

# Following the online trail of veterans and improving clinical PTSD therapy

---

Irene Michelle López Carrón

Advisor: PhD. Rosa Arriaga

Georgia Institute of Technology

Ubiquitous Computing Laboratory

Master in Informatics Engineering

Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona

*March, 2019*

# Abstract

Understanding veterans' needs is an important task that has so far been addressed using qualitative methods. In the first part of this work we analyze Reddit's *r/veterans* over a 10-year period in order to discover trends and behaviors in this online community. We use *Latent Dirichlet Allocation* (LDA) to extract the main topics of the subreddit and analyze the emotional and linguistic content of the posts. The topics found highlight the issues that are known to be important to veterans offline (e.g., sharing stories, mental health issues). The analysis of historical trends shows a transition from a marketing and news site to an interactive one. Veterans now ask questions about military benefits and about navigating civilian life. We also find that contributors are not likely to seek help from outside the *r/veterans* community and their interactions outside this subreddit were mainly about non-military topics (e.g., entertainment, news).

Having confirmed that *Post-Traumatic Stress Disorder* (PTSD) is the most common mental health issue discussed by online veteran communities, in the second part of this document we explore current issues with PTSD therapy and propose a system to monitor the development and progress of group therapy sessions. We propose to extract audio from the group sessions, to process and aggregate the data in order to monitor the behavior of each patient and the different type of interactions that take place between the members of the group. The biggest challenges in this section lie in maintaining the anonymity of patients and in respecting the privacy of those that might not agree to be monitored during these sessions.

# Resumen

Comprender las necesidades de los veteranos es una tarea importante que hasta los momentos ha sido llevada a cabo usando métodos cualitativos. En la primera parte de esta memoria analizamos *r/veterans* de Reddit durante un período de 10 años para descubrir las tendencias y comportamientos de esta comunidad virtual. Usamos *Latent Dirichlet Allocation* (LDA) para extraer los principales temas discutidos en el subreddit y analizar el contenido emocional y lingüístico de las publicaciones. Los temas encontrados resaltan los problemas que sabemos son importantes para los veteranos en la vida real (compartir historias, salud mental, etc). El análisis de tendencias históricas muestra una transición de una página donde se compartían noticias y publicidad, a una más interactiva; los veteranos hacen preguntas sobre beneficios militares y sobre la reinserción a la vida como civiles. También encontramos que los contribuyentes no suelen buscar ayuda fuera de la comunidad de *r/veterans* y que sus interacciones fuera de este subreddit son principalmente sobre temas no militares como entretenimiento y noticias.

Habiendo confirmado que el *Trastorno por Estrés Postraumático* (PTSD, por sus siglas en Inglés) es uno de los problemas de salud mental más discutidos en comunidades de veteranos virtuales, en la segunda parte de esta memoria exploramos los problemas actuales con la terapia para PTSD y proponemos un sistema para monitorizar el desarrollo y progreso de las sesiones de terapia grupal. Proponemos capturar una grabación de audio de las sesiones grupales y procesarlo para monitorizar el comportamiento de cada paciente junto con los distintos tipos de interacciones entre los miembros del grupo. Los principales retos en esta sección corresponden a la necesidad de mantener la anonimidad de los pacientes y de respetar la privacidad de los participantes que no hayan accedido a ser monitorizados durante las sesiones.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>3</b>
<b>3</b>	<b>Exploratory Analysis of Veterans in Social Media</b>	<b>6</b>
3.1	Data Collection and Pre-processing . . . . .	7
3.1.1	Corpus Construction . . . . .	7
3.2	Topic Modeling . . . . .	8
3.2.1	Topics of the Veterans Subreddit . . . . .	10
3.2.2	Topics' Validation . . . . .	10
3.2.3	Relation Between Topics . . . . .	12
3.3	Data Analysis . . . . .	13
3.3.1	User Interactions . . . . .	13
3.3.2	Emotion Analysis . . . . .	15
3.3.3	Historical Trends . . . . .	16
3.3.4	Veteran Activity in Other Subreddits . . . . .	21
3.3.5	Military and mHealth Subreddits . . . . .	23
<b>4</b>	<b>Improvement of Psychotherapy Through Sensorization</b>	<b>25</b>
4.1	Problem Analysis . . . . .	25
4.1.1	Requirements . . . . .	26
4.2	System Proposal . . . . .	27
4.3	Development and Practical Issues . . . . .	30
<b>5</b>	<b>Conclusions</b>	<b>32</b>
5.1	Exploratory Analysis . . . . .	32
5.2	Therapy Improvement . . . . .	33
5.3	Future Work . . . . .	33
	<b>Bibliography</b>	<b>35</b>

## List of Figures

3.1	Architecture for the label validation web-app . . . . .	10
3.2	Validation web-app requesting the name of the user . . . . .	12
3.3	Validation web-app showing a post that has to be labeled . . . . .	12
3.4	Number of submission over time . . . . .	16
3.5	Number of comments per submission . . . . .	17
3.6	Standard topic progression over time . . . . .	17
3.7	Negative topic progression over time . . . . .	18
3.8	Topic progression over time of 2018 most popular topics . . . . .	19
3.9	Progression of the question and upvote ratios . . . . .	19
3.10	Score of LIWC personal pronoun variables over time . . . . .	20
3.11	Score of LIWC summary variables over time . . . . .	20
3.12	Dates of user milestones . . . . .	21
4.1	ReSpeaker Core v2.0 . . . . .	27
4.2	Proposed architecture for the <i>Group Session Monitor</i> . . . . .	30

## List of Tables

3.1	Number of tokens per type of contribution . . . . .	8
3.2	Topics and the data associated with posts in which they were the dominant label . . . . .	11
3.3	Top-10 most popular topic combinations by number of posts labeled . .	13
3.4	Average metrics of posts according to their classification as questions .	14
3.5	Top-10 Subreddits where users made their first contribution after joining the site . . . . .	22
3.6	Top-10 Subreddits where <i>r/veterans</i> users also posted . . . . .	22
3.7	Possible veteran presence and references on different subreddits . . .	23
4.1	Events detected by the <i>Group Session Monitor</i> . . . . .	29

# Introduction

Military veterans face many challenges during re-adjustment to civilian life. About one third of returning service-members report mental health issues [40], which are compounded by trouble in navigating their educational and medical benefits.

Facilitating access to support groups is important [19] since they could help veterans overcome the challenges of reinsertion. In fact, positive interactions in social networks have been found to improve wellness. Researchers have studied the ways in which information technologies affect veterans[36] but they did so using qualitative approaches. [35].

An open questions is whether and how veterans are using online forums but much of the analysis done in social networks and forum content for specific groups is qualitative or done with a small sub-sample of the data. Here we aim to bridge that gap by taking a fully data-driven approach with unsupervised topic modelling using *Latent Dirichlet Allocation* to investigate the veteran population in Reddit's *r/veterans* forum. We seek to understand the behaviour of this specific online community made for veterans, their needs, the way the conversation in the forum has changed in the past decade, and how they interact with the rest of Reddit.

After modeling and understanding veteran interactions in a social network, we also aim to improve psychological therapy for *Post-Traumatic Stress Disorder* (PTSD). An estimated of 20% of Operations Iraqi Freedom and Enduring Freedom veterans, 12% of Gulf War veterans, and 15% of Vietnam veterans are thought to have PTSD<sup>1</sup> and it is believed that less than half receive any treatment at all, while of those receiving treatment less than one-third receive evidence-based care [27].

Among the evidence-based treatments currently applied for PTSD, Prolonged Exposure (PE) therapy has proven to be the best [34, 31]. Nevertheless, access and utilization of PE is low because of the challenges associated with it [29, 38].

A common component of PTSD therapy and psychological therapy in general are group sessions, where patients can perform the corresponding exercises and have discussions. To improve this process we propose the *Group Session Monitor*, a system

---

<sup>1</sup>[ptsd.va.gov/understand/common/common\\_veterans.asp](https://ptsd.va.gov/understand/common/common_veterans.asp) (2019)

that would record audio during the sessions and process it to extract different metrics corresponding to the behavior, engagement and progress of each patient; it would also track the interactions between patients and with the therapist. The results would later be shown in a dashboard to the clinician who could then track the progress of patients between different sessions and notice details that might have gone missing otherwise.

This audio monitoring tool would have to adhere to IRB (*Institutional Review Board*) regulations, keeping the privacy of the patients and only monitoring those who agree to participate in the study; any external API provider would also have to be approved in that regard.



## Background

Previous research shows that receiving social feedback is helpful to keep individuals engaged and to improve behaviour [6]. More specifically, social networking has proven to be an effective way of increasing social *connectedness* [10] and to manage existing health issues [13]. As such, work has been done using forum data to understand the needs of communities with special needs such as autism [14].

Reddit is a website with a forum format, based around communities dedicated to specific subjects known as “subreddits”, where registered users can make posts called submissions, as well as comment and vote on other users’ contributions; submissions and comments on the site can appear as *[removed]* by the user who posted them or *[deleted]* by moderators, user accounts can also be deleted so the contributions remain but all the user data associated with them is no longer accessible. Most subreddits are open and can be read by anyone without registration making Reddit the 5th most visited site in the US according to Alexa, with millions of monthly active users and thousands of subreddits<sup>1</sup>.

There are also other public and private social networks for veterans, but the other open social networks do not seem to be as active as r/veterans and the private ones might not be accessible to all of them or the additional identity verification required could be a deterrent for veterans that want more power over their anonymity.

For example the Department of Veterans Affairs (VA) has a mobile application called *DoD Veteran Link*<sup>2</sup> which works as a verified forum for veterans and service-members, but the usage conditions indicate that veterans must be eligible for VA benefits, something based on factors such as military service, type of discharge, among others<sup>3</sup>.

Previous research has been done based on the changes over time of other Reddit communities [32] in order to draw insights from specific populations. Topic modelling has been applied to studies using Reddit [11] as well as other mixed data sources [41, 1]. Although there are many topic-modeling techniques the one we use, *Latent Dirichlet Allocation* (LDA), is a widely used and validated [12, 39, 20]. In the first

---

<sup>1</sup>redditinc.com (2019)

<sup>2</sup>mobile.va.gov/app/va-dod-veteran-link (2019)

<sup>3</sup>va.gov/health-care/eligibility (2019)

section of this thesis, we seek to draw insights about veterans behavior on Reddit using topic modelling with LDA, emotional and linguistic content analysis, and historical trends' analysis.

For the second part of this thesis we focus on the challenges and limitations of current psychological therapy, which is constrained by the subjective data of the clinician's observations and the patients' self-report. We focus specifically in therapy for Post-Traumatic Stress Disorder, since it is the most common mental health issue that affects military veterans.

Post-Traumatic Stress Disorder (PTSD) is defined by the Veterans Administration (VA) of the United States of America as "an anxiety disorder that can occur following the experience or witnessing of a traumatic event" which is "complicated by the fact that people with PTSD often may develop additional disorders such as depression, substance abuse, problems of memory and cognition, and other problems of physical and mental health" <sup>4</sup>.

To date, the most effective treatment for PTSD is Prolonged Exposure (PE) therapy [34, 31]. PE is a type of exposure-based therapy that reduces the patient's excessive anxiety through confronting anxiety-provoking or avoided thoughts, situations, activities and people that are not realistically threatening [18].

Many efforts already exist to improve medical practice through technology using mHealth (mobile health) applications, whose aim is to expedite the gathering and use of information, promote patient engagement, and improve the delivery of treatments. Many mHealth applications have already been developed to aid in the diagnosis and treatment of PTSD symptoms [28], some were even designed with the objective of helping with the actual clinical treatment, like the well known PE Coach.

PE Coach [25] is a patient-facing smartphone application designed to be used during PE therapy for PTSD with the guidance of the clinician in charge. The application guides the patients through the exercises assigned by the therapists, allows the patients to keep track of their progress and to record an audio of their sessions, helps users practice techniques such as controlled breathing to decrease the level of distress, and keeps track of upcoming therapy sessions<sup>5</sup>.

A study found that many clinicians are already using or intend to use PE Coach in the future and that the perception of the application was generally favorable regarding its relative advantage over current practices, compatibility with clinicians' values and

---

<sup>4</sup>[mirecc.va.gov/cih-visn2/Documents/Patient\\_Education\\_Handouts/Handout\\_What\\_is\\_PTSD.pdf](https://mirecc.va.gov/cih-visn2/Documents/Patient_Education_Handouts/Handout_What_is_PTSD.pdf) (2019)

<sup>5</sup><https://mobile.va.gov/app/pe-coach-2> (2019)

needs, complexity, trialability, and observability [16]. In a different study, soldiers rated the application positively and reported higher levels of satisfaction during PE with PE Coach as compared with PE alone [26].

Studies have also been done to expose how the rise of ubiquitous computing and personal sensing has improved mHealth applications [21]. For this reason in the second part of this work we explore how the PE Therapy could be improved by acquiring and processing data from the sessions, more specifically audio data from the group sessions.

There have already been projects which explore the use of audio-monitoring to improve meetings and give feedback on the participants' behavior [33]. Our proposed system applies a similar idea to the specific case of psychological group therapy.

We must keep in mind that when dealing with audio analysis some of the most relevant metrics are:

- Duration, usually in seconds.
- Number of syllables.
- Number of pauses: A pause is defined as an interval during which there is no voice (phonation) on the audio track and it is distinguished from the normal gap between syllables by defining the *pauses* over a certain time and noise threshold levels.
- Phonation time: Total time during which someone is speaking on the audio track. It is defined as the duration of the track minus the sum of all pause intervals.
- Speech rate: It is defined as the number of syllables divided by the duration of the track.
- Articulation rate: It is defined as the number of syllables divided by the phonation rate.

Other interesting insights can also be drawn from the content of the audio, such as the emotion and linguistic characteristics of the speech.

## Exploratory Analysis of Veterans in Social Media

Within the frame of the project to improve psychological therapy for war veterans we developed an exploratory and data-driven analysis of the social media interactions, needs and behaviors of our target population, for which we focused on Reddit's *r/veterans*, the most popular veteran's subreddit as of January 2019<sup>1</sup>.

Reddit is a website with a forum format, based around communities dedicated to specific subjects known as “subreddits” and named by following the convention “*r/theme\_of\_subreddit*”. Most subreddits are open and can be read by anyone without registration making Reddit the 5th most visited site in the US according to Alexa, with millions of monthly active users and thousands of subreddits<sup>2</sup>; registered users can make posts called submissions, as well as comment and vote on other users' contributions (both on comments and submissions), the votes can be positive or negative and they accumulate to make the overall score of a submission which, along with the number of comments and other factors determines how high on the main page the post will appear, and thus how likely it is to be seen by more users, being a form of group moderation.

Inside a subreddit, submissions are organized in different order within five categories based on the following sorting criteria:

- **New:** Time of posting, with most recent posts at the top of the ranking.
- **Top:** Overall score from user votes, from highest to lowest.
- **Controversial:** Favours posts with an even amount of upvotes and downvotes, along with many comments.
- **Hot:** Posts with high score and many recent comments are placed higher and they are ranked lower as recent activity (votes and comments) stops.

---

<sup>1</sup>[reddit.com/r/veterans](https://www.reddit.com/r/veterans) (2019)

<sup>2</sup>[redditinc.com](https://www.redditinc.com) (2019)

- **Rising:** Similar to hot, but only takes into account whether a post has a lot of recent activity, be it comments or votes.

The decision to focus on Reddit and specifically on *r/veterans* was made based on the fact that it was the most active public-access veteran community online, with over 33 thousand subscribers and many posts per day as of January 2019.

## 3.1 Data Collection and Pre-processing

Reddit data can be freely accessed using its API<sup>3</sup> which defines objects such as “submission”, “comment” and “redditor”, however there are limitations on the amount of posts and comments than can be retrieved and from how far back. To work around this and be able to obtain all the data from our target subreddits we used *Pushshift*<sup>4</sup>, a public big-data project that stores a copy of Reddit objects, to get the IDs of every submission and the comments made to it, then we queried the official API directly with the IDs to obtain the complete and most updated object data.

Submissions and comments on the site can appear as *[removed]* by the user who posted them or *[deleted]* by moderators, user accounts can also be deleted so the contributions remain but all the user data associated with them is no longer accessible. This is something important to be taking into account when querying and processing Reddit data, since some parts of it might stop being available short time after an initial data crawling.

### 3.1.1 Corpus Construction

The data retrieved from the veterans subreddit contained 132.591 documents corresponding to 18.836 submissions and 113.755 comments; we eliminated those that had been either removed or deleted because they did not contain the full original text, an action which reduced the dataset to 13.829 submissions and 108.806 comments.

On the remaining data we removed links and emails, substituting them for the tokens [LINK] and [EMAIL] respectively. We then tokenized each document using the gensim library “simple\_preprocess” function, then we calculated the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the number of tokens in the submissions and comments which can be seen in Table 3.1.

---

<sup>3</sup>[praw.readthedocs.io](http://praw.readthedocs.io) (2019)

<sup>4</sup>[github.com/pushshift/api](https://github.com/pushshift/api) (2019)

Based on the length of the contributions and on the fact that unlike comments, submissions do not usually need a context to be understood we decided to discard the comments for the topic modelling and only kept the submissions, so the remaining dataset contained only 13.829 documents.

	Tokenized		Processed	
	$\mu$	$\sigma$	$\mu$	$\sigma$
<b>Submissions</b>	77	137	37	65
<b>Comments</b>	48	64	22	30

**Tab. 3.1:** Number of tokens per type of contribution

We proceeded to create bigrams and trigrams from the tokens in each document in order to obtain concepts like “post\_traumatic\_stress” and “vet\_center”. To create the n-grams, different words that co-occur together with certain frequency are joined and considered as one, with the purpose of providing natural language processing algorithms with more information to extract concepts and similarities.

We also lemmatized the tokens to obtain the dictionary form of the words, for example, the words “walk”, “walked” and “walking” would all be transformed to “walk” and thus counted as repetitions of the same token.

With all the pre-processing done we generated a corpus of 21.428 unique tokens and filtered the extremes keeping only words with a frequency of appearance higher than 5 and that appeared in less than 30% of the corpus, this led to a corpus of 7.471 unique tokens. As a final step we converted the documents into a bag-of-words to be used by the topic-modeling algorithm.

A bag-of-words is a model for feature extraction from text that discards all information about the order of the words and only counts the occurrences, meaning that it only detects whether words from the corpus are present or not in a given text and how many times.

## 3.2 Topic Modeling

Topic modeling is a statistical modeling technique to find the distribution of topics within a set of documents, where each topic is formed by a group of words picked through a metric of similarity.

In this section we use *Latent Dirichlet Allocation* (LDA), a generative probabilistic model of a corpus, where documents are represented as random mixtures over

latent topics, each topic is characterized by a distribution over words and the topic distribution is assumed to have a sparse Dirichlet prior[3].

This model cannot discover the optimal number of topics for a given corpus of documents, so it must be explicitly provided. The usual way to estimate the number of topics is through model fit metrics followed by expert inspection of the topics.

There are different metrics commonly used to evaluate model fit but the most frequent ones are perplexity and coherence.

Perplexity or held-out likelihood is useful for evaluating how well the model would perform when applied to new documents, but does not address the more explanatory goals of topic modeling[5]. The perplexity measure does not reflect the semantic coherence of individual topics learned by a topic model [23] and thus models with better predictive capabilities may not be the most interpretable for the current corpus.

The alternative metric used is coherence, which measures whether the words in a topic tend to co-occur together and is usually defined as the average or median of pairwise word similarities formed by top words of a given topic[30]. We can thus apply an elbow method to select the number of topics with the highest coherence.

It is also worth noting that modeling with a high number of topics would generate redundancy, while a low number would cluster different themes into the same topics, so analysis of the generated topics using domain knowledge is also useful to select the best fit for alternatives with similar coherence.

There are two popular libraries which provide efficient implementations of LDA, Gensim<sup>5</sup> and Mallet<sup>6</sup>, one of the key differences between them is that the former uses Variational Bayes while the later uses Gibbs Sampling.

We computed the coherence score of both models using our pre-processed corpus and a range of topics from 10 to 30, the highest value was obtained using Mallet and 17 topics [17]. We also analyzed manually the models obtained using the two libraries and the topics returned by Mallet seemed to be the most interpretable and descriptive.

---

<sup>5</sup>[radimrehurek.com/gensim/models/ldamulticore](http://radimrehurek.com/gensim/models/ldamulticore) (2019)

<sup>6</sup>[radimrehurek.com/gensim/models/ldamallet](http://radimrehurek.com/gensim/models/ldamallet) (2019)

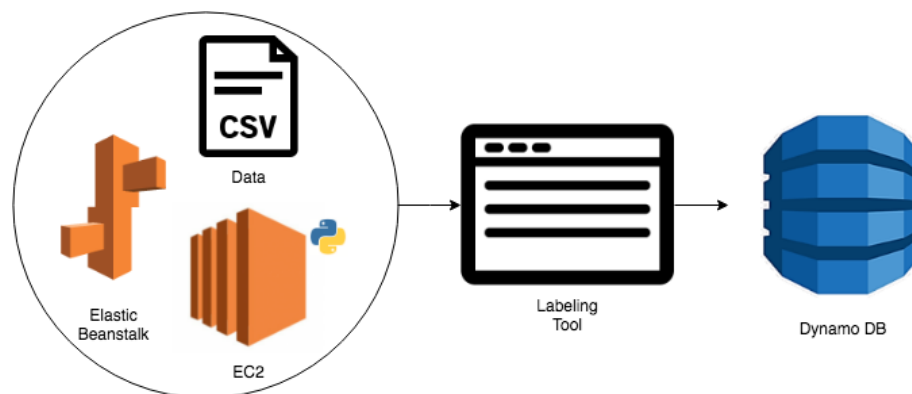
### 3.2.1 Topics of the Veterans Subreddit

We obtained the best model using the coherence metric followed by manual analysis of the topics generated. The selected model provided 17 topics, two of which were discarded since they were formed by common words that appeared in very unrelated submissions and were not useful to understand latent themes in the corpus. The final list of 15 topics can be seen in Table 3.2; the topics were manually named based on the top keywords and an analysis of the posts corresponding to each one of them.

### 3.2.2 Topics' Validation

The accuracy of the topic labels obtained with LDA was verified empirically by checking it against random samples of the data, nevertheless this was only done by one person so it was not enough to draw strong conclusions about the quality of the model applied. In order to get a more reliable metric of labeling accuracy and in line with the method used in other similar works we decided to find volunteers to verify the labels and compute an estimated accuracy based on the results obtained.

To simplify the label verification task we developed a label validation web-app deployed using Amazon Web Services (AWS) Elastic Beanstalk and DynamoDB, with the architecture from Figure 3.1.



**Fig. 3.1:** Architecture for the label validation web-app

The application was developed in Python using Django and a subset of 100 random posts was extracted from the LDA-labeled dataset into a CSV file. The application asked the user's name when it was first opened, then it started showing the posts from the files one by one. For each post the users were asked to write the dominant and secondary label that they felt better fit the text, then after pressing submit this information along with the post ID and the name of the user was stored in



Topic label	Top-10 Keywords	Posts	Emotion
Sharing Stories/Thoughts	vet, war, soldier, father, serve, marine, post, story, honor, country	2039	Positive
Mental Health/Vet Issues	veteran, ptsd, resource, life, government, care, veterans_affair, fight, suicide, fund	1744	Neutral
Education	school, gi_bill, class, college, pay, semester, post_gi_bill, university, credit, student	1114	Positive
Offers and Research	veteran, military, link, service, experience, study, veterans_day, research, interested, free	1104	Positive
Claiming Disability	claim, disability, file, rating, rat, ptsd, exam, appeal, question, deny	917	Positive
Rants	people, feel, life, fuck, friend, day, shit, ptsd, iraq, love	884	Negative
Housing/Family	live, home, move, work, family, wife, house, place, good, loan	822	Positive
Benefits Questions	benefit, question, find, state, cover, anyone_know, answer, info, offer, website	752	Positive
Discharge/Deployment	year, army, military, active_duty, service, time, discharge, navy, leave, join	744	Positive
Job Search	job, work, program, apply, post, voc_rehab, plan, advice, year, career	728	Positive
Health and VA	issue, appointment, doctor, va, medical, care, schedule, give, problem, show	644	Neutral
Payments/Money	month, pay, receive, bill, money, start, payment, check, benefit, back	625	Positive
Information: Paperwork	call, send, letter, office, mail, contact, form, wait, today, change	612	Positive
Health Issues	pain, back, bad, treatment, medication, med, hurt, diagnose, time, leave	554	Negative
Opportunities/Programs	experience, training, company, provide, information, include, business, civilian, support, area	442	Positive

**Tab. 3.2:** Topics and the data associated with posts in which they were the dominant label

DynamoDB and a new post was retrieved from the CSV file. Screen captures of the web application can be seen in Figure 3.2 and Figure 3.3.

This application was deployed on the cloud and at the time of writing could be found [here](#). Some of the advantages of hosting it on the cloud were ease of use for the

labelers, a consolidated database without need for the users to send back a resulting file and the ability to make changes to the data or the application itself without having to resend an executable.

Despite how easy it was to send and use the validation web-app, it was impossible to find enough volunteer-labelers and the accuracy metrics could not be computed.

Example 0/100

Title  
Here the title of a Reddit submission will be displayed

Body  
If said submission has a body it will be displayed here

**UNDERSTOOD**

**Fig. 3.2:** Validation web-app requesting the name of the user

7/100

Title  
CDL schools in TX

Body  
Does anyone know of any VA approved CDL schools in TX that's NOT tied to a company? Specifically the DFW area. Thanks. Edit: Never mind. Found 3. Thanks all.

Input a label that best fits the text shown above

**SUBMIT**

**Fig. 3.3:** Validation web-app showing a post that has to be labeled

### 3.2.3 Relation Between Topics

Unlike other techniques such as K-means that assign every document to a single cluster, LDA provides the mixture of topics corresponding to each document with an associated probability distribution. Since the posts were in average short texts centered in a specific subject we decided to follow a procedure similar to the one from Toulis Andrew and Golab Lukasz [41] and label post with its “Dominant topic”, the one that had the highest probability, a “Secondary topic” was also assigned to documents where a second topic had at least half the probability of the dominant one, this was useful to separate posts that could be labeled by just one topic from those that could be a mixture of more. For the submissions where the dominant topic was one of those discarded, the secondary topic was selected instead, if there was no secondary topic we considered that the text could not be accurately classified and was discarded; the two removed topics were not very prevalent in the dataset so only 50 posts were discarded.

Since the posts labeled with a given dominant and secondary topic were very similar to those labeled with the inverse combination, they were grouped together into one, after this the topic combinations that had less than 14 topics were removed, because they represented less than 0.1% of the dataset. The final result were 124 possible

topic combinations the most popular of which, by number of posts labeled, can be seen on Table 3.3.

Topic Combination	Posts
Sharing Stories/Thoughts mHealth/Veteran Issues	702
mHealth/Veteran Issues Offers and Research	447
Sharing Stories/Thoughts Offers and Research	429
Education	387
Reinsertion: Rants	318
Payments/Money Education	315
Sharing Stories/Thoughts Reinsertion: Rants	309
Sharing Stories/Thoughts	307
Claiming Disability	295
Education Job Search	257

**Tab. 3.3:** Top-10 most popular topic combinations by number of posts labeled

## 3.3 Data Analysis

In order to better understand *r/veterans* and its users we performed different analyses over the topic-labeled dataset of valid submissions.

### 3.3.1 User Interactions

We found that 75% of all posts made on *r/veterans* elicited at least one comment in return. The popular moderation bot “AutoModerator” appeared for the first time in the subreddit on October 2018, so we know that most comments are likely to be real interactions instead of automatic moderation replies.

In order to understand why some posts generated more interaction we decided to classify them based on whether they contained a question or not. Since an empirical analysis of the subreddit showed that the posts tended to be well written we classified them based on the presence of a question mark “?” finding that 6.309 posts could be considered as questions, this amounted to 47% of the dataset.

When we analyzed which posts elicited comments the result was that 97% of those classified as questions had comments while only 57% of those not classified as questions did. We also found that the amount of comments in submissions classified as questions was on average over two times higher than those not classified as such, but that at the same time they got a lower average user score<sup>7</sup> than those not containing questions as can be seen in Table 3.4; the upvote ratio, a metric of how many upvotes against downvotes the post has, is also lower for submissions which contains questions. Taking those three metrics into account it is clear that posts which contain questions are on average more controversial.

Type of post	Comments		Score		Upvote ratio	
	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$
<b>Question</b>	10.5	11.8	4.9	8.8	0.68	0.16
<b>Not a question</b>	4.1	8.8	7.2	14.6	0.75	0.2

**Tab. 3.4:** Average metrics of posts according to their classification as questions

In order to understand the response generated by posts labeled with different topics we made ranking ordering all possible combinations by total number of posts, the ratio of how many were questions, ratio of how many got comments, average user score, average number of comments and average upvote ratio<sup>8</sup>. We then applied Kendall's Tau, a measure of the correspondence between two rankings, but most of the results obtained had a p value over 0.05 meaning that the rankings obtained were uncorrelated. The only exception was between the upvote ratio and the question ratio, where we found a slightly positive correlation of 0.12 ( $p=0.006$ ), meaning that posts which included questions were more likely to generate mixed opinions between voting users.

This was not completely surprising, since it explains the structure of Reddit, where posts can be ordered by "Top" based on their user score or by how "Controversial" they are if they have a high number of comments but a balanced upvote ratio.

Based on the Reddit terminology we still wanted to understand which topics were the "top", the most "controversial", which ones generated more interactions and which ones were more often associated with questions. An analysis of the corresponding rankings allowed us to conclude the following:

The hottest topics, by user score, were the ones that involved rants, thoughts and stories, physical and mental health issues, and the VA.

<sup>7</sup>In Reddit the users can vote a post up or down and the final score is the sum of all the votes received, a metric that identifies how much users like or agree with the post

<sup>8</sup>Metric from 0 to 1, that represents the amount of upvotes vs downvotes so if there is an equal number of votes the upvote ratio will be 0.5

The most interactive topics, by number of comments, were those associated with rants, education, finding jobs, payments, and health issues.

The most controversial topics, by upvote ratio, were those about discharge or deployment, payments, the VA, claiming disability, and finding jobs.

The topics which labeled questions most frequently were the ones related to education, payments, claiming disability, finding jobs, and discharge or deployment.

Finally it is worth reviewing from Table 3.2 that the most popular topics by number of posts were those containing thoughts and stories, veteran issues, education, claiming disability, and research or offers.

This shows that there is indeed a difference in what topics engage users, which ones they like or agree with and the ones that fill the subreddit the most, but the previously found correlation between posts with questions and controversial ones remains.

### 3.3.2 Emotion Analysis

To better evaluate the content of the posts labeled with different topics we extracted the emotion associated with each one using Vader[15], a rule-based sentiment analysis tool, better suited for social media contexts than other popular alternatives such as LIWC.

Vader provides a measure of *positive*, *neutral* and *negative* emotion associated with a text in a range from 0 to 1, along with a *compound* score that ranges from -1 to 1, representing an overall emotion from negative to positive.

We used the *compound* score obtained from the algorithm to classify each post as positive, neutral or negative, selecting the following threshold values recommended in the literature:

*positive* :  $score \geq 0.05$

*neutral* :  $0.05 > score > -0.05$

*negative* :  $score \leq -0.05$

With this data we also made a ranking, ordering all possible combinations by the average compound emotion score. We then applied Kendall's Tau against the

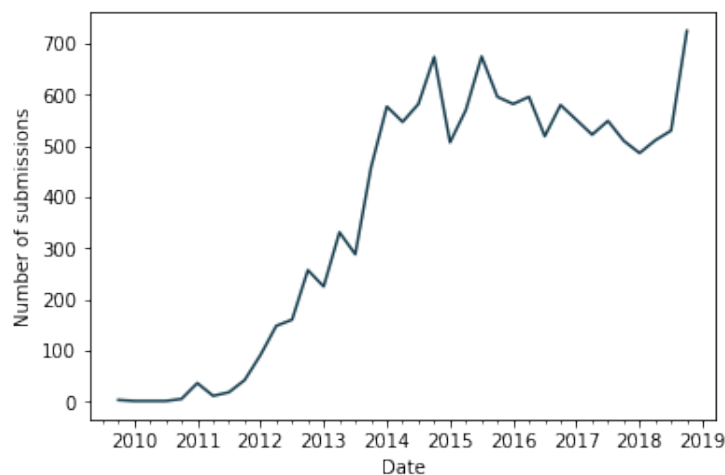
previously obtained rankings, finding a slightly positive correlation of 0.11 ( $p=0.012$ ) between the emotion score and the number of posts which received comments; the rest of the rankings proved to be uncorrelated once again with a  $p$  value over 0.05. These results would indicate that posts with a positive emotion are less likely to go unanswered but do not necessarily generate more interactions.

We were also able to identify that the topics most frequently associated with a negative sentiment were those involving rants and health issues. While the ones associated with a positive sentiment were those about finding jobs, education and other benefits.

### 3.3.3 Historical Trends

The analyses done so far have been based on average values over the whole dataset, but given that *r/veterans* has been active since 2009 it is likely that some insights can only be found while examining the changes in the subreddit over time, which is the goal of the present section.

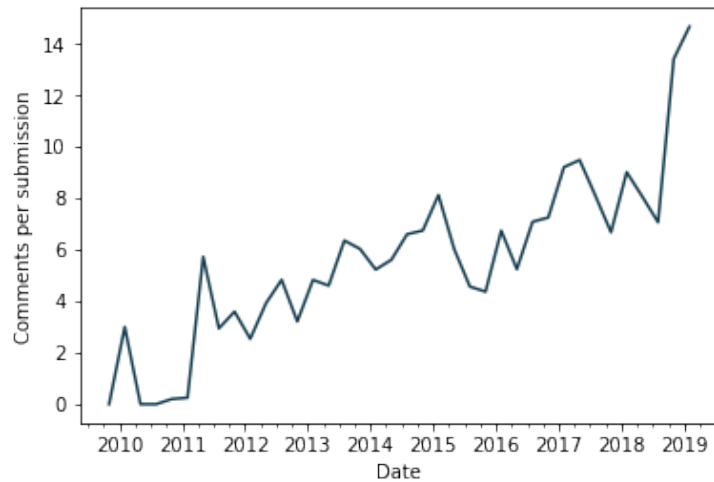
Despite *r/veterans* being created on 2009, in that year only 3 valid posts remain, a value that increases to 31 for the year 2010, but which is still too low to draw conclusions, we will therefore focus the following analyses on the trend-lines observed and on the changes starting from 2011 when the number of submission started becoming more relevant.



**Fig. 3.4:** Number of submission over time

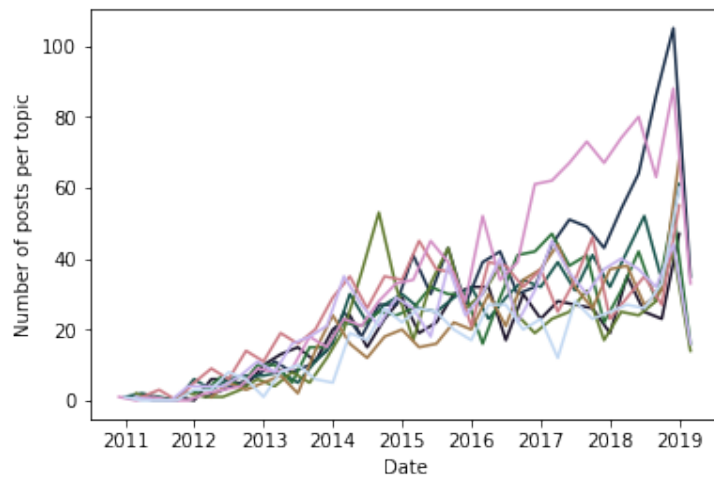
We observed in Figure 3.4 that the number of posts in *r/veterans* had increased over time, in line with the rising popularity of the site, but we also wanted to visualize how the interactions in the subreddit changed over the years and for this we plotted in Figure 3.5 the average number of comments generated per post, where we can

see how the amount of comments generated also increased, which means that posters are more likely to get feedback as time progresses and that the community is becoming more active.



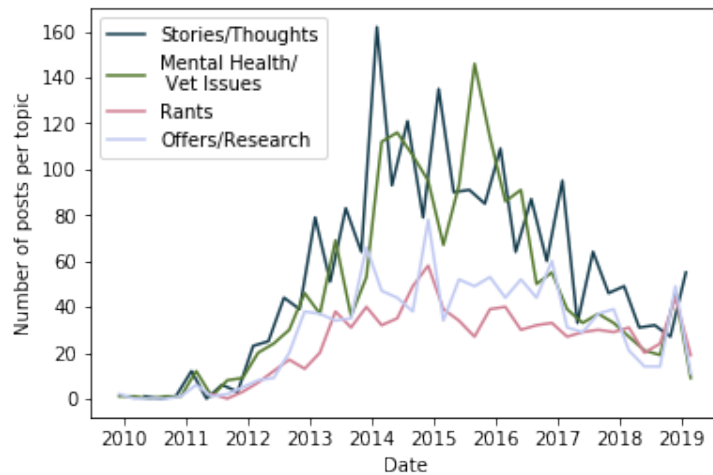
**Fig. 3.5:** Number of comments per submission

We also analyzed the distribution of topics over the years and for eleven of the fifteen topics, the number of documents labeled with them increased over time as expected and as shown in Figure 3.6. The remaining four had a peak between 2014 and 2016 but then started declining in usage as can be seen in Figure 3.7.



**Fig. 3.6:** Standard topic progression over time

Going back to Table 3.3 we can see that the topics from Figure 3.7 tend to appear together in documents, which might explain why their change in popularity followed a similar trend. An example of a post labeled with one of these topics and made between 2014 and 2016 can be seen below:



**Fig. 3.7:** Negative topic progression over time

“On average; 22 veterans commit suicide each day; according to the Iraq and Afghanistan Veterans of America...”

These topics were associated with a low number of questions and many of them seemed to be news or facts like the one above. On the other side, the posts labeled with one of these topics starting from 2017 look like the following example:

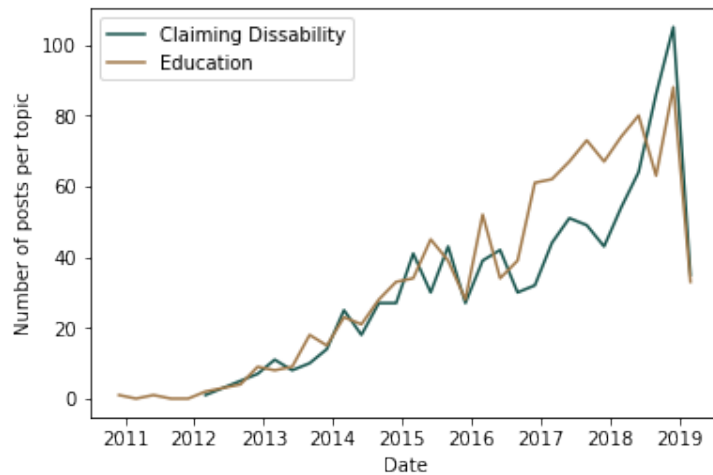
“What would it take to end veteran suicides?. The rise of veteran suicide rates has become a hidden epidemic across our nation...”

As the examples show not only did the number of posts labeled with more 'news-related' topics decreased, but even the type of post in the categories changed, evidenced also in the fact that from the posts of 2014 and 2015 only 11% had questions while 28% of the posts from the same topics starting 2017 did.

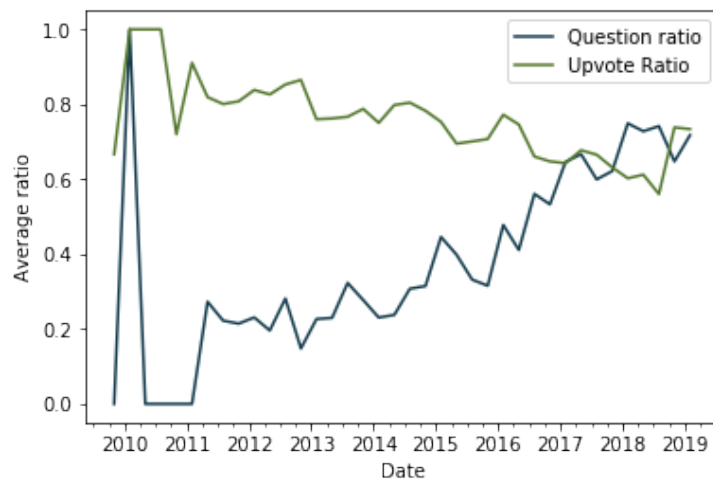
In 2018 the topics with the highest number of posts were those related to “Claiming Disability” and “Education” which were of a very interactive nature and mostly composed by questions (72% and 82% respectively). The trend followed by these two topics can be seen in Figure 3.8 and it has been generally positive since they first started being used, although this did not happen until 2012 and 2011 respectively.

The subreddit seems to be growing away from being a place to get news and information, in order to become a more interactive help and discussion forum, an observation that is also backed by Figure 3.9 where we see how the number of posts containing questions increased while at the same time they became more controversial, a trend which is also in line with the positive correlation of those variables found in section 3.3.1.





**Fig. 3.8:** Topic progression over time of 2018 most popular topics



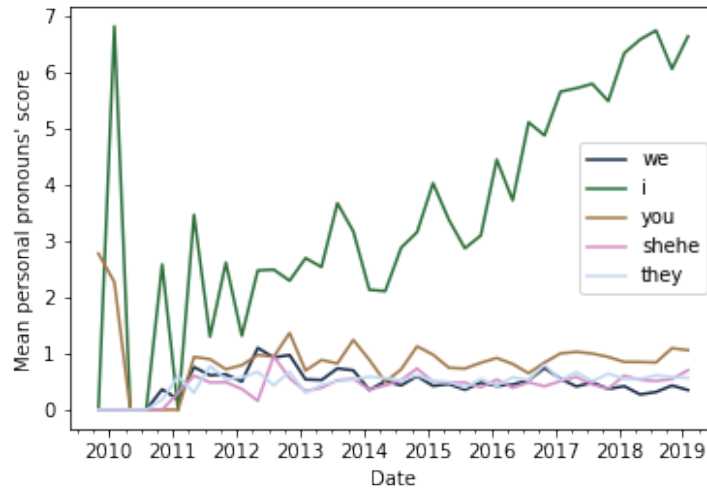
**Fig. 3.9:** Progression of the question and upvote ratios

Since reviewing examples manually is not feasible or in line with our data-focused approach we decided to analyze the language of the posts using LIWC2015 (Linguistic Inquiry and Word Count), an extensively validated tool for these types of tasks which counts word usage and distributes it over pre-established categories, with values on a 100-point scale ranging from 0 to 100. From all the categories provided by the LIWC tool we analyzed the summary variables (Analytical, Authentic, Clout, Tone) and personal pronouns (i, we, you, heshe, they). The summary variables are associated with the following traits:

- Analytical: Formal, logical, and impersonal.
- Authentic: Honest, personal and disclosing.
- Clout: Expertise and confidence.

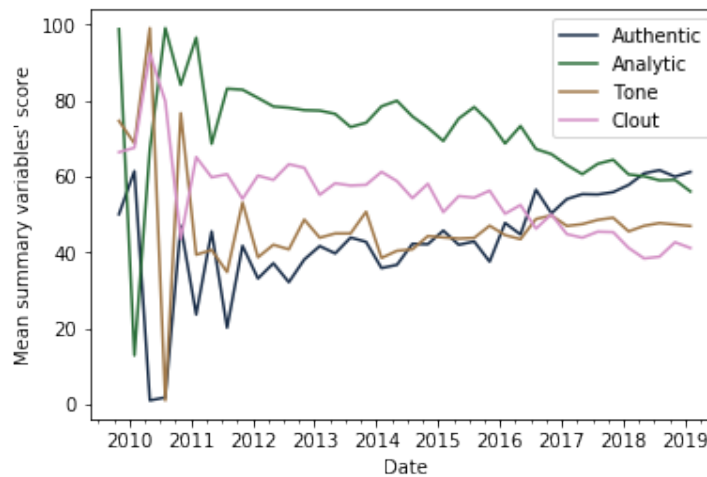
- Tone: Higher values reflect a positive style, around 50 represents ambivalence.

Figure 3.10 shows that the usage of the first person singular (I, me, mine) is much higher than other personal pronouns and that it has increased over time.



**Fig. 3.10:** Score of LIWC personal pronoun variables over time

On Figure 3.11 we can see how the Analytic and Clout variables have decreased and the Authentic one has increased, which corresponds to the content of the posts becoming more informal, personal, honest and tentative.



**Fig. 3.11:** Score of LIWC summary variables over time

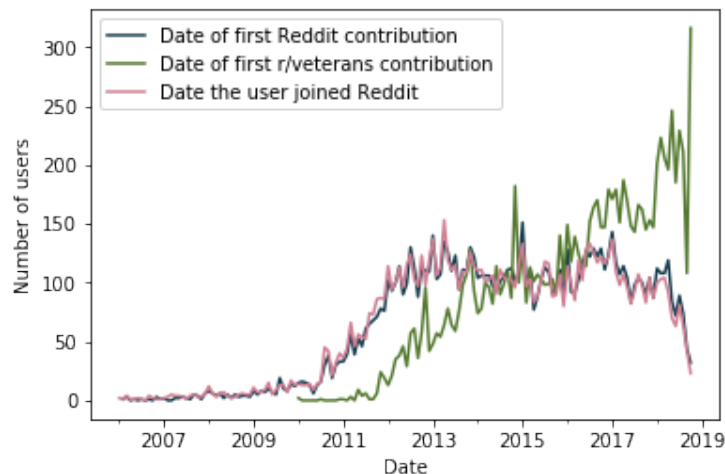
The observations derived from the LIWC variables supports the previous conclusion that the subreddit is becoming a place where veterans ask question related to their personal issues and interact with each other.

### 3.3.4 Veteran Activity in Other Subreddits

An empirical analysis of the veterans subreddit showed that most posts and comments seemed to be from people who identify as veterans, with the exception of posts with news and about research studies or offers. Since we wanted to track the real veterans on the site we used the topics obtained previously to remove the users that had posted submissions tagged with the “Offers and Research” topic and applied our previous pushshift approach to crawl the data associated with each one of the remaining users, obtain information about all their other contributions on Reddit; this allowed us to discover the date and subreddit where they became active on the website (made their first contribution) as well as the other subreddits where they were involved.

Our dataset had contributions from 14.978 different authors, of which 12.533 still had active accounts on the site at the time of the study, we removed from the list of authors those that had made posts labeled with topic 15, since they were the least likely to be veterans, leaving us with 11.420 users whose personal pages we crawled in order to find when and where their first contribution on the site happened as well as all the other subreddits where they had interacted.

Figure 3.12 shows when users first joined the site and when their first contributions occurred, it is evident from the plot that users seem to be making a first contribution on the site shortly after creating an account, which is to be expected since an account is not needed to read unless the person wants to subscribe to the subreddit or vote.



**Fig. 3.12:** Dates of user milestones

Further analyzing this data we found out that users made their first post 49 ( $\sigma = 172$ ) days after joining the site and the first contribution on the *r/veterans* after 565 ( $\sigma = 654$ ) days. Table 3.5 contains the top 10 subreddits where users made

Subreddit	Users
Veterans	13.9%
AskReddit	9.9%
pics	2.9%
funny	2.4%
reddit.com	1.6%
IAmA	1.5%
politics	1.4%
WTF	1.3%
army	1.1%
gaming	1.0%

**Tab. 3.5:** Top-10 Subreddits where users made their first contribution after joining the site

Subreddit	Users
AskReddit	30.5%
funny	22.5%
pics	19.2%
aww	15.6%
Showertoughts	13.7%
Military	13.5%
gaming	12.3%
videos	11.9%
todayilearned	11.2%
news	11.1%

**Tab. 3.6:** Top-10 Subreddits where *r/veterans* users also posted

their first contribution on the site while Table 3.6 shows other subreddits where *r/veteran* posters also made submissions. Since only 14% of the users made their first contribution on *r/veterans* it is likely that the rest of them did not join the site to get help or to contribute on veteran subjects, but rather that they were already users of the site drawn initially to it for other reasons, which based on Tables 3.5 and 3.6 seems to be mostly for entertainment and general information.

It is also worth adding that for 73% of the users whose first contribution was on *r/veterans*, said contribution was a submission. This stands in opposition to the users whose first contribution was on another subreddit, since in this case only 32% were submissions, which might indicate that they did not join the site to solve a question or get help but rather to contribute on a subreddit where they were already readers. The activity as readers cannot be confirmed with the information that is publicly available but seems likely given the standard deviation of the time between creating an account and contributing to the community.

For the almost 14% of users whose first contribution was on *r/veterans* it seems likely that they created an account for this subreddit; verifying the posting dates

Subject	Subreddit	Total Posts	Valid Posts	Veteran mentions	Cross-posters from <i>r/veterans</i>
Military	AirForce	22.344	11.386	92 (0.4%)	647 (5.6%)
	Army	19.304	9.240	264 (1.4%)	1.153 (10.0%)
	Military	31.676	17.229	1.132 (3.6%)	1.430 (12.3%)
	Navy	20.796	12.541	229 (1.1%)	796 (6.9%)
	USMC	20.157	12.465	449 (2.2%)	1.022 (8.8%)
	<b>Veterans</b>	<b>18.836</b>	<b>13.829</b>	<b>4865 (35%)</b>	-
mHealth	Anxiety	23.511	12.800	27 (0.1%)	91 (0.8%)
	Bipolar	22.333	13.054	22 (0.1%)	79 (0.7%)
	Depression	27.906	11.990	31 (0.1%)	80 (0.7%)
	MentalHealth	19.934	13.452	34 (0.2%)	66 (0.6%)
	PTSD	13.848	9.211	416 (3%)	304 (2.6%)
	Rape	8.346	4.399	6 (0.1%)	8 (0.1%)
	RapeCounseling	12.549	7.566	14 (0.1%)	32 (0.3%)
	Relationships	27.265	9.159	12 (0%)	134 (1.2%)
	StopDrinking	23.279	12.228	26 (0.1%)	56 (0.5%)
	SuicideWatch	29.720	12.899	32 (0.1%)	62 (0.5%)

**Tab. 3.7:** Possible veteran presence and references on different subreddits

there is an increasing trend meaning that more people seem to be posting here as their introduction to Reddit.

### 3.3.5 Military and mHealth Subreddits

After analyzing all the data gathered in relation to *r/veterans* we proceeded to crawl some of the most relevant mental health, substance abuse, relationships and military subreddits in order to check if there were any explicit mentions or self-identification from veterans there.

We extracted all the submissions from each subreddit where a variation of the word *vet* or *veteran* was present, taking out those that could refer to “veterinarians”. The results can be seen on Table 3.7.

The only mental health subreddit with a significant number of mentions about veterans was *r/PTSD*. This is not surprising given that as many as 20% of Operations Iraqi Freedom and Enduring Freedom veterans, 12% of Gulf War veterans, and 15% of Vietnam veterans are thought to have PTSD [22] with many other illnesses and addictions diagnosed on the military being linked to it.

We also analyzed how many of the users considered possible veterans from *r/veterans* also posted or commented on the military and mHealth subreddits analyzed, the results obtained can be seen on the rightmost column of Table 3.7, which shows

that cross-posters were most common among military subreddits and that the most frequented mHealth subreddit was *r/ptsd*, which was to be expected. Of all the cross-posters only 1.553(13%) users made submissions and of those 233(2%) were on one of the ten mHealth subreddits analyzed.

To understand if there was a gap between users who posted about mental health issues on *r/veterans* and those who reached out to the specialized subreddits, we extracted the posts with explicit mentions of keywords related to depression, suicide, addiction, rape, PTSD, anxiety and bipolar disorder. The result were 1.265 posts (over 9% of the dataset) by 932 users (6% of the total), of which only 46 (0.4%) had also posted on an mHealth subreddit.

Given the small percentage of users from *r/veterans* participating on mental health subreddits compared to those explicitly discussing about that same topic in the subreddit, it can be inferred that the possible veterans are not as likely to seek assistance outside of their own group.

# Improvement of Psychotherapy Through Sensorization

## 4.1 Problem Analysis

As shown in section 3, reinsertion into civilian life is hard for military veterans and one of the most common challenges many have to overcome is Post-Traumatic Stress Disorder (PTSD).

Prolonged Exposure (PE) therapy has proven to be the best evidence-based treatment for PTSD [34, 31], but its use is still not widespread [29, 38] and faces challenges derived from the fact that psychotherapy is subjective and constrained to the perception of the physician delivering it. It can be ineffective for some patients that might leave the therapy if the way of approaching them is not corrected and it is time-consuming to train new therapists since they have no objective metrics to gauge their process and must do so by acquiring experience.

There are two components in Prolonged Exposure Therapy which aim to modify trauma-related fear structures in order to break the cycle of fear activation and avoidance [8]:

- **Imaginal exposure:** Patients revisit the memories of the traumatic event by recounting them aloud in vivid detail with the clinician and then, between sessions, listening to recordings of the most recent session's recounting.
- **In-vivo exposure:** Patients engage with real-world stimuli and situations that have become associated with the trauma memory and which the patient typically avoids to reduce the negative affect associated with the avoided situation

Intensive Outpatient Programs (IOP) provide PE to patients, who have to come to a facility for a 2-week period. On the first day patients are introduced to the PE therapy process, then they are assigned to a clinician and to a cohort along with other patients. During the next days patients are on a fixed daily schedule that

includes one-on-one PE therapy with the clinician, group therapy, and individual PE practice. These events are divided in the following phases:

- **Phase 1:** Individual therapy session during which the patient revisits the traumatic event through imaginal exposure with the clinician.
- **Phase 2:** Group therapy session where the patients meet with the rest of the cohort and a group leader, to do in-vivo exposures and discuss progress.
- **Phase 3:** It is developed outside the clinical setting. Patients listen to the recording of the morning's imaginal exposure and complete an in-vivo exposure exercises.

Although there is currently no sensorization at any point of the process to improve the perception of the clinician or to track progress, some mHealth applications like PE Coach [25] are being used by the patients, mostly for help with Phase 3.

#### 4.1.1 Requirements

Mental health treatments can be improved by enhancing the current therapies with data acquired during the session. This could be done in any of the three phases of the process, but since the patient has the full attention of the clinician during Phase 1 and there are already mHealth applications for Phase 3, we have decided to focus on Phase 2, the group therapy, where a group of patients interact and the clinician has to keep track of all of them.

As previously mentioned, there is currently no objective data being collected during group sessions to evaluate their quality or impact on the efficacy of the therapy process. To fill this gap we propose the implementation of the *Group Session Monitor*, an audio monitoring tool which should be able to:

- Differentiate between the participants of the session.
- Track the verbal behaviour of each patient. The number of interventions they make and the amount of time they speak, their speech's speed, the overall emotion associated with the interventions, among others.
- Monitor the interactions between participants.
- Detect disruptions during the session.



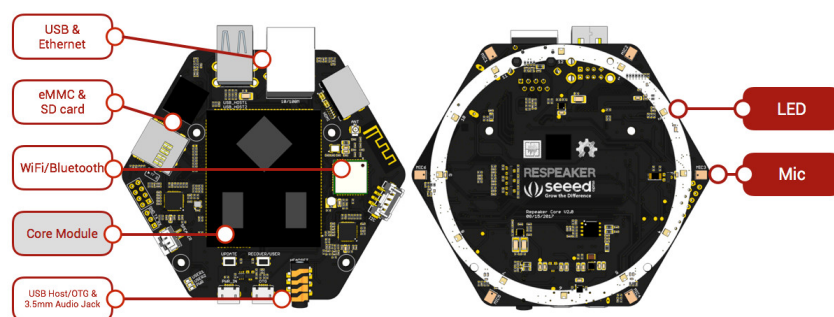
- Generate aggregated measures with the purpose of evaluating progress, generating improvements between sessions and detecting anomalies in the interaction between participants.

## 4.2 System Proposal

The *Group Session Monitor* has to be able to acquire an audio of the session and detect the portions corresponding to the different speakers in order to separate their audio tracks and to monitor their interactions. The process of splitting an audio file into segments corresponding to different speakers is known as *speaker diarization* [2, 42] and it is a complicated issue that is usually approached using clustering algorithms [44] or neural networks [43]. Nevertheless, the error rates associated with these methods would not be acceptable for our system, since they could negatively affect the therapy process and they would violate the privacy of users who did not consent to being recorded by keeping their session audio, albeit under another patients name.

To achieve speaker diarization we proposed the use of a microphone array as the audio-capture device. Since the group sessions were an organized type of meeting, with seating participants of known number, the microphone array could be installed in the ceiling in the center of the chair circle and it would detect the different places where the audio came from as different people speaking.

After investigating the available commercial options of microphone arrays we chose the ReSpeaker Core v2.0<sup>1</sup> which can be seen in Figure 4.1.



**Fig. 4.1:** ReSpeaker Core v2.0

The ReSpeaker Core v2.0 is a development board with a microphone array, Ethernet connectivity and a seemingly good library which provides filtered audio and

<sup>1</sup>[cnx-software.com/2018/03/13/respeaker-core-v2-is-a-6-mic-array-audio-development-kit-powered-by-rockchip-rk3229-processor](https://cnx-software.com/2018/03/13/respeaker-core-v2-is-a-6-mic-array-audio-development-kit-powered-by-rockchip-rk3229-processor)

DoA (*Direction of Arrival*), a metric corresponding to the direction of the incoming sound measured in degrees from the array's center. It also has the following specifications:

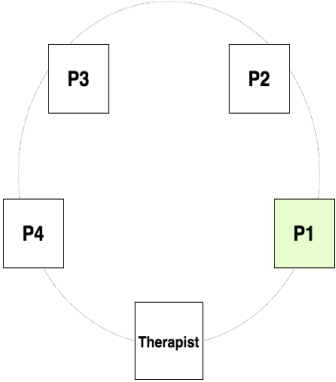
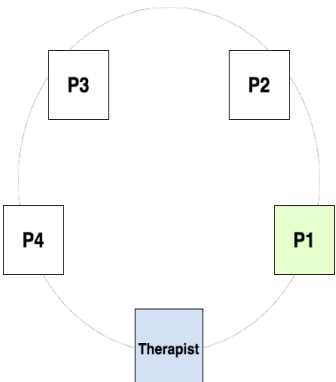
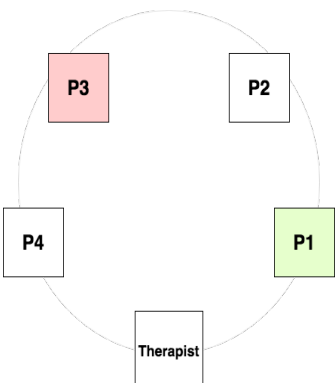
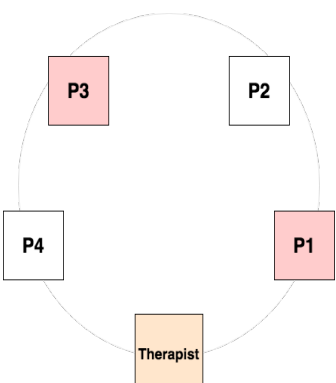
- SoC – Rockchip RK3229 quad core Cortex A7 processor @ up to 1.5 GHz with Arm Mali-400MP2 GPU.
- System Memory – 1GB DDR3 RAM.
- 6x Microphone Array with 5 meters detection range.
- 8 channel ADCs for 6 microphone array and 2 loopback (hardware loopback).
- 3.5mm audio jack.
- 10/100M Ethernet, WiFi 802.11b/g/n and BLE 4.0 (AP6212 module).
- 12 programmable LEDs ring, power LED, 2 user LEDs.
- Dimensions of 100x96x18 mm.

The type of events that had to be identified using the microphone array have been defined in Table 4.1.

After the audio had been captured and split into parts belonging to the different participants it had to be processed in order to extract useful information about the session, for this we intended to process the audio data using two different methods:

1. Audio analysis using Praat [4], an open-source program for speech analysis which would allow us to obtain for each audio segment its duration in seconds, the total number of syllables and pauses, the phonation time, speech rate and articulation rate.
2. Natural Language Processing (NLP) applied to the transcription of each audio segment to retrieve information such as the mood that could be associated with the interventions, its authenticity, usage of personal pronouns or different temporal conjugations, among others.

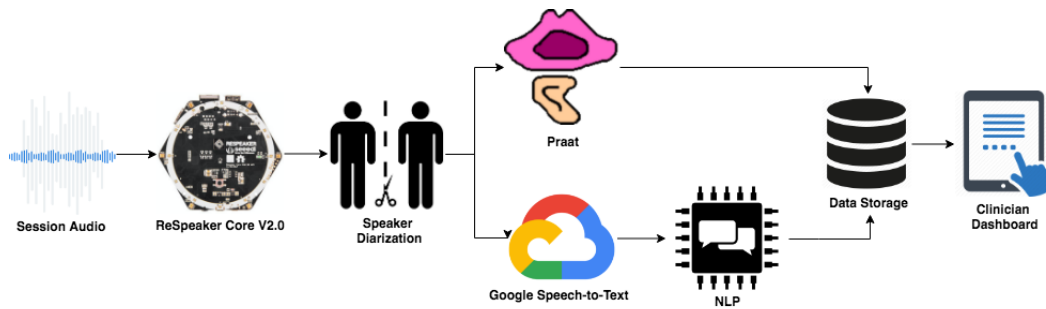
For the transcription of the different audio segments we proposed to use Google's Cloud Speech-to-Text. Since the clips would be longer than one minute, each

<p><b>Intervention</b></p> <ul style="list-style-type: none"> <li>• Monitor the time that the patient is speaking during the session.</li> <li>• Monitor patients' speech patterns for tone and speed changes, as well as emotion estimation. <ul style="list-style-type: none"> <li>– Aggregated measure for every intervention of the patient.</li> <li>– Thresholds can be set to detect uncommon speech patterns.</li> </ul> </li> </ul>	
<p><b>Interaction</b></p> <ul style="list-style-type: none"> <li>• Monitor the interaction between the different participants of the sessions to track behaviours triggered. <ul style="list-style-type: none"> <li>– Count by pairs.</li> </ul> </li> </ul>	
<p><b>Interruption</b></p> <ul style="list-style-type: none"> <li>• Detect when a participant is speaking and is cut off by another person. <ul style="list-style-type: none"> <li>– Count by occurrences per patient.</li> </ul> </li> </ul>	
<p><b>Disruption</b></p> <ul style="list-style-type: none"> <li>• Detect when there's more than one participant speaking at the same time. <ul style="list-style-type: none"> <li>– Can be considered a prolonged interruption, after passing over a certain time threshold.</li> </ul> </li> </ul>	

**Tab. 4.1:** Events detected by the *Group Session Monitor*

audio segment would need to be uploaded to the Google Cloud in order to apply asynchronous speech recognition to it. Once we had the transcript of each part of the session we could use tools such as LIWC [24] in order to extract the semantic content of it.

Finally all of this information would have to be stored in a database in order to be retrieved by the clinician dashboard, a tool that would allow the therapist to monitor the progress of each patient and to notice trends in the changes of behavior. The final system architecture can be seen on Figure 4.2.



**Fig. 4.2:** Proposed architecture for the *Group Session Monitor*

## 4.3 Development and Practical Issues

The different parts of the system started being developed and tested individually, in order to verify that all the modules would fit together, achieving their purpose while respecting the restrictions imposed by the sensitive nature of the data to be treated.

The audio analysis using Praat was successful and accurate enough to be useful, as were the transcript insights provided by LIWC. Nevertheless, the audio monitoring tool had to adhere to IRB (*Institutional Review Board*) regulations, this meant that the participants had to be informed of the recording and the information that would be extracted from it, so they could give explicit consent. If a patient did not agree, we would have to avoid recording them in order to protect their privacy, this was another reason why the microphone array approach seemed like a better option than the machine learning methods, since the audio coming from the direction of users who did not consent could be discarded upon detection.

After testing it, we realized that the DoA provided by the ReSpeaker Core v2.0 was not accurate enough to do this, even with corrections, tuning and under ideal noise conditions there were jumps in the detected direction of the sound that would have

made it impossible to be completely certain that the wrong speaker was not being recorded. This issue could have been solved with a hybrid approach of DoA+ML where both the direction of the incoming sound and its features would be used to classify to which user it belonged to, but it would have meant to record all the users and it would have likely not gotten IRB approval.

Another potential issue was with the audio transcription. Having to upload the sessions to the cloud in order retrieve the transcription was considered riskier than doing offline processing, but all the offline transcription tools tested were much worse and the results were not accurate enough to extract good linguistic features from them.

After taking everything into consideration and discussing it with the experts, the *Group Session Monitor* was deemed too invasive and it was decided not to continue with its development.

# Conclusions

## 5.1 Exploratory Analysis

Our aim was to understand the needs of veterans via the data-driven analysis of *r/veterans*, as opposed to qualitative methods [9, 7, 37]. We have three main contributions. First, we show that the top 10 topics highlight the issues that are known to be important to veterans offline. This includes the need to share their experiences in the military (e.g., topics related to Stories/Thoughts and Rants), the need to navigate military programs (e.g., topics on Education, Claiming Disability, Benefit Questions, Discharge/Deployments), and the need to adjust to civilian life<sup>1</sup> (topics on Housing/Family, Job Search, Opportunities and Programs). The other strand is issues related to health, physical or mental [22, 40, 27].

The second set of findings highlight how *r/veterans* has changed over the past decade. There is now less publicity and more organic content which might be due to an improvement in moderation as well as the growth of Reddit. We provide evidence that as the site became more interactive (as noted by more questions being posed) it also turned more controversial with a rising upvote ratio. The shift toward more organic content was reflected in the increase of first person versus third person language; and in the trend toward more "Authentic" and less "Analytic" language.

Our third contribution is in establishing that this community has become the de facto place for veterans (or those interested in finding out about veteran issues) to come. We found that around 12% of the users seem to have joined Reddit in order to get help or information from *r/veterans*, with the rest likely being existing users from the site who also happened to be veterans. This suggests that this subreddit is not a primary resource used by the general public.

The implications of our findings may be of interest to those providing services for military veterans and their families. We saw that an increasing number of posts contain questions and first person speech. The fact that Reddit has become an alternative to other private forums (e.g., the ones promoted by the VA) may suggest that the VA could include moderators that answer VA related questions on Reddit.

---

<sup>1</sup>[mentalhealth.va.gov/communityproviders/docs/readjustment.pdf](https://mentalhealth.va.gov/communityproviders/docs/readjustment.pdf) (2019)

This is especially important since we know that many of the topics were specifically about physical/mental health and benefits.

In conclusion, this study provides quantitative insights about how veterans (and related users) behave on Reddit. It also provides implications for ways in which this community can be supported.

## 5.2 Therapy Improvement

After analyzing the current state of psychological therapy and specifically of PTSD therapy we noticed that some level of sensorization is needed in order to extract important information from the sessions, which can be leveraged in order to generate improvements on the delivery made by clinicians, to detect which techniques work better with each patient and to monitor their progress in an objective manner.

We also explained that Prolonged Exposure (PE) therapy is currently the most effective type of treatment for PTSD and that it is formed by three phases, the second one of which involves a group session with other patients and a therapist. We decided to focus on improving these group sessions with a proposed tool by the name of *Group Session Monitor*.

The individual parts of the *Group Session Monitor* worked well in isolation. We were able to get information from the audio tracks, transcribe them and extract linguistic content from them. The biggest issue with the system occurred in the speaker diarization using a microphone array; it was concluded that the diarization was not accurate enough to comply with IRB regulations and the development of the audio-monitor was cancelled.

## 5.3 Future Work

Regarding the exploratory analysis of veterans there are a number of areas for future investigation. One shortcoming of our study is that we did not distinguish between veterans and other stakeholders (e.g., family members). Doing so might provide insights into their specific needs. We established the increase in questions and answers, but we did not establish the quality of the answers that are being provided. This may in fact be more easily done through qualitative means [14]. Finally, our data shows that users who contribute on *r/veterans* do not seem to reach out for help outside of the military subreddits. This is expected since they probably want feedback from people in a similar situation, as found by Dosono et al. [7].

However, this is a potentially isolating behavior that limits the help they can get. One improvement would be to provide suggestions for subreddits most frequently associated with the same keywords that a user is employing when making a new post on *r/veterans*.

Regarding the improvement of psychological therapy, in order to improve the process sensorization would be helpful, but it must be done in such a way that complies with regulations and respects patients' privacy. The *Group Session Monitor* could work better if machine learning based diarization was combined with the microphone array information, this could drive the error rates down enough for the system to be considered acceptable; the audio-tracks corresponding to unwilling participants could be deleted immediately after isolating them from the rest but this would have to be approved.

Regarding the transcription of the sessions, it was suggested that it would be more acceptable for it to be done by a human instead of using an external transcription service since they would require online access of the information. If offline transcription tools were to improve in the future they could be used instead and it would increase the level of acceptance of the system.



# Bibliography

- [1] Allard J. van Altena, Perry D. Moerland, Aeilko H. Zwinderman, and Sílvia D. Olabarriaga. “Understanding big data themes from scientific biomedical literature through topic modeling”. In: *Journal of Big Data* 3.1 (Dec. 2016), p. 23. DOI: 10.1186/s40537-016-0057-0.
- [2] Xavier Anguera, Simon Bozonnet, Nicholas Evans, et al. “Speaker diarization: A review of recent research”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 356–370.
- [3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”. In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022.
- [4] Paul Boersma and David Weenink. “PRAAT, a system for doing phonetics by computer”. In: *Glott international* 5 (Jan. 2001), pp. 341–345.
- [5] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. “Reading Tea Leaves: How Humans Interpret Topic Models”. In: *Proceedings of the 22Nd International Conference on Neural Information Processing Systems. NIPS’09*. Vancouver, British Columbia, Canada: Curran Associates Inc., 2009, pp. 288–296.
- [6] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L. Pappa. “The Effect of Social Feedback in a Reddit Weight Loss Community”. In: *Proceedings of the 6th International Conference on Digital Health Conference - DH ’16*. New York, New York, USA: ACM Press, 2016, pp. 99–103. DOI: 10.1145/2896338.2897732.
- [7] Bryan Dosono, Yasmeen Rashidi, Taslima Akter, Bryan Semaan, and Apu Kapadia. “Challenges in transitioning from civil to military culture: Hyper-selective disclosure through ICTs”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), p. 41.
- [8] Edna B. Foa, Elizabeth Hembree, and Barbara Rothbaum. *Prolonged Exposure Therapy for PTSD: Emotional Processing of Traumatic Experiences, Therapist Guide*. New York, NY, Jan. 2015.
- [9] Pedro Gamito, Jorge Oliveira, Pedro Rosa, et al. “PTSD elderly war veterans: A clinical controlled pilot study”. In: *Cyberpsychology, Behavior, and Social Networking* 13.1 (2010), pp. 43–48.
- [10] Suparna Goswami, Felix Köbler, Jan Marco Leimeister, and Helmut Krcmar. “Using Online Social Networking to Enhance Social Connectedness and Social Support for the Elderly”. In: (June 2010).
- [11] Reilly Grant, David Kucher, Ana M. Leon, Jonathan Gemmell, and Daniela Raicu. “Discovery of Informal Topics from Post Traumatic Stress Disorder Forums”. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Nov. 2017, pp. 452–461. DOI: 10.1109/ICDMW.2017.65.

- [12] David Hall, Daniel Jurafsky, and Christopher D. Manning. “Studying the History of Ideas Using Topic Models”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Honolulu, Hawaii: Association for Computational Linguistics, 2008, pp. 363–371.
- [13] John D Hixson, Deborah Barnes, Karen Parko, et al. “Patients optimizing epilepsy management via an online community: the POEM Study.” In: *Neurology* 85.2 (July 2015), pp. 129–36. DOI: 10.1212/WNL.0000000000001728.
- [14] Hwajung Hong, Eric Gilbert, Gregory D. Abowd, and Rosa I. Arriaga. “In-group Questions and Out-group Answers”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. New York, New York, USA: ACM Press, 2015, pp. 777–786. DOI: 10.1145/2702123.2702402.
- [15] Clayton J. Hutto and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *ICWSM*. 2014.
- [16] Eric Kuhn, Jill J Crowley, Julia E Hoffman, et al. “Clinician characteristics and perceptions related to use of the PE (prolonged exposure) coach mobile app.” In: *Professional Psychology: Research and Practice* 46.6 (2015), p. 437.
- [17] Andrew Kachites McCallum. “MALLET: A Machine Learning for Language Toolkit”. <http://mallet.cs.umass.edu>. 2002.
- [18] Carmen P McLean and Edna B Foa. “Prolonged exposure therapy for post-traumatic stress disorder: a review of evidence and dissemination”. In: *Expert review of neurotherapeutics* 11.8 (2011), pp. 1151–1163.
- [19] Neil Mehta and Ashish Atreja. “Online social support networks”. In: *International Review of Psychiatry* 27.2 (Mar. 2015), pp. 118–123. DOI: 10.3109/09540261.2015.1015504.
- [20] Hemant Misra, Olivier Cappe, and Francois Yvon. “Using LDA to detect semantically incoherent documents”. In: *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Manchester, England: Coling 2008 Organizing Committee, 2008, pp. 41–48.
- [21] David Mohr, Mi Zhang, and Stephen Schueller. “Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning”. In: *Annual review of clinical psychology* 13 (Mar. 2017). DOI: 10.1146/annurev-clinpsy-032816-044949.
- [22] National Center for PTSD. U.S. Department of Veterans Affairs, “How Common is PTSD in Veterans? - PTSD: National Center for PTSD”. Feb. 2019.
- [23] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. “Evaluating topic models for digital libraries”. In: *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*. New York, New York, USA: ACM Press, 2010, p. 215. DOI: 10.1145/1816123.1816156.
- [24] James W Pennebaker, Martha E Francis, and Roger J Booth. “Linguistic inquiry and word count: LIWC 2001”. In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.
- [25] GM Reger, J Hoffman, D Riggs, et al. “The PE coach smartphone application: an innovative approach to improving implementation, fidelity, and homework adherence during prolonged exposure.” In: *Psychological services* 10.3 (2013), pp. 342–349.
- [26] Greg M Reger, Nancy A Skopp, Amanda Edwards-Stewart, and Eder L Lemus. “Comparison of prolonged exposure (PE) coach to treatment as usual: A case series with two active duty soldiers.” In: *Military Psychology* 27.5 (2015), p. 287.

- [27] Miriam Reisman. "PTSD Treatment for Veterans: What's Working, What's New, and What's Next." In: *P & T : a peer-reviewed journal for formulary management* 41.10 (Oct. 2016), pp. 623–634.
- [28] Carolina Rodriguez-Paras, Kathryn Tippey, Elaine Brown, et al. "Posttraumatic Stress Disorder and Mobile Health: App Investigation and Scoping Literature Review". English (US). In: *JMIR mHealth and uHealth* 5.10 (Oct. 2017), e156. DOI: 10.2196/mhealth.7318.
- [29] C. S. Rosen, M. M. Matthieu, S. Wiltsey Stirman, et al. "A Review of Studies on the System-Wide Implementation of Evidence-Based Psychotherapies for Posttraumatic Stress Disorder in the Veterans Health Administration". In: *Administration and Policy in Mental Health and Mental Health Services Research* 43.6 (Nov. 2016), pp. 957–977. DOI: 10.1007/s10488-016-0755-0.
- [30] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. "Evaluating topic coherence measures". In: *CoRR* abs/1403.6397 (2014).
- [31] Barbara Olasov Rothbaum and Ann C. Schwartz. "Exposure Therapy for Posttraumatic Stress Disorder". In: *American Journal of Psychotherapy* 56.1 (Jan. 2002), pp. 59–75. DOI: 10.1176/appi.psychotherapy.2002.56.1.59.
- [32] Mattia Samory and Tanushree Mitra. "Conspiracies Online: User Discussions in a Conspiracy Community Following Dramatic Events". In: *undefined* (2018).
- [33] Samiha Samrose, Ru Zhao, Jeffery White, et al. "CoCo: Collaboration Coach for Understanding Team Dynamics during Video Conferencing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (2018), p. 160.
- [34] Ulrich Schnyder, Anke Ehlers, Thomas Elbert, et al. "Psychotherapies for PTSD: what do they have in common?" In: *European journal of psychotraumatology* 6 (2015), p. 28186. DOI: 10.3402/ejpt.v6.28186.
- [35] Elizabeth M Seabrook, Margaret L Kern, and Nikki S Rickard. "Social Networking Sites, Depression, and Anxiety: A Systematic Review." In: *JMIR mental health* 3.4 (Nov. 2016), e50. DOI: 10.2196/mental.5842.
- [36] Bryan C. Semaan, Lauren M. Britton, and Bryan Dosono. "Transition Resilience with ICTs: 'Identity Awareness' in Veteran Re-Integration". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: ACM, 2016, pp. 2882–2894. DOI: 10.1145/2858036.2858109.
- [37] Bryan Semaan, Lauren M. Britton, and Bryan Dosono. "Military Masculinity and the Travails of Transitioning: Disclosure in Social Media". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: ACM, 2017, pp. 387–403. DOI: 10.1145/2998181.2998221.
- [38] Brian Shiner, Leonard W. D'Avolio, Thien M. Nguyen, et al. "Measuring Use of Evidence Based Psychotherapy for Posttraumatic Stress Disorder". In: *Administration and Policy in Mental Health and Mental Health Services Research* 40.4 (July 2013), pp. 311–318. DOI: 10.1007/s10488-012-0421-0.
- [39] Jafar Tanha, Jesse de Does, and Katrien Depuydt. "An LDA-based Topic Selection Approach to Language Model Adaptation for Handwritten Text Recognition". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, 2015, pp. 646–653.
- [40] Terri Tanielian, Lisa Jaycox, Terry Schell, et al. *Invisible Wounds: Mental Health and Cognitive Care Needs of America's Returning Veterans*. RAND Corporation, 2008. DOI: 10.7249/RB9336.

- [41] Toulis Andrew and Golab Lukasz. “Social Media Mining to Understand Public Mental Health”. In: *Data Management and Analytics for Medicine and Healthcare*. Ed. by Wang Fusheng Begoli Edmon and Luo Gang. Munich, Germany: Springer International Publishing, 2017, pp. 55–70.
- [42] Sue E Tranter and Douglas A Reynolds. “An overview of automatic speaker diarization systems”. In: *IEEE Transactions on audio, speech, and language processing* 14.5 (2006), pp. 1557–1565.
- [43] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno. “Speaker diarization with lstm”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5239–5243.
- [44] Chuck Wooters and Marijn Huijbregts. “The ICSI RT07s speaker diarization system”. In: *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 509–519.