Direct Message Extraction for Automatic Emotional Inference and Drug Detection

Grant Fong
grant_fong@brown.edu

Advisor: Jeff Huang
Reader: Carsten Eickhoff

April 15, 2019

Contents

1 Abstract

We introduce the Sochiatrist Data Extractor, a multi-platform system that retroactively consolidates and anonymizes an individual's direct messages for clinical and research uses. Aimed at non-technical audiences, this tool has been used by clinical RAs to extract data over 350 times, demonstrating its effectiveness. The utility of this data is demonstrated through an affective modeling and a drug detection task. First, mood was predicted using messaging data from 25 college undergraduates with engineered features and features derived from ELMo embeddings. A hierarchical linear model using the latter was able to predict negative affect with an RMSE of 3.9 on a 50 point scale. These affective models were then tested on a group of 12 clinical participants, demonstrating their generalizability. For the second task, seeded topic modeling was used to automatically detect drug use within messaging data with seeds derived from clinical surveys. Modeling topics for each individual drug category resulted in a recall of 91%. Finally, the ethical considerations and best practices when using sensitive messaging data are discussed. The Sochiatrist Data Extractor and the findings generated by the tool provide an accessible way for clinicians to gather messaging data and better understand patients outside of in-person sessions.

2 Introduction

With the advent of social media and digital private messaging, the way we communicate with those that we are close to has drastically changed. Given the importance of communication and support networks in mental health, we explore how we can help mental health professionals better understand their patients by collecting and analyzing patients' messaging data in a safe, anonymized manner with their consent. Clinicians now recognize the importance of understanding a patient's changing affective [63] and emotional [16] states outside of in-person sessions, which have been shown to relate to the onset of long term mental health issues [14]. Consequently, some mental health professionals have tried collateral information gathering from a patient's social media to inform treatment strategies [19]. However, in the context of the historical surveillance and institutionalization of those with mental illness [21, 24], methods must also recognize patients' agency over their data. Research in analyzing messaging data has largely gone unstudied because of the sensitive nature of the data and the technical difficulty in collecting direct messaging data. Given the passive nature of this data generation, this type of data has the potential to retroactively provide quantitative diagnostic information to clinicians.

In order to make this data more accessible, I have continued to build a tool in collaboration with mental health clinicians that collects and analyzes participants social media and direct messaging data. Sochiatrist (a portmanteau of Social and Psychiatrist) is a data extractor designed from a clinical psychology research perspective that allows non-technical audiences to collect their patients' messaging data with their informed consent. Over the past year, the system has been rebuilt from the ground up and and data was analyzed from studies using the extractor.

The system retroactively downloads direct messages from the most used social media and messaging platforms as identified by a 2018 Pew Research Center study, an important contribution given that prior research has largely focused on a single platform even though most people use multiple social media formats [56]. Past tools have also generally required an individual to be physically present in the lab for the length data capture, constraining how much real-world data can be collected [23, 45]. To protect patients' privacy, Sochiatrist removes potential personally identifiable or sensitive information immediately after extraction.

This work discusses the work done 1) in creating and expanding Sochiatrist, 2) building models from the data collected to detect changes in affect (short term mood), and 3) exploring methods of message topic classification to detect drug use. As demonstrated by these different tasks, there is great utility in using messaging data in a clinical space. However, using this private data also comes with its own ethical concerns. At the end of this paper, these considerations are explored as well as how they informed the system design.

3 Sochiatrist Data Extractor

3.1 Introduction

In order to make use of this rich data at a clinical level, there needs to be a robust and scalable way to collect and clean it. The Sochiatrist system enables researchers to collect social media data from multiple web and mobile-based platforms through the use of an automated script. The data extractor can be used to collect direct messages from Facebook, Instagram, Twitter, Kik, WhatsApp and SMS (iOS's Messages) on both Android and iOS phones. These platforms are the most used messaging platforms in the US as identified by a 2018 Pew study [56]. Participants must be physically present to input their login information for each messaging platform, building in participant consent into the core of the extractor. Monitoring of participants outside of the extraction is not possible with this system.

3.2 Related Work

Previous studies collecting messages used a variety of data gathering methods, including the manual loading of pages that contain messages [23] and having participants chat in real-time while recording all actions [45]. These past methods required an individual to be physically present in the lab for data collection, restricting the quantity and variety of data available.

Furthermore, prior research has generally focused around one social media platform [23, 5], but individuals tend to use more than one. A 2018 Pew Research Center survey of American adults found that 73% of the

Timestamp	Status	Message	Messaging Partner	Platform
7/1/17 12:32	sent	I haven't been feeling that great this past week	6af224	Facebook Messenger
7/1/17 12:32	received	Do you want to talk about it 7ac988?	6af224	Facebook Messenger
7/2/17 18:01	received	Are you feeling better?	72c7e0	Instagram DM

Table 1: Sochiatrist Data Extractor example output, demonstrating how messages are collected across platforms and how names are anonymized. These are example messages, not actual participant data.

public used more than one of the eight social media platforms assessed in the survey [56]. This multi-platform use demonstrates the necessity of a tool to collect and analyze many different forms of social media data. It is unknown whether mood can be inferred based on both the metadata (platform, conversation participant, and timestamp data) and the content (text and emoji data) of informal messages, not just broad public social media posts. The work done in this paper explores these questions regarding the gathering of data across multiple platforms and capturing message content as well as metadata.

3.3 Tool Design

In order to extract messaging data from these platforms across the different devices participants may have, multiple extraction strategies must be used. In this section, the setup and design of the extractor will be discussed in depth.

The only assumption required for installation is that there is an internet connection and a version of python is installed on the machine. The script handles its own setup on launch and displays step by step instructions, allowing for non-technical research assistants to set up the system and run extractions. It achieves this by prepackaging required binaries such as Android Debug Bridge (ADB) and creating a python virtual environment on the host machine in order to isolate the system from the system python environment.

