et al., 2020). To evaluate performance improvements from fine-tuning with a relatively small model, the CNN-RNN model was re-trained on two labels (survey helpfulness and suicidal time frame) separately, once from scratch and once from a multi-task training checkpoint. Results are shown in Table 5.

Fine-tuning slightly improved results for both labels, compared with single- or multi-task training. Interestingly, the helpfulness label performed slightly better trained alone, compared to only being trained in a multi-task setting, while the time-frame label performed slightly worse alone. A likely explanation is that the texter survey helpfulness label was given less attention during the optimisation due to its relatively small contribution to the loss function, being only one of a large number of labels and also only being present in a relatively small subset of conversations. As a result, the intermediate feature vectors would contain less signal for predicting helpfulness. That said, the label clearly benefits from multi-task training, as seen in the improved metric in the fine-tuned setting. By contrast, the time-frame label is present in nearly all conversations. Therefore, the benefit from multi-task training leads to a slightly better metric, even in the multi-task-only setting.

Notably, the difference between metrics in all three cases is slight. While fine-tuning provides benefits, it also requires training separate models for each label. Therefore, compromises can be made, such as fine-tuning several related labels simultaneously, fine-tuning only the final layers of the model and not the entire model to reuse specific layers, or simply using the multi-task model where its performance is sufficiently strong to outweigh the benefit from fine-tuning.

4.1.3. BERT

Using BERT significantly improves performance across all labels, which is mainly a testament to the task's difficulty. As discussed in the background (Section 1.4), BERT's Transformer architecture can model long-term relationships far better than CNNs or RNNs. BERT's semi-supervised approach allows for better generalisation.

As explained in the previous section, BERT was first fine-tuned with a Masked Language Modelling task. The loss curves on a training and hidden validation set, shown in Fig. 7, show a smooth loss and a good balance between bias and variance. BERT can also be manually evaluated by selecting random conversation segments from the test set,

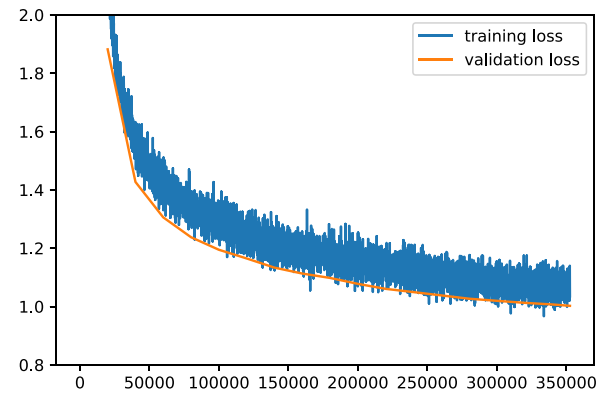


Fig. 7. Training and Validation loss curves for Masked Language Model training (distilRoBERTa).

masking random tokens, and generating predictions of those masked tokens. As expected, outputs were often correct (F1 score varying from 0.46 to 0.92) and matched the tokens a human should also predict for the same masked tokens.

Returning to supervised learning, the end-to-end trained ToBERT model outperforms the other two BERT methods on nearly all labels. This is expected because ToBERT has more model weights from the two additional layers. Crucially, unlike ordinary BERT, ToBERT has access to the entire conversation at once and can, therefore, track long-term linguistic relationships across a conversation; for labels with less relevant long-term relationships, such as demographic labels, the Transformer method will likely provide less benefit than a larger model with comparable training time. Attention over BERT proves to be an effective method for combining BERT outputs. It improves performance compared to voting on several labels, with virtually no change in the BERT model or training loop. Interestingly, for the Heterosexual/Straight label, voting outperforms Attention or Transformer over BERT. This may be an artefact of multi-task learning, where more difficult-to-predict labels may receive less training and disappear with label fine-tuning.

Voting and Attention over BERT methods are compared in Table 6. For voting, the weighted and unweighted variations produced very similar results, with the weighted method slightly outperforming the unweighted variation but within a reasonable margin of error. For Attention over BERT, the weight after the softmax method generally outperforms the alternatives. As expected, Attention over BERT outperforms the simpler Voting method.

