# Customized Training of Pretrained Language Models to Detect Post Intents in Online Health Support Groups

Tootiya Giyahchi, Sameer Singh, Ian Harris, and Cornelia Pechmann

**Abstract** Online support groups offer low-cost and accessible health and mental health support, but low engagement challenges their effectiveness. To encourage engagement and provide evidence-based responses appropriate to participants' needs, we propose an intent detection model for online support groups based on state-of-the-art natural language processing methods. Such a model enables a chatbot that can increase interactions and improve discussion quality. Posts in social media are often short and noisy, especially in group chat. Furthermore, many intents lack data, overlap and/or have specific priorities. We create a human-annotated dataset of posts with intent labels from 45 three-month online support groups for quitting smoking. We then train and examine models to predict the intent behind each post. To reduce the effect of noisy and sparse data, we fine-tune a massive pretrained language model. Also, to represent the unique relationships between intents, we design customized loss functions for training. Empirical evaluations show significant performance improvements with the proposed method; our best model obtains 95.5% accuracy. We also use a fine-grained set of intents and obtain higher accuracy compared to prior models on online health forums and communities. Accurate detection of fine-grained intents opens up new opportunities to improve online self-help support groups.

**Key words:** Text Classification, Online Support Groups, Chatbot, Smoking Cessation

Tootiya Giyahchi
University of California, Irvine, CA, USA e-mail: tgiyahch@uci.edu

Sameer Singh
University of California, Irvine, CA, USA e-mail: sameer@uci.edu

Ian Harris
University of California, Irvine, CA, USA e-mail: harris@ics.uci.edu

Cornelia Pechmann
University of California, Irvine, CA, USA e-mail: cpechman@uci.edu

## 1 Introduction

Social media-based health interventions are increasing in the medical field to provide low-cost and accessible support at scale. However, a significant concern about such support groups is low engagement, i.e., infrequent interactions [1]. Many studies have shown a correlation between engagement and better outcomes in online health support groups [2, 3, 4]. Research [5] also suggests if participants in an online group do not receive responses to their posts promptly enough, they are likely to drop out. Using novel methods or tools to increase interactivity in such groups can positively affect user engagement and outcomes [6, 3, 4].

An accurate intent detection model facilitates solutions to increase engagement and interactivity in the group while requiring minimal human effort. For example, an effective intent detection model enables a chatbot to actively respond to posts with relevant content to address participants' concerns and provide evidence-based health information and support. Previous intent detection models in the health domain have either focused on 1:1 conversations or a limited set of intents and have generally concentrated only on the original thread of posts. However, text classification for online support groups requires detection of the main topic discussed by the group at each point in time. Furthermore, the discussion topics and desired categories for detection are particular to the group's purpose; hence, there is often a lack of data for the application-specific task. Further, the available datasets are often extremely imbalanced, with many important labels being rare. Additionally, online group chats are extra noisy due to parallel conversations and talk-turn interruptions. Lastly, there are application-specific relationships between the labels: labels may be symmetrically or asymmetrically related, and instances of them may be ambiguous.

This paper introduces customized training of pretrained language models for automated prediction of user intents in online support groups. We create an annotated dataset of 45 online support groups for quitting smoking from the Tweet2Quit study [7] as our empirical task. The dataset consists of 82000 posts labeled with 24 expert-identified intent codes. We experiment with fine-tuning a BERT language model to address the noisy dataset's problems and reduce the need for structured and high-incidence labeled data for the domain-specific task. To improve the accuracy of detecting important infrequent intents, we balance class weights for training. Finally, we take the unique relationships between the labels into account and propose an adaptive modeling approach by designing customized loss functions and adjusting the evaluation metrics. Overall, we show that these techniques provide considerable improvements in recognition of specific intents in an online health support group with an accuracy of 95.5%.

## 2 Background and Related Work

**Chatbots in Health** Using chatbots in health domains is increasing because they can provide online assistance through interactive conversations at users' convenience and

for a very low cost. Many chatbots have been proposed for affordable and scalable promotion of well-being or helping users cope with mental or physical problems [8, 9]. Understanding the user's intent is essential for such chatbots to reply with helpful content. [10] Studies suggest that using a chatbot that can accurately predict users' mood and respond appropriately improves engagement. Researchers [11, 12] also report that chatbots with greater accuracy are more effective at comforting users and improving their mood.

However, all the mentioned chatbots interact 1:1 with users. They also give users several fixed options to choose from, and reply based on users' preformatted responses. It is rare for them to detect a user's mood from sensor data [10]. For example, Woebot [9] is a self-help chatbot for mental health that relies mainly on pre-written questions and answers. It only uses a natural language algorithm to interpret freely written text in limited contexts: to detect self-to-harm and crisis language. It then asks the user for confirmation and, if the user confirms, it offers resources.

For chatbots to expand into online self-help support groups, they must be able to detect intents based on freely written text. They cannot rely on pre-set questions and answers to identify intents since this would be unnatural and disrupt the ongoing and dynamic group discussions that are the hallmark of online communities. Moreover, most of the important intents are specific to the health topic of the group and infrequent in other corpora, and so using only pretrained models is not an option [13].

**Classification of Medical/Mental Health Conversations** Intent detection of posts or conversations in medical and mental health contexts has been a topic of emerging interest in recent years. One important application has been to identify intents in health forums. Zhang et al. [14] try to understand 3 user intents in an online health forum to help users find useful information within the unstructured datasets. Using a support vector machine (SVM) classifier, their best model achieves 52.47 F1. McRoy et al. [15] study online health forums for breast cancer patients and survivors to detect posts expressing information needs that could be used to improve forum resources and materials. They develop Naïve Bayes, SVM, and Random Forest classifiers to detect expressed information need in 8 categories and their best model obtains 63 F1. Huh et al. [16] use a Naïve Bayes classifier in an online health community (WebMD) to detect posts that require an expert moderator's intervention. Their proposed model detects 4 intents for classification with the best model performing at 54 F1.

Another emerging health application is to detect intents in face-to-face medical sessions between patient and provider. Park et al. [17] experiment with different classifiers to detect 27 patient-provider conversation topics in primary care office talk-turns to more efficiently understand the patient's most significant complaint among all those expressed. Their best model attains 61% accuracy. Xiao et al.[18] classify patients' utterances in psychotherapy sessions based on domain-specific behavioral codes (8 therapist codes and 3 client codes) using a Recurrent Neural Network (RNN) model to provide guidance for the therapists, and they achieve 75% accuracy in therapist code prediction.

Intent detection in online support communities and groups involves unique challenges as the posts often consist of various group- or subgroup-level discussions

**Table 1** Tweet2Quit Intent Labels by Category

| Category | Intent labels |
|---|---|
| **medical regimen** | nrt_dontwork, nrt_dreams, nrt_howtouse, nrt_itworks, nrt_misuseissues, nrt_od, nrt_skinirritation, nrt_stickissue, quitdate, ecigs |
| **empathy for negative events** | fail, scared, stress, tiredness, cravings |
| **empathy for positive events** | smokefree, smokingless, support |
| **facts** | cigsmell, savingmoney, health, weightgain |
| **greetings** | greetings |
| **nonresponse** | nonresponse |

that transpire concurrently but may be on different topics, leading to exceptionally unstructured and noisy data. Intent codes may overlap even within a single post. Moreover, domain-specific data may be sparse, and the distribution of important intents may be especially imbalanced with some rare intents being very important to predict accurately such as self-harm or crisis events.

**Multi Party Dialogues** While virtually all prior work has focused on detecting intents in one-on-one conversations or question-answer discussion forums, our context is a group chat setup where approximately 20 peers talk to each other as part of an online self-help group. Intent classification in such groups is crucial to design a helpful chatbot but it involves numerous challenges. Intent detection must not be focused on one person but rather on the group or subgroup discussions. Also, in most cases, the chatbot should only reply to posts that align with the group's health or mental health goals, and not to inappropriate tangential posts that could be detrimental to group member satisfaction and outcomes. The chatbot cannot prompt with clarifying questions to understand the intents, as any irrelevant input could disturb the group's ongoing conversations and damage engagement. Moreover, the model must be able to accurately identify the most dominant and relevant intent being discussed in the group at any one time, so the chatbot contributes to the ongoing conversation meaningfully and appropriately, without being disruptive or derailing the dialogue which would be counterproductive.

Currently, there are very few chatbots designed to engage in an online group discussion. Savage et al. [19] and Kim et al. [20] both use a chatbot to improve collaboration among group members, but their goal is to ensure equal participation by members, which is just one of the many goals we have for improving engagement. Seering et al. [21] report improvement in an online community's engagement using a chatbot but they examine a gaming community, not a self-help group for health. Seering et al. [22] suggest chatbots have great potential to serve online groups and communities and can help to address challenges regarding maintaining and moderating group members and they urge more research in the area.

**Table 2** Description and examples of important post intents for Tweet2Quit online support groups

| Intent Code | Description | Examples |
|---|---|---|
| **nrt_howtouse** | Asks question or gives instructions about how to use NRT products | How often should I use the lozenges?<br>Chew it for a little bit and put between cheek and gum. |
| **nrt_misuseissues** | States NRT gum/lozenge has bad taste, irritates throat, causes sense of burning or spicy, makes nauseous or gag | The gum has a strong nicotine taste.<br>The gum burns my mouth.<br>It hurts my throat.<br>I want to throw up. |
| **nrt_od** | States NRT is too strong and causes overdose | The 21mg patches are making me sweat and feel bad.<br>I am doing better with the 14 MG. |
| **nrt_itworks** | States NRT works | The patches work if you're determined.<br>Lozenges are good.<br>I still use the gum for cravings. |
| **tiredness** | States feeling tired | I feel like I am about to fall asleep.<br>I am out of energy. |
| **smokefree** | States success in being smoke free | It has been 13 days since my last smoke.<br>I've had some severe cravings today but worked through them. |