To extract SMS, WhatsApp, and Kik data, the script moves data from the phone onto a lab computer since there is no data export or API alternative provided by these platforms. The Sochiatrist system locates the databases storing messages by searching for `sqlite` files, then reads from tables that are known from previous work in computer forensics to contain messaging data [49, 11]. For Android phones, the phone file-system is accessed by enabling developer mode on the phone and mounting the device with ADB. On iOS, additional security features do not allow for devices to be mounted like with the Android extraction. As a result, a backup is made on the computer and files are accessed through the backup [11]. Extracting these messages does not require rooting or jail-breaking, a hard-to-reverse procedure that may damage a participant's phone.

For web-based social messages, Sochiatrist downloads messages directly from the browser-based website, either as a data download provided by the website (Facebook, similar to Saha et al. [53]), from an API (Instagram), or directly from the website through the use of a web scraping script (Twitter).

After extraction, the messages are compiled in a consistent format, and the user is prompted to specify the time range of messages for their final Comma-Separated Value (CSV) file. The participant also has the ability to request certain conversations be removed from the data dump. Only data within the specified time range is saved, and identifying names in the data are anonymized via the method specified below. At the end of the process, all backups, data downloads, and unanonymized files are deleted from the lab computer leaving one final csv file. This protocol can be seen in Figure 1.

Table 1 shows an example output of the Sochiatrist Data Extractor. For each message, the dataset includes the timestamp, the text of the message, whether the message was sent or received, a hashed id representing the sender's name, and what social media platform the message was exchanged on.

3.4 Robustness

Due to the large number of dependencies and requirements to interact with secure files on the system (iOS) and external devices (Android), the Sochiatrist system must be run locally and thus needs to work on multiple operating systems and machines. The biggest challenge with making the extractor robust is the variation between machines, coupled with the number of requirements to run the extractor and the non-technical audience of the users.

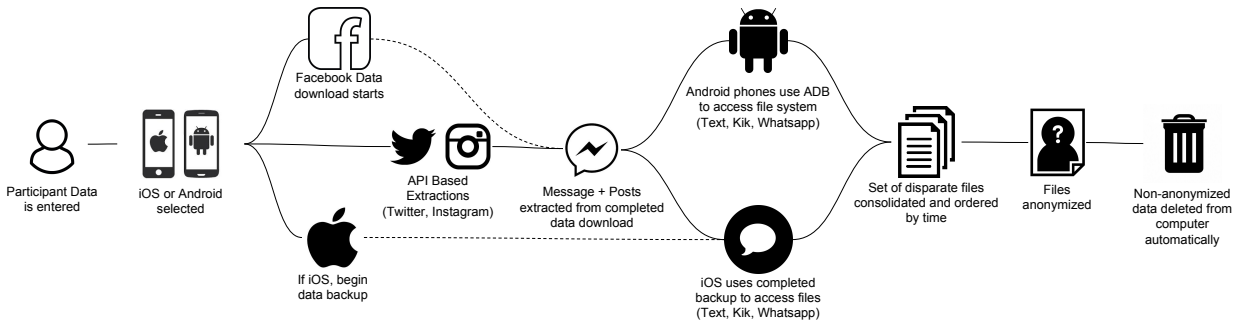

Figure 1: The protocol that is followed when using the Sochiatrist data extractor. The time-intensive Facebook data dump and iOS backup are started in the background, notated by the dotted lines, in order to minimize the amount of time needed for extraction.

Because of the reliance of the extractor on social media platforms to gather data, one of the largest problems with maintaining the extractor is detecting and compensating for changes in data format. These disparate approaches to extracting data have been chosen with this robustness in mind. The most robust platforms are those that interact with mobile backends, since app and platform data are serialized within the request, along with deprecation notices. As a result, changes can be detected before they happen, and updates can be pushed behind the scenes to maintain functionality. Website scraping is also fairly robust, barring large redesigns of websites. The most fragile form of data gathering is through data downloads, as they are subject to change format at any time without notice. Accessing data on phones is also fairly stable, barring major refactors and major updates of apps and operating systems, which are all relatively transparent and easy to debug. Thus, their use is kept to a minimum and only Facebook currently requires a data download to access data. This tradeoff was chosen due to the difficulty in accessing Facebook data through a spoofed mobile app/their website. The tool also handles multi-factor authentication, which occasionally occurs if a login is flagged as being suspicious when a new device accesses the account.

Over its lifetime the data extractor has been used by clinical research assistants without computational backgrounds to successfully extract data over 350 times across 150+ individuals. With the current timeline of studies that use the extractor, there will be continued support and updates for at least the next 5 years.

3.5 Anonymization

To protect the identity of conversation partners, multiple methods of pseudonymization are used to remove names and potential identifiers. These methods detect names and numbers to be replaced respectively with hashes or substitute symbols. This anonymization increased the level of comfort participants had when sharing their messaging data and removed sensitive information that could be contained in messages such as addresses and account numbers.

In order to anonymize numbers, a simple regular expression search is run over the messages, replacing numbers with the # symbol. This preserves the context and the general form of the numbers, so one is able to still interpret the type of a specific number, such as a phone number or a dollar amount.

In order to anonymize names within messages, a few different approaches can be taken. These methods label certain works within the sentences as names, which can then be combined through voting (a majority of the labeling schemes identify the word as a name) or through a threshold (a certain number of methods detect the word as a name). This ensemble of anonymizers can compensate for the shortcomings of other anonymizers. When names have been recognized, they are replaced by a pseudo-random string that is internally consistent, mapping the same name to the same value across platforms.

The simplest method of anonymization simply searches for a set of pre-programmed names that can be customized by the end user. We will refer to this method as List Anonymization. While this method is straightforward, it is able to take in feedback from past extractions with uncommon names that are not detected by automatic approaches, such as non-anglicized names.

Names can also be identified by matching against the names of the participant's friends on Facebook

through the Facebook Anonymizer. With this approach, a blacklist of names is also used in order to prevent commonly occurring words that are occasionally names from being removed. One limitation is that this technique misses cases where people use nicknames for people that they messaged (e.g. Auntie, Honey) or when people exchanged messages with someone who had a common English word as one of their names (e.g. April, Hope). Numbers are also anonymized by substituting in the '#' character.

For these methods that use a set of names, simple regular expression searches were used to anonymize different forms of the name within the metadata and messages such as possessives and different capitalization. Such an approach was chosen due to the messy nature of messaging communication since only the spelling of the name needs to be preserved.