4.1.4. Big bird

While Big Bird performs better than the baseline models on most labels, it is universally outperformed by the other BERT-based models implemented in this project. The former observation is expected to highlight the complexity of the task at hand, requiring a more sophisticated model. What is unexpected is that the CNN-RNN model outperforms Big Bird on the self-harm and suicidal desire labels. There are two possible explanations for this. Firstly, as described previously in Section 4.1.2, multi-tasking may lead to compromises during training — where performance on some labels is preferentially optimised at the expense of others. As such, the CNN-RNN model may have preferentially optimised the self-harm and suicide risk labels at the expense of other labels. This is supported by the CNN-RNN model's much poorer performance on substance use and depression labels. In contrast, the prediction performance of the Big Bird model is relatively more balanced across the multiple labels.

Secondly, another reason for the better performance of the CNN-RNN model could be the availability of more data. While the Big Bird model was trained exclusively on conversations with complete labels,

Table 6

Comparison over methods for combining BERT forward passes. In this example, we abbreviated the Shout Volunteer as SV.

	Age: 21 or under	Heterosex- ual/ Straight	Topic: Self harm	Topic: Depression	Texter: Helpful	SV: Suicidal capability
Voting unweighted	0.723	0.204	0.748	0.453	0.486	0.737
Voting weighted	0.723	0.211	0.747	0.454	0.480	0.741
Attention unweighted	0.711	0.210	0.761	0.449	0.486	0.764
Weight after softmax	0.737	0.199	0.755	0.454	0.505	0.753
Weight before softmax	0.731	0.210	0.766	0.452	0.480	0.760

Table 7

Results from the final ToBERT-base model, i.e. Transformer over RoBERT-Base. A greater selection of metrics is shown to understand model performance better.

	MCC	Avg precision	F1 score	AUC-ROC	Accuracy
Topic:					
Self Harm	0.777	0.894	0.815	0.974	0.938
Depression	0.480	0.741	0.662	0.829	0.762
Substance Use	0.559	0.611	0.570	0.960	0.978
Suicide risk:					
Desire	0.758	0.895	0.838	0.956	0.893
Capability	0.793	0.875	0.822	0.981	0.949
Texter survey:					
Helpful?	0.531	0.985	0.943	0.908	0.900
Age > 21	0.761	0.943	0.858	0.957	0.884
Heterosexual	0.204	0.709	0.658	0.652	0.610
Race: White	0.172	0.943	0.922	0.673	0.859
Gender: Male	0.384	0.443	0.463	0.803	0.863

the CNN-RNN model was introduced on the whole data set — omitting missing labels. Since most conversations are labelled with topics and suicide risks, these labels should benefit from having more data. In contrast, only the conversation helpfulness label would have similar data. This may explain better performance on the self-harm and suicidal desire labels, but not its much poorer performance on the depression and substance use labels. It is worth noting, however, that even with this advantage, the CNN-RNN model may still be “handicapped” by the many more labels (489) it is modelling. In contrast, the Big Bird model was only used to model a small subset of nine labels, the CNN-RNN model. The number of weights the CNN-RNN model can allocate to each label could be much fewer than Big Bird.

There are three explanations for the poorer performance of the Big Bird model relative to the other BERT-based models. Firstly, this could point to the utility of semi-supervised learning in exploiting unlabelled data to improve predictive performance. Considering that the complete dataset of conversations used for the other BERT-based models is roughly three and a half times the size used for Big Bird, this provides significantly more information for the other BERT-based models to work with and learn from. Secondly, and more simply, the other BERT-based models – ToBERT and ToBERT-base in particular – may generally be superior to model these conversation data.

Lastly, the sequence truncation to 2048 tokens may have affected Big Bird’s performance. In contrast, the other BERT-based models can cover the entire conversation length. However, since extremely long conversations were eliminated from the analysis (see Section 3.1), it is unlikely that the BERT-based models would have access to significantly more information. Moreover, given that the conversation stages involve exploring the texters’ problems earlier in the conversation, labels such as topics and suicide risk would be apparent in the earlier part of the conversation. They should not be too affected if the latter part of the conversation is truncated. Indeed, performance on only the conversation helpfulness label is likely to suffer since determining whether the conversation was helpful or not may only be apparent towards the end of the conversation. Interestingly, Big Bird’s performance on this label is comparable to, or not much poorer than, the other BERT-based models, which provides evidence against the idea that the Big Bird model outperforms the other BERT-based models because the latter has the advantage of longer sequence length.

