ALJI: Active Listening Journal Interaction

Patrick R. Sullivan

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Bert Huang, Chair
Lee D Cooper
Tanushree Mitra

26 September 2019
Blacksburg, Virginia

Keywords: Machine Learning, Expressive Writing, Computational Social Science

ALJI: Active Listening Journal Interaction

Patrick R. Sullivan

(ABSTRACT)

Depression is a crippling burden on a great many people, and it is often well hidden. Mental health professionals are able to treat depression, but the general public is not well versed in recognizing depression symptoms or assessing their own mental health. Active Listening Journal Interaction (ALJI) is a computer program that seeks to identify and refer people suffering with depression to mental health support services. It does this through analyzing personal journal entries using machine learning, and then privately responding to the author with proper guidance. In this thesis, we focus on determining the feasibility and usefulness of the machine learning models that drive ALJI. With heavy data limitations, we cautiously report that with a single journal entry, our model detects when a person's symptoms warrant professional intervention with a 61% accuracy. A great amount of discussion on the proposed solution, methods, results, and future directions of ALJI is included.

ALJI: Active Listening Journal Interaction

Patrick R. Sullivan

(GENERAL AUDIENCE ABSTRACT)

An incredibly large number of people suffer from depression, and they can rightfully feel trapped or imprisoned by this illness. A very simple way to understand depression is to first imagine looking at the most beautiful sunset you've ever seen, and then imagine feeling absolutely nothing while looking that same sunset, and you can't explain why. When a person is depressed, they are likely to feel like a burden to those around them. This causes them to avoid social gathering and friends, making them isolated away from people that could support them. This worsens their depression and a terrible cycle begins. One of the best ways out of this cycle is to reveal the depression to a doctor or psychologist, and to ask them for guidance. However, many people don't see or realize this excellent option is open to them, and will continue to suffer with depression for far longer than needed.

This thesis describes an idea called the Active Listening Journal Interaction, or *ALJI*. ALJI acts just like someone's personal journal or diary, but it also has some protections from illnesses like depression. First, ALJI searches a journal entry for indicators about the author's health, then ALJI asks the author a few questions to better understand the author, and finally ALJI gives that author information and guidance on improving their health. We are starting to create a computer program of ALJI by first building and testing the detector for the author's health. Instead of making the detector directly, we show the computer some examples of the health indicators from journals we know very well, and then let the computer focus on finding the pattern that would reveal those health indicators from any journal. This is called machine learning, and in our case, ALJI's machine learning is going to be difficult because we have very few example journals where we know all of the health indicators. However, we believe that fixing this issue would solve the first step of ALJI. The end of this thesis also discusses the next steps going forward with ALJI.

Dedication

In memory of Taylor Rydahl.
You have changed me.
You are missed.

Acknowledgments

I would like to acknowledge my generous and patient advisors, who withstood my overly-ambitious ideas and guided me towards the realm of real and possible positive impacts.

An extra thanks is due to Alyssa Gatto of the Psychology Department for both sharing in my motivation and not letting it take me astray.

And a great final thanks to my parents, who undoubtedly introduced me to the virtues of kindness and determination.

Contents

List of Figures

List of Tables

List of Abbreviations

ALJI Active Listening Journal Interaction

ALR "Ask, Listen, Refer"

CGI-S Clinical Global Impressions - Severity scale

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

EWT Expressive Writing Therapy

LIWC Linguistic Inquiry and Word Count

ML Machine Learning

NLP Natural Language Processing

PTSD Post-Traumatic Stress Disorder

QPR "Question, Persuade, Refer"

Chapter 1

Introduction

Depression is ranked as the single largest contributor to global disability, according to the World Health Organization. 322 million people suffered from depression in 2015, showing an increase of 18.4% over 10 years [24]. Its symptoms are fundamentally linked to suicide, the second leading cause of death in 15–29-year-olds [25]. There is no single cause of depression, which complicates both the identification and treatment of people who suffer. It is particularly damaging for people who are disconnected from social support and experiencing unhealthy thinking patterns. But these people with higher risks can be very capable of writing out their thoughts.

Personal journals (or diary entries) have the potential to give extraordinary insight on the author's well-being. Journals can be used to reflect on recent events, organize ideas, experience emotions, and design future actions. Starting as just blank paper and a pen, we recognize the great accessibility and freedom of expression that is available to the author. Over a long period of time, they even showcase the personal growth and changing values of the author [1]. However, there are additional considerations to be made when focusing on completely private journals. The author will receive no feedback or guidance when they are in a crisis, and they may not even understand the severity of their circumstances. The increased privacy could also impact the writing style, language, and topics contained in the journal [6]. By having no outside audience, we expect private journal to have fewer incentives to misrepresent themselves than in social media posts and other writing mediums.

We believe that mental healthcare professionals can estimate the author's mental wellness from a journal entry. But reading every personal journal entry could be an unproductive use of the counselor's time and raises costs to the patient. Instead, this task may be suited for software that can analyze the text of a journal and detect concerning mental health indicators. A machine learning model could learn to detect these indicators based off of journal examples that are provided by the counselors. Successful models are then able to identify the mental health indicators on later journals that are not seen by a counselor.

The overall goal of ALJI is to leverage personal journals to promote the well-being of the author. However, a much more precise case is necessary in order to test and validate the feasibility of this concept. Therefore, we focus on creating a machine learning model capable of analyzing the personal journal entry of an adolescent student and then predicting their overall mental health as well as any symptoms of depression. From this prediction, we can inform and guide the author towards mental health support services.

Chapter 2

Related Work

2.1 Mental Health Assessment and Intervention

Monitoring mental health has been a core component of modern clinical settings. The Diagnostic and Statistical Manual of Mental Disorders (DSM) is an evolving diagnostic tool that has become the mental healthcare professional's standard for psychiatric diagnosis in the USA [2].

The Clinical Global Impressions-Severity (CGI-S) scale was developed by mental healthcare professionals as a simple measure for the severity of a patient's symptoms [13], and has been used as a measurement tool for cases of depression in the past [14]. It has since been evaluated for its value in clinical settings with positive results [5]. It is a simple scale from 1-7, where higher ratings signify more severe symptoms and negative effects on the patient's health.

Forming the initial response to someone's crisis or mental illness is also critically important. Today's counselors and psychotherapists continually develop the best response and treatment of a person at risk of suicide [34]. The key role that the general public can play in this scenario is encourage people at high risk towards qualified mental health support services. In communities of young students, suicide prevention training programs such as "Ask, Listen, Refer" have been implemented to increase public awareness of risk factors and to give instruction on how to refer people to professional care [16]. Another variant, "Question, Persuade, Refer" (QPR), was founded for the purpose of empowering trainees to be active, rather than passive, when encouraging someone else towards support [22, 31]. While no study today has been able to directly measure the number of prevented suicides through QPR training, approximately 90% of completed suicides are people lacking treatment for mental health disorders (including depression) [9], and it is widely agreed that proper treatment would save lives [23].

2.2 Expressive Writing Therapy

Emotional confession and disclosure have been an interest for understanding the psychological healing process [28]. This has been standardized into expressive writing therapy (EWT),

a writing task that has been studied in a multitude of settings and populations with results that vary [29]. EWT's resemblance to a personal journal is striking due to the expressive freedom and introspective writing focus. Groups with depressive symptoms who engage in EWT report decreased depressive symptoms, and increased life satisfaction [11]. While EWT does not necessarily reduce intrusive thoughts, it can moderate the impact of intrusive thoughts on depressive symptoms [17]. The reduction of depressive symptoms has also been reported in college students after a time delay [10].

2.3 Natural Language Processing

Natural language processing (NLP) is the transformation of normal human language into measures that can be directly used by computers. It has commonly been used to translate one human language to another, or to interpret a user's query in a valid format for a computation. The linguistic characteristics a person uses (verbal or written) can even give insight on their psychological state [12] and on their feelings towards a topic [41]. "Linguistic Inquiry and Word Count" (LIWC) is an example of a NLP tool purposed for quantifying text along 93 different linguistic and psychological measures, such as "positive emotion" and "certainty" [30, 36]. For college students writing essays, some of the LIWC measures correlated to whether they were currently depressed, formerly depressed, or had never been depressed [33].

2.4 Machine Learning

Machine learning (ML) is a problem solving method of computers to complete a complex task without relying on specific instructions. This typically requires input data that has some relationship (however complex) to the solution of a problem. The relationship between the input data and the solution can be computed or inferred through many different approaches (The models we will use are detailed in Section 4.4.1).

Useful data for understanding a person's health is widely available on a person's computer, and collecting this data can be done without interfering with the device's normal use [18]. Machine learning has been utilized on the data of a person's smartwatch to identify health-related behaviors with 93% accuracy, an improvement over smartphone data yielding 77% accuracy [40]. This serves as some evidence that the more connected a source of data is to the target (the problem's solution), the more accurate predictions will be given by the machine learning model. Another study produced machine learning models that use a person's Twitter usage and post content to predict the *future onset* of depression for those users with 70% accuracy [7]. These examples show the usefulness (and some limitations) of using machine learning for the purposes of understanding and improving mental healthcare.

Chapter 3

Proposed Solution

It is beneficial to establish an ideal solution as a reference when making the design decisions for ALJI. The enables us to prioritize the key aspects over minor concerns and anticipate conflicts that would diminish our possible impact. ALJI is meant to be an alternative (or addition) to a personal journal. At all times, the author needs to remain aware that ALJI is not a replacement or an alternative to a counselor or psychologist.

3.1 Human-Centered Design

ALJI must be created in a way that doesn't sacrifice the benefits that EWT provides (see Section 2.2). For this reason, we should match the natural journaling task and environment of an author as closely as possible. Journal authors expect complete control over the privacy and accessibility of their journal. ALJI needs to win an author's trust by being completely transparent and under the control of the author.

The best way to ensure this is to create ALJI as a standalone, normal computer program without the capacity to communicate with any other device. At any time, authors can erase all traces of ALJI and their writings in the same way they can burn a journal. An online service would allow journal entry communications to be intercepted, and the service itself can be a target for data breaches that are commonplace on the internet. Relying on an online service would not be acceptable for journals that need to never be copied, accessed, or shared without the author's permission. Authors also need to see the entire inner workings of the ALJI program and verify our claims of privacy, which leads us to make ALJI as open source software and prefer open source dependencies. The entire history and current code for running ALJI is publicly accessible at this GitHub repository: `https://github.com/sublime09/ALJI`.

If ALJI were to fail in these requirements, then authors would understandably lose trust in ALJI. A breach of trust fundamentally changes how the author would write journal entries, viewing them more as social media posts or an interview environment where their character is being openly judged. This change in environment can remove many benefits of EWT and negatively impact the power and accuracy of ALJI's predictions of the author's mental wellness.

3.2 Interaction Sequence

Figure 3.1 shows the basic interaction sequence of the ALJI. ALJI is composed of three modules that interact with the author (or user): the Machine Learning Module, the Crisis Criteria Module, and the Referral Module.

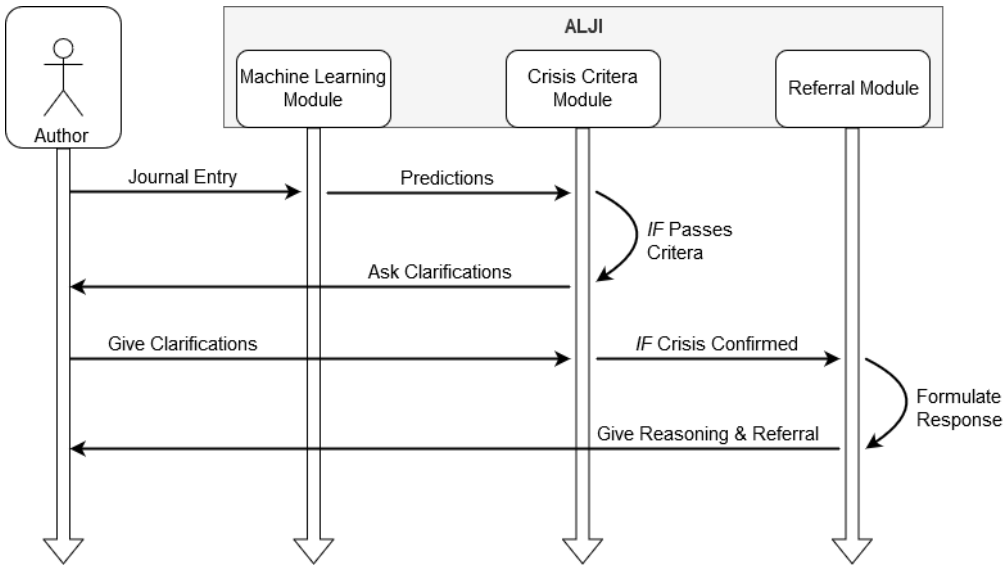


Figure 3.1: Basic interaction sequence between a journal author and ALJI

3.2.1 Machine Learning Module

This module in fig. 3.1 contains the machine learning model that is best suited for quantifying and predicting the author's mental health. This model will accept the text of a journal entry, standardize the text (e.g. segmenting, parts-of-speech tagging, stemming) to an acceptable form, analyze the text for basic psychological markers, and then output mental health predictions. Note that the machine learning model is already trained at this step.

The kinds of predictions that are made depends on what the model is trained to predict: tiredness, weight changes, substance abuse, insomnia, catastrophic thinking, thoughts of death, self-harm, suicidal ideation, suicide planning, or any metric that is deemed important towards understanding the mental health of someone. Uncertainty around a mental health indicator needs further investigation, and so that mental health indicator should be included in the predictions. However, these predictions represent the *possibility* of mental health indicators, and should not be taken as truth.

3.2.2 Crisis Criteria Module

This module from Figure 3.1 assists in confirming the accuracy (or inaccuracy) of the Machine Learning Module. Meeting the "crisis criteria" of this module is done though a single question: if the previous predictions were taken as truth, would this justify an intervention on behalf of the author? It is vitally important that mental health professionals determine this criteria, not software designers or the general public. If the criteria are not matched, then ALJI has no further work, since all mental health indicators *possibly* detected did not warrant an intervention.

Meeting the crisis criteria means ALJI must reduce the uncertainty around the author's mental health before proceeding. Human-computer miscommunication is incredibly common, especially when the text input from journals allows for the author's expressive freedom. ALJI gives direct yes/no questions to the author, asking if they are currently experiencing the mental health indicators that were predicted. It may be possible to quote meaningful passages in the author's writings as a reference for why ALJI is asking these health questions. These clarification questions are a defense against misunderstandings due to sarcasm, humor, hyperbole, non-literal language, and fictional events in the journal entry. The author's response to these yes/no questions is taken as absolute truth, and compared against the same crisis criteria. If the criteria are matched again, then the Referral Module is triggered to intervene and guide the author towards mental health support services.

No software is perfect, but in the context of ALJI, we prefer certain errors over others. Consider a false positive prediction incoming from the Machine Learning Module: the author is healthy, but the prediction raises concerns. This can easily be corrected by the clarifications in the Crisis Criteria Module. Even if it was not, the author is given a referral to mental health professionals that are capable of understanding the error. A much worse scenario happens when a false negative error in prediction occurs: an author that needs guidance towards support, but it is not given to them. Thankfully, many models are able to leverage a trade-off between a critical error and a non-critical error by assigning weights to certain samples [3, 4]. A crude version of this could be achieved by simply duplicating samples that we wish to have more weight within the Machine Learning Module.

3.2.3 Referral Module

This module in Figure 3.1 provides the author with the intervention and guidance towards appropriate mental health support services. It would be most persuasive if this guidance presented reasoning or new information to the author about the concern. Which support services it shows depends on the mental health indicators that are confirmed by the author. For example, indicators for consistent depressed mood and recent loss of interest could result in a response of what depression is, how treatable it is, and how to make an appointment at a counseling service nearby. Some mental health concerns require an immediate response,

so every step of ALJI must be swift.

An appropriate response to suicide planning and the tools for suicide would be providing the contact information of a 24-hour national suicide prevention hotline and the medical emergency telephone number. It is tempting to design ALJI so the most serious mental health concerns are immediately and automatically report to emergency services, with no input from the author. However, this would break the author's trust of privacy and non-judgement, leading to the same consequences outlined in Section 3.1 and lessening those positive impacts of ALJI. It also ignores how receptive the author would be of outside aid at that moment, or how they may need a moment to decide for themselves. Forcing aid upon someone who does not want it can have absolutely dire consequences. Designers of ALJI should always be aware that ALJI is an alternative to pen-and-paper journals, and that journal authors could easily switch to a writing medium that respects privacy. The more that ALJI is removed from the environment of a private journal, the fewer journal authors will consider using ALJI.

Chapter 4

Methods

Implementing the full proposed solution of ALJI requires a massive amount of effort and validation. Chapters 4 to 6 will focus on the investigating the feasibility of the Machine Learning Module detailed in Section 3.2.1. Without the Machine Learning Module, ALJI would require a radical redesign, so it is prudent to prioritize this component.

4.1 Gathering Journal Data Source

The unique writing environment of a private personal journal is less understood due to its invisible nature. This means we cannot expect the language and structure of other writing mediums (e.g. news articles, social media, books) to be present in private personal journals. Collecting a dataset of private personal journals will be difficult, but will be invaluable for understanding the writings and what they mean in terms of mental health. This dataset will be the foundation for the Machine Learning Module proposed in Section 3.2.1 and is necessary for the machine learning training process.

We collected data from the "Personal Essay & Memoir" category of the Scholastic Awards website on February 4, 2019 to use as journal entries (`https://www.artandwriting.org/explore/online-galleries`). For our purpose here, we will refer to these documents as the journals (see Section 6.1 for discussion on the validity and biases of this data). They were written between 2010 and 2018 by 12^{th} grade high school students in the USA, responding to the following prompt:

> A non-fiction work based on opinion, experience, and/or emotion that explores a topic or event of importance to the author ... Essays in which humor is the key element should be submitted to the humor category ... Limits: 500–3,000 words

There are 531 entries fitting our selection criteria on the website. To gather them into a usable format, we write a Python script to control Selenium, a web browser automation tool [39], BeautifulSoup, a parser helper [32], and "lxml", a parser for HTML [20]. Selenium is necessary due to the webpage's dynamic loading of the journals through JavaScript, which isn't handled by many webpage scraping tools. Parsing is quite necessary since journals had different formats when submitted which caused them to be located in slightly different

locations in the HTML structure. Details can be seen in the "Scholastic Pull" folder of ALJI's code repository.

4.2 Measuring Journal Language

To have a basic knowledge of the journal data set, we can begin with finding some basic language metrics and emotion markers. We use Python's TextBlob library to discard punctuation and select every word in each journal [19]. We then estimate the part of sentence they belong to (e.g. noun, adverb) and lemmatize each word to their standard form (e.g. "driving" becomes "drive"). This puts words into their most basic form, making them more recognizable to later textual analysis tools. We also are able to segment a journal into sentences, which may aid us in finding the specific language markers that foreshadow an author's crisis.

Some emotional markers of this text could be produced by using LIWC (see Section 2.4), however the commercial license and proprietary nature of surrounding LIWC could compromise several goals of ALJI presented in Section 3.1. It is unclear to this researcher on the legality of using the LIWC software, or the dictionaries involved when ALJI could reach beyond an academic context. Distributing a full ALJI prototype that includes LIWC's software or dictionaries poses additional questions as well. A suitable alternative is found in Empath, an open-source lexical analyzer of text that has been evaluated with results comparable to LIWC [8]. Empath produces 194 quantifiable metrics, a number of which have clear psychological ties. Refer to Section 5.1 to view the primary results of this analysis.

4.3 Journal Labeling Participation

Arguably the most important information to us is the perspective and impression that mental health professionals have of these journals. Here we will describe the process of gaining these impressions, and the revision that was required. Mental healthcare professionals, participants, and experts all represent the same people, so the terms are interchangeable, but the most applicable term is used based on context.

There are many people that have formal education and experience in mental healthcare. For our purposes, we consider those with some graduate level clinical psychology knowledge to be uniquely qualified to analyze these journals for mental health indicators. We recruit these people though email mailing lists that have many members of this description. We advertise and ask for voluntary participation in our study, giving a description of ALJI's goals and of their role. The following is the primary advertisement communication:

> Help is needed to identify, inform, and refer young people with depression to men-

> tal health support services. Academics and professionals in mental health care can evaluate personal journal entries for mental health indicators of the author. Machine learning is then used to provide an immediate and widely accessible response to those suffering on making steps towards recovery. Contributing to this research is expected to take under 20 minutes of your time, and it can be completed anywhere. This research is being conducted by Patrick Sullivan under direction from Dr. Lee Cooper. To participate, contact Patrick through email: sublime@vt.edu

Participants were informed that their participation confidentiality and ability to withdraw at any time without penalty. It was vitally important to inform them of the risks that the journal could pose to them, so we give clear warning that the journals could describe graphic and harmful content including: sexual abuse, domestic violence, substance abuse and hate crimes. Their health is always a higher priority than the ALJI project.

Labeling Task

The mental healthcare professionals are providing material for the machine learning model to learn, so they are considered the "experts" on evaluating the journals. The task that the experts are given is designed to give us the most necessary information to determine if the author requires intervention. We chose to rely on the standard which would be most familiar to the experts: the CGI-S rating scale (refer to Section 2.1). This scale can be viewed in Appendix A.1. There is also an advantage in the CGI-S scale since it clearly defines that a score of 4 or above warrants an intervention to help the author.

In addition to CGI-S rating, experts can specify specific symptoms of depression that are evidenced in the journal, thus giving that journal a "concern label". The default concern labels given are: all nine symptoms of Major Depressive Disorder, three central symptoms of Post-Traumatic Stress Disorder, and one central symptom of Generalized Anxiety Disorder as stated in the DSM-5 [2]. In the scenario that an expert has a specific concern about the journal's author, but there are no default concern labels to assign it, the expert creates a custom concern label and apply it to the journal. Appendix A.2 shows the examples that are included in the task instructions that the experts could use as reference when evaluating the journals.

It is doubtful that experts could label all 531 journals, so there is need of some prioritization for which journals are most useful to label. The Empath scores serve as a good starting point for this prioritization. It is unlikely that a journal with high amounts of positive words, and low amounts of negative words, would raise the concern of the expert. So we will start our labeling with the journals containing the most negative language, and label the more positive journals later.

Attempt using Label Helper

In anticipation of creating a prototype for ALJI, we created a "Label Helper" program that could accomplish two goals: gather the labels from experts and provide a foundation for the journaling interface of the final ALJI prototype. However, it became clear that these two goals can directly conflict with each other. The Label Helper program was built using PyQt5 as a graphics framework and interface designer [15] [37] [35]. PyQt5 was selected due to its cross-platform availability and non-commercial license option (both enable ALJI's accessibility). Screens of the Label Helper interface are seen in Figure 4.1.

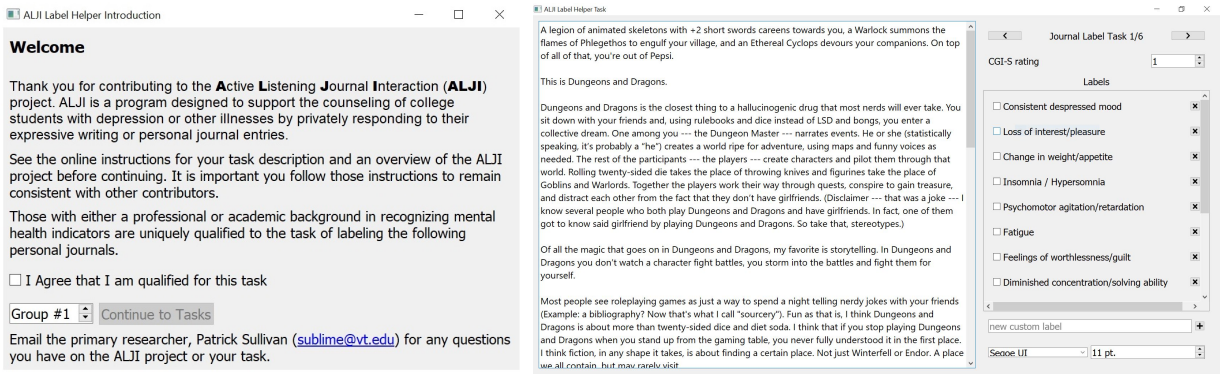


Figure 4.1: Label Helper task introduction and labeling interface

However, there were many pitfalls stemming from this choice leading to a major revision and the disuse of the Label Helper program:

1. The Label Helper program required detailed instructions on its use, since it was unfamiliar to many of the experts.

2. The setup process of the Label Helper program was a barrier to the experts who wished to begin the task right away

3. One labeling task was deemed to be roughly 6 journals. While this is an acceptable amount of time at a normal reading speed, experts needed to carefully analyze the text for meaningful mental health indicators. This is simply not possible at normal reading speeds.

4. Some journals were much longer than the typical journal, but this was not taken into account for time estimates.

All of the above issues compounded into a central issue: a greatly increased time commitment asked of the participant. The time commitment for a voluntary task is typically under 30 minutes to respect participant time, so our aim is a task that lasts under 30 minutes as well.

While the researcher's rough task duration estimates based on reading speed and their own setup time could achieve this task within the target time frame, this was later found to be in error. A Test run of the Label Helper task with a mental health expert showed the setup and task to take up to 2 hours to complete! The setup process is more difficult to replicate that the researcher realized, and the process of analyzing the journal for mental health indicator uses takes more effort than simply reading the text.

The advertising of the 2-hour long task impacted public interest to a high degree. This participation friction allowed the production of only one response of interest (note: *not* task completion) when advertising to a large mailing list of qualified applicants. This extremely low participation would generate an insufficient amount of data that potentially be incredibly biased, crippling the effectiveness of future steps. To combat this friction and raise participation levels, a revision to the journal labeling process was necessary. This revision unfortunately has no use of ALJI Label Helper program, so it remains archived for future opportunities.

Revision using Google Forms

Each of the issues listed within Section 4.3 has the following potential solutions:

1. Use software that is familiar to the mental health experts and has a well-tested and mature interface.
2. Use an online website that puts experts into a short introduction followed immediately by the labeling task (eliminating all setup steps).
3. Shorten the task length to a more manageable number of journals.
4. Add some selection preference for shorter journals that don't require as much time to label. Time spent on meaningless or filler text is otherwise lost.

The online Google Forms service matches the first two solutions quite well. It is familiar to a great many people and is relatively simple to operate. One downside encountered with using Google Forms was how each journal task required a separate Google Form. Google Forms may be unproductive in large future studies since the forms, responses, and participation would need to be manually organized and managed. There is also little control given to designers over the input processing and user interaction (making single-sentence annotation not possible). But it serves well in ALJI's current case when labeling a whole journal.

The selection criteria were also able to account for journal length due to the word count metric already captured. With many journals to choose from, a small amount of selection pressure towards shorter journals effectively halved the amount of reading required of the experts, and it would be spent on more emotional language rather than unnecessary prose.

Each task was also reduced from 6 to 3 journals. The sum of these changes resulted the new task time estimate to be under 30 minutes, resulting in a much higher participation. We can see these results in Section 5.2.

4.4 Machine Learning Process

Equipped with the journal texts, Empath scores, and the experts' impressions for mental health, we can construct a machine learning model to mimic the labeling tasks that experts completed.

4.4.1 Model Selection

Selecting the correct machine learning model is difficult, especially when the problems are complex. For this reason, we opt to attempt many models that are promising, and then compare their performance metrics to determine a winning model. We choose models from the scikit-learn library for machine learning in Python [26]. To aid in data transformations and management, we rely on the Numpy and Pandas libraries [21, 38]. All of these libraries are open-source and are tested extensively. Table 4.1 shows the models we test in ALJI.