The final method uses the Stanford Named Entity Recognizer (NER) to flag names by labeling locations, people, and organization and removing detected names [18]. This method benefits from not having an explicit set of names to be searching for while also taking into account the syntactic information embedded within the messages. However, this method is significantly slower than the regular expression based anonymization strategies, especially when run on older equipment that is used in hospitals. Additionally, from observation during data extractions, this method struggles with non-anglicized names and can have a high rate of false positives.

While these methods provide a way to anonymize messages without much effort, even the best automated anonymization strategies will still have errors. As a result, some clinical collaborators find it necessary to hand anonymize this information after being passed through the anonymizers. In order to facilitate this and encourage the use of the anonymizers, research assistants have the option of saving the mapping of names to random strings to fix errors that might occur.

3.6 Summary

With the Sochiatrist Data Extractor, non-technical research assistants can extract data from the most used messaging platforms, enabling research to be performed on data that was previously difficult to collect. Additionally, the built in anonymizers run immediately after extraction in order to preserve a maximum amount of privacy for the research subjects.

4 Affective Prediction Based on Messaging Contents

4.1 Introduction

One of the key research questions we seek to answer is whether messaging data can predict changes in patients' affect, a problem well grounded both in clinical psychology as well as the HCI community [38, 30]. Before working with clinicians' data, 25 college undergraduates were used to develop predictive models. By gathering positive and negative affect over two weeks, we demonstrate the utility of features inferred from messages by showing their predictive power on affect. An additional classification task for when participants are feeling more positively than negatively is explored. To derive features from messaging data, two featurization methods are explored: a classic feature engineering approach and a method deriving features from a pretrained ELMo model, a deep contextualized word representation. We then evaluate these models using data from an additional 12 clinical patients from an adolescent inpatient psychiatric unit. Using this clinical group, the use of private and public data is also compared. To contextualize these findings and understand just how emotional states can be inferred solely from past mood, we also perform a time series analysis of emotional state without the use of message data. Based on these analyses, we demonstrate the utility of messaging data in retroactively estimating affect, which provides clinically significant information in the assessment of long term diagnoses [14, 63].

4.2 Related Work

4.2.1 Clinical Psychology in Context

The evaluation of a participant's affect has often been predicated on self-reported data. To reduce potential negative recall bias, Ecological Momentary Assessments (EMA) are used as a method intended for participants

to record their feelings in a structured format [12, 61]. This work adopts an EMA technique where research participants are prompted to fill out mood assessment surveys at semi-random times on their smartphone [61]. While a step in accuracy above retrospective participant recall in reporting mood, EMA can also potentially be burdensome for participants, and rates of compliance with EMA can be affected by a host of external factors [31, 39].

To mitigate the participant burden associated with EMAs, and to add more nuance to clinical assessments of emotional state, several researchers have suggested using mobile data, including social network data [48, 62]. Through the analysis of this data, basic survey-based EMA data can be augmented with deeper and more accurate representations of participant experience that do not overburden the participant [48, 28]. However, these approaches simply augment the EMA protocol, and thus cannot be applied retroactively.

4.2.2 Mobile Personal Informatics for Prediction

Much work has focused around affect prediction from passively generated metadata. Moodscope, a mood inference tool developed by LiKamWa et al., attempted to infer the daily mood average of 32 participants through communication-based metrics, application use, browser history, and location data [38]. While initial results only had a 66% accuracy, accuracy increased to 93% over 2 months of training. A subsequent attempt by Asselbergs et al. to replicate the results of the Moodscope study with 27 students failed to perform better than chance. Jaques et al. used Multitask Learning and Domain Adaptation to create personalized models that predicted mood utilizing physiological data, location, surveys, and communication logs. The Multitask Learning model was able to achieve an RMSE of 13.05 in predicting mood on a 1–100 scale [30].

Servia-Rodriguez et al. tried operating at a more granular level by using location data, microphone data, accelerometer measurements, and call/SMS logs to predict self-reported mood state. They used a deep neural network on a dataset of 18,000 users over three years, with self-report assessments happening twice a day. The maximum accuracy found was approximately 70% on weekends but varied between 50% and 65% on other days [55]. Also attempting to predict granular affective results, Ameko et al. focused predictive efforts on negative affect, framing negative affect as a proxy for mental health in adults [2]. The dataset was comprised of location data, accelerometer data, phone call metadata, and SMS metadata from 65 undergraduate students over the course of two weeks, with six affect assessments each day. A behavioral group based predictive approach achieved an RMSE = 22.05 in predicting negative affect on a 1–100 scale. Such approaches using phone metadata have required sophisticated models to produce strong results, and still require these individual elements to be tracked through systems loaded onto the phone. As a result, these methods are also unable to retroactively estimate mood.

4.2.3 Social Media Analysis of Sentiment and Mood

In parallel to work done in mobile sensing and mood, social computing research demonstrates that emotional state and mood can be inferred from social media data, even after considering the social and cultural factors that influence affect and mood. Andalibi et al. have found that Instagram posts and comments contain a strong sense of community and social support to those who self-disclose mental health challenges [3]. Saha et al. found that mood instability could be predicted from public Twitter Tweets with 96% accuracy by training semi-supervised models on EMA data [53]. It is well accepted that public social media content is reflective of mood, but little work has explored the usage of private social media such as messaging for this task.

4.2.4 Message Embeddings

In order to use textual data in such models, the text must be featurized. Past work in social computing has largely relied on engineered features [23]. However, representation learning in text has become a highly studied field with word2vec and GloVe being two major approaches [50, 46]. More recently, embeddings from language models (ELMo) has become a popular embedding method which use deep learning to generate contextualized word and sentence embeddings [51]. ELMo is used due to its state of the art performance on word embedding metrics to generate features.

There exist ways to combine individual embeddings together to represent larger documents, or in this case periods of time. Paragrah2vec provides an additional method to develop representations of short amounts of text. Developed by Le et al., this method builds off of the skip-gram method explored in word2vec [33].

However, this approach is ill-suited for messaging data, since it requires retraining if there are unseen word groups. Other methods exist to combine word embeddings to sentence embeddings, the most basic but also commonly used is by combining sentence embeddings through max, min, or mean functions [33].