Table 8

Results on final Transformer Over BERT model on Suicide Risk Ladder labels.

	F1-Score	AUC-ROC	Avg precision	MCC
Desire (thoughts)	0.834	0.955	0.895	0.752
Intent (plan)	0.837	0.978	0.898	0.797
Capability (means)	0.814	0.981	0.871	0.783
Timeframe	0.774	0.985	0.824	0.756

4.1.5. Irreducible error

The last model shown in Table 4 is ToBERT-base, a ToBERT model built upon a pre-trained RoBERTa-base model rather than distilRoBERTa-base, with several metrics shown in Table 7. A slightly larger model was used to explore how larger models will improve performance on each relevant task. It is worth noting that the difference between the two underlying BERT models used here is the number of Transformer layers rather than the embedding dimension between layers (768), so it is unlikely to reduce label interference effects discussed in Section 4.1.2.

Interestingly, while ToBERT-base generally outperforms ToBERT, this is not universal. Since the ToBERT models are trained for just three epochs, it is unlikely that poorer ToBERT-base performance is due to model overfitting. A possibly better explanation is the presence of *irreducible error* - intrinsic label noise that cannot be modelled in terms of the relevant inputs, no matter how powerful the model is.

One source of irreducible error is labelling errors and inconsistencies. For example, the topic label of ‘depression’ is only correctly identified by the model in a mere 76.2% of test conversations. One might presume this task would be easier to identify from the conversation than the texter helpfulness review rating, which was correctly predicted in 90.0% of test cases. The most likely explanation for this failure to recognise ‘depression’ in some conversations is noise due to variations in how Shout Volunteers determine the presence of depression and label the conversations. If this is the case, it is likely that reducing label noise, for example, by evaluating gold standard labels identified by the Clinical Supervisors, may produce superior model performance. A proper analysis of noise from mislabelling will be explicitly addressed for the Suicide Risk Ladder in Section 4.2.

4.2. Predicting texter suicide risk

Results from evaluating the ToBERT model on Suicide Risk Ladder labels can be seen in Table 8. Upon initial review, these metrics may either seem very high, relative to the difficulty of the task at hand, or unacceptably low, given the cost of errors in assessing risk. Due to the large quantity of data and complexity of the ToBERT model, it is unlikely that significant improvements upon these metrics can come from Machine Learning solutions. Admittedly, the model used is considered “small” in terms of the number of parameters trained compared to other state-of-the-art deep learning models, and for these metrics, label fine-tuning did not take place. Using a larger model and fine-tuning the relevant labels would be expected to afford some very slight performance improvement, though at a significant computational cost. Nonetheless, the likely explanation for the relatively low performance is an irreducible error.

Table 9

A performance comparison for metrics developed in this section, using the ToBERT model. Performance is measured on a hidden test set. Raw model outputs are used for texter reviews without applying the decision function.

	Pearson Corr. Coeff	Explained variance	# ground truth
Texter reviews:	0.562	0.219	30,085
Avg. texter review:	0.509	0.252	134,509
Shout volunteer experience	0.724	0.524	233,882

4.2.1. Suicide risk: Discussion

Following this review, there are several considerations relevant to error analysis. First, model performance is likely considerably higher than the test set metrics indicated. One reason is that human Shout Volunteers and even expert reviewers may be more prone to oversight than an AI. Reviewers have limited time and attention capabilities and were generally looking for the laddering up stages in a conversation; if this laddering up did not follow the format they expected or was not performed correctly, they were more likely to miss the present risk. By contrast, the AI is trained with neural dropout, ensuring it learns a robust algorithm that considers the entire conversation in decision-making, resulting in outputs that likely reduce noise in underlying labels.