Model Category	**Classifiers**	**Regressors**
Dummy	DummyClassifier	DummyRegressor
Support Vector Machine	SVC	SVR
Ensemble	RandomForestClassifier GradientBoostingClassifier AdaBoostClassifier	RandomForestRegressor GradientBoostingRegressor AdaBoostRegressor
Neighbors	KNeighborsClassifier	KNeighborsRegressor
Linear Models		Ridge Lasso ElasticNet

Table 4.1: Machine learning models tested

Short descriptions of the models are as follows:

- **Dummy Classifier** and **Dummy Regressor** are simple models that disregard all input information and only predict the class with the most probable output. They are similar to a student who would mark "C" on all multiple choice questions on a test without reading them.

- **SVC** and **SVR** are the Support Vector Machine classifier and regressor, respectively. Support Vector Machine models first select the data points that are nearest to data points in a different class, and then construct a boundary between them to serve as a border between classes.
- **K Neighbors Classifier** and **K Neighbors Regressor** are models that store the entire training dataset, find the k nearest neighbors to a new data point, and predict that new data point's class based on the majority vote of the k nearest neighbors.
- **Random Forest Classifier** and **Random Forest Regressor** are models created by repeatedly taking a random subset of the training data, and then constructing a decision tree for making predictions based on that subset. Together, these trees form a "forest" whose many predictions are averaged into a final prediction for a new data point.
- **AdaBoost Classifier** and **AdaBoost Regressor** (aka adaptive boosting) are models that produce many weak decision tree models in a similar fashion as Random Forest, but these models are created *in sequence* (rather than independently) so that they gain more focus on misclassified training data as the training process continues. The predictions of a new data point are then combined via a weighted majority vote into a final prediction.
- **Gradient Boosting Classifier** and **Gradient Boosting Regressor** are models very similar to the AdaBoost models except that they do not increase focus on the misclassified training data, and instead focus on reducing the remaining error from the previous predictions.
- **Ridge** is a linear regression model that minimizes the sum of square errors while also penalizing large coefficients (so that a select few features do not overtake the entire learning process of the model).
- **Lasso** is a linear regression model that also minimizes the sum of square errors while also completely disregarding as many features that lack information as possible (eliminating the noise caused by those features).
- **ElasticNet** is a linear regression model that combines the goals of Ridge and Lasso together so that it simultaneously reduces error, avoids feature exaggeration, and disregards features with low information.

4.4.2 Preprocessing and Tuning

As our training data, we focus on the Empath scores of journals (normalized by dividing by the word count of the journal). We also need to standardize the Empath features by removing the mean and scaling to unit variance. This ensures that no feature is dominant

over the others in magnitude, confusing the machine learning models. Several models also require this preprocessing in order to correctly assign penalties and adjust the model learning process in an optimal way.

Several models have parameters that can be tuned, so we will also perform a grid search for the parameter that is provides the optimal performance of each model. The searched parameters are: *C*, *Alpha*, *Kernel*, *N Neighbors*, and *N Estimators*. A short explanation of these parameters and their applicable models are as follows:

- **Kernel** is a parameter for SVC and SVR that describes the method for building the class boundary. Attempted methods were "linear" (straight boundary), "poly" (a polynomial boundary that is curved) and "RBF" (radial basis function forms boundary)
- **C** is the penalty parameter which applies to SVR and SVC models. It defines how much the model adjusts to avoid a misclassification during training. Some misclassification may actually be desirable, since it can produce a simpler, "smoother" model which is not overfitting on the training data (and increasing the performance on the testing data).
- **Alpha** is the regularization strength that is applied in the Ridge and Lasso models. It is inversely related to the *C* parameter of SVR and SVC, and is similarly used to increase or decrease the model's misclassification penalities.
- **N Neighbors** is the number of neighbors included in the voting mechanism for the KNeighborsClassifier and KNeighborsRegressor models.
- **N Estimators** is the parameter controlling the number of estimators (decision trees) used in every model within the "ensemble" model category.

4.4.3 Model Evaluation

The evaluation of ALJI's machine learning models greatly depends on the kinds of predictions they are making. The CGI-S scale in Appendix A.1 is ordinal data (an ordered range), ranging from 1 to 7. However, the two main kinds of machine learning models are classifiers (for categorical data predictions) and regressors (for numerical continuous data predictions). For our case, we will use regressors and view the CGI-S ratings as continuous numerical data. While this does mean our models may give predictions that are not defined or are nonsensical (e.g. 3.67 or -4), our learning models will have some understanding of the "distance" between CGI-S ratings (e.g. the distance between 4 and 5 is less than the distance between 2 and 6). This measurement of distance is a key component in determining the relative magnitude of errors of our regressor models, which has effects on the learning process of our regressor.

However, the precise CGI-S rating may not be necessary to produce a useful learning model. A primary goal of ALJI is to determine whether or not the journal author is in crisis and

therefore needs professional mental health intervention. This is a binary decision (intervention versus non-intervention) and can therefore be viewed as categorical data. The criteria for making this decision could be complex, but we fortunately can rely on the CGI-S scale again, which specifically defines a rating of 4 as

> symptom level warrants intervention. By simply encoding the original CGI-S ratings into one of two classes (3 and below versus 4 and above) we can apply machine learning classifiers instead of regressors. While this sacrifices the precision that a precise CGI-S prediction would give us, our classifiers will never give nonsensical predictions and the prediction target is simpler and may be easier for models to learn.
>
> Each concern label is a binary class (present or absent) which can be individually targeted directly with classifier models. There is also the option to group these labels together and apply multi-class classifiers to predict all concern labels at once. However, some labels may prove to be impossible to predict in a useful manner if our dataset has zero actual information about that label or there are zero occurrences of a concern label in the dataset. This reasoning leads us to use classifiers to predict each concern label individually, and to evaluate them individually as well.
>
> We will focus on the *balanced accuracy* metric for comparing classifier models. Balanced accuracy is a simple performance metric of classifiers that modifies the traditional accuracy metric to account for unbalanced datasets. We do not expect the journal entries we collect to be uniformly distributed among the possible CGI-S ratings or among the concern labels. This imbalance causes the regular accuracy metric to be less helpful since classes that are a minority of samples are outnumbered (and thus outweighed) by classes in the majority.
>
> Our regressors will primarily be trained and evaluated based on the r^2 (aka coefficient of determination) metric. The r^2 metric defined as the variance found by the regressor divided by the total variance of the data. $r^2 = 1$ is the maximum possible score of any model, and denotes a regressor that gives absolutely perfect predictions. $r^2 = 0$ is the maximum theoretical score of a regressor that consistently predicts the most average prediction regardless of input (i.e. our DummyRegressor). This metric gives us a sense of a regressor's progress towards perfect predictions.
>
> To validate our models on how they could perform in a real scenario, we do a 5-fold cross-validation of each model. This repeatedly splits our original dataset, trains the model on a training subset, tests the model on a testing subset, and finally averages the testing score the model achieves on each fold. This kind of validation can be a strong defense against abnormalities in the performance metrics, since it essentially does 5 different training/testing setups. So when a

single model has a favorable train/test split though pure chance, that model can still be diminished by not performing well on the other 4 folds of evaluation.

Chapter 5

Results

5.1 Basic Journal Results

It may be useful to gain some basic knowledge about the journal dataset we have collected to use as reference. Figure 5.1 depicts the word counts and a compound Empath metric of the 531 journals. This compound metric is simply the *positive emotion* score minus the *negative emotion* score generated by Empath's analysis. We see a wide variety of positive to negative journals, a central cluster, and some outliers. It is also apparent that the word limits of the original writing prompt were not strictly enforced. By ignoring the word counts, we can also see the overall emotion of the majority of authors would fit into a bell curve.

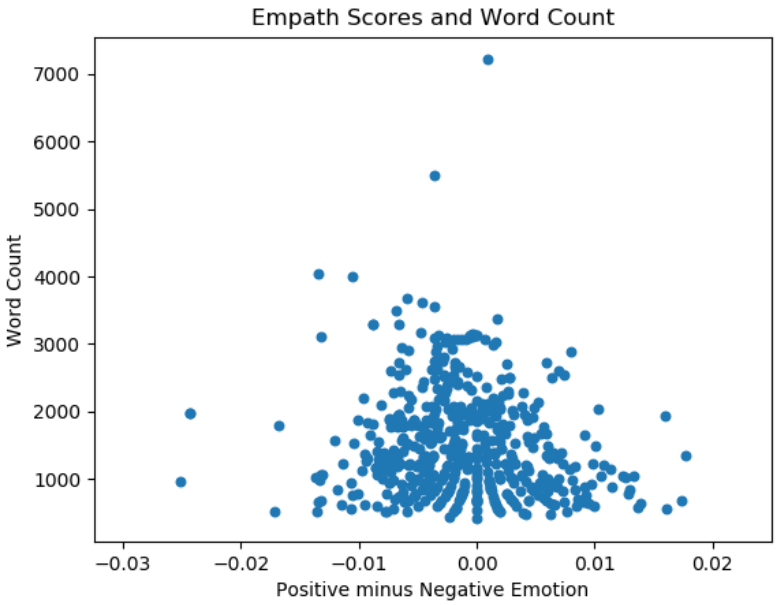


Figure 5.1: Word count and Empath scores of journals

A reading of some of these journals revealed that several do contain extremely personal and powerful events and thoughts. These included child abuse, sexual abuse, suicidal ideation, self-harm, and eating disorders. One common theme

that a mental health professional spotted was a focus on overcoming great adversity. This could be due to the author seeking justification for the awards and recognition of the Scholastic Arts & Writing Awards organization. This biases and effects are discussed in detail in section 6.1

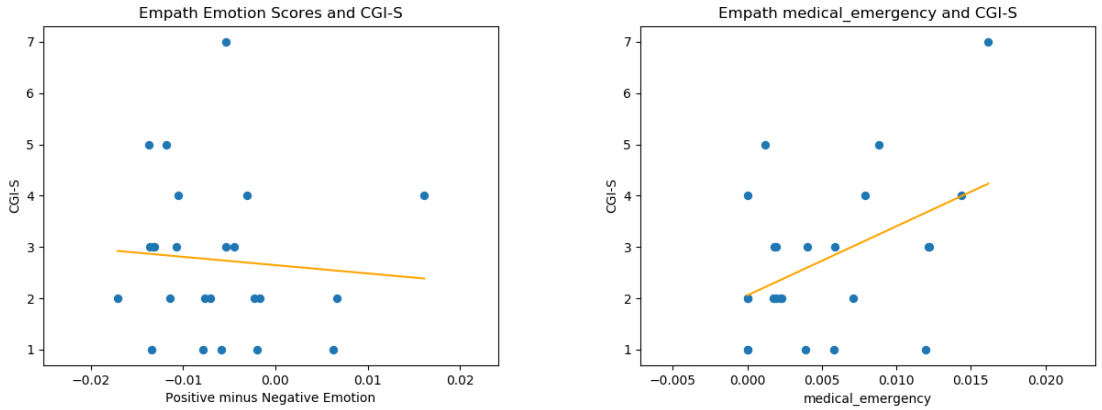
5.2 Journal Labeling Results

The first attempt using the Label Helper had disappointing, but understandable, response and task completion rates. No meaningful labeling data was generated under these circumstances. However, the participation through Google Forms provides us with 24 journals that are labeled both with CGI-S ratings and the expert's concern labels. While this small of sample size is not able to withstand tests of statistical significance, and our machine learning models will likely be unreliable, there are possibly some insights we can gain. There is also a possibility that segmenting each journal into sentences with their own Empath features could give our models enough samples to learn from. There is a total of 897 sentences in the 24 journals that were evaluated, however we expect this to introduce more noise since each sentence wasn't given an individual CGI-S rating (the sentences will inherit the CGI-S rating and concern labels of the entire journal).

Figure 5.2 shows how the CGI-S ratings of the journals compare to several Empath measures. Each figure includes a trend line to give some indication for how Empath features and CGI-S could be related. Note that in both subfigures, the data points do not show a consistent relation between CGI-S and these Empath features. This is somewhat expected, as the expert's labeling of the journals is far beyond the simple task of finding words that match an Empath category. Although our sample size limits may also contribute to why we do not see a noticeable relationship. The emotion categories and *medical emergency* category were chosen as the most likely to have a relationship, but perhaps the relationship is made clear using a much more complex combination of the 194 Empath features. If so, then the machine learning models will likely capitalize on this relationship in making accurate predictions.

Table 5.1 shows the occurrences of the concern labels in our expert-evaluated journals. Concern labels not shown had zero instances across the 24 journals. "Stressful traumatic event" is the most common label, and is assigned to nearly *half* of all journals evaluated. The EWT studies that inspire our ALJI concept similarly noted that people were surprisingly willing to disclose extremely personal and painful events and thoughts in their writings [29].

Experts also had several custom concerns assigned among the 24 evaluated journals. None of the custom concern labels had multiple occurrences. We do note that some of the custom concerns overlap with the default concern labels provided



(a) Empath's positive minus negative emotion (b) Empath's medical_emergency measure

Figure 5.2: Empath characteristics compared to expert CGI-S ratings

Concern Label	**Instances**
Stressful traumatic event	11
Negative alterations in cognition / mood	10
Intrusive memories of past trauma	8
Consistent depressed mood	8
Feelings of worthlessness / guilt	5
Excessive Anxiety and worry over 6 months	4
Recurrent thoughts of death	3
Psychomotor agitation / retardation	2
Insomnia / Hypersomnia	1
Diminished concentration / solving ability	1

Table 5.1: Instances of concern labels across 24 evaluated journals

(e.g. “previous traumatic memories” is likely equivalent to “Intrusive memories of past trauma”). They also cover a wide range of specificity: from a general emotion (anger) to a quite precise mental disorder (trichotillomania). The custom concerns given by mental health experts are listed below:

- anger
- Isolation
- depression
- potential PTSD
- suicide attempt
- low self-esteem
- excessive crying
- Moderate Anxiety
- trichotillomania
- transitional issues
- feelings of hopelessness
- previous traumatic memories
- existential depressive symptoms likely present

5.3 Learning Models Results

Before comparing models to each other, we first should establish that the learning process is working correctly. We can do this by training on the entire dataset and then testing on the same dataset, showing that our models’ predictions are strong when in placed in the same scenario they are trained in. Figure 5.3 shows this learning validation, where each color represents one of our attempted regressor models. Points in this figure show the relationship of a model’s predicted CGI-S rating to the true CGI-S rating given by experts. These results are purely indicative that our basic learning process is sound, it is *not* evidence of the predictive power of our models. Do note that Figure 5.3 includes a DummyRegressor, which only ever predicts the average CGI-S rating, forming a vertical strip of predictions. Many of the models form overlapping points along the diagonal, which indicates that the learning process has achieved predictions close to the true values. A few learning models are showing relatively low performance for predicting precise CGI-S values, but their classifier version may still be useful when targeting the intervention category.

5.3.1 Predicting CGI-S

To evaluate and estimate a model’s usefulness outside of experimental settings, we now split our whole dataset into training and testing subsets. This allows us to evaluate our models’ performance based on “unseen” data (since the testing subset is withheld from the model during the training/learning process). We can repeat this process several times with different splits and favor the reliable learning models over lucky ones. This evaluation process is the standard setup of K-fold cross-validation.

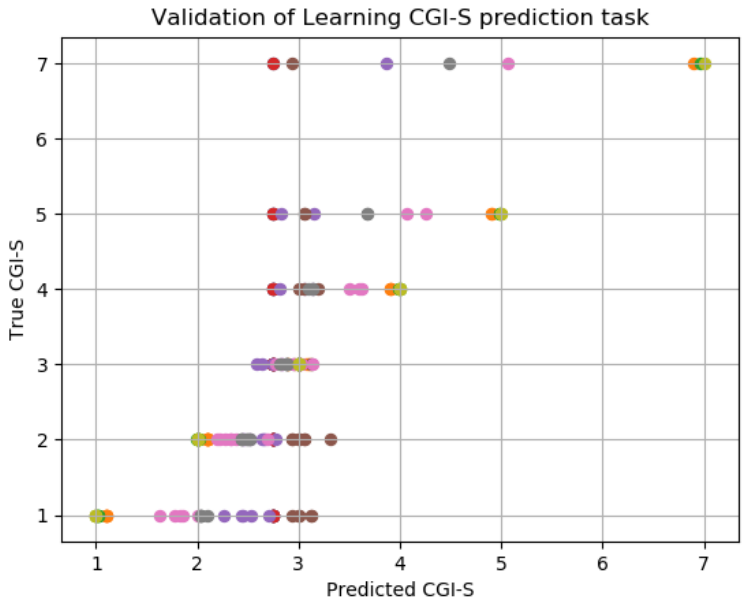


Figure 5.3: Learning validation of best-in-class models

Table 5.2 shows the performance of the best regression models at predicting the expert's CGI-S rating. *Mean test score* is the average r^2 score on testing datasets of the model over all 5 of the cross-validation folds. Only the top performer of each estimator is shown here, as there are many combinations of parameters which do not lead to effective learning. An SVR model with an RBF kernel is tested as the strongest model for predicting a precise CGI-S rating. It remains far from the absolute perfect predictor, which achieves the score of $r^2 = 1$. The DummyRegressor is included for a simple baseline comparison, which has a theoretical maximum score of $r^2 = 0$. We can see that many models did not outperform the DummyRegressor, likely due to the noise and small sample size of our dataset. However, the cross-validated average testing score of the SVR model gives support to the idea that the Empath features contain some amount of information that can be used to give predictions of CGI-S rating.

Table 5.3 shows the sentence segmentation approach to predicting precise CGI-S ratings. The *mean test score* again denotes the regressor's average r^2 metric across 5 folds of cross-validation. Here we see that DummyRegressor, Lasso, and ElasticNet models were tied for best performance. This means that none of the "smart" models could separate good information from the noise in the Empath features for each sentence. Noise and bias issues are discussed at length in Sections 6.1 to 6.4. Splitting the journals by sentence could be losing contextual information, such as negative language in a journal that accumulates across many sentences. While this does increase our dataset's sample size, it clearly comes at

Mean Test Score	Estimator	Parameters
-0.647070	Ridge	{'alpha': 0.9}
-0.522193	GradientBoostingRegressor	{'alpha': 0.7, 'n_estimators': 5}
-0.469705	AdaBoostRegressor	{'n_estimators': 625}
-0.272230	RandomForestRegressor	{'n_estimators': 625}
-0.271280	ElasticNet	{'alpha': 0.9}
-0.156135	KNeighborsRegressor	{'n_neighbors': 16}
-0.154268	Lasso	{'alpha': 0.9}
-0.143785	DummyRegressor	{}
0.043504	**SVR**	{'C': 10, 'kernel': 'rbf'}

Table 5.2: Learning model performance for predicting journal CGI-S

the cost of predictive power.

Mean Test Score	Estimator	Parameters
-1.649514	Ridge	{'alpha': 0.9}
-0.477072	KNeighborsRegressor	{'n_neighbors': 16}
-0.439528	RandomForestRegressor	{'n_estimators': 625}
-0.116175	AdaBoostRegressor	{'n_estimators': 5}
-0.024836	GradientBoostingRegressor	{'alpha': 0.7 ,'n_estimators': 5}
-0.020065	SVR	{'C': 0.01, 'kernel': poly}
-0.009323	**DummyRegressor**	{}
-0.009323	**Lasso**	{'alpha': 0.3}
-0.009323	**ElasticNet**	{'alpha': 0.6}

Table 5.3: Learning model performance for predicting sentence CGI-S

5.3.2 Predicting Intervention

While predicting CGI-S is preferable so that the level of severity is understood, it may be easier to simply learn the decision boundary for intervention (as described in Section 4.4.3). A CGI-S rating of 3 does not warrant intervention, while a CGI-S rating of 4 does, we choose any predicted CGI-S rating above 3.5 as a positive for intervention action. This changes our regression problem into a binary classification problem. The metric we use for comparing the performance of classifiers is *balanced accuracy*, which is the average accuracy of the model across all possible classes. This is especially useful for comparing models using small and imbalanced datasets, one of our constraints.

Table 5.4 is our performance results of using Empath's features to predict whether to intervene for the journal author. The DummyClassifier's method of always predicting the most prominent class caused it to have a balanced accuracy of 35.8%. But this is clearly overshadowed by the actual machine learning models, which consistently had a higher balanced accuracy. The KNeighborsClassifier earned a balanced accuracy of 61% in tests, which stands as evidence that Empath's features have importance to determining an intervention for the journal author.

Mean Test Score	Estimator	Parameters
0.358333	DummyClassifier	{}
0.408333	AdaBoostClassifier	{'n_estimators': 125}
0.450000	GradientBoostingClassifier	{'n_estimators': 5}
0.500000	SVC	{'C': 0.01, 'kernel': 'rbf'}
0.500000	RandomForestClassifier	{'n_estimators': 5}
0.608333	**KNeighborsClassifier**	{'n_neighbors': 1}

Table 5.4: Learning model performance for predicting journal intervention

Table 5.5 shows our models attempting to predict the intervention action from a single sentence in a journal. In the same manner as predicting CGI-S from sentences, we find that predicting intervention from sentences to be too noisy for the machine learning models. The DummyClassifier is alongside the KNeighborsClassifier as the best performers with a 52% mean balanced accuracy score across 5 fold cross-validation. This shows that the costs of sentence segmentation outweigh the performance benefits of simplifying the target down to a binary intervention decision.

Mean Test Score	Estimator	Parameters
0.492206	AdaBoostClassifier	{'n_estimators': 5}
0.496936	RandomForestClassifier	{'n_estimators': 5}
0.497052	GradientBoostingClassifier	{'n_estimators': 5}
0.505948	SVC	{'C': 10, 'kernel': 'rbf'}
0.511359	**KNeighborsClassifier**	{'n_neighbors': 8}
0.511857	**DummyClassifier**	{}

Table 5.5: Learning model performance for predicting sentence intervention

5.3.3 Predicting Concerns

Below are the results of attempting to classify each of the default concerns that experts could assign to a journal (each concern is a depression symptom). Some concern labels were never used by experts, meaning that no analysis can be done on how to detect their presence from the journals. Each are summarized as the best model found, its balanced accuracy score, and its comparison to the DummyClassifier.

- "**Consistent depressed mood**" is best classified by AdaBoostClassifier, scoring 73.3% in balanced accuracy; the DummyClassifier scores a 72.5%.
- "**Loss of interest/pleasure**" was never applied to any journals as a concern label, so there are no examples in our dataset to learn about detecting it.
- "**Change in weight/appetite**" was never applied to any journals as a concern label, so there are no examples in our dataset to learn about detecting it.
- "**Insomnia / Hypersomnia**" is best classified by RandomForestClassifier and KNeighborsClassifier, scoring 90% in balanced accuracy; the DummyClassifier scores an 85%.
- "**Psychomotor agitation/retardation**" is best classified by AdaBoostClassifier, scoring 96% in balanced accuracy; the DummyClassifier scores an 85%.
- "**Fatigue**" was never applied to any journals as a concern label, so there are no examples in our dataset to learn about detecting it.
- "**Feelings of worthlessness/guilt**" is best classified by DummyClassifier, scoring 69% in balanced accuracy.
- "**Diminished concentration/solving ability**" is best classified by the AdaBoost, RandomForest, and KNeighbors classifiers, all scoring a 90% in balanced accuracy; the DummyClassifier scores an 85%.
- "**Recurrent thoughts of death**" is best classified by KNeighborsClassifier, scoring 80% in balanced accuracy; the DummyClassifier scores a 63%.
- "**Excessive Anxiety and worry over 6 months**" is best classified by GradientBoostingClassifier, scoring 65% in balanced accuracy; the DummyClassifier scores a 63%.
- "**Stressful traumatic event**" is best classified by AdaBoostClassifier, scoring 55% in balanced accuracy; the DummyClassifier scores a 48%.
- "**Intrusive memories of past trauma**" is best classified by SVC (linear kernel), scoring 67% in balanced accuracy; the DummyClassifier scores a 31%.
- "**Negative alterations in cognition / mood**" is best classified by AdaBoostClassifier, scoring 63% in balanced accuracy; the DummyClassifier scores a 50%.

Some of the concern labels have classifier have models that outperform the DummyClassifier, which can indicate that the Empath features found in journals have a relationship to that symptom. However, we should remain cautious and aware of the sample size of our journal data and of the chance a concern label was overlooked by an expert when it should have been applied. We omit the results generated by the sentence-segmented journals for reading convenience and how its effects on results were previously shown to be unhelpful.

Each custom concern label had only one occurrence among our 24 evaluated journals. We believe that it would be impossible to build a useful learning model with such a small sample size and imbalance among the classes. There is also the possibility that some would have much more use if they were included in the default concern labels. We should remain mindful of the custom concern labels for future labeling processes and datasets, but machine learning models will not be productive here.

Chapter 6

Discussion

A great deal of discussion can be had on the current state of ALJI. The concept remains to be fully developed and tested, and the methods and results described here have primarily focused on the feasibility of the machine learning module. Our methods had some strict limitations which could impact the viability of ALJI's results. There are also several unanswered questions left by the results.

6.1 Journal Data Source

Private personal journals would be the most valuable, unbiased source of data, since they are the precise setting that ALJI aims to be deployed. Unfortunately, these private journals are designed to not be easily found or accessed (by definition). Using the submitted essays to the Scholastic Awards website introduces several influences that could bias our models from being applicable in real-world scenarios. However, they could still prove to be a suitable proxy for personal journals and the problem at hand. The essay prompt seen in Section 4.1 reflects many core values that we expect to find in journals. Journal authors are likely to write about events or topics of personal significance, which is the exact essay category. We should be aware of possible sources of bias in our dataset, including:

- **Visibility**: The mere knowledge that the readers and audience of the essay is not limited to the author likely has a large impact on the language they use and the topics they write about. They are also more likely to be misrepresent or be untruthful on their experiences in order to manage judgements coming from others.
- **Expressive Freedom**: Since the essay writers choose a category and are given a prompt, they are already in a more structured writing environment than a personal journal. This reduces the range of topics they would normal write about in an unstructured writing environment. The essay writers are also given word count limitations, unlike personal journals.
- **Rewards from Writing**: There is the possibility of recognition and awards from these writings, so authors may craft them towards the mindset of a typical reader and seek to appease the judges more so than write of their own experience and feelings.

- **Formality**: The essays are likely to be carefully proofread and checked for errors that could be normally left in their personal journal. Spelling errors could cause the meaningful Empath features to stay undetected.

The 531 essays that we collected forms a dataset that could be useful in a wide variety of NLP problems. Even if ALJI's concept is eventually proven to be infeasible, there is value gained through this data collection.

6.2 NLP and Empath

Word lemmatization was found to be necessary in analyzing the author's writings, as several instances of significant differences were spotted between the Empath features of a sentence and the lemmatized version of that same sentence. There is a slight chance that the lemmatized version of the word is in error however, so perhaps a hybrid of the two approaches could be beneficial:

$$FinalEmpathFeatures = \frac{Empath(text) + Empath(lemmatize(text))}{2}$$

While Empath gives us 194 lexical categories when analyzing text, other textual analysis tools like LIWC can give more features that seem valuable and fewer features that seem noisy (e.g. LIWC's "Past focus" versus Empath's "ocean"). We should also be aware that reducing a journal down to the default Empath features is removing information from the journals. Punctuation can carry meaning, and emoji/emoticons can express an author's emotions. Empath has the ability to add custom lexical categories to its analyzer, so perhaps only an extension would suffice instead of an entirely new analyzer.

6.3 Journal Labeling Process

The ALJI Label Helper program was largely unsuccessful due to an opposing conflict between the program's two goals. Using the labeling process as a testing ground for the user interface greatly increased the burden on the volunteer experts of the labeling task. The task of evaluating a journal for mental health indicators is also more demanding than originally estimated. With no usable amount of data generated, we redesigned the labeling process. Switching to Google Forms removed the setup process and was familiar to experts. Preferring to evaluate shorter journals over longer journals also helped support the expert's time efficiency. While 24 labeled journals fall short of the nearly universal recommendation of 50 samples for many machine learning models, we can still gain

preliminary insights on the information within these journals. We will not be able to make any statistical guarantees, especially when our data source abounds with freedom and expressiveness.

There is a major presumption that we follow when creating ALJI: that a mental health professional can accurately understand a person's mental condition from a personal journal. There might be a large enough disconnect between a person's writings to their mental health that a mental healthcare professional would need to meet the author in person before being sure of their health. In addition, the evaluation of journals may not be as objective of a task for the experts as we would hope. Personal biases and experiences could be having an effect on evaluating the journals, but a consensus between experts can serve as a validation technique. Our participation rates were not high enough for us to repeatedly evaluate journals until a consensus was reached, and we are limited to rely on the one evaluation per journal.

The selection of journals to be labeled also carries an element of bias. This narrows our evaluation of journals to those with the most obviously negative language, perhaps missing on the covert meanings behind phrases such as "Everybody was happy, except me". Even the preference of journals with fewer words would filter out journals that have a significant amount of "filler" text surrounding an intense and emotionally insightful passage. These biases will make our models less useful in practice outside than compared to our testing conditions.

Single-sentence labeling was not possible in Google Forms version, so we attempted to pass down journal labels to each sentence and train our machine learning model on those labels. But sentences cannot simply inherit the same labels as a journal because this introduces an overwhelming amount of noise and obscures the journal information. We suspect this is the primary cause behind the failures in the models that targeted sentence-segmented journals. If single-sentence labeling was used, noise levels would be much lower, possibly lower than noise levels in the whole-journal dataset due to the high "resolution" of the sentence data.

6.4 Learning Model Outcomes

With the limit of just 24 evaluated journals to work with, our machine learning models are not likely to achieve high accuracy scores. However, we still apply a full 5-fold cross-validation procedure to measure our model's performance, so that our accuracy scores are not invalid or circumstantial.

The precise CGI-S rating is a good target for our regression models, and would be an invaluable measure for the mental health of authors. However, all of our models were outperformed by the DummyRegressor except for our SVR model.

Even then, our SVR model only barely surpassed the maximum DummyRegressor's r^2 score. With roughly the same effectiveness of a DummyRegressor, our best "smart" regressor here should never be put to real use. It adds a minuscule amount of useful information on top of our Empath features, and nothing more (its predictive power can humorously be considered similar to "*I'm confident that this journal has a CGI-S rating below a 7!*"). This small amount of information could be useful in other ways (as we can see in predicting intervention), just not for predicting a precise CGI-S rating.

Simplifying our problem from the precise CGI-S rating down to the binary intervention decision shows much more promise. Every "smart" model outperformed the DummyClassifier's 36% balanced accuracy by a noticeable margin. The KNeighborsClassifer's respectable 61% balanced accuracy should be kept within context of our small dataset however. Adding more expert-evaluated journals to our dataset could cause this metric to fluctuate before stabilizing.

Whether our target was for the precise CGI-S rating or the intervention decision, segmenting the journals into sentences only served to add enough noise into the data to confuse the learning models. We still consider sentence segmentation has potential to be extremely beneficial if the sentences are individually labeled, rather than in our case of inheriting labels from the entire journal. Gathering the individual labels for each sentence is our primary barrier from answering this question. However, we can also imagine situations where sentence segmentation and individual labels could negatively impact our models (three borderline concerning sentences in a row may "accumulate" into a major concern). So we should also be prepared to find that a personal journal is more than a single sentence, or the sum of its sentences.

Concern label prediction models are extremely varied, likely due to our data limitations and imbalance. Some concern labels had zero instances across the 24 evaluated journals, and only 4 labels had greater than 5 occurrences. This makes the metrics for evaluating our machine learning models more unreliable. In such small datasets, it's typically simple to surpass the DummyClassifier by finding a unique feature (that has no actual connection to the target) and classifying using that feature. The most notable concern label in terms of predictability is "Intrusive memories of past trauma", and it remains below a 70% in balanced accuracy. It is almost certain that the custom concern labels would have even lesser outcomes.

We should note that the CGI-S rating required some selection action, whereas the concern labels were unselected by default. So it is possible that some experts either skipped or overlooked considering these labels for the journal they were reading. In addition, we should not assume that our few instances of concern labels in the 24 evaluated journals can broadly represent every way that these symptoms are presented in the journals of the greater public.

Chapter 7

Future Work

There is a numerous number of new directions and improvements that can strengthen the impacts of ALJI. These are just several that warrant more attention.

7.1 Expanded ALJI solution

While Chapter 3 showed ALJI's core proposed solution, there are many other considerations to include for implementing and distributing an ALJI prototype. The privacy of the journal data absolutely warrants the encryption of this data with password protection. Building trust with authors who are not versed in programming or open-source software is also a worthwhile endeavor. This could be supported through publicly showcasing a formal "bug bounty", inviting and rewarding security experts that find security issues in ALJI. While the technical details may elude the layperson, they can clearly understand that ALJI's security must remain competitive in order to survive.

ALJI also could have the ability to use the author's responses to update the machine learning model and tune the model for the individual author. If ALJI's questions seem too intrusive or overbearing, then this tuning would make ALJI less intrusive to the author over time. It may also help ALJI determine a baseline of the author's behavior in writing their journal over a period of time.

7.2 Journal Source Improvements

The necessary improvements to the data source is reducing bias and expanding audience. While collecting actual personal journals as data would be difficult, there are too many factors in "proxy" journals that can change their writing style and language. There are many anonymous personal journals that are accessible online, but those journals likely have zero demographic information. We should expect that different age groups have wildly different journal entries, so focusing on one group at a time is hugely beneficial to understanding their mental health. This study was only able to investigate the mental health of high school seniors

in the USA. We should also be considering how the author's sex, gender, race, and sexual orientation are important pieces of an author's self-awareness and self-esteem, leading to differences in journals.

There is also a possibility of allowing authors to voluntarily donate their private personal journals to the continuing research efforts of ALJI. This may be implemented within the first full ALJI prototype, although the design decisions of Section 3.1 should be remembered. Allowing authors to be able to anatomise their journals before sending them would be beneficial. This would be among the highest quality data that is achievable for ALJI's goals.

7.3 Enhanced NLP

Note that Empath, like many textual analysis tools, views each word as semantically independent. This does not account for special meanings that cross word boundaries, such as "sheep pen" or "lie down". Journals could instead be analyzed as Bigrams or N-grams, capturing more contextual information from passages such as: "There is no hope". And of course punctuation can carry important meaning in a journal (e.g. "I'm leaving!"). Much more complex NLP methods can be applied to the journals to better understand their semantics and hidden mental health indicators. Lexical dictionaries and software such as LIWC may be possible to use, but the ALJI design principles stated in Section 3.1 should be carefully considered when implementing such additions. Previous work by De Choudhury crafted both a depression lexicon and an antidepressant lexicon [7], and these appear to be very applicable to the goals of ALJI.

7.4 Journal Labeling Expansions

Our journal labeling process is far from perfection as well. The largest positive impact would be seen through increases of expert participation or more efficient tasks given to experts, leading to more evaluated journals. Gaining and managing larger studies is difficult, but can be done through networking, preparation, a streamlined task, incentives, and a longer time period of gathering responses. With enough participation, we can also begin to validate journal evaluations based on the consensus of the experts' reviews. Consensus among experts will protect the dataset from the biases of a single expert on our dataset. When conflicting annotations between experts occur on the same journal, we should begin applying *reliability measures* that are commonly used in psychology studies to verify the objectivity of the expert responses.

We attempted to use sentence segmentation by blindly applying all labels of the journal to all of its sentences. The failures in our sentence-segmented results are almost certainly caused by this, and labeling each sentence individually would drastically reduce this noise and enable our models to learn the information. Google Forms does not seem to support these kinds of actions in a survey, so a heavyweight surveying tool or a custom website would be needed.

A different way of increasing the precision of the expert's evaluation of a journal is by providing them with a confidence rating of their evaluations. The expert's confidence of a CGI-S rating or label ultimately help inform our model of the weight of each instance label holds. Cases where the author is experiencing borderline symptoms that are concerning is where these confidence ratings can help manage an expert's indecision.

A final labeling expansion is to look into other mental illnesses that could be detectable in personal journals. The work of ALJI here was mostly targeted towards depression in adolescents, but other illnesses can be expressed in writing. PTSD is an interesting target since journals are commonly a place of processing past events. Bipolar individuals will likely have especially large variations of positive and negative language over many entries. If experts begin to evaluate and label journals for the features of these illnesses, then ALJI can attempt to detect them.

7.5 Machine Learning Modifications

As described in Section 4.4.3, the ordinal data of the CGI-S ratings are not preserved for the machine learning process. However the regressors of the scikit-learn library are able to be wrapped inside a custom regressor which could account for ordinal data (transforming the predictions to fall within the CGI-S scale). Pedregosa-Izquierdo appears to have created such a regressor model for ordinal predictions that is compatible with scikit-learn's library [27].

7.6 Clinical Study of ALJI Prototype

A clinical study of an ALJI prototype will solidify answers to a multitude of questions: How does ALJI's impact compare to baseline EWT on paper? To a computerized version of EWT? Do authors feel benefited by using ALJI? Can personal journal writing tasks be transformed into a typing task? How much trust do authors have in ALJI? Do authors feel judged by the concern questions? Is ALJI's first interpretation of an author's writings typically correct or do they constantly need to be clarified?

if the writing task can become a typing task without sacrificing the benefits of expressive writing therapy. It will also allow tests of ALJI's interpretations of journal entries (by assessing how often the author corrects ALJI's first interpretation during clarification). Most importantly, it would be beneficial to test the efficacy of ALJI against a baseline of traditional EWT. It would also be of interest to test the comparison between EWT written on paper, and EWT typed on a computer.

7.7 Clinical Variation

There is also an interesting use-case of ALJI that could be explored: rather than acting as a personal journal alternative/expansion, ALJI could be assigned by a psychologist or counselor as "homework" to a patient. This would allow the patient to have some consistent mental health assessment while not using up valuable provider-patient meeting time. The results of the clinical study of ALJI compared to EWT could either support or deter this idea since EWT has been accepted as a therapy tool in the past.

Chapter 8

Conclusion

We have conceptualized the Active Listening Journal Interaction (ALJI), a personal journal extension that uses machine learning to assess the mental health of the author. Many design decisions were justified through the previous work of Expressive Writing Therapy and on exploring the setting and nature of private personal journals in order to maximize ALJI's possible positive impacts. Through trials and with effort, we gathered and annotated 24 journal-like writings for many mental health indicators. This very low number of samples can still be useful in determining the feasibility of the machine learning models that ALJI requires. A K-Nearest Neighbors classifier was able to predict from just a single personal journal entry if the author requires a professional mental health intervention with a 61% balanced accuracy. Even with very strict data limitations, we see this model shows good promise towards becoming an accurate and useful predictor of an author's mental health. These findings motivate us to improve our methods and eventually clinically test ALJI as a real application for improving mental health and assessment.

Bibliography

[1] J. E. ASHBURY, B. M. FLETCHER, and R. V. BIRTWHISTLE. Personal journal writing in a communication skills course for first-year medical students. *Medical Education*, 27(3):196–204, 1993. doi: 10.1111/j.1365-2923.1993.tb00257.x.

[2] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[3] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

[4] Leo Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[5] Joan Busner and Steven D Targum. The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)*, 4(7):28, 2007.

[6] Mia Davis. Who am i writing for? an exploration of the influences of the private and public sphere. *Plan II Honors Theses-Openly Available*, 2019.

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[8] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[9] Centers for Disease Control et al. *Youth suicide prevention programs: A resource guide*. The Department, 1992.

[10] Eva-Maria Gortner, Stephanie S Rude, and James W Pennebaker. Benefits of expressive writing in lowering rumination and depressive symptoms. *Behavior therapy*, 37(3):292–303, 2006.

[11] Eva-Maria Gortner, Stephanie S Rude, and James W Pennebaker. Benefits of expressive writing in lowering rumination and depressive symptoms. *Behavior therapy*, 37(3):292–303, 2006.

[12] Louis August Gottschalk and Goldine C Gleser. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press, 1969.

[13] WBRR Guy. CGI. clinical global impressions. *ECDEU assessment manual for psychopharmacology*, 1976.

[14] David J Hellerstein, Philip Yanowitch, Jesse Rosenthal, Lisa W Samstag, Martin Maurer, Karen Kasch, Lara Burrows, Meredith Poster, Marc Cantillon, and Arnold Winston. A randomized double-blind study of fluoxetine versus placebo in the treatment of dysthymia. *The American journal of psychiatry*, 1993.

[15] David Hess and Mark Summerfield. PyQt whitepaper. Technical report, Riverbank Computing, 2013. URL https://www.riverbankcomputing.com/software/pyqt/whitepaper.

[16] Margo Leitschuh. Ask Listen Refer, 2016. URL http://www.asklistenrefer.org/.

[17] Stephen J Lepore. Expressive writing moderates the relation between intrusive thoughts and depressive symptoms. *Journal of personality and social psychology*, 73(5):1030, 1997.

[18] Monika N Lind, Michelle L Byrne, Geordie Wicks, Alec M Smidt, and Nicholas B Allen. The effortless assessment of risk states (EARS) tool: An interpersonal approach to mobile sensing. *JMIR Ment Health*, 5(3): e10334, Aug 2018. ISSN 2368-7959. doi: 10.2196/10334. URL http://mental.jmir.org/2018/3/e10334/.

[19] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing, 2019. URL https://textblob.readthedocs.io/.

[20] lxml dev team. lxml, 8 2019. URL http://lxml.de/. [Online; accessed 2019-09-04].

[21] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[22] Sharon L Mitchell, Mahrin Kader, Sherri A Darrow, Melinda Z Haggerty, and Niki L Keating. Evaluating question, persuade, refer (QPR) suicide prevention training in a college setting. *Journal of College Student Psychotherapy*, 27(2):138–148, 2013.

[23] World Health Organization. *The World Health Report 2001: Mental health: new understanding, new hope*. World Health Organization, 2001.

[24] World Health Organization. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

[25] World Health Organization et al. *Preventing suicide: A global imperative*. World Health Organization, 2014.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[27] Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI: from practice to theory*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2015.

[28] James W Pennebaker. Confession, inhibition, and disease. In *Advances in experimental social psychology*, volume 22, pages 211–244. Elsevier, 1989.

[29] James W Pennebaker. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166, 1997.

[30] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[31] Paul Quinnett. QPR gatekeeper training for suicide prevention: The model, rationale, and theory. *Retrieved July*, 28:2008, 2007.

[32] Leonard Richardson. Beautiful Soup documentation — Beautiful Soup 4.4.0 documentation, 2007. URL https://www.crummy.com/software/BeautifulSoup/bs4/doc/index.html.

[33] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[34] John Sommers-Flanagan and Sidney L Shaw. Suicide risk assessment: What psychologists should know. *Professional Psychology: Research and Practice*, 48(2):98, 2017.

[35] Mark Summerfield. *Rapid GUI programming with Python and Qt: the definitive guide to PyQt programming*. Pearson Education, 2007.

[36] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[37] Phil Thompson. *PyQt5 Reference Guide*. Riverbank Computing, 2019. URL https://www.riverbankcomputing.com/static/Docs/PyQt5/.

[38] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011. doi: 10.1109/MCSE.2011.37. URL https://aip.scitation.org/doi/abs/10.1109/MCSE.2011.37.

[39] Fei Wang and Wencai Du. A test automation framework based on WEB. In *2012 IEEE/ACIS 11th International Conference on Computer and Information Science*, pages 683–687. IEEE, 2012. URL https://docs.seleniumhq.org/.

[40] Gary M Weiss, Jessica L Timko, Catherine M Gallagher, Kenichi Yoneda, and Andrew J Schreiber. Smartwatch-based activity recognition: A machine learning approach. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 426–429. IEEE, 2016.

[41] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE international conference on data mining*, pages 427–434. IEEE, 2003.

Appendices

Appendix A

First Appendix

A.1 Clinical Global Impressions —Severity scale

CGI-S	Description
1	NORMAL - Not at all ill, symptoms of disorder not present past seven days
2	BORDERLINE ILL - Subtle or suspected pathology
3	MILDLY ILL - Affected but functioning reasonably well
4	MODERATELY ILL - Overt symptoms causing noticeable but modest, functional impairment or distress; symptom level warrants intervention
5	MARKEDLY ILL - Intrusive symptoms that distinctly impair social/occupational/school function or cause intrusive levels of distress
6	SEVERELY ILL - Disruptive pathology, behavior and function are frequently influenced by symptoms, may require assistance from others
7	AMONG THE MOST EXTREMELY ILL PATIENTS - Pathology drastically interferes in many life functions, may be hospitalized

Table A.1: The Clinical Global Impressions —Severity rating scale as described by Busner [5]

A.2 Concern Label Examples

A.2.1 Concerning Journal Examples

> … Day after day, I somehow wake up more tired than the day before. The air resists me like a syrup, and yet I can only stare at the ceiling while yearning for sleep to come over me. …

Reasoning: The reference to tiredness may represent insomnia, psychomotor retardation, and restlessness. These labels can be checked for this journal.

> … By my third year in Newton, I had lost all hope of feeling normal again. Every memory of my mother was now stained with her death …

Reasoning: While these feelings are focused on past events, they are coded for the purposes of this exercise as they may continue through the present. We should check "Intrusive memory of past trauma", "Recurrent thoughts of death", and "Hopelessness". ALJI's final prototype can confirm if an author still has these thoughts and feelings.

> … I joked about my traumatic stay at a foster home filled with verbal abuse when I was nine. But they only stared back instead of laughing. So I quickly told them how I had mostly forgotten about it (a lie). …

Reasoning: While the events happened in the past, it is unclear if the author's feelings are resolved about this today. "Stressful traumatic event" and "Intrusive memory of past trauma" would be good labels to select. Since there is an incongruence between trauma (verbal abuse) and affect (joking), an additional code could be added to specify this (e.g. "Incongruent mood and behavior")

A.2.2 Safe Journal Example

> … I became used to my little brother's panicked yells in the night. Each time I thought that the night terrors would make him think twice about watching another scary movie, but he would always ask to rent another one. …

Reasoning: Since it is the brother instead of the author experiencing the night terrors, there are no labels to check based off of these sentences. It would be checked if these behaviors were upsetting the author.