The study of representation learning specifically in social media has also been explored in the past. Boom et al. studied how very short texts seen in social media data could be used to build word representations using word embedding aggregation [7]. Using Wikipedia and Twitter data, they weighted individual words using tf-idf before averaging them to arrive at a vector representing a single social media post or message. However, they found this method to be not particularly well suited for Twitter data due to the short nature of texts. Lang et al. explored using user-user, user-message, and message-message embeddings to predict re-tweeting behaviors on Twitter [36]. Such representation learning work has enabled promising increases in performance over engineered features, but has not been studied in this context.

4.3 Study Procedure

To test how well direct messages predict mood, a study was designed to collect affect and messaging data at Brown University. Study recruitment occurred via posters and Facebook posts in groups created for undergraduate students in different class levels. College students were studied due to the high number of mental illness diagnoses [26] and the heavy usage of messaging platforms in this population [20]. 129 students responded with interest in the study, and 25 people participated for the full two-week length of the study.

Participants were required to have an Android or iOS phone to join the study, and use at least one Sochiatrist-supported messaging application. Each day, to measure a baseline score of negative and positive affect for the model to predict, participants were asked to complete the PANAS, a self-reported questionnaire that measures positive and negative affect [64]. The PANAS has been externally validated and is known as an internally consistent [64] assessment of affect [63, 4]. The version of the PANAS used for this study had 20 questions, with 10 questions measuring positive and negative affect via a Likert scale, for a potential maximum score of 50 each, and a potential minimum score of 10.

Participants were prompted to complete the PANAS survey three times each day. The survey was sent to participants at a random point in each third of their day, based on their reported sleep and wake times. Surveys were sent via email or text, depending on the participant's preference, with surveys administered by an online form that was prefilled with the participant's unique identifier. After receiving a prompt, participants had one hour to complete the survey and would receive a reminder prompt after 30 minutes if the survey was still not submitted. If participants did not complete a survey within the hour after they received the survey, they were not compensated for it. Extra assessments that were completed without being prompted were disregarded from the analysis.

After completing the two weeks, participants were asked qualitative questions about their experience. The Sochiatrist Data Extractor was then used to gather and consolidate the messaging data they produced over the course of the study. Then, participants were compensated a maximum of $60 for their participation. Compensation was based on: completing at least 95% of the PANAS surveys within the prompted 1-hour window ($35), the provision of some amount of social data ($5), and wearing the Microsoft Band ($20) The Microsoft Band data was not used for this analysis due to the focus on messaging data.

During the course of the study, two participants discontinued their participation for privacy reasons and one for undisclosed reasons. Two participants were unable to extract their text messages but were able to gather data from their other other social media accounts.

To test the generalizability of the models, an additional 12 participants' data was acquired from a larger existing clinical research dataset created using the Sochiatrist system at the Bradley Hasbro Children's Research Center. These participants filled out a similar PANAS assessment up to six times a day. Their scores have been transformed to match the 1–50 scale of this study. Data was collected over three weeks following a psychiatric discharge from December 29, 2016, to February 7, 2018. While the age demographics are similar to that of the college population, this provides a check to see how such a model could work in a clinical setting.

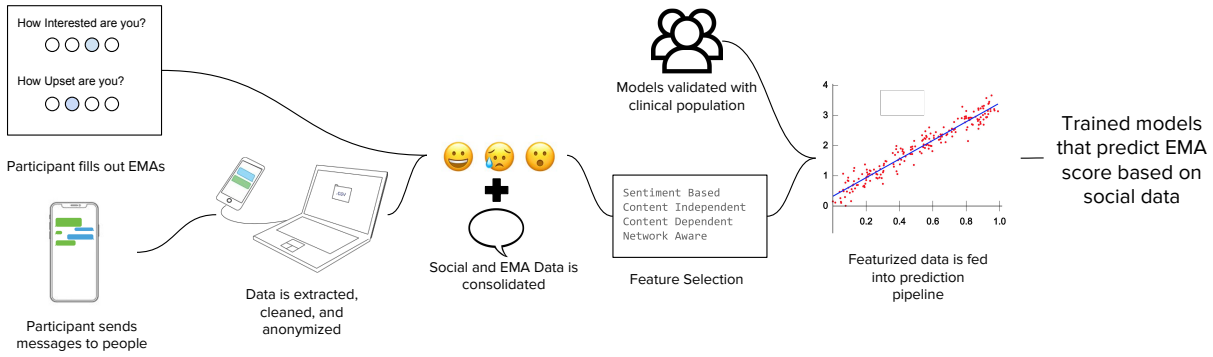

Figure 2: After data collection, features were derived using the undergraduate population. The models trained on this data were then validated using the data from clinical patients to test generalizability.

4.3.1 Data Collection

The study with undergraduate students was approved by the Brown Human Subjects Office, and the clinical population was approved by the Lifespan IRB. Data from the undergraduate population was collected between November 14, 2017 and December 17, 2017, and from the clinical population between December 29, 2016 and February 7, 2018.

4.3.2 Data Summary and Cleaning

Of the 25 participants we collected data from, 20 were female, and 5 were male. 6 participants were Android users, and 19 were iOS users. The age of participants ranged from 17–22 years old, with the average being 19 years old. Over the course of the study, we collected a total of 74,586 messages and 1,009 complete EMA entries, 55 of which were incomplete and discarded. The most commonly collected message types were text messages or iMessages ($n = 23$) and the least common was Twitter Direct Messages ($n = 1$). EMA compliance ranged from 81–100% with a median compliance rate of 98%. Specific data about the messages we collected, as well as the rates of compliance with PANAS prompts, can be seen in Table 2.

In the clinical population used for validation, 5 were female, 3 were male, and 4 lacked gender information. All participants used a provided Android phone and were between the ages of 13–18. In total, 29,197 private messages, 847 public posts, and 780 EMA entries were collected from this population. For the EMA validation, only the private messages were used. EMA compliance was lower, ranging from 1–99% with a median compliance rate of 29%. In both populations, the median number of messaging platforms used was 2, highlighting the importance of having a comprehensive multi-platform extractor since communication happens on several different platforms. The distribution of messages across both populations can be seen in Table 3.

Besides pseudonymization, messages that were created by a non-human source were removed through an iterative scripting process. These removed messages included those created by embedded Facebook Messenger games as well as the text messages from the survey prompter. Additionally, if individuals completed surveys unprompted or after the allocated time window, the block was removed from the analysis. We then cleaned the resulting EMA periods, the lists of messages associated with a certain EMA value. 23 out of the 954 EMA periods in the undergraduate population were removed as they had no messages sent in them. In the clinical population, 346 EMA periods that had no messages were also removed.

4.4 Affective Modeling Methods

Using the participants' messaging data and self-reported affect scores, features based on the messaging were required to model the relationship between the two. Two approaches were used, one with a feature engineering approach and another using ELMo to generate features. In constructing models, ground truth was defined as

	Mean	Max	Min	Median	σ
College Population:					
Message Statistics					
All Messages	3,068	16,697	358	1,897	3,531
Sent Messages	1,269	7,141	25	603	1,572
Received Messages	1,797	9,826	296	1,123	1,973
Messaging Applications Used	2.24	3	1	2	0.71
EMA Statistics					
Compliance Rate (%)	96	100	81	98	6.3
Positive Affect Score	24	50	10	24	8.2
Negative Affect Score	17	49	10	14	7.3
Clinical Population:					
Message Statistics					
All Messages	2,502	17,123	93	391	4,707
Sent Messages	1,254	8,083	46	218	2,291
Received Messages	1,237	9,040	0	176	2,444
Messaging Applications Used	2.08	4	1	2	0.95
EMA Statistics					
Compliance Rate (%)	40	99	1	29	30
Positive Affect Score	25	50	10	23	10.7
Negative Affect Score	15	50	10	13	5.9

Table 2: Message and PANAS Assessment Statistics across participants in both populations. With a median of 2 applications used in both populations, it is necessary to have a multi-platform data extractor in order to capture the full messaging behaviors of participants.

the numerical sum of the affect based PANAS questions for the positive and negative category respectively. This is consistent with how the PANAS is used to assess overall affect.

Because some designed features were count-based, EMA periods occasionally contain outliers. To reduce the positive skew of these count-based features, 1 was added to each count, and then log transformed before calculating correlation coefficients and inputting the features into regression models.

A pre-trained ELMo model which was trained on the 1 Billion Word Benchmark [9] was downloaded from the Tensorflow hub and tuned on the messaging data. 1000 dimensional embeddings for each individual message were generated by passing messages through the model which were combined through mean aggregation [7]. PCA was used to decrease the dimensionality to match the number of engineered features in order to limit the issues with fitting highly dimensional data [67].

For analysis, EMA data was fit to messaging features using a linear model (LM) due to its simplicity and interpretability, a hierarchical linear model (HLM) to capture the unique communication style of each individual, and a support vector regression (SVR) which does not assume a linear relationship between PANAS scores and the features. Neural nets were also explored but it did not have enough training data to perform better than these tested models. To compare these models, we calculated the cross-validated root mean squared error (CV RMSE) of the model trained on all the data to determine the goodness of fit. Leave-one-out cross-validation was chosen to accommodate a small sample size, folding across participants ($k = 25$) to test generalizability across participants. We additionally explored the impact of using only features based off of sent messages and tested the generalizability of these models by evaluating them on the clinical population. To understand the impact of past affect on current assessments and to contextualize these results, we performed an Autoregressive Integrated Moving Average (ARIMA) analysis on the isolated PANAS assessment data without taking into account any message data.

Both EMA scores were combined into an affect classification task to determine when participants felt more negatively than positively at a given time. A simple comparison if the negative score was greater than or equal to the positive score was used to generate these labels. Such a task decreases some of the noise inherent within the affect scores given the direct interpretation of feeling more negatively than positively. Positive affect was dominant at 75% of the time points. Using the same features, a logistic regression, random

Platform:	**SMS**	**Facebook**		**Instagram**		**Twitter**		**WhatsApp**
	Messages	Messages	Posts	Messages	Posts	Messages	Tweets	Messages
College Population	33,779	39,029	0	47	0	196	0	1,535
Clinical Population	28,657	391	219	149	67	0	561	0

Table 3: The number of messages sent and received on each platform. While the most popular platforms for the college population were Facebook and text messages, text messages were the dominant form of communication for the clinical population. While other direct messaging platforms were used less frequently, those who did use them communicated at roughly equal amounts on all platforms. Note that public data (Facebook posts, Instagram posts, and Twitter tweets) were not collected in the college student population. Public posts were removed from the clinical validation set when predicting affect.

forest, and AdaBoost were tested. As with the regression problem, leave-one-out cross-validation was used, folding over each participant. This was additionally tested against the clinical group, where positive affect was dominant 74% of the time.

4.5 Results

4.5.1 Feature Selection

To create models for predicting emotion, features were generated based on simple metrics and clues from prior literature. These features were designed to measure values in aggregate to limit the effect of noise due to being derived from casual conversations while also being interpretable. Five types of features emerged:

Sentiment Based Features: These features are derived from a compound score of sentiment, ranging from +1.00 to -1.00, as calculated via the Valence Aware Dictionary and sEntiment Reasoner (VADER) which provides lexicon and rule-based sentiment analysis [27]. VADER was chosen as it was specifically designed for sentiment analysis in social media contexts and has been shown to outperform other gold standard models for this task [52].
Content Independent Features: These features are calculated solely based on messaging metadata (e.g. number of messages sent). They do not require the actual text of a message or any data about who the messages are being sent to or received by.
Content Dependent Features: These features are dependent on the content of the message and include counts of specific words and phrases within the messages.
Network Aware Features: These features are dependent on the people that the user communicates with and the frequency of the communication with those users.
Time Sensitive Features: These features depend on the timing of responses.

A full list of features that were used in the models is presented in Table 4. In choosing features, specific design choices were made:

- The "sentilength score" was calculated by multiplying the number of words in a message with the magnitude of the sentiment of that message. This equalizes sentimental conversations with short messages with sentimental conversations with fewer, longer messages. The 99th percentile of sentilength scores was the cutoff for exceptional sentiment, the count of which was used as a feature. This feature captures the number of highly emotional and long messages that are generated.

- Social network-based features were limited to the five most messaged people over the course of the two weeks based on Dunbar's [15] work and MacCarron et al.'s supporting research [42] theorizing that an individual is able to maintain a close relationship with approximately five people at a given time.

- To create a corpus of emotion related language to be used for content-dependent features, we applied a technique based on De Choudhury et al.'s use of data from Reddit communities to study online mental health language [13], where nouns, adjectives, trigrams, and four-grams were scraped from online forums related to mental health issues. The words "sorry," "I,", and "feel" were counted as predictors of states of

Features	Pos. Corr.	Neg. Corr.	Max	Min	Mean	Median	σ
Sentiment Based Features							
Number of sent messages with exceptional sentiment scaled by message length (sentilength, sent)	*0.28	-0.11	0.42	-0.29	0.03	-0.01	0.20
Number of received messages with exceptional sentiment × message length (sentilength, received)	-0.02	0.06	0.43	-0.27	0.01	0.00	0.16
Content Independent Features							
Messages sent or received by the participant	0.87	-0.37	0.32	-0.35	0.04	0.03	0.19
Messages sent by the participant	-0.98	0.62	0.27	-0.23	0.03	0.05	0.15
Unique users communicated with	0.60	0.41	0.41	-0.37	0.01	0.01	0.20
Messages received by the participant	-0.35	0.13	0.33	-0.29	0.01	0.00	0.17
Content Dependent Features							
Emotion words used in messages sent by the participant	0.59	0.13	0.31	-0.28	0.03	0.04	0.16
Emotion words used in messages received by the participant	0.22	0.12	0.40	-0.35	0.00	-0.01	0.17
Times "sorry" was used in messages sent or received by the participant	0.22	0.32	0.38	-0.39	0.02	0.01	0.16
Times "I" was used in messages sent or received by the participant	-0.27	0.40	0.37	-0.39	0.02	0.01	0.16
Times the word stem "feel" was used in messages sent or received by the participant	-0.30	0.21	0.41	-0.32	0.06	0.05	0.18
Mental health support phrases used in messages sent and received by the participant	0.00	0.35	0.36	-0.34	0.04	0.07	0.17
Characters in all messages during an assessment period	0.00	0.01	0.36	-0.33	0.04	0.06	0.17
Network Aware Features							
Messages sent by the participant that were not to the top 5 communication partners	0.06	0.09	0.36	-0.28	0.01	0.00	0.18
Messages received by the participant that were not from the top 5 communication partners	-0.34	0.45	0.42	-0.27	0.02	0.00	0.18
Messages sent by the participant to the top 5 communication partners	-0.11	0.17	0.28	-0.32	0.03	0.03	0.16
Messages received by the participant from the top 5 communication partners	0.39	*-0.65	0.29	-0.44	0.00	0.01	0.18
Emotion words used in messages with the top 5 communication partners	-0.18	0.32	0.40	-0.35	0.04	0.04	0.20
Emotion words used in messages not with the top 5 communication partners	0.06	-0.55	-0.38	-0.41	0.00	0.00	0.21
Time Sensitive Features							
Mean response time	**-0.02	-0.02	0.32	-0.25	0.00	-0.01	0.12
Deviation from overall response rate	0.00	1.47	0.31	-0.27	0.00	-0.01	0.13

Table 4: List of Features (* and ** represent $p < 0.05, 0.01$ respectively). The correlation coefficient compared features and PANAS values from from the HLM in order to control for inter-personal effects as well as from each participant individually to highlight the differences in communication style. The standard deviations (σ) show a variability in the level of correlation between emotional state and messaging data for individual participants. Max and Min refer to the largest and smallest correlation coefficient for the individual. Features that did not use any content from received messages are italicized.

emotional distress, as Al-Mosaiwi's et al. found that those experiencing depression, anxiety, or suicidal ideation are more likely to use first person singular pronouns, and words directly related to feeling [1].

4.5.2 Feature Analysis

A Pearson correlation for each feature was calculated for each individual as well as across all 25 participants using the HLM to normalize for interpersonal differences, the results of which can be seen in Table 4. As expected, there were stronger correlations to negative affect with engineered features than positive affect. This may demonstrate that participants were more likely to talk about negative topics in messages, as supported by past research where direct messaging content was more intense and negative than public social media [5]. Building on Bazarova et al.'s work, these findings show that there is a correlation between specific messaging patterns and negative affect.

Few features were significant within the HLM. For predicting negative EMA scores, the number of messages received from the top 5 friends ($p < 0.05$) was significant. The number of messages received from the 5 closest friends had a large negative correlation to negative affect ($r = -0.65$), suggesting that active support networks and higher levels of connection are an important predictor to mood. For predicting positive EMA scores, the sentilength of sent messages was significant ($p < 0.05$) with a coefficient of 0.28 suggesting that a higher number of messages sent with exceptional sentiment are correlated with higher positive affect. Mean response time was also significant ($p < 0.01$), but had a relatively small coefficient of -0.02. These different classes of features that are significant for predicting positive and negative affect suggest that specialized models and features need to be created for each task.

The highest correlation seen for an individual participant was the count of emotion words in messages sent to the participant by the five people that they communicate with most often ($\rho = 0.45$). This could possibly be explained by co-rumination, where people tend to repeatedly discuss negative feelings with their peers. High levels of co-rumination has also been shown to be a predictor for depressive episodes [59]. Surprisingly, the same feature also had the lowest correlation ($\rho = -0.45$) suggesting these participants mainly talked to their close peers while they were not feeling negative. These opposing results for the same feature demonstrate the significance of Sochiatrist for helping clinicians distinguish between patients' behavior patterns to better

inform treatment strategies.

Nonetheless, while there is variability on an individual level, the correlations of the data in aggregate suggest that certain affective patterns exhibit themselves on a cross participant level, such as talking with close friends in response to (or as a result of) an increase in negative affect. This also suggests that with any general model, there is high variability in how well we can infer emotion, with predictive accuracy potentially being dependent on external factors, such as the rate of use of social media, or the personality and coping style of a specific individual. This is further addressed by the time-series analysis section.

4.5.3 Affective Modeling Results

In order to asses model performance, the cross validated root mean squared error (CV RMSE) was calculated for each type of model. The additional 12 participants from the clinical context were used to test the generalizability of these models. To do so, every model was trained on all of the college dataset and then run against the clinical participants. Predicting negative affect had better performance, with the HLM having the best performance when using the engineered features (RMSE = 7.2 ± 2.4, 4.0 ± 2.0, 5.3 ± 3.6 for the LM, HLM, and SVR respectively). In general predicting positive affect was less accurate, with the HLM marginally having the best performance (CV RMSE = 5.0 ± 1.7). For the models using ELMo, performance was improved across the board when predicting both positive and negative affect, but only by a slight margin. A full list of results can be seen in Table 5.

On a 50-point scale, this generates a good estimate if participants are feeling strongly, moderately, or weakly negative based on messaging features alone. While the HLM initially seems to have the best performance, there is a risk of overfitting on personal features as the model is heavily indexed towards the study sample. Prediction of negative PANAS scores had lower error compared to positive affect. This is consistent with Bazarova et al.'s finding that individuals tend to describe more negative emotions in messages [5] and Sun et al.'s finding that the use of phrases with positive sentiment is not necessarily related to positive affect [60].

The linear model ultimately produced a weak fit when using the engineered features ($R^2 = 0.21$, RMSE $= 7.2 \pm 2.4$ for negative affect and $R^2 = 0.20$, RMSE $= 7.4 \pm 2.2$ for positive affect). While all features used resulted in a weak to moderate correlation to negative affect scores, only a select number of features resulted in significant standardized coefficients. Likewise, some features that had a high level of correlation with the aggregated dataset, such as the number of emotion words sent, were insignificant in the linear regression model. This suggests that while correlated on a small-scale, they have little predictive power for specific negative PANAS scores. When trained on the features generated using ELMo, the performance increased marginally for both negative (CV RMSE: 6.0 ± 2.8) and positive affect (CV RMSE: 7.5 ± 2.3). However the proportion of explained variance remained low (negative = 0.16, positive = 0.11). Thus, a linear model likely makes poor assumptions for modeling affect.

Using the HLM with engineered features to control for individual differences and find overarching messaging and mood patterns, we found a fixed intercept of 15.6 with a standard deviation of 7.3 across users for negative affect and a fixed intercept of 17.6 with a standard deviation of 5.6 for positive affect. This suggests while people's baseline negative PANAS scores tend to fall around this point, it is not unusual for there to be significant differences between the baseline PANAS scores of different individuals. The varied intercept was highly significant ($p < 0.001$), suggesting that an accurate estimation of baseline mood is necessary for affect prediction. When tested on the clinical data, there was a decrease in performance, which could be explained by over-fitting to personal effects within the undergraduate dataset. Given the different demographics of the clinical group and the over reliance on personal effects, it is unsurprising that the HLM failed to generalize as well as other models.

The HLM has a low proportion of variance explained by the fixed factors ($R^2 = 0.023$ for negative affect), which describe factors that generalize across multiple participants. However, the proportion of variance explained by both the fixed and random factors was much greater ($R^2 = 0.62$, RMSE $= 4.0 \pm 2.0$ negative affect). This took into account both the generalized factors as well as the personal factors that could affect EMA scores. In context, this result demonstrates that the features explain a small proportion of the variation that is seen, while when taking individual's random factors into account drastically increases the explained variance of this model. This further emphasizes that individual traits, such as messaging habits and baseline mood, have a large effect on inferring negative affect from messaging data. When trained on the features

	All Messages		Sent Messages	
Model:	**CV RMSE**	**Clinical RMSE**	**CV RMSE**	**Clinical RMSE**
LM	7.2 ± 2.4	8.1	6.0 ± 2.7	7.6
HLM	4.0 ± 2.0	8.0	4.0 ± 2.0	8.0
SVR	5.3 ± 3.6	6.5	5.0 ± 3.2	7.8
ELMo LM	6.0 ± 2.8	6.7	6.7 ± 2.7	6.7
ELMo HLM	3.9 ± 2.0	8.5	3.9 ± 2.0	8.5
ELMo SVR	4.2 ± 3.1	7.0	4.25 ± 3.0	7.1

Table 5: In models using engineered features, the HLM has the lowest cross-validated root mean squared error (CV RMSE) for predicting negative affect in the College group, but SVR outperforms other models for the clinical group. Similar results are seen using ELMo derived features, where the HLM had the best performance on the collage group, and the linear model on the clinical group. By using features from sent messages and evaluating the models on the clinical data, the model performance either improved or was only marginally impacted, suggesting that the received features add noise.

derived from ELMo, there was little improvement compared to the engineered features (negative CV RMSE: 3.9 ± 2.0). This provides more evidence that predicting the individual differences and EMA baseline for individuals is incredibly important for EMA modeling.

For the SVR, an eps-regression with an RBF kernel was chosen to determine the maximum potential performance of the model by not bounding the number of support vectors. This is done at the cost of generating a more complicated model. From training, the model identified 836 of the data points as support vectors and achieved an RMSE $= 5.3 \pm 3.6$ for negative affect. A similar number of support vectors was selected when modeling positive affect and ELMo features. The high number of support vectors suggest that there is little regularity within the dataset, which is to be expected given the high number of features and irregularity within the textual messaging data itself. The high number of support vectors is also unsurprising when considering the noisy nature of the data. Training a SVR on the ELMo features increased performance for both negative (CV RMSE: 4.2 ± 3.1) and positive (CV RMSE: 5.3 ± 1.6) EMA. When tested on the clinical population, the performance of the SVR was similar to the metrics generated from the college student population. Given the formulation of the SVR, this suggests that there are a set number of characteristic messages that predict emotion moderately well, and that characteristic messages are applicable to multiple groups.

4.5.4 Affective Classification

As demonstrated by the HLM, baseline affect plays a significant role in determining the performance of the predictive models. This affective classification for when participants feel more negatively than positively helps control for these varied baselines by simply comparing the relative values of positive and negative affect. In order to asses the performance of these models, the overall accuracy and F1 score are calculated.

Basic logistic regression using the engineered features performed moderately well, achieving a cross-validated accuracy of 0.73 ± 0.23 with a moderately high F1 score of 0.81. The random forest had lower cross-validated accuracy (0.64 ± 0.17), and a lower F1 score (0.73), suggesting that it overfits to the training data. AdaBoost had similar performance to the logistic regression (CV RMSE: 0.70 ± 0.20) with a similar F1 score (0.80). In general, the classifiers trained on the ELMo features had similar accuracy to the models using engineered features. The best model tested using all features ELMo was the logistic regression, achieving a cross-validated accuracy of 0.71 ± 0.23. These results can be seen in Table 6.

The classification task suffered from degraded accuracy on the clinical population, with accuracy dropping roughly 10% across the board. Like in the college group, the logistic regression and AdaBoost performed similarly with 62% accuracy and the random forest resulted in the lowest accuracy (52%). These comparisons can be seen in Table 6. As with the classification task, the best performance on the clinical group was achieved by the logistic regression using ELMo derived features (accuracy of 72%). This degradation seems to suggest that such a classification task is more sensitive to personal messaging habits within groups. Given the noisy nature of mood data and the unique messaging styles of individuals, these initial results are promising.

	All Messages			Sent Messages		
Model:	**CV Accuracy**	**F1**	**Clinical Accuracy**	**CV Accuracy**	**F1**	**Clinical Accuracy**
Logistic Regression	0.73 ± 0.23	0.81	0.62	0.74 ± 0.23	0.82	0.68
Random Forest	0.64 ± 0.17	0.73	0.53	0.65 ± 0.18	0.77	0.54
AdaBoost	0.70 ± 0.20	0.80	0.62	0.70 ± 0.23	0.80	0.64
ELMo Logistic Regression	0.70 ± 0.23	0.79	0.67	0.72 ± 0.23	0.80	0.71
ELMo Random Forest	0.67 ± 0.20	0.75	0.61	0.64 ± 0.21	0.73	0.62
ELMo AdaBoost	0.65 ± 0.19	0.73	0.63	0.64 ± 0.20	0.73	0.66

Table 6: The best performance on the classification task was achieved by the logistic regression trained with engineered features based off only sent messages. In general, limiting analysis to sent only messages improved performance. While the models based on ELMo features had similar performance to those with engineered features, they suffer in terms of interpretability of features.

4.5.5 Models Based off Sent Features

While also tested with positive affect, we will only discuss the results models on sent features for negative affect prediction given the better performance in predicting negative affect. Results were similar when using positive affect. For the regression task, models limited to features derived only from sent messages performed roughly equally (CV RMSE $= 6.0 \pm 2.7$, CV RMSE $= 4.0 \pm 2.0$, CV RMSE $= 5.0 \pm 3.2$ for the LM, HLM, and SVR respectively). The same change in performance can be seen within the classification task, with the best classifier being the logistic regression using the engineered features from only sent messages (CV accuracy 0.74 ± 0.23). Comparisons for each task can be seen in Table 5 and Table 6.

When evaluating models based off of solely sent messages on the clinical population's data, the models' performance either remained the same or increased for the regression task. For the classification task, the accuracy of clinical predictions increased across the board. This can be seen in Tables 5 and 6. This finding suggests that basing models off of sent features could be more generalizable. It is likely that the variation in received messaging behavior and the lack of correlation to EMA scores results in additional noise, which impacts performance. However, there is a risk that this is simply due to the small sample size of the validation set.

This small discrepancy between just using data originating from the participants and using all available messaging data for both tasks suggests that solely sent messages are sufficient to model affect. Thus, future work with affective modeling with messaging data could just utilize sent messages, avoiding some of the ethical concerns discussed later.

4.5.6 Time-Series Analysis

To isolate the effect of previous emotional state on future emotional states and to contextualize the relationship between messaging data and affect, we performed a time-series analysis on the isolated PANAS data. An autoregressive moving average (ARIMA) model was used to see if past PANAS values could accurately predict future scores. As part of the analysis, each individual had their own ARIMA fit generated for them with optimized parameters for that person. This identifies a bound of using just using past scores in models as well as any cyclical patterns in EMA scores. The results of this analysis suggest that only a moderate amount ($R^2 = 0.37$) of the variation in PANAS score can be predicted by previous values, showing the inefficiencies of using previous scores alone.

Emotional state is subject to sudden changes and did not seem to have a clearly identifiable pattern over the course of two weeks of data collection. Indeed, the best time-series forecast, an ARIMA model with parameters ($p = 1, d = 0, q = 0$), resulted in a simple linear model that predicted PANAS scores based on the single prior observation. While affect is dependent on previous assessments, this demonstrates the need to utilize additional feature sources. Intuitively this makes sense due to the influence of external factors on affect [40]. Similar results were obtained for positive affect.

4.6 Discussion

Although accurately predicting the exact PANAS is challenging in a general model, the performance of the models based off of direct messaging data gathered using the Sochiatrist system demonstrate the viability of using this data in challenging social computing tasks. In addition, we have shown that these models can generalize across both college and clinical groups.

The Sochiatrist system and associated experimental use of collected data demonstrate the feasibility and value of working with direct messages. On a broader level, the streamlining of this form of data collection with user consent could lead to more computational and feature-based methods of analyzing participant affect, thereby augmenting health-care professionals' traditional in-person clinical assessments of mood and affect.

Besides reporting predicted EMA, engineered features that have significant correlations could be reported to clinical psychologists, along with traditional EMA based methods of self-reported affect. Instead of an automated algorithm, a clinician could use their interpretation of these features, along with the corresponding correlations with an individual's emotional state, to better inform treatment strategies [63, 16]. The emotional and affective states of patients outside of the office have been demonstrated to reflect the onset of long term mental health issues [14]. While EMAs have been traditionally used to predict emotional state in the past, the burdensome nature of having to continuously take various questionnaires often results in low compliance rates [31, 39]. By utilizing the Sochiatrist system, clinicians can learn about the emotional and affective states of patients outside of the office, which have been demonstrated to reflect the onset of long term mental health issues, in a sustainable way without impeding patients' normal lives [14].

Separately, message-based statistics, such as the number of unique users messaged since the last psychotherapy session, could be used by clinicians as a less invasive mechanism to explore a patient's past experience and serve as a method of reducing the impact of difficulties in memory recall [66]. Messaging statistics could also alert a clinician to a patient withdrawing from social connection, experiencