Routledge
Taylor & Francis Group

Journal of Technology in Human Services

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/wths20

Detecting Intensity of Anxiety in Language of Student Veterans with Social Anxiety Using Text Analysis

Morgan Byers, Mark Trahan, Erica Nason, Chinyere Eigege, Nicole Moore, Micki Washburn & Vangelis Metsis

To cite this article: Morgan Byers, Mark Trahan, Erica Nason, Chinyere Eigege, Nicole Moore, Micki Washburn & Vangelis Metsis (2023) Detecting Intensity of Anxiety in Language of Student Veterans with Social Anxiety Using Text Analysis, Journal of Technology in Human Services, 41:2, 125-147, DOI: 10.1080/15228835.2022.2163452

To link to this article: https://doi.org/10.1080/15228835.2022.2163452

Published online: 17 May 2023.

JOURNAL OF TECHNOLOGY IN HUMAN SERVICES
2023, VOL. 41, NO. 2, 125–147
https://doi.org/10.1080/15228835.2022.2163452

Taylor & Francis
Taylor & Francis Group

Detecting Intensity of Anxiety in Language of Student Veterans with Social Anxiety Using Text Analysis

Morgan Byers[a], Mark Trahan[a], Erica Nason[a], Chinyere Eigege[b], Nicole Moore[b], Micki Washburn[c] and Vangelis Metsis[a]

[a]Texas State University, San Marcos, Texas; [b]University of Houston, Houston, Texas; [c]University of Texas Arlington, Arlington, Texas

ABSTRACT

Approximately one-third of the veteran population suffers from post-traumatic stress disorder, a mental illness that is often co-morbid with social anxiety disorder. Student veterans are especially vulnerable as they struggle to adapt to a new, less structured lifestyle with few peers who understand their difficulties. To support mental health experts in the treatment of social anxiety disorder, this study utilized machine learning to detect anxiety in text transcribed from interviews with patients and applied topic modeling to highlight common stress factors for student veterans. We approach our anxiety detection task by exploring both deep learning and traditional machine learning strategies such as transformers, transfer learning, and support vector classifiers. Our models provide a tool to support psychologists and social workers in treating social anxiety. The results detailed in this paper could also have broader impacts in fields such as pedagogy and public health.[1]

ARTICLE HISTORY
Received 4 June 2022
Accepted 24 December 2022

KEYWORDS
Machine learning; deep learning; datasets; social anxiety; text analysis; topic modeling

Introduction

Post-traumatic stress disorder, or PTSD, affects approximately one quarter of the veteran population and is often accompanied by social anxiety disorder (Fulton et al., 2015; McMillan, Sareen, & Asmundson, 2014). About 8–10% of returning combat veterans have PTSD, and of those, another 7–13% also have social anxiety disorder (Trahan et al., 2019). Student veterans are particularly vulnerable because of the difficulty involved with transitioning to life on a college campus. Often, student veterans struggle with the lack of structure in college classes and have difficulty relating with their younger, less experienced peers (Morris, Powers Albanesi, & Cassidy, 2019). Furthermore, student veterans with social anxiety may avoid social scenarios, such as crowded walkways or

CONTACT Vangelis Metsis vmetsis@txstate.edu Texas State University, San Marcos, TX 78666, USA.

classes, and eating facilities, such as restaurants, and may be more likely to isolate at home (Trahan et al., 2019). A radical lifestyle shift, from service to campus, in combination with limited access to resources, may negatively impact a student veteran's quality of life.

To improve the support network for student veterans, an interdisciplinary team of computer scientists, graphic designers, social workers, and psychologists developed a virtual reality environment to assist mental health experts in quantifying and treating social anxiety disorder (Nason, Trahan, Smith, Metsis, & Selber, 2020). Previous research from the team provided qualitative information about social anxiety experienced by patients in various settings and situations (Trahan et al., 2019). The purpose of this study was to evaluate the language of study participants in describing both the emotions associated with social anxiety in the context of stimulating environments and the intensity of those emotions as it relates to specific language. The information extracted using the current analysis is subsequently used to inform and modify the design and development of Virtual Reality (VR)-based therapeutic interventions (Metsis et al., 2019; Nason et al., 2020).

The team conducted ten qualitative interviews with student veterans about their experiences with social anxiety, transcribed them, and divided them into 1,187 responses. Three independent coders, including one psychologist, one licensed mental health professional with a Ph.D. in social work, and a doctoral student with a history of psychological training, listened to audio recordings of interviews and rated anxiety levels on a scale of zero to three based upon indications of stress from voice as delineated by qualifiers within the interviews (Ogunfunmi, Togneri, & Narasimha, 2015). Ultimately, coding resulted in four target classes, with class zero corresponding to no stress present in a response and class three corresponding to high stress present.

We employed both deep learning and traditional machine learning in order to provide quantitative data by ranking anxiety levels in text. In order to properly evaluate our models, we developed two baselines. The first baseline is a blind guess, which would result in about a 25% accuracy score since there are four classes. The second baseline accuracy is 46.2%, which is yielded by predicting only the most common class.

We experimented with several deep learning methods, including a bidirectional long short-term memory (LSTM) network with attention, transfer learning with pre-trained language models BERT and ELMo, and a Transformer. In addition, we used traditional machine learning models like decision trees, logistic regression, naive Bayes, and a support vector classifier (SVC) with a radial basis function (RBF) kernel to address issues present in our deep learning models as well as for feature selection. Notably, using traditional machine learning classifiers allowed us to

overcome limitations imposed on us by the small sample size of our data. Our support vector classifier achieved a ten-fold cross-validation accuracy of 59.1% when classifying among the four levels of anxiety.

In addition to classifying anxiety levels, we used feature selection and topic modeling to uncover more information about the language used to reveal social anxiety and the experiences of student veterans. Using several feature selection techniques, we were able to highlight words or phrases that correlate with anxiety. Further, we used Latent Dirichlet Allocation (LDA) to provide a list of topics common among each student veteran who was interviewed. Our results bring insight into sources of anxiety for veterans, such as visiting restaurants or feeling seen.

The main findings of this study can be summarized as follows:

- Natural language processing and text analysis methods can be valuable supporting tools complementing the judgment of human experts and revealing cues and common themes about social anxiety.
- Understanding and rating the levels of social anxiety experienced by patients is very subjective, as indicated by the medium level of agreement in ratings provided by experts.
- Deep learning models are not very effective with small dataset sizes like the one used in this study, even when pre-trained language models and transfer learning are used.
- Traditional machine learning pipelines can produce better classification accuracies in small datasets, although that accuracy is still negatively affected by the subjectivity of the ground truth labels and the imbalanced class distribution of the data.
- Feature selection and topic modeling seemed to be the most useful practical outcome of this analysis, as it revealed terms and thematic topics that can be interpreted by humans to inform their interventions. Since most of these methods are unsupervised, they are not affected as much by imprecise data labeling.

Related work

Labeling of emotions has been proposed within the context of developing systems of classification of the coding of facial expressions and proposed within the framework of six primary manifestations of emotion in facial coding, anger, fear, sadness, joy, and surprise (Ekman, 1999). Fear and anxiety have common features, although they have been distinguished within the literature into two distinctive states. Fear is the brief and present-moment appraisal of a perceived threat, while anxiety is nonspecific, future-oriented cognitive appraisal of threat (Sylvers, Lilienfeld, & LaPrairie,

2011). While they are distinguishable, these two states overlap, which allows the study of the manifestation of fear as anxiety in the reflections of veterans with social anxiety disorder.

While emotions may be identified within a complex set of physiological symptoms, there is no specific settled theory around labeling emotions solely from auditory clues (Sethu, Epps, & Ambikairajah, 2015). Classifying emotions may include low level features, such as pitch, loudness, or energy. High-level features may include more voice quality features, such as a shimmer, stuttering, or using a "bag of words" (BOW) feature with multiple words uttered simultaneously (Sethu et al., 2015). (Van Puyvelde, Neyt, McGlone, & Pattyn, 2018) used a variety of speech parameters and causal factors to analyze stress. However, despite the lack of a consensus on classification, humans can interpret paralinguistic information in speech (Sethu et al., 2015).

Detecting emotion from text has been a task of interest for a long time, and much research has been dedicated to training classifiers capable of detecting a range of affective states. With the advent of popular social network sites such as Twitter and Reddit, many researchers have turned to the internet to collect data (Fatima et al., 2019; Gruda & Hasan, 2019; Jere & Patil, 2020; Low et al., 2020; Rajabi, Shehu, & Uzuner, 2020; Shen & Rudzicz, 2017) and have achieved promising results. However, recognizing emotion from text alone is still quite challenging, so text data is sometimes accompanied by other supplemental data like audio, resulting in even greater levels of classification accuracy (Chuang & Wu, 2004; Kim, 2020).

Our project differs from most other attempts to classify emotion because we analyze the nuances in a single emotion, anxiety, rather than a range of different emotions, including anger, fear, enjoyment, sadness, disgust, and surprise (Ekman, 1999). While fear is emotion manifested as a stimulus-response to threat, anxiety is cognition inducing the perception of perceived future threat (Watson, Clark, Simms, & Kotov, 2022). Social anxiety is based upon the perceived threat that social interaction, specifically social engagement Although some work exists in detecting anxiety through text (Gruda & Hasan, 2019), the work is limited in scope, analyzing only tweets containing the words "work" and "feeling." In addition, our approach pairs anxiety recognition with qualitative data in order to provide additional support for psychologists and social workers interested in treating social anxiety. This capacity for human identification of paralinguistic information informs the method of coding for this research project and offers an opportunity to evaluate whether coders may reach inter-rater reliability around coding anxiety. Reaching a consensus on coding anxiety is one of the aims of this research, while using deep learning approaches may also provide

helpful information about language associated with anxiety in this specific population.

We explore several deep learning approaches to this task. Initially, we chose to work with a long short-term memory network with attention, since this architecture has been shown to perform well on text classification tasks (Feng, Wei, Pan, Qiu, & Ma, 2020). The sentences in our data set vary greatly in length, so in an effort to capture the more long-term dependencies in each instance, we applied a Transformer (Vaswani et al., 2017). A challenging aspect of our data set is the limited occurrences of words in the corpus. Often a word will appear only once, or a word will be used multiple times in different contexts. We utilized pre-trained ELMo and BERT word representations in order to model the lexical complexities that may be missed by other embedding schemes (Devlin, Chang, Lee, & Toutanova, 2018; Peters et al., 2018).

When working with traditional machine learning options, we first chose to explore support vector classifiers both because of their performance on text classification (Fatima et al., 2019) and their ability to perform well on small data sets (Chuang & Wu, 2004). In addition to the support vector classifier, we also implemented both Gaussian and Multinomial Naive Bayes (NB) classifiers as they typically serve as a baseline in text classification problems (Xu, 2018).

Text analysis problems are often quite highly dimensional, so choosing effective feature selection techniques is imperative for creating adequate models. We implement select feature reduction techniques discussed in (Kou et al., 2020) as well as other techniques not mentioned in that paper. Further, in deep learning, the word vector representation or embedding used can have a significant impact on classifier accuracy (Maas et al., 2011). Popular word embedding methods include word2vec (Maas et al., 2011), which we use in this paper, and GloVe (Pennington, Socher, & Manning, 2014).

Methods like structural topic modeling and network analysis are often used to analyze sentiment and can be used to understand areas of stress or anxiety for populations (Jo, Lee, Kim, & Park, 2020). We chose to use Latent Dirichlet Allocation (LDA), because of the powerful visualization tools associated with the model (Sievert & Shirley, 2014).

Methodology

Data collection

This project is a secondary data analysis from qualitative interviews conducted to design the virtual reality intervention of a grocery store with cues to stimulate social anxiety (Trahan et al., 2019). The design of the

intervention is outlined in previous studies (Metsis et al., 2019; Nason et al., 2020). In the original study, twelve student veterans were interviewed about their experiences with social anxiety; however, due to problems with audio, only ten were usable for secondary analysis of transcribed interview text. Two professional social work faculty members, both with experience in research and practice with veterans, interviewed student veterans (n = 12) about their experiences with social anxiety on a university campus (Trahan et al., 2019). The audio of these interviews (n = 10) was then analyzed using Dedoose Version 9.0.17.

Four coders, including a professional psychologist with a background in research on PTSD, a social work assistant professor with a background in anxiety and virtual reality research, and two doctoral students, coded the data using the audio. Every interview was coded by at least three coders. The coders listened to recordings of each interview and rated the transcribed responses for anxiety following a scale of zero (no anxiety present in the sentence) to three (high anxiety present in the sentence). Prior to coding, coders conferred to discuss common features of anxiety-related indicators, including increased pace, reduced volume, increased stammers, and BOW that would indicate greater levels of stress. Using both low and high features, coders listened to recorded audio interviews of student veterans discussing their experiences with social anxiety. After listening to the audio of each interview, the coders reviewed the text sentence by sentence and assigned low, medium, and high labels to the level of anxiety demonstrated in each sentence. Two sentences were not coded by our judges, bringing our final sample size down to 1,185 responses.

We measured the amount of agreement between our three coders by using the intraclass correlation coefficient (ICC), which is a measure of inter-rater reliability. In other words, the ICC quantifies how much our different judges agree with each other. We used *ICC*(3, 3) because each judge rated each instance, and we considered the mean of their ratings to be our target (Shrout & Fleiss, 1979). Our results are found in Table 1. According to the "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology" (Cicchetti, 1994), the following are indicative ranges of the rating correlation: less than 0.40–poor; between 0.40 and 0.59–fair; between 0.60 and 0.74–good; and, between 0.75 and 1.00–excellent. An ICC score of 0.678 suggests a reasonable amount of agreement between our three coders but

Table 1. ICC(3,3) statistic for our data set.

Type	ICC	F Statistic	p-value	95% CI
ICC(3,3)	0.678	3.103	0	[0.64, 0.71]

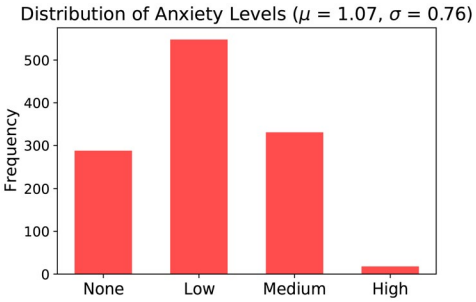


Figure 1. The distribution of average anxiety ratings across all responses.

not a perfect agreement, which highlights the subjectivity and difficulty of rating perceived human emotions even by the experts.

After each judge rated every target, their ratings were averaged and rounded to the nearest integer, which then became our final target class. The distribution of average ratings is right-skewed, with only 18 instances of high-stress responses out of 1,185 total rated instances. The distribution of our target classes can be seen in Figure 1.

Materials

We used Sci-kit learn for feature selection, topic modeling, and traditional machine learning, and a combination of TensorFlow, Keras, Simple Transformers (Rajapakse, 2020), and PyTorch to implement our deep learning pipelines. Since our target is ordinal, we experimented with the use of a custom loss function that reflects the nature of our target. The loss function, called ordinal categorical cross-entropy (Hart, 2017), adds weight to the cross-entropy function. Given an instance's target class y and the predicted class $\hat{y}$, the weight w used to modify the traditional categorical cross-entropy function is given by:

$$w = \frac{|y - \hat{y}|}{n-1}$$

Where n is the total number of classes. Then the loss, L, is computed by:

$$L = (w+1)H(y, \hat{y})$$

where categorical cross-entropy is given by the function H.

Methods

Because of its ability to process sequential data, we first chose to implement a bidirectional long short-term memory network with attention.

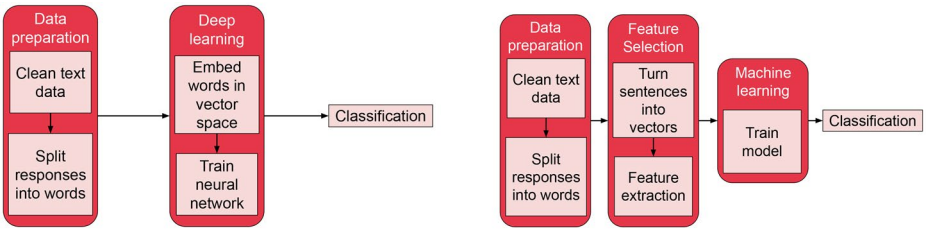
Attention is a popular mechanism in emotion recognition tasks (Feng et al., 2020) because of its ability to add focus to important words in a sentence. Further, the input lengths of our responses vary greatly; attention mechanisms help to mitigate the loss of information as it gets compressed along each time step (Bahdanau, Cho, & Bengio, 2014). We also chose to experiment with a transformer model because of its ability to generalize well to a broad range of natural language processing tasks (Vaswani et al., 2017). In addition, we experimented with transfer learning by using pre-trained BERT and ELMo models. Since our data set is small and the individual sentences lack context, using robust word representations can help to increase our predictive accuracy (Peters et al., 2018) and overcome the limitations of the small training set.

We began preparing our data by stripping each response of punctuation and converting all letters to lowercase. Then, we tokenized each response into unigrams and encoded the words as integers.

For the LSTM networks, we experimented with two different embeddings: word2vec and Keras' built-in embedding layer. Our decision to try word2vec in addition to the built-in embedding layer was motivated by word2vec's ability to create meaningful vector representations of words. The embeddings it creates reflect sentiment as well as semantics (Maas et al., 2011). Since our data set is small, we trained our own word2vec model and then used the learned embeddings as input for the rest of our deep learning model. Our deep learning pipeline is illustrated in Figure 2a. For our Transformer, we embedded our features using Keras' built-in embedding layer.

After we prepared our data, we trained both the embedding and deep learning model on our data set and tested using five-fold cross-validation.

Since our target is ordinal, we also approached our problem as a regression task. We used a long short-term memory network with a word2vec embedding layer. The model was again trained and tested using five-fold cross-validation. Since this was a regression task, we did not round the average anxiety rating of each response to an integer class, instead leaving



(a) Our deep learning pipeline. (b) Our traditional machine learning pipeline.

Figure 2. Supervised learning pipeline.

them as floating-point values. After the model was trained, we accumulated its predictions on the test set across each fold.

To provide an equal basis for comparison between our regression and classification models, we rounded both the actual and predicted values found by our regressor to integers. We used these rounded values to generate a classification report and confusion matrix. The actual target values were rounded following standard convention to yield four integer classes. We optimized the rounding threshold for our predicted values in order to achieve the highest classification accuracy possible. In our case, for a prediction $\hat{y}$ the corresponding rounded value $\hat{\hat{y}}$ is given by:

$$\hat{\hat{y}} = \begin{cases} 0, & \text{if } \hat{y} < 0.85 \\ 1, & \text{if } 0.85 \le \hat{y} < 1.23 \\ 2, & \text{if } 1.23 \le \hat{y} < 2.01 \\ 3, & \text{otherwise} \end{cases}$$

Preparing text for traditional machine learning also begins by cleaning the text of punctuation and converting all letters to lowercase. After cleaning, the data was tokenized with n-grams ranging from unigrams up to trigrams. An n-gram is a contiguous sequence of n tokens from a given sample of text or speech. The tokens can be phonemes, syllables, letters, or words depending on the application. In our case, whole words were considered as tokens. The reason for using n-grams is that previous text analysis studies have shown improved text classification accuracy compared to using each token separately (Peng & Schuurmans, 2003). After the text is cleaned and tokenized, each sentence must also be vectorized. We used the bag of words (BOW) vectorization method. Our traditional machine learning pipeline is illustrated in Figure 2b, and it is similar to the one described in (Sethu et al., 2015).

Due to the limitations of our small data set, we also experimented with an array of popular traditional machine learning algorithms, which generally require less training data compared to deep learning algorithms. We used the bag-of-words approach to convert each separate text sentence into a token count vector using the Scikit-learn 'CountVectorizer' with n-gram range 1 to 3. We prepared a pipeline to perform feature selection (detailed in the next paragraph) and then trained the classifiers on the extracted feature set. A cross-validated grid search algorithm was used to tune the particular hyperparameters of our pipeline, such as the number of features to select and the particular hyperparameters of each classification algorithm. After the best parameters for our pipeline were selected,

we used ten-fold cross-validation to confirm our model's accuracy. We collected the test set predictions across each fold and aggregated them to produce a confusion matrix and classification report. Our support vector classifier and Gaussian naive Bayes classifier were the best performers on this task. Other traditional machine learning algorithms we tested include logistic regression, naive Bayes multinomial, decision tree, and random forest classifiers. The detailed evaluation results can be found in Table 3 and in Figure 6. In order to properly evaluate each model's performance, we conducted five trials of ten-fold cross-validation in which each fold had a random distribution of classes. We then used the results of the repeated trials to construct confidence intervals for the test accuracy of each of our models.

We chose to implement several feature selection algorithms, both to improve the accuracy of our models and also to provide qualitative data to psychologists and social workers about the importance of different words to the prediction outcome. The algorithms we implemented are:

- χ^2
- Mutual information
- LASSO
- Recursive feature elimination
- L1 feature selection
- Tree-based feature selection

Preparing text data for feature selection follows the same steps as preparing data for traditional machine learning. Each feature selection algorithm used provides some metric of attribute relevance, e.g., a p-value for Chi-squared. Thus, for each feature selection algorithm, we saved each keyword and its corresponding relevance metric in order to provide as much information as possible to our social work colleagues.

In addition to feature extraction, topic modeling allows us to find latent or hidden topics present in each interview. The results provided by topic modeling differ from those of feature selection since topic modeling works by aggregating related terms together, while feature selection does not. However, while topic modeling draws attention to words associated with a particular subject, the topics presented are not labeled and require human interpretation. Since topics found with machine learning can sometimes be difficult to interpret, we used a visualization library that presents topics in a more intuitive way (Sievert & Shirley, 2014).

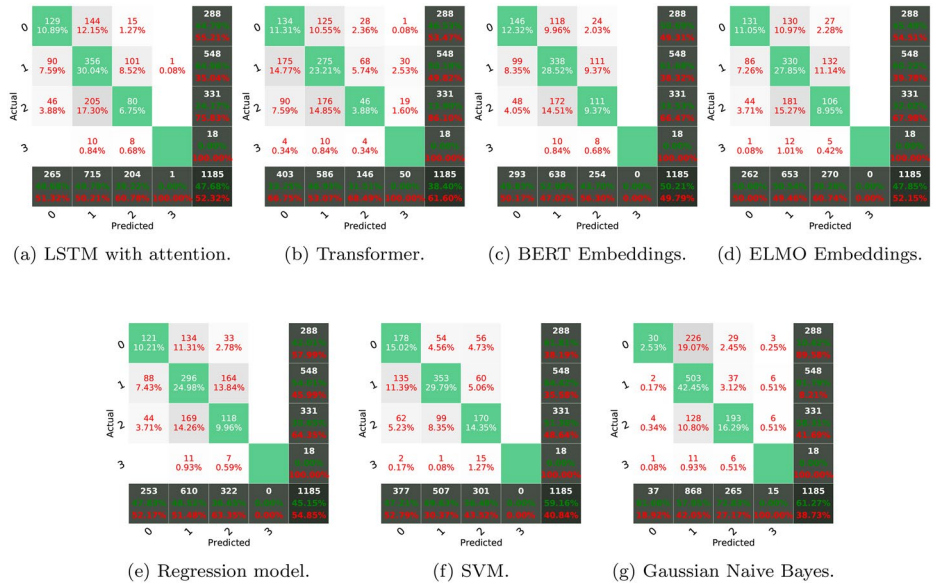
The topic modeling algorithm we used is Latent Dirichlet Allocation (LDA), which is a probabilistic model. LDA is a popular topic modeling technique to extract topics from a given corpus. LDA looks at a text input (generally a document) to determine a set of topics that are likely to have

generated that collection of words. The goal is to retrieve latent information that is not immediately obvious within a single document but is shared among multiple documents. The topics are learned as a probability distribution over the words that occur in each document. Each document, in turn, is described as a mixture of topics. In our case, each instance of text, as segmented by the coders, is considered a document. For more information about LDA, the reader is referred to (Blei, Ng, & Jordan, 2003).

We prepared our data by first cleaning it, then tokenizing the responses into n-grams as we did for traditional machine learning. Including only unigrams and bigrams provided more interpretable results, so trigrams were excluded from our final model. We vectorized the text using a BOW approach and then also removed stop words. Although we did not remove stop words for traditional machine learning or feature selection, we chose to do so for topic modeling as their removal provided results much richer in meaning.

Results

Our bidirectional LSTM, with attention and a word2vec embedding layer, achieved a cross-validated accuracy of 47.7% (±9.4%), which is above our baseline. However, this model struggled to overcome the imbalanced nature of our data set, even when the classes were weighted (Figure 3a). This is likely due to the limited number of training samples available to us. The addition of an attention mechanism provided a reasonable increase to our



(a) LSTM with attention. (b) Transformer. (c) BERT Embeddings. (d) ELMO Embeddings.

(e) Regression model. (f) SVM. (g) Gaussian Naive Bayes.

Figure 3. The Confusion matrices produced by different classification algorithms.

model's performance. An interesting observation is that the attention layer actually decreased our model's overall accuracy, but the f1-score for classes zero and two increased. The highest accuracy (48.7%) was observed when using a word2vec embedding layer without attention, although at the cost of a lower f1-score (Table 2).

Similarly, our LSTM regression model achieved an accuracy of 45% and a mean squared error of 0.445 (±0.127) when using a word2vec embedding layer. Like the classifier, our regression model was also impacted by our skewed data and mislabeled a majority of all samples except for those from class one (Figure 3e).

Our Transformer achieved a cross-validated accuracy of 38.4% (±8.9%), which is well below the average performance of our other deep learning models. Our Transformer was also one of the only deep learning models to predict the third class label (Figure 3b), which, when considered in conjunction with the low evaluation accuracy, indicates an overfitted model. This overfitting is likely a result of our small dataset. Prior research has shown that Transformers tend to overfit on small datasets (Ye, Guo, Gan, Qiu, & Zhang, 2019).

We found better results when working with transfer learning. Both of our models that used BERT and ELMo embeddings performed about the same in terms of accuracy, with our implementation of BERT earning a 50.2% (±8.0%) five-fold cross-validated accuracy, while ELMo earned a 47.8% (±6.9%) accuracy. A notable aspect of our results is that BERT outperformed every other deep learning model (Figure 6). This is likely due to the depth and size of the model (Devlin et al., 2018) compared to other embeddings like word2vec. As with most of our deep learning models, both of our transfer learning models struggled to overcome our imbalanced data set (Figure 3c and d).

The support vector classifier, on the other hand, achieved a cross-validated accuracy of 59.1% (± 7.2%). A notable aspect of the resulting confusion matrix (Figure 3f) is that although no model was able to identify instances of the third class correctly, the SVC labeled nearly all instances of class three as class two. Because of the ordinal nature of our data set, this misclassification is not as severe an error as misclassifying most instances of class three as class one, for instance. In order to ensure the accuracy of our SVC, we performed repeated trials of ten-fold cross-validation to yield a 95% confidence interval for the model's accuracy (Figure 5b).

Table 2. A comparison of the classification accuracy (%) of our different LSTM embedding schemes and LSTM with Attention.

Embedding layer	LSTM	LSTM + Attn.
Keras basic	42.0	44.1
word2vec	48.4	47.7

Our Gaussian naive Bayes model performed similarly to our SVC, achieving a cross-validated accuracy of 61.2% (±8.6%) when trained on the same feature set that was used to train the SVC. As in the case of our support vector classifier, we again performed multiple runs of ten-fold cross-validation in order to yield a 95% confidence interval for our model's accuracy (Figure 5a). Although this model achieved a high classification accuracy, it is important to note that our model only managed to correctly predict instances of class one and two over half the time. Classes zero and three had very low f1 scores (Figure 3).

Our logistic regression, multinomial naive Bayes, decision tree, and random forest classifiers all performed moderately well at this task. These models were again evaluated with ten runs of ten-fold cross-validation. The average 95% confidence intervals of each traditional machine learning model are detailed in Table 3. Unlike the SVC and naive Bayes models, these classifiers did not outperform deep learning approaches.

Feature selection greatly improved the accuracy of our models (Figure 4). LASSO performed the worst out of all the feature reduction methods we tried in terms of its effect on classifier accuracy. This is possibly due to the fact that if multiple features are correlated, LASSO will only include one of those features in the final feature set. Chi-squared ended up providing the greatest boost to our model's accuracy, which is consistent with other findings that show Chi-squared to be an effective feature selection technique for text analysis on small data sets (Kou et al., 2020).

Further, feature selection provided phrases correlated with anxiety levels. These phrases give much insight into areas of stress for student veterans. Mutual information (MI), Chi-squared, and recursive feature elimination (RFE) yielded the most intuitive results (Table 4).

We used a visualization library called pyLDAvis to display the results from our topic modeling. The visualization provides an easy way to interface with results (Sievert & Shirley, 2014) and evaluate the effectiveness of our model. Our final model yielded five topics, each of which is distinct (Figure 7). For choosing the optimal number of topics, we did a grid search on the number of components, and we chose the optimal number by looking at the perplexity value and by looking at the top 30 words of each topic. N = 5 seemed to be the value with low perplexity that also

Table 3. A comparison of the 95% confidence intervals for our traditional machine learning models.

Model	95% CI
SVC	(56.2, 62.8)
Logistic Regression	(49.9, 56.4)
Decision Tree	(40.1, 46.5)
Random Forest	(46.6, 52.8)
Gaussian Naive Bayes	(60.7, 66.7)
Multinomial Naive Bayes	(42.1, 49.8)

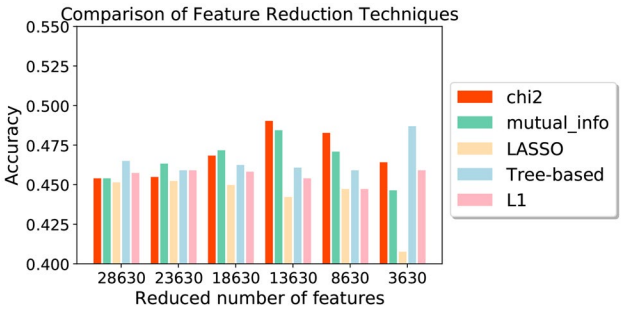


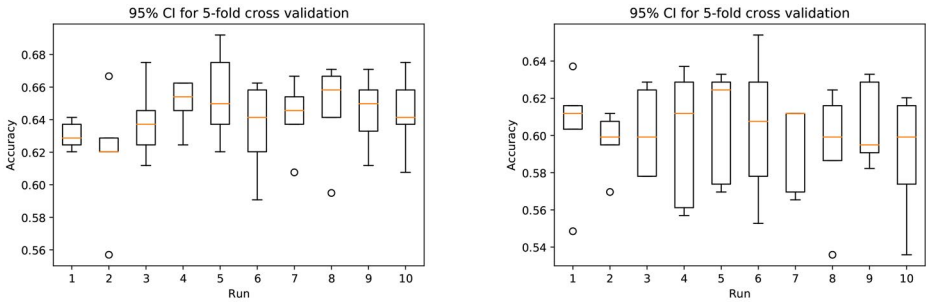
Figure 4. The cross-validated accuracy of our deep learning and traditional machine learning models compared to our two baselines. Our baselines are explained in Section "Introduction".

Table 4. Select top results from our best feature selection algorithms.

χ^2	RFE	MI
Grocery store	people	crowds do
Grades	sound	restaurant during prime
Hectic	the restaurant	large classrooms
Breathing techniques	suspicious person	noises

Table 5. A sample of topics and some prominent words.

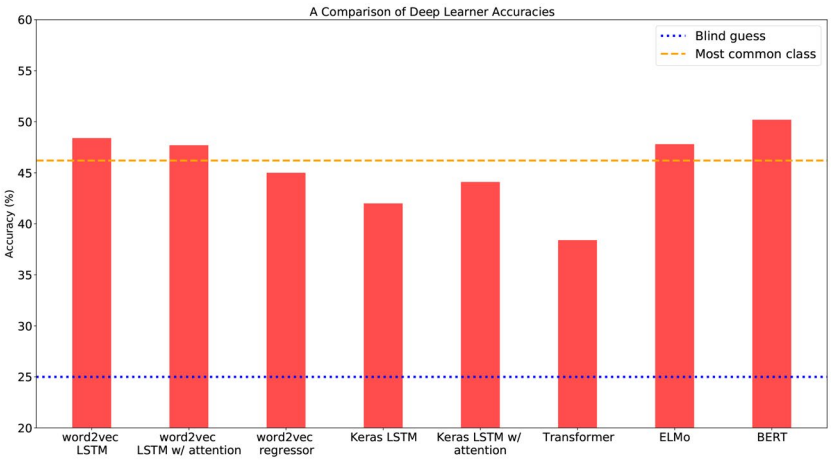
Topic #	1	2	3	4
	best friend	staring	beliefs	surrounded
	talking	eye contact	big store	busses
	chill	aware	group strangers	common sense
	familiar	avoid people	class	actually panic
	don't feel	miserable	sound	concerts



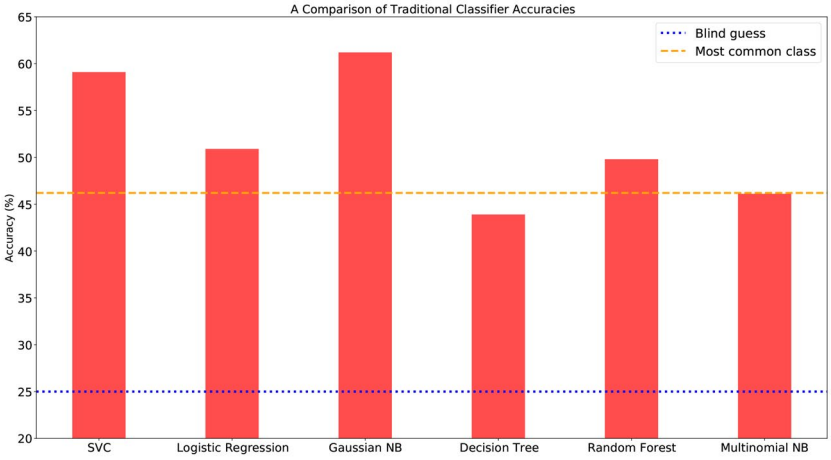
(a) Gaussian Naive Bayes model. (b) SVC model.

Figure 5. 95% Confidence Intervals (CI) yielded from each run of ten-fold cross-validation.

made more sense when looking at the prominent subjectively. We use the combination of perplexity and human judgment because previous studies have shown that predictive likelihood (or equivalently, perplexity) and human judgment are often not correlated and even sometimes slightly anti-correlated (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009).



(a) Deep learning models.



(b) Traditional machine learning models.

Figure 6. The effects of feature selection techniques on the classification accuracy of our SVC. Each group on the horizontal axis shows the number of features that were retained and the accuracy obtained for each feature selection algorithm for that number of features. We start with the entire dictionary of 28,630 tokens and decrease by 5000 at each step.

Select results from our model are highlighted in Table 5, and visualizations for each topic can be found in Appendix. Words like "looking", "friends", "class", "starring", and "people", appear to have a high overall term frequency in the dataset, as evident in Figure 7. Topic 1 in Table 5 appears to be associated with things that provide comforting feelings, whereas topics 2–4 appear to be associated with factors that cause different levels of anxiety, with 2 being a low level, 3 intermediate, and 4 high-anxiety situations.

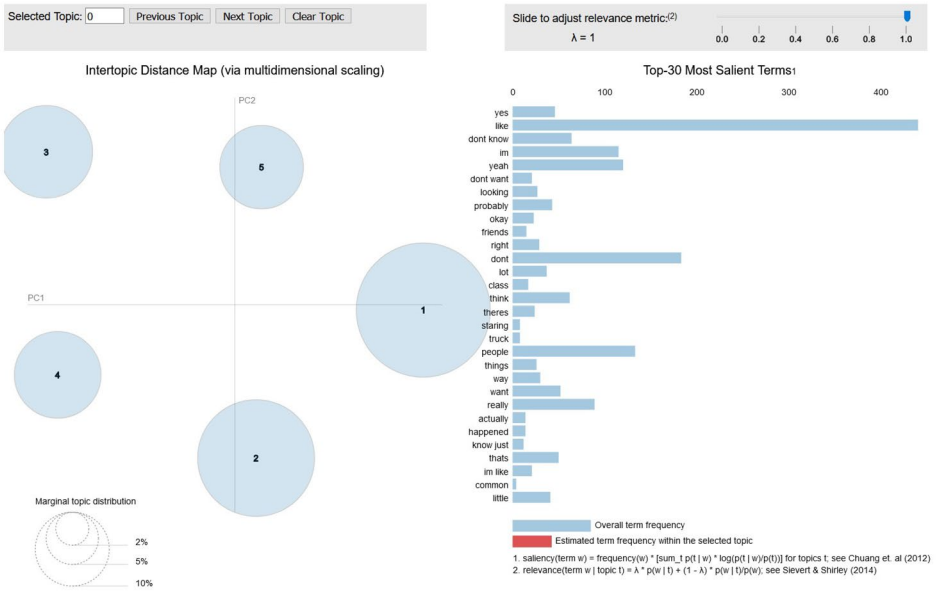


Figure 7. A visualization of corpus-wide topic distribution.

Discussion

Our findings show promise as well as significant challenges in the field of text-based stress detection. The limited sample size, imperfect labeling, and skewed distribution of our target classes affected the performance of our deep learning models, although each model performed better than a blind guess, and a number of our models achieved an accuracy higher than our second baseline (Figure 6a and b).

Coders listened to the recorded interviews while rating each response for anxiety, which resulted in a few instances of sentences having the same text content with different anxiety levels. For example, the lone word 'yeah' appears as a response in our data set 68 times, and the anxiety level of the response varies between instances. Labeling emotional content within the context of audio recordings is still a relatively new field of study and does not benefit from a systematic and validated standard of classifiers.

Further, the performance of our support vector classifier and Gaussian naive Bayes models indicates the potential for greater success if a larger and more robust data set is provided for training. Deep learning models are known to require large training sets for successful training.

The qualitative results provided by our models also offer much insight into areas of language associated with anxiety in student veterans with PTSD and/or social anxiety, and could be regarded as the most valuable outcome

of this study. Our topic model visualization provides an intuitive method for exploring each topic discovered by our LDA model. This interactive visual benefits social workers and psychologists because of its ease of use.

Since the goal of our project was to create tools for treating anxiety, the interpretation of our findings is paramount. The practice of treating social anxiety, specifically for the veteran population, is vital to the functioning of veterans with a history of PTSD. By identifying language associated with varying degrees of assessed anxiety, the work provides an indication of language cues that practitioners may use to evaluate the discourse of veterans with social anxiety. However, because classification in the auditory evaluation of emotions does not have a strong foundation of research evidence currently, this may only serve as an initial exploration of how anxiety and stress could be evaluated for labeling purposes. Limitations to this exploration include the lack of a formal classification system and the potential that anxiety is underestimated in the auditory review of interviews. Further research is needed with larger data sets and more concrete specifics around evaluation cues for anxiety in the verbal delivery of information.

Conclusion

This study used state-of-the-art text analysis tools to assess and treat the problem of social anxiety in patients with PTSD. Our results demonstrate that detecting and classifying anxiety levels from text alone is possible, although still challenging. In addition, we have shown that feature selection and topic modeling are viable methods of producing qualitative data about social anxiety and areas of stress for student veterans. Our work produced tools that can help to support the decision-making of psychology and social work professionals who are treating social anxiety.

Beyond the immediate scope of our problem, our research is applicable to many other areas as well. For instance, automatic emotion recognition can help refine text-to-speech synthesis (Alm, Roth, & Sproat, 2005) or aid in identifying subject areas in which students don't feel confident (Feng et al., 2020). The qualitative data produced by our models also provides a more intuitive look at stressful topics that can be applied and interpreted by a wide audience.

Note

1. The code of the experiments of this study can be found in: https://github.com/imics-lab/text-analysis

Funding

This material is based upon work supported by the National Science Foundation under REU grant #1757893, and by the Texas State University MIRG Grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Notes on contributors

Morgan Byers, holds a Bachelor of Science in Mathematics and Computer Science from Texas State University and is pursuing a Master of Science in Computer Science at the University of Colorado, Boulder.

Mark Trahan, PhD, LCSW, is an Associate Professor at the School of Social Work, Texas State University. Dr. Trahan's research interests include fathering engagement and the use of technology, specifically virtual reality/augmented reality, to enhance paternal and social self-efficacy.

Erica Nason, PhD, is an Assistant Professor at the School of Social Work, Texas State University. Her primary research and clinical expertise are in trauma and posttraumatic stress disorder.

Chinyere Eigege, is a Doctoral Student and Researcher at the Graduate College of Social Work (GCSW), University of Houston.

Nicole Moore, Med, LPC, is a Clinical Researcher at the Graduate College of Social Work, University of Houston, Houston, Texas, USA.

Micki Washburn, PhD, LMSW, LPC-S, is an Assistant Professor at the School of Social Work, University of Texas at Arlington. Her research interests include Health disparities related to mental health and substance use in historically underserved communities.

Vangelis Metsis, PhD, is an Associate Professor of Computer Science at Texas State University. His research interests span the areas of Machine Learning and Computer Vision, focusing on AI-powered applications of Smart Health, Pervasive Computing, and Human-AR/VR Interaction.

ORCID

Micki Washburn http://orcid.org/0000-0003-1486-2501

References

Alm, C., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 579–586).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations,* ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, January). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada (pp. 288–296). Curran Associates, Inc.

Chuang, Z., & Wu, C. (2004). Multi-modal emotion recognition from speech and text. *International Journal of Computational Linguistics and Chinese Language Processing*, *9* (2). (August 2004: Special Issue on New Trends of Speech and Language Processing. 2004.)

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 1, pp. 4171–4186).

Ekman, P. (1999). Basic emotions. In M. P. Tim Dalgleish (Ed.), *Handbook of cognition and emotion*, (Vol. 98, p. 16). John Wiley & Sons.

Fatima, I., Abbasi, B. U. D., Khan, S., Al-Saeed, M., Ahmad, H. F., & Mumtaz, R. (2019). Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, *36*(4). doi:10.1111/exsy.12409

Feng, X., Wei, Y., Pan, X., Qiu, L., & Ma, Y. (2020). Academic emotion classification and recognition method for large-scale online learning environment—Based on A-CNN and LSTM-ATT deep learning pipeline method. *International Journal of Environmental Research and Public Health*, *17*(6), 1941. doi:10.3390/ijerph17061941

Fulton, J. J., Calhoun, P. S., Wagner, H. R., Schry, A. R., Hair, L. P., Feeling, N., … Beckham, J. C. (2015). The prevalence of posttraumatic stress disorder in Operation Enduring Freedom/Operation Iraqi Freedom (OEF/OIF) veterans: A meta-analysis. *Journal of Anxiety Disorders*, *31*, 98–107. doi:10.1016/j.janxdis.2015.02.003

Gruda, D., & Hasan, S. (2019). Feeling anxious? Perceiving anxiety in tweets using machine learning. *Computers in Human Behavior*, *98*, 245– 255. doi:10.1016/j.chb.2019.04.020

Hart, J. (2017). Keras ordinal categorical crossentropy loss function.

Jere, S., & Patil, A. P. (2020). Deep learning-based architecture for social anxiety diagnosis. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1–6). IEEE.

Jo, W., Lee, J., Kim, Y., & Park, J. (2020). Online information exchange and anxiety spread in the early stage of the novel coronavirus (Covid-19) outbreak in South Korea: Structural topic model and network analysis. *Journal of Medical Internet Research*, *22*(6), e19455. doi:10.2196/19455

Kim, J.-S. (2020). Multimedia emotion prediction using movie script and spectrogram. *Multimedia Tools and Applications: An International Journal*, *80*(26), 34535–34551.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, *86*, 105836. doi:10.1016/j.asoc.2019.105836

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, *22*(10), e22635. doi:10.2196/22635

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Portland, OR: Association for Computational Linguistics.

McMillan, K. A., Sareen, J., & Asmundson, G. J. G. (2014). Social anxiety disorder is associated with PTSD symptom presentation: An exploratory study within a nationally representative sample. *Journal of Traumatic Stress*, *27*(5), 602–609. doi:10.1002/jts.21952

Metsis, V., Lawrence, G., Trahan, M., Smith, K. S., Tamir, D., & Selber, K. (2019). 360 video: A prototyping process for developing virtual reality interventions. *Journal of Technology in Human Services*, *37*(1), 32–50. doi:10.1080/15228835.2019.1604291

Morris, P., Powers Albanesi, H., & Cassidy, S. (2019). Student-veterans' perceptions of barriers, support, and environment at a high density-environment enrollment campus. *Journal of Veterans Studies*, *4*(2), 180–202. doi:10.21061/jvs.v4i2.102

Nason, E. E., Trahan, M., Smith, S., Metsis, V., & Selber, K. (2020). Virtual treatment for veteran social anxiety disorder: A comparison of 360 video and 3d virtual reality. *Journal of Technology in Human Services*, *38*(3), 288–308. doi:10.1080/15228835.2019.1692760

Ogunfunmi, T., Togneri, R., & Narasimha, M. (2015). *Speech and audio processing for coding, enhancement and recognition*. New York, NY: Springer. doi:10.1007/978-1-4939-1456-2

Peng, F., & Schuurmans, D. (2003). Combining naive bayes and n-gram language models for text classification. In *European Conference on Information Retrieval* (pp. 335–350). Springer.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (Vol. 1, pp. 2227–2237). Association for Computational Linguistics (ACL).

Rajabi, Z., Shehu, A., & Uzuner, O. (2020). A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)* (pp. 303–306).

Rajapakse, T. (2020). Simple transformers.

Sethu, V., Epps, J., & Ambikairajah, E. (2015). Speech based emotion recognition. In T. Ogunfunmi, R. Togneri, & M. (sim) Narasimha (Eds.), *Speech and Audio Processing for Coding, Enhancement and Recognition* (pp. 197–228). New York: Springer. doi:10.1007/978-1-4939-1456-2_7

Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology–From Linguistic Signal to Clinical Reality* (pp. 58–65).

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70).Baltimore, MD: Association for Computational Linguistics.

Sylvers, P., Lilienfeld, S. O., & LaPrairie, J. L. (2011). Differences between trait fear and trait anxiety: Implications for psychopathology. *Clinical Psychology Review*, *31*(1), 122–137. doi:10.1016/j.cpr.2010.08.004

Trahan, M. H., Ausbrooks, A. R., Smith, K. S., Metsis, V., Berek, A., Trahan, L. H., & Selber, K. (2019). Experiences of student veterans with social anxiety and avoidance: A qualitative study. *Social Work in Mental Health*, *17*(2), 197–221. doi:10.1080/15332985.2018.1522607

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, *9*, 1994. doi:10.3389/fpsyg.2018.01994

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Transformer: Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017* (pp. 5998–6008). Long Beach, CA, USA.

Watson, D., Clark, L. A., Simms, L. J., & Kotov, R. (2022). Classification and assessment of fear and anxiety in personality and psychopathology. *Neuroscience & Biobehavioral Reviews*, *142*, 104878. doi:10.1016/j.neubiorev.2022.104878

Xu, S. (2018). Bayesian naive bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59. doi:10.1177/0165551516677946

Ye, Z., Guo, Q., Gan, Q., Qiu, X., & Zhang, Z. (2019). Bp-transformer: Modelling long-range context via binary partitioning. arXiv preprint arXiv:1911.04070.

Appendix. Topic modeling visualizations

The following visualizations show examples of the common topics that emerged from the analyzed text using Latent Dirichlet Allocation (LDA) and the top 30 relevant terms associated with each topic.

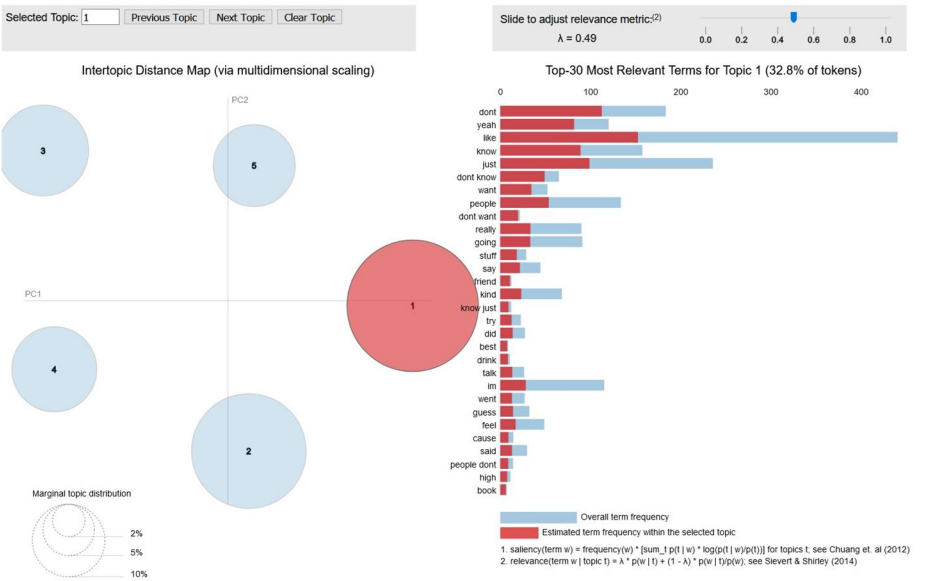


Figure A1. The corpus-wide term distribution given topic 1.

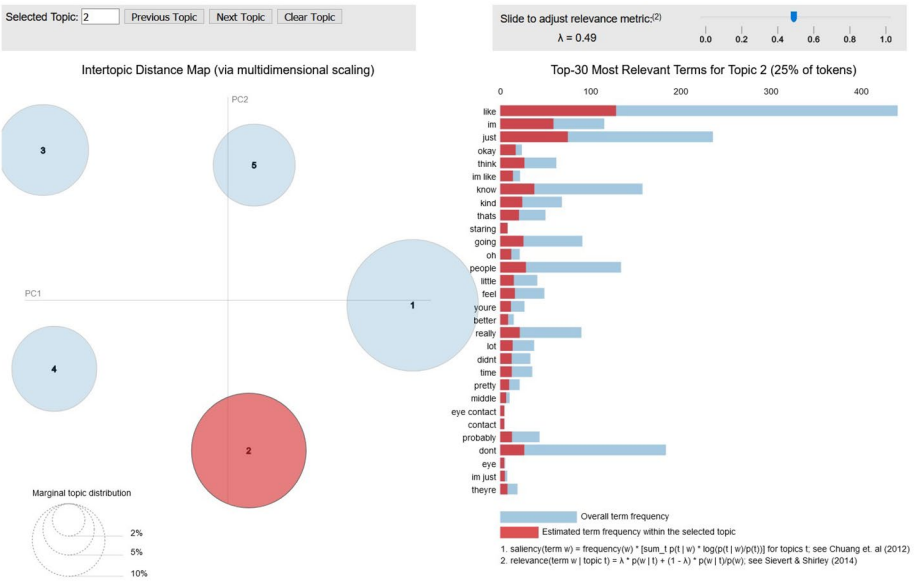


Figure A2. The corpus-wide term distribution given topic 2.

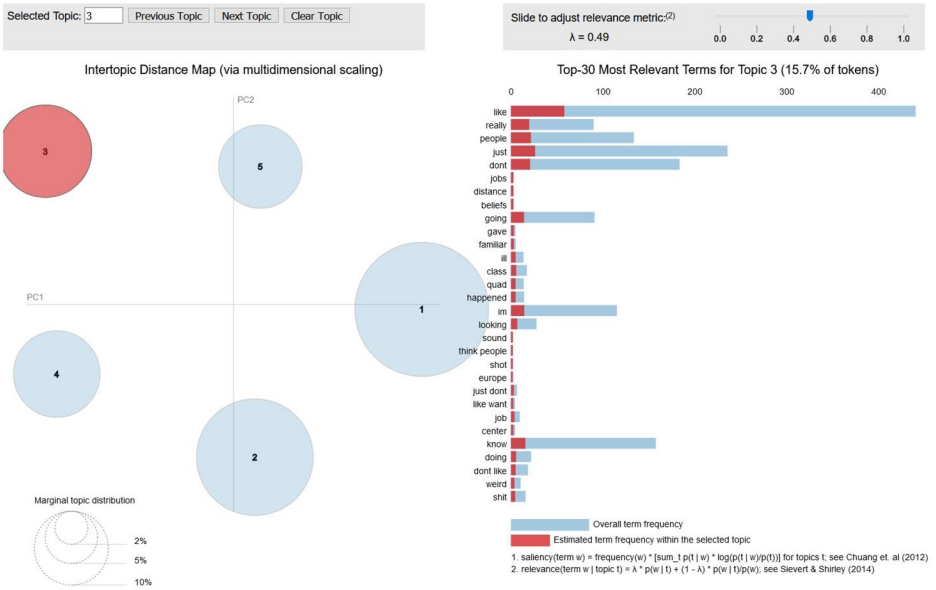


Figure A3. The corpus-wide term distribution given topic 3.

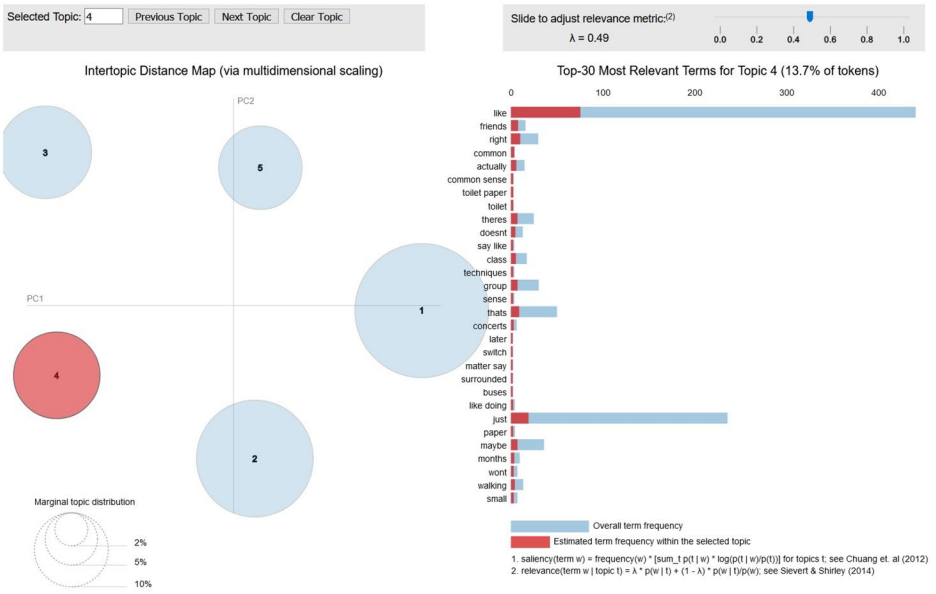


Figure A4. The corpus-wide term distribution given topic 4.

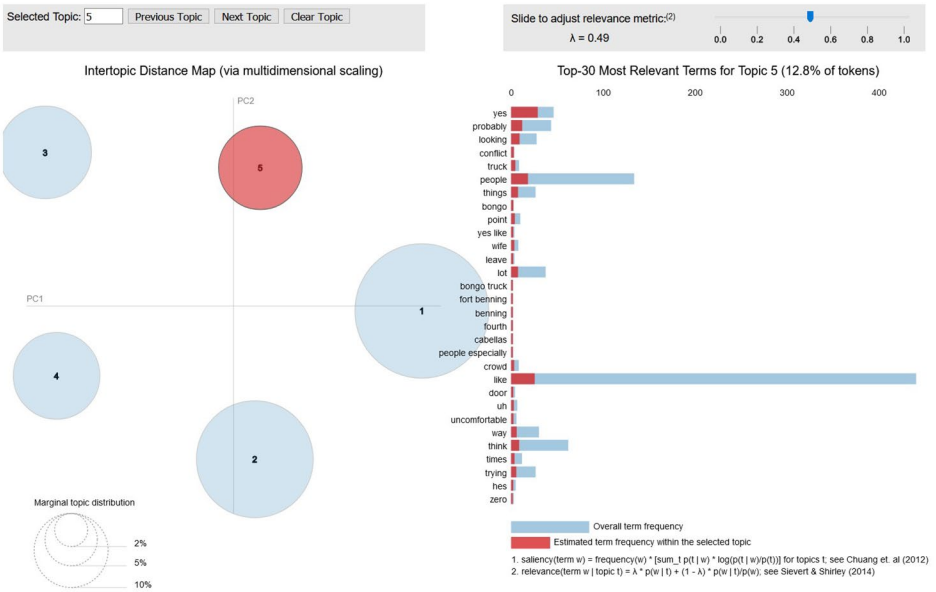


Figure A5. The corpus-wide term distribution given topic 5.