Second, significant subjectivity is likely present in determining suicide risk at a given level. Shout Volunteers must decide where to start laddering up based on the risk initially volunteered by the texter; determining the risk indicated by the texter and whether clarification is needed may involve a degree of subjectivity, with different experts interpreting a message differently. This subjectivity may decrease with experience level, with very experienced Shout Volunteers agreeing more often with each other than less experienced Shout Volunteers would. Beyond subjectivity, an intrinsic disparity may exist between Shout Volunteers and Clinical Supervisors in assessing risk: coaches we spoke with were more worried about false positives than false negatives, but Shout Volunteers expressed the opposite concern.

Model outputs may, therefore “average out” subjectivity, outputting what an average Shout Volunteer would predict in a given conversation. While that may be desirable, leading to greater consistency, the average Shout Volunteer label may not correspond well with a “gold standard” risk assessment. In future work, a set of “gold standard” risk assessments may be fine-tuned to produce model outputs less biased by the average Shout Volunteer and more inclined towards desirable outcomes. Alternatively, risk assessments may be trained with a weighted loss function, considering more experienced Shout Volunteer assessments than newer Shout Volunteers, assuming that experienced Shout Volunteers label more closely to a “gold standard”.

4.3. Conversation success metrics

This section discusses the development and measurement of three metrics for conversation success using the application of the previously developed deep learning model, namely (1) how helpful would the texter have found this conversation?; (2) how helpful would one expect that texters generally find this Volunteer, based only on this conversation?; (3) how experienced does this Volunteer appear, based only on this conversation? These metrics are summarised in Table 9.

4.3.1. Texter reviews

The texter conversation helpfulness label was successfully predicted with reasonably high performance: 0.531 MCC, 0.943 F1 Score and 90% Accuracy. Predictions for whether the texter-reported level of conversation helpfulness was greater than 1, 2, 3, or 4 had MCC values of 0.51, 0.49, 0.45, and 0.39, respectively. Therefore, These predictions can be used as a one-conversation success metric, answering the question: how helpful would the texter have found this conversation?

Caution is warranted in treating these reviews directly as success metrics, as these review metrics are likely noisy. As discussed in Section 1.1, how helpful the texter finds their state of mind can influence

the conversation at the time of the conversation rather than the Shout Volunteer’s skill or the inherent value of the conversation. Additionally, not all texters may interpret the questions about conversation helpfulness or the 1–5 scale similarly. This latter point may explain the significant disparity in MCC between predicting different helpfulness levels. This and many other factors affecting a texter’s rating are not predictable based on conversation text and are, therefore, unlikely to be accounted for in model predictions. Model predictions may ‘average out’ texter biases, predicting how an average texter would rate a conversation rather than the particular texter. The unpredictable texter noise that affects ratings introduces considerable irreducible error to predictions of texter reviews; as a result, an MCC lower than 1.0 is to be expected – and even wished for – as it indicates that this noise may be being averaged out to reveal something more fundamental about the conversation itself.

4.3.2. Volunteer success metric: from the perspective of an “average texter”

Another conversation success metric centres around the performance of the Volunteer in the conversation as perceived by an “average texter”. While model predictions of a texter rating may reduce some noise, model predictions can be influenced by predictable factors related to the texter. For instance, texters discussing physical abuse have been found to find conversations more helpful (Gould et al., 2022). The model may, therefore, look for and use the conversation topic to predict the texter review.

To escape that noise, the model is trained to predict the average review for a Shout Volunteer based on a single conversation rather than the single review corresponding with that conversation itself. This model output corresponds with the answer to the question “how helpful would one expect that texters generally find this Volunteer, based only on this conversation?” An excellent response to this question will pay less attention to the texter’s issue and state of mind and more attention to the factors differentiating a Shout Volunteer that consistently receives high reviews and one that consistently receives low reviews.

To train a model to answer this question, Shout Volunteers were assigned a “skill” rating for each conversation, using the average of all reviews before the conversation of interest. Rather than using the numerical scale, which may add noise due to the interpretation of the scale, only the binary “Did you find this conversation helpful?” is used. Additionally, this label is only used once a Shout Volunteer has received at least five reviews (on average, after about 50 completed conversations) to minimise bias from a single texter’s review. Lastly, because assessing the relative Shout Volunteer skill or success is the goal rather than the underlying fraction of helpful conversations, these ratings are replaced with corresponding quantiles relative to other Shout Volunteers.

It is important to note that using texter reviews as a proxy for Shout Volunteer skill or conversation success assumes that an average texter review reflects a Shout Volunteer’s abilities in some way. One may argue that the Shout Volunteer skill may not be reflected at all by the texter survey — it is the job of a Shout Volunteer to bring texters from a “hot moment to a cool calm”, to use Shout’s phrasing, which might not be reflected in texter helpfulness ratings. Therefore, the average texter review metric must be interpreted only as what it is — not a measure of skill per se, but a measure of the texter helpfulness reviews the Shout Volunteer who took this conversation will likely get. If this

is not exactly a skill, it may be something related or independently valuable. This measure also assumes that conversations (e.g. type of issue and severity) that each Shout Volunteer handles are generally evenly distributed. Bias can be introduced, for example, if many Shout Volunteers only work at night and reviews are usually better or worse for nighttime conversations. Additionally, a source of bias may arise from top Shout Volunteers and Clinical Supervisors being referred to the most severe discussions, which may have less representative and harsher reviews. Ultimately, due to the size of the dataset and the number of unique Shout Volunteers, the effect of this bias is unlikely to be significant.

For this metric, 25.2% of the ground truth label variance is accounted for in model prediction. While it is hard to identify the extent of the irreducible error, it is logically unlikely that much more than this degree of label variation could be successfully modelled. A Shout Volunteer's ground truth label (average texter helpfulness rating) may vary significantly between two adjacent conversations, especially if the Shout Volunteer has recently received just one outlier texter survey review. The level of noise is, therefore, likely high. The explained variance score is insufficient to demonstrate that this metric is *useful*, which would require time for Shout staff to use in conversation review and determine its value to their work. Still, it is sufficient to show that the model can predict this metric successfully.

4.3.3. Shout Volunteer level of experience

Within the Shout platform, Shout Volunteers are assigned experience levels based on completed conversations, with more experienced Shout Volunteers granted privileges such as participating in multiple conversations simultaneously. Experience is, therefore, implicitly and justifiably proxied for skill. However, caution should be maintained when asserting that discussions with more experienced Shout Volunteers are better since Shout coaches suggest that the correlation between Shout Volunteer experience and conversation success is possibly non-linear. While new Shout Volunteers are likelier to adhere to prescribed structures, intermediate Shout Volunteers may deviate – for better or worse – as they develop their conversational style. Nonetheless, as experience is currently proxied for skill, developing a label that answers the question “based on this conversation, how much experience do you expect that the Shout Volunteer has?” has value, especially when compared with the previously described ones to determine a conversation's success and Shout Volunteer performance.

The explained variance score achieved between predictions and labels is 52.4%. This is much higher than average texter responses and is expected because the Shout Volunteer experience ranking changes more slowly and smoothly than average reviews and contains far less noise. However, similarly to the other two labels, a very high coefficient of determination is unlikely desirable. The goal is to identify the likely skill level of a Shout Volunteer in a conversation, not the exact number of conversations they have had. Therefore, failure to perfectly identify the ground truth is to be expected. Again, the usefulness of this metric to Shout will be determined following sufficient longer-term use in Shout Volunteer and conversation review.

4.3.4. Conversation success metric: Discussion

Evaluating the three conversation success metrics is difficult because all three metrics are trained specifically for noise reduction. As such, perfect accuracy between model predictions and corresponding ground truths would not be a testament to the value of these labels but rather the contrary; even very skilled Shout Volunteers may occasionally have conversations indicative of a low-skill Shout Volunteer, and likewise, Shout Volunteers with minimal experience may sometimes speak in the manner of a very experienced counterpart. Nonetheless, a non-existent correlation between ground truth labels and predicted labels is undoubtedly a sign of failure. Hence, the correlation values presented are at least “sanity checks” for the model training success.

The next step in evaluating these metrics is ongoing and is based on their usefulness. Each metric is used separately to identify the best and worst conversations from select Shout Volunteers; these conversations will then be passed on to coaches to read and provide feedback to the Shout Volunteers. The coaches will then be able to give feedback on how useful the metric they used was for identifying “good” and “bad” conversations. This feedback can be used to improve these metrics and ensure they are of high quality for later use with real-time models.

5. Ethical concerns and limitations of using AI technologies in mental health applications

Digital provision of mental health support is increasingly important for care to be timely, accessible, and feasibly delivered at scale. This study demonstrates the potential for AI tools to be helpfully employed for rapid service evaluation and timely insights into mental health risks from the large datasets generated by digital mental health services. However, AI tools must be employed with abundant care and in deep consideration of ethical implications, such as mitigating bias within AI models, ensuring transparency in decision-making, and maintaining human oversight to balance machine predictions with professional judgment. This study took extensive precautions to ensure the dataset was anonymised and highly secure, addressing key data privacy concerns. Any future work to employ deep learning models alongside human supporters in mental healthcare must also carefully consider these ethical dimensions and centre the desires and rights of service users.

The deep learning models in this project, primarily attention-based and BERT-based, face limitations like data sparsity, which hampers performance in data-scarce settings while also coming with high computational costs, making them resource-intensive. Additionally, these models have limited decision windows, often struggling with long-range context. In terms of interpretability, their black-box nature is a challenge, though methods like LIME can be used to provide some insight into their decision-making processes. While effective in image processing, the CNN-based models are less suited for complex NLP tasks as they struggle with capturing long-range dependencies and hierarchical relationships in text. On the other hand, BigBird, designed to handle longer sequences efficiently, mitigates some of these issues but still faces challenges with model interpretability and requires fine-tuning to specific domains for optimal performance.

6. Conclusion and future work

This project is the first of its kind to develop a state-of-the-art deep learning model to understand the conversation content of a distress text messaging service and apply it to gain insights relevant to increased understanding of individuals experiencing mental distress and seeking support from non-traditional services and improvement of the text messaging service itself.

This objective was successfully met, with the best-performing model, ToBERT, performing well across all metrics when making predictions about texter demographics, suicide risk level, and conversation helpfulness. Moreover, ToBERT performs comparably to its larger counterpart, ToBERT-base - even outperforming it on some labels. This is desirable because, being smaller, ToBERT has the added speed advantage when making predictions. Our model evaluation highlighted the need for a Transformer-based model and the complexity of BERT to meet these objectives. Our results demonstrate the benefit of semi-supervised learning through pre-training and multi-task learning. Fine-tuning on specific labels after multi-task training was also shown to improve model performance, especially in the relatively small models we initially explored.

Indeed, one could expect that fine-tuning labels generally improves performance, especially on harder-to-predict or less frequently occurring labels. However, fine-tuning many different labels would require

producing many models, drastically increasing memory requirements and time for training and inference. For this reason, the multi-task model was considered here to be the optimal compromise between accuracy and computational demand. In future research, in addition to using a bigger model, further compromises could be made between separately using a multi-task-only model or several fine-tuned models. For example, early model layers could be frozen after multi-task training, and later layers could be differentiated by task. Alternatively, labels could be fine-tuned in groups that correspond to similar features.

As discussed in Section 3, the Transformer over BERT model adjusts the methodology from Pappagari et al. (2019), training the Transformer and BERT simultaneously, end-to-end. The benefits of training end-to-end were not evaluated here due to time constraints, but it required a little additional computational time. For the Transformer over the BERT model, the large Transformer acting on the BERT outputs dominates training time due to Transformers' $O(N^2)$ computational time requirement, with N representing sequence length. As a result, the cost of training end-to-end compared to training BERT first and then the Transformer was negligible. In future research, the difference in performance may be evaluated. Another methodology that could be considered for future experiments is reducing sequence length between Transformers, for example, using a Maximum-Pooling layer to minimise training time significantly. This was unnecessary here because the maximum sequence length was not excessively large.

Regarding future work for this project, one significant limitation in using texter reviews is the limited number of texter-reviewed conversations. As mentioned, averages are only taken once a Shout Volunteer has received five reviews, meaning many conversations are labelled based on only five or slightly more texter reviews. It is expected that the greater the number of conversations averaged, the less noise there will be due to any single outlying conversation or review; one solution to this problem is to use model-predicted reviews in addition to – or instead of – actual texter reviews. A first model would predict individual texter reviews, and a second independent model would then use these model outputs in place of texter reviews. A combination approach may also be used, with the model-predicted reviews weighted less. This methodology may produce significantly better results but requires sufficient success in predicting texter reviews to avoid compounding errors.

Additionally, it is worth mentioning that the metrics provided in this section are produced from a multi-task-only model. Methods that may improve this metric include those described earlier, such as using a larger model and label fine-tuning. As explained in Section 4.1, with the CNN-RNN, label fine-tuning for the texter helpfulness rating brought an absolute MCC increase of 0.0385 compared to the multi-task-only model. However, the irreducible error for this label may be significant, and thus, a much higher metric is unlikely to be achievable or desirable for even a much bigger model. Moreover, as noted in Section 2.2, a few long discussions were excluded from this study as they would require additional work to segment the conversations. This could be performed as a future work project, where conversations are first passed through a stage segmentation filter and then analysed by our model.

Finally, a methodology not assessed here that may prove helpful involves training a model to predict texter reviews, average reviews, or Shout Volunteer experience *independently* of confounding variables. Confounding variables include conversation topic, time of day, texter demographics, risk of suicide/self-harm, or any other conversational factors irrelevant to Shout Volunteer skill. Still, they may influence texter reviews or correlate with a Shout Volunteer's experience level. A promising way to do this involves adversarial learning in which a model is trained to predict a target without influence from confounding variables. A method for adversarial training can be found in Ganin et al. (2016), in which confounding variables are predicted by the model but trained with inverted gradients so that the model learns the correct task without knowing the features to predict the confounding variables.

CRediT authorship contribution statement

Daniel Cahn: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Sarah Yeoh:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Lakshya Soni:** Writing – original draft. **Ariele Noble:** Resources, Conceptualization. **Mark A. Ungless:** Funding acquisition, Conceptualization. **Emma Lawrance:** Writing – review & editing, Validation, Supervision, Conceptualization. **Ovidiu Șerban:** Writing – original draft, Validation, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Emma Lawrance's work on this project is partially funded by the Rayne Foundation. The authors would like to thank the team at Mental Health Innovations, the organisation supporting the operation of the Shout platform, for their help and support during this project, which includes access to the data and financial support to run the experiments on their servers. We would also like to thank all the Clinical Supervisors and Shout Volunteers for their hard work in supporting this platform: without their hard work and dedication, this project would not have been possible. The authors thank Edward Conway, Ann Farrant, Emily Meli and Podd Basquille for their technical support during the project. We would also like to thank Louise Gagan and Karen-Anne Fisher for their help reviewing the suicide risk results. Finally, the authors would like to thank Professor Lucia Specia for her marking and feedback on Daniel's and Sarah's MSc projects and Dr Robert Peach and Professor Mauricio Barahona for their input on the early versions of this work.

References

- Ainslie, J., Ontanon, S., Alberti, C., Cvicek, V., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q., Yang, L., 2020. ETC: Encoding long and structured inputs in transformers. arXiv preprint arXiv:2004.08483.
- Aljohani, T., Cristea, A.I., 2021. Learners demographics classification on MOOCs during the COVID-19: author profiling via deep learning based on semantic and syntactic representations. *Front. Res. Metr. Anal.* 6.
- Althoff, T., Clark, K., Leskovec, J., 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Trans. Assoc. Comput. Linguist.* 4, 463–476.
- Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M., 2018. Simply the best: minimalist system trumps complex models in author profiling. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 143–156.
- Bayram, U., Benhiba, L., 2021. Determining a person's suicide risk by voting on the short-term history of tweets for the CLPsych 2021 shared task. In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics, pp. 81–86. <http://dx.doi.org/10.18653/v1/2021.clpsych-1.8>, Online, URL <https://aclanthology.org/2021.clpsych-1.8>.
- Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv:2004.05150.
- Cesare, N., Grant, C., Nsoesie, E.O., 2017. Detection of user demographics on social media: A review of methods and recommendations for best practices. arXiv preprint arXiv:1702.01807.
- Chamberlain, A., Smart, M., 2017. Give to Get: A Mechanism to Reduce Bias in Online Reviews. Technical Report. Glassdoor Research Report.
- Child, R., Gray, S., Radford, A., Sutskever, I., 2019. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.
- Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning*. pp. 160–167.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41 (6), 391–407.

- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT Conference*. pp. 4171–4186.
- Duan, Y., Wu, O., 2016. Learning with auxiliary less-noisy labels. *IEEE Trans. Neural Netw. Learn. Syst.* 28 (7), 1716–1721.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17 (1), 2030–2096.
- Gould, M.S., Pisani, A., Gallo, C., Ertefaie, A., Harrington, D., Kelberman, C., Green, S., 2022. Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. *Suicide Life-Threat. Behav.*
- Halko, N., Martinsson, P.-G., Tropp, J.A., 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* 53 (2), 217–288.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE*.
- Kaiser, L., Sutskever, I., 2016. Neural GPUs learn algorithms. In: Bengio, Y., LeCun, Y. (Eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Liu, X., Gao, J., He, X., Deng, L., Duh, K., Wang, Y.-y., 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 912–921.
- Liu, X., He, P., Chen, W., Gao, J., 2019a. Multi-task deep neural networks for natural language understanding. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. RoBERTa: A robustly optimized BERT pretraining approach. [arXiv:1907.11692](https://arxiv.org/abs/1907.11692).
- Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. [arXiv preprint arXiv:1711.05101](https://arxiv.org/abs/1711.05101).
- Macavaney, S., Mittu, A., Coppersmith, G., Leintz, J., Resnik, P., 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. pp. 70–80.
- Mohammadi, E., Amini, H., Kosseim, L., 2019. CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Association for Computational Linguistics, Minneapolis, Minnesota*, pp. 34–38. [http://dx.doi.org/10.18653/v1/W19-3004](https://dx.doi.org/10.18653/v1/W19-3004), URL <https://aclanthology.org/W19-3004>.
- Morales, M., Dey, P., Kohli, K., 2021. Team 9: A comparison of Simple vs. Complex models for suicide risk assessment. In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access. Association for Computational Linguistics*, pp. 99–102. [http://dx.doi.org/10.18653/v1/2021.clpsych-1.11](https://dx.doi.org/10.18653/v1/2021.clpsych-1.11), Online, URL <https://aclanthology.org/2021.clpsych-1.11>.
- Nam, J., Kim, J., Mencía, E.L., Gurevych, I., Fürnkranz, J., 2014. Large-scale multi-label text classification—revisiting neural networks. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer*, pp. 437–452.
- Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T., 2013. “How old do you think I am?” A study of language and age in Twitter. In: *Proceedings of the International AAAI Conference on Web and Social Media*. 7, pp. 439–448.
- Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G., 2016. Ordinal regression with multiple output cnn for age estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4920–4928.
- Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N., 2019. Hierarchical transformers for long document classification. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, IEEE*, pp. 838–844.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pisani, A.R., Gould, M.S., Gallo, C., Ertefaie, A., Kelberman, C., Harrington, D., Weller, D., Green, S., 2022. Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. *Suicide Life-Threat. Behav.*
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1–67.
- Ragheb, W., Aze, J., Bringay, S., Servajean, M., 2021. Negatively correlated noisy learners for at-risk user detection on social networks: A study on depression, anorexia, self-harm and suicide. *IEEE Trans. Knowl. Data Eng.*
- Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pp. 1715–1725.
- op Vollenbroek, M.B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M., 2016. Gronup: Groningen user profiling notebook for PAN at CLEF. In: *CLEF 2016 Evaluation Labs and Workshop—Working Notes Papers*.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on EMNLP: System Demonstrations. Association for Computational Linguistics*, pp. 38–45.
- Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al., 2020. Big bird: Transformers for longer sequences. In: *NeurIPS*. pp. 17283–17297.
- Zhang, M.-L., Zhou, Z.-H., 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. Data Eng.* 18 (10), 1338–1351.
- Zirliky, A., Resnik, P., Uzuner, O., Hollingshead, K., 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. pp. 24–33.