## 3 Tweet2Quit Dataset

To explore the problem of modeling a post intent predictor in an online health support group, we use data from Tweet2Quit [3]. Tweet2Quit is a social media intervention for quitting smoking, where in addition to receiving free nicotine replacement therapy (NRT), smokers seeking to quit are assigned to a private 20-person Twitter-based support group to interact with peers and exchange information. A previous study on Tweet2Quit [4] has shown that engagement is a challenge because it is often low; but the more a participant posts the more likely they are to maintain abstinence from smoking ($p < 0.001$). Other main challenges with smoking cessation support groups are participants' hesitancy to use NRT and their other struggles with medical regimen compliance [23]. While sending daily auto-messages with relevant questions to encourage peer-to-peer discussions has been effective in Tweet2Quit [3], the abstinence rate is still below 40% [4].

We believe that an automated intent detection system can be used to design a chatbot that can create a better interactive environment and encourage medical regimen compliance by contributing to the group discussion based on the immediately preceding post intent to improve engagement and ultimately successful smoking abstinence. The next section explains data collection and annotation and discusses how we identified the intent labels.

### 3.1 Data Collection

We collected a dataset of posts from 45 groups in two prior Tweet2Quit studies. Eight groups came from the first study conducted from 2012-2014 [3] and 36 groups plus one pilot came from the second study conducted from 2016-2019. Each group ran for a three month period with 20 members, and the mean number of posts per group was 1822. Overall we collected more than 82000 posts by Tweet2Quit participants.

### 3.2 Identification of the Intents to Annotate

Tweet2Quit researchers identified intents that were desirable or useful posts, and arranged for these intents to be annotated, so that a chatbot will be able to accurately detect and respond to such intents. The ultimate goals are to increase the number of desirable posts, enhance engagement and improve abstinence. Intents are considered desirable for the online groups if they are relevant to quitting smoking or important to the proper functioning of the support groups, and they have meaningful frequency.

First, Tweet2Quit researchers identified a set of important intents related to clinical practice guidelines about the medical regimen for quitting smoking [24]. Medical regimen intents included use of Nicotine Replacement Therapy (NRT) both its efficacy and side effects, setting a quit date, and e-cigarettes. Studies on Tweet2Quit [3] affirm that posts about medical regimen compliance, e.g., setting a quit date or use of nicotine patches, relate significantly to smoking abstinence. Facts about quitting smoking were also included as intents including health benefits, money saved, weight gain and second-hand cigarette smell, because clinical practice guidelines recommend that doctors discuss these topics with smokers.

In addition, intents were identified that relate to online community building, emotional bonding and self-disclosure among the group members based on the literature on online community functioning [6]. Community-building intents included greetings to group members, and empathy for both positive and negative events related to members' quit-smoking attempts (e.g., successes and failures) which were self-disclosed. Tweet2Quit research has shown that bonding through self-disclosure increases the strength of social ties within the support group which significantly enhances smoking abstinence [25]. The final intent category included all posts that were tangential to the support group goal of quitting smoking, and were labeled nonresponse because the chatbot should not respond to such posts. Our recent analysis shows that, as anticipated, such posts are statistically unrelated to smoking abstinence.

Overall we identified 24 intent codes that were annotated as shown in Table 1 with their corresponding categories. Twenty-three of the intents are designed to be triggering intents that will cause the chatbot to contribute to the group chat when recognized, while one ("nonresponse") represents all other intents that are designed to be non-triggering, meaning the chatbot will not respond. Table 1 provides descriptions and examples of some of the more important intent labels.

### 3.3 Reliability

The overall reliability of the annotation process was determined based on Cohen's Kappa measure as 0.93, which is considered high agreement between the annotators [26]. The reliability of each label was calculated considering the number of times two research assistants agreed on the intent compared to the total number of posts with that intent based on the final annotation. Per label reliability scores ranged from 87.3% for "nrt_itworks" to 97.2% for "nonresponse".

### 3.4 Annotation Process

The annotation of intents took place from October 2019 through February 2021 with 25 research assistants working on over 82000 posts. The majority of the research assistants were undergraduates with two being high school seniors. The students' majors included Public Health, Nursing, Biology, Biomedical Engineering, Urban Studies, Cognitive Science, Education, Sociology, Business Administration, Economics and Informatics and the students were from the United States primarily but also Brazil, Germany and China. All posts were annotated.

During the training of the research assistants, the project manager discussed each intent in detail and provided formal definitions, key words and examples. After this initial training, the research assistants were given a practice set of posts and were required to achieve an 80% or higher accuracy in terms of selecting the correct intent. For continued learning and training, a database of posts for each intent was maintained, shared and referred to as needed.

The training meetings were conducted in person in the beginning, then transitioned to Microsoft Teams once the team expanded to included international researchers and due to COVID-19. Meetings were held weekly until formal training was completed and then became biweekly. At later meetings, posts that were difficult to annotate were discussed by the team to determine the most suitable codes to use and why. During the post-training or annotation phase, each post was reviewed by two trained research assistants working independently to determine what intent fit the post best, including whether the post fell into the "nonresponse" (irrelevant) category. If the two reviewers disagreed, a third more highly expert research assistant was brought in to review the post. This third reviewer worked independently as well, i.e., without seeing the intents assigned by the first two. Whenever two research assistants agreed on the intent, that was the final intent that was annotated; otherwise the third research assistant who had the greatest domain-specific expertise determined the final annotated intent.

The annotation process resulted in an extremely unbalanced dataset with more than 56% of the posts labeled as "nonresponse", i.e., support group irrelevant. Figure 1 shows the distribution of intents in the training and validation dataset aggregated by their categories.

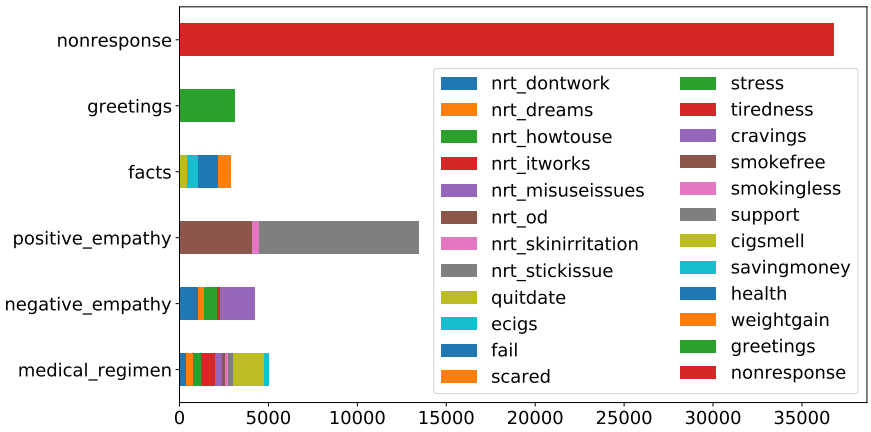


**Fig. 1** The distribution of intent labels in the combined training and validation dataset grouped by domain. 56.2% fell into the "nonresponse" (support group irrelevant) class.

## 4 Models

We use the annotated dataset to solve a supervised text classification problem to develop an intent detection model for our online support groups. We explore and compare two different NLP model strategies: 1- Bag-of-Words: Using Tf-Idf (Term Frequency Inverse Document Frequency) vectorization and a Random Forest model. 2- Language Models: Transfer Learning of a BERT model.

### 4.1 Random Forest (baseline)

Random Forest (RF) [27] is one of the best performing classifiers for text classification and is specifically suitable for noisy and high-dimensional data such as social media posts [28, 29]. RF creates a set of decision trees that each decides for a random collection of features. We evaluate the performance of an RF model for our text classification task as our baseline.

### 4.2 Pretrained Language Models

In recent years, pretrained large neural language models [30, 31, 32] have demonstrated state-of-the-art performance for all types of downstream tasks. Being trained on a massive unlabeled corpus of text, these models offer powerful universal language representations that can significantly improve the results with proper fine-tuning on a target task. A major advantage of transfer learning with pretrained large neural

language models is their ability to tackle dataset sparsity; more than 56% of our labeled dataset consists of "nonresponse" posts unimportant for the intervention.

Among pretrained language models, BERT (Bidirectional Encoder Representations from Transformers) [30] has obtained some of the bests results on popular NLP benchmarks such as GLUE [33]. BERT is based on a deep bidirectional transformer and is trained for masked word prediction (MLM) and next sentence prediction. Thus BERT generates token-level and sentence-level representations for both left and right contexts. We investigate the effect of fine-tuning a BERT [30] on task performance and study if it can address issues regarding noise and sparsity in the dataset.

## 5 Adapting to the Labels' Relationships

Although we first experiment assuming our task is a standard multi-class classification, intents are usually not completely independent in an online support group. Also, different mispredictions have different levels of impact on the group, depending on actual and predicted classes. For example, suppose the model mispredicts a "fail" labeled post (stating a failure to quit smoking) as "smokefree" (meaning success in quitting), and the chatbot responds wrongly with praise. Not only would that be an irrelevant and disruptive message, but it could also have a counter-productive effect on the group by conveying that smoking is ok. But if the model mispredicts a "nrt_skinirritation" post (comment that the NRT patch causes skin irritation) as "nrt_howtouse", the chatbot's response would not be perfect but still would be helpful and relevant to the topic to some degree. These relationships may be asymmetric, i.e., misprediction of intent X as intent Y may be tolerable, but misprediction of intent Y as intent X may be unacceptable and harmful to the support group.

To understand the special relationships between the intents in our dataset, we asked domain-expert Tweet2Quit researchers to answer if every possible misprediction would be acceptable or not. Figure 2 shows the final tolerable mispredictions and demonstrates the asymmetries in the intents' relationships. For instance, when users post seeking empathy for a positive or negative experience, a general empathetic supportive response should be acceptable if the model cannot recall the specific label with high confidence. In contrast, when users seek help with using NRT, the bot should respond with suitable NRT use guidance.

### 5.1 Customized Loss Functions

To represent the unique relationships between the labels in our model and adjust the training process based on that, we define multiple customized Negative Log Likelihood (NLL) loss functions and use them for fine-tuning the pretrained model. A standard NLL loss for a neural network is formulated as:

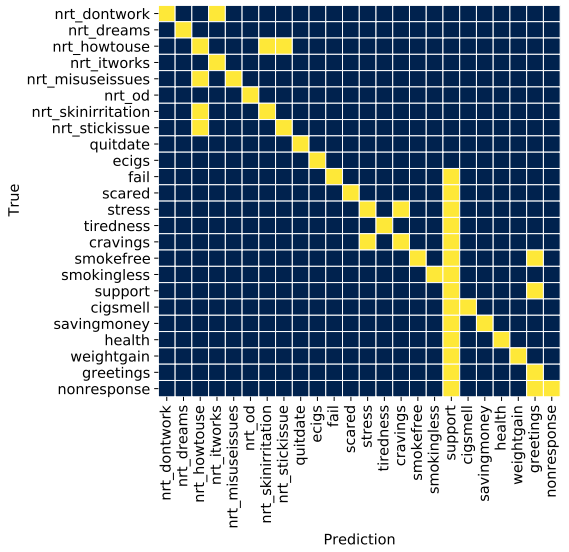


**Fig. 2** Acceptable mispredictions: For each true label (y-axis), its acceptable predictions are marked (x-axis). The asymmetry in the labels' relationships implies that the classes should not be combined for training or evaluation.

$$\begin{aligned} l_n &= -w_{y_n} x_{n,y_n} \\ \ell(x,y) &= \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N} w_{y_n}} l_n \end{aligned} \tag{1}$$

Where $N$ is the batch size, $x$ is the predicted vector, $y$ is the target vector, and $w$ is the class weight.

To adjust our loss function based on the label's relationship, for each label $y$, we consider a vector $z$ of size $C$ that represents all the acceptable labels for $y$. Then we use $z$ to mask out all $y$-acceptable indices of $x$ and sum over the rest to calculate the loss. We experiment with different versions of adjusted loss functions, as shown here:

- Customized Non-balanced Loss:

$$\begin{aligned} v_n = z_{y_n} \circ x_{n,y_n} \qquad & l_n = -\sum_{i=1}^{C} v_{n_i} \\ \ell(x,y) &= \sum_{n=1}^{N} \frac{l_n}{N} \end{aligned} \tag{2}$$

  While this loss function focuses on non-acceptable mispredictions to penalize the model during training, it doesn't consider the class weights. We introduce the following loss that uses the original true label's class weight for balancing the loss function.

- Customized Weighted Loss:

$$\begin{aligned} v_n = z_{y_n} \circ x_{n,y_n} \qquad & l_n = -\sum_{i=1}^{C} (v_{n_i}) w_{y_n} \\ \ell(x,y) &= \frac{\sum_{n=1}^{N} l_n}{\sum_{n=1}^{N} w_{y_n}} \end{aligned} \tag{3}$$

  Since we mask out the tolerable predictions to calculate the loss, using only the true label's class weight for balancing may not be the best option. For the following

loss function, we introduce $u_n$ that aggregates (uses the mean of) the weights of all acceptable classes for prediction to compute a balanced loss.

- Customized Balanced Loss:

$$\begin{gathered} W = \{w_y, \ldots, w_{yc}\} \qquad t_n = z_{y_n} \circ W \\ u_n = \frac{\sum_{i=1}^{C} t_{n_i}}{|t_n|} \qquad l_n = -\sum_{i=1}^{C} (v_{n_i}) u_n \\ \ell(x, y) = \frac{\sum_{n=1}^{N} l_n}{\sum_{n=1} u_n} \end{gathered} \tag{4}$$

## 6 Experiments

In this section, we present experiments to evaluate the proposed methods in sections 4 and 5. We compare the results of different models and discuss our observations.

### 6.1 Experimental Setup

From the 45 labeled groups, the seven chronologically latest groups are set aside for testing, and the remaining 38 groups' posts are randomly split into training and validation (development) sets using a 75%-25% ratio, respectively. We use stratified sampling to make sure the split is inclusive for the imbalanced dataset. We perform training using the training dataset, find the best version of the model using the validation set, and then report the model's performance using the test dataset.

**Random Forest (baseline)** To train an RF model, we first use common preprocessing techniques to clean the text data before extracting features for vectorization; we eliminate uninformative noisy data from the text such as mentions and links, convert the contracted form of verbs, and decode emojis. We also use Tf-Idf to vectorize text data for 1 to 3-grams, and remove English stop words tokens along with setting a threshold to identify and remove corpus-specific stop words. Table 3 contains the performance results after training an RF classifier on our training dataset.

As a result of our imbalanced dataset, aggregated recall and F1-score are very low for our baseline model, and the model does not detect many important labels properly. High precision scores and poor recall and F1-score scores for the infrequent intents indicate the model's poor recognition of the labels. The RF model does not recognize labels like "nrt_don'twork", "nrt_howtouse" and "smokingless".

### 6.2 Pretrained Language Models

To fine-tune a BERT model on our training dataset, we use the pretrained parameters as our starting point and fine-tune all parameters while appending a dense layer and

**Table 3** BERT vs RF performance with the original metrics.

| Intent Label | Precision | | Recall | | F1 | | Support |
|---|---|---|---|---|---|---|---|
| | RF | BERT | RF | BERT | RF | BERT | |
| nrt_dontwork | 0.0 | 56.4 | 0.0 | 43.7 | 0.0 | 49.2 | 71 |
| nrt_dreams | 82.6 | 57.8 | 35.8 | 69.8 | 50.0 | 63.2 | 53 |
| nrt_howtouse | 0.0 | 46.3 | 0.0 | 42.3 | 0.0 | 44.2 | 104 |
| nrt_itworks | 50.0 | 56.0 | 1.5 | 59.8 | 2.9 | 57.9 | 132 |
| nrt_misuseissues | 50.0 | 53.1 | 2.4 | 63.4 | 4.7 | 57.8 | 41 |
| nrt_od | 0.0 | 30.4 | 0.0 | 70.0 | 0.0 | 42.4 | 10 |
| nrt_skinirritation | 100.0 | 63.9 | 2.9 | 65.7 | 5.6 | 64.8 | 35 |
| nrt_stickissue | 100.0 | 84.2 | 2.4 | 75.3 | 4.6 | 79.5 | 85 |
| quitdate | 76.1 | 80.9 | 54.7 | 75.3 | 63.6 | 78.0 | 320 |
| ecigs | 100.0 | 78.2 | 17.8 | 93.2 | 30.2 | 85.0 | 73 |
| fail | 83.3 | 69.1 | 3.3 | 61.4 | 6.3 | 65.1 | 153 |
| scared | 100.0 | 77.6 | 7.2 | 62.7 | 13.5 | 69.3 | 83 |
| stress | 87.5 | 85.1 | 23.6 | 77.0 | 37.2 | 80.9 | 148 |
| tiredness | 71.4 | 56.3 | 11.6 | 62.8 | 20.0 | 59.3 | 43 |
| cravings | 55.1 | 53.0 | 16.0 | 65.8 | 24.8 | 58.7 | 269 |
| smokefree | 66.4 | 74.2 | 35.4 | 77.2 | 46.2 | 75.7 | 821 |
| smokingless | 0.0 | 56.5 | 0.0 | 78.8 | 0.0 | 65.8 | 33 |
| support | 85.6 | 78.1 | 53.9 | 78.3 | 66.1 | 78.2 | 2167 |
| cigsmell | 65.8 | 70.2 | 32.9 | 86.8 | 43.9 | 77.6 | 76 |
| savingmoney | 64.7 | 55.0 | 24.4 | 78.9 | 35.5 | 64.8 | 90 |
| health | 74.1 | 54.5 | 15.0 | 77.2 | 24.9 | 63.9 | 267 |
| weightgain | 70.0 | 56.7 | 29.5 | 74.1 | 41.5 | 64.2 | 166 |
| greetings | 82.3 | 85.2 | 69.9 | 78.7 | 75.6 | 81.8 | 634 |
| nonresponse | 76.4 | 91.2 | 97.6 | 89.1 | 85.7 | 90.1 | 10518 |
| **macro avg** | 64.2 | 65.4 | 22.4 | 71.1 | 28.5 | 67.4 | 16392 |
| **weighted avg** | 76 | 84.7 | 77 | 84 | 72.7 | 84.3 | 16392 |

**Table 4** Evaluating loss functions using adjusted metrics for BERT. Best performance in **bold**.

| Loss Function | Macro Average | | | Weighted Average | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| RF (baseline) | 66.3 | 23.8 | 30.5 | 83.3 | 78.5 | 80.8 |
| NLL | 68.1 | 73.5 | 70.0 | **87.7** | 87.5 | **87.6** |
| Nonbalanced | **80.4** | 78.9 | **77.6** | 70.0 | **95.5** | 80.8 |
| Weighted | 71.6 | 82.1 | 73.6 | 59.6 | 94.5 | 73.1 |
| Balanced | 71.8 | **88.8** | 77.1 | 58.9 | 94.3 | 72.5 |

a softmax layer specific to our task. As our dataset is imbalanced, we calculate a compensating balanced set of class weights to improve the model's prediction for the rare labels. We fine-tune the model for 15 epochs and evaluate the model after each training epoch on the validation dataset to pick the best performing model. In this stage, we use validation accuracy (weighted average recall) to choose the best epoch. Table 3 compares the results for training the RF and the BERT. As we expect, using pretrained language models causes significant improvement in every

aggregated metric compared to the RF model. In addition, performance scores of the infrequent labels dramatically improve compared to RF. RF scores are better solely for "nonresponse" recall and for precision with some less frequent labels, but overall show low recall and F1. These results indicate the ability of pretrained BERT to address dataset sparsity.

### 6.3 Loss Functions and Adjusted Metrics

The conventional performance metrics for multi-class classification assume all classes are completely independent and treat all mispredictions the same. However, as we explained earlier in Section 5, this is not the case in our problem. Here we describe how we adjust our metrics and evaluate customized loss functions which consider the special relationships between classes in our application.

To adjust the evaluation metrics we move the acceptable mispredictions from the false negative and false positive counts to the corresponding true positive counts. True negative counts of a label continue to indicate that the misprediction is unacceptable. To examine the proposed customized loss functions, we re-evaluate the performance of our RF and BERT (original loss) models with the new adjusted metrics for comparison. In addition, besides adjusted accuracy, we determine the macro average F1-score as the conclusive metric to pick the best model during the validation phases. We pick the macro average since the majority class (nonresponse) is least important in our problem, and we do not want it to have more weight in our calculation.

For customized training of BERT with the non-balance (equation (2)), weighted (equation (3)), and balanced (equation (4)) loss functions, we follow the same configuration and process explained in section 6.2. We train the model for 15 epochs for each of the proposed customized losses, and pick the best model based on the macro-F1 in the validation phase.

Table 4 summarizes the results for fine-tuning BERT with the customized loss functions. For all the macro averaged metrics, every customized training performs better than the original loss function (+ 4-15%) demonstrating the effectiveness of the proposed loss functions. Furthermore, using the suggested loss functions, the accuracy increases to 94.3-95.5% (from 87.5% using NLL loss). Although the original loss displays better weighted precision and F1-scores, given our highly imbalanced dataset, its significantly lower macro recall demonstrates how customize training improves the model's recognition of the infrequent labels towards the desired performance. Overall the non-balanced loss method scores the best, excelling on macro-precision, macro-F1 and accuracy. Weighting or balancing on weight class may not work as well in our application since we accept predictions from classes with different weights, e.g., a minority class may be acceptably predicted as an instance of a majority class.

## 7 Conclusion and Discussion

This paper seeks to address the low engagement and interactivity issues in online health support groups via a post intent detection model. Subtle detection of post intents is the first step to effectively intervening with engaging strategies such as a Chatbot to improve the quality of discussions and the health outcomes of the groups.

While previous work focuses on 1:1 conversations or question-answer forums, our context is a live group chat where approximately 20 peers post. Accurate detection of intents is more critical in a group chat setup where irrelevant interruptions are likely to disrupt group conversations and functioning and discourage participation. However, intent detection in such an environment is exceptionally challenging due to extremely noisy turn-taking and sparsity of certain important intents in the dataset. Furthermore, intent labels are often overlapping with asymmetric relationships.

We present an expert-annotated dataset with a fine-grained set of 24 intents from support groups for quitting smoking and use it to explore the problem of intent detection. We propose fine-tuning a pretrained language model (BERT) with a customized loss function representing the relationship between the labels as a promising solution that obtains 95.5% accuracy in our application. To our knowledge, no prior intent detection model in online health communities has performed this well.

Although our experiments are limited to samples from online support groups for smoking cessation, given the large and fine-grained set of intents that we recognize compared to other related works, our method may well have a bearing on different online support groups. As this paper aims to identify the most dominant intent of each post to respond appropriately, future research could usefully explore multi-label intent detection in online support groups where a single post contains multiple unrelated intents. A further study could investigate ways to utilize contextual information from the whole group discussions to improve intent recognition and to distinguish between labels involving similar words. Recognition of intents expressed as jokes, memes, or particular references (e.g., inter-group incidents, movies, books, etc.) is another interesting topic for future work. Finally, more research is needed to test and evaluate the effectiveness of our work for increasing engagement in online support groups.

## References


1. Jaime Arguello, B. Butler, Elisabeth Joyce, R. Kraut, Kimberly S. Ling, C. Rosé, and Xiaoqing Wang. Talk to me: foundations for successful individual-group interactions in online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2006.
2. Amanda Richardson, Amanda L Graham, Nathan Cobb, Haijun Xiao, Aaron Mushro, David Abrams, and Donna Vallone. Engagement promotes abstinence in a web-based cessation intervention: Cohort study. *J Med Internet Res*, 15(1):e14, Jan 2013.
3. Cornelia Pechmann, Li Pan, Kevin Delucchi, Cynthia M Lakon, and Judith J Prochaska. Development of a twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages. *J Med Internet Res*, 17(2):e50, Feb 2015.
4. Cornelia Pechmann, Kevin Delucchi, Cynthia M Lakon, and Judith J Prochaska. Randomised controlled trial evaluation of tweet2quit: a social network quit-smoking intervention. *Tobacco Control*, 26(2):188–194, 2017.
5. Elisabeth Joyce and Robert E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.
6. Anatoliy Gruzd and Caroline Haythornthwaite. Enabling community through social media. *J Med Internet Res*, 15(10):e248, Oct 2013.
7. C. Pechmann, D. Calder, Connor Phillips, K. Delucchi, and J. Prochaska. The use of web-based support groups versus usual quit-smoking care for men and women aged 21-59 years: Protocol for a randomized controlled trial. *JMIR Research Protocols*, 9, 2020.
8. Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet Interventions*, 10:39–46, 2017.
9. Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study. *J Med Internet Res*, 23(3):e24850, Mar 2021.
10. Asma Ghandeharioun, Daniel McDuff, M. Czerwinski, and Kael Rowan. Emma: An emotion-aware wellbeing chatbot. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019.
11. Mauro de Gennaro, Eva G Krumhuber, and Gale M. Lucas. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, 10, 2019.
12. T. Bickmore and Daniel Schulman. Practical approaches to comforting users with relational agents. In *CHI Extended Abstracts*, 2007.
13. Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 10 2018.
14. Thomas Zhang, Jason H. D. Cho, and ChengXiang Zhai. Understanding user intents in online health forums. *IEEE Journal of Biomedical and Health Informatics*, 19:1392–1398, 2015.
15. Susan McRoy, Majid Rastegar-Mojarad, Yanshan Wang, Kathryn J Ruddy, Tufia C Haddad, and Hongfang Liu. Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval. *JMIR Cancer*, 4(1):e10, May 2018.
16. Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, 46(6):998–1005, 2013. Special Section: Social Media Environments.
17. Jihyun Park, D. Kotzias, Patty B Kuo, IV RobertLLogan, Kritzia Merced, Sameer Singh, Michael J Tanana, Efi Karra Taniskidou, J. Lafata, David C. Atkins, M. Tai-Seale, Zac E. Imel, and Padhraic Smyth. Detecting conversation topics in primary care office visits from transcripts

of patient-provider interactions. *Journal of the American Medical Informatics Association : JAMIA*, 26:1493 – 1504, 2019.
18. Bo Xiao, Dogan Can, James Gibson, Zac E. Imel, David C. Atkins, P. Georgiou, and Shrikanth S. Narayanan. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *INTERSPEECH*, 2016.
19. Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 813–822, New York, NY, USA, 2016. Association for Computing Machinery.
20. Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. *Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.
21. Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. *It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.
22. Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. *Beyond Dyadic Interactions: Considering Chatbots as Community Members*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2019.
23. Amy N. Kerr, Barbara A. Schillo, Paula A. Keller, Randi B. Lachter, Rebecca K. Lien, and Heather G. Zook. Impact and effectiveness of a stand-alone nrt starter kit in a statewide tobacco cessation program. *American Journal of Health Promotion*, 33(2):183–190, 2019. PMID: 29747516.
24. Jane E. Anderson, Douglas E. Jorenby, Walter J. Scott, and Michael C. Fiore. Treating tobacco use and dependence: an evidence-based clinical practice guideline for tobacco cessation. *Chest*, 121(3):932–941, March 2002.
25. Cornelia Pechmann, Kelly Yoon, Denis Trapido, and Judith Prochaska. Perceived costs versus actual benefits of demographic self-disclosure in online support groups. *Journal of Consumer Psychology*, forthcoming, 10 2020.
26. Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, October 2012.
27. Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.
28. Thiago Salles, Marcos André Gonçalves, Victor Rodrigues, and L. Rocha. Improving random forests by neighborhood projection for effective text classification. *Inf. Syst.*, 77:1–21, 2018.
29. Md Zahidul Islam, Jixue Liu, Jiuyong Li, Lin Liu, and Wei Kang. A semantics aware random forest for text classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1061–1070, New York, NY, USA, 2019. Association for Computing Machinery.
30. J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
31. Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. A robustly optimized bert pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Liu Kang, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Chinese Computational Linguistics*, pages 471–484, Cham, 2021. Springer International Publishing.
32. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
33. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 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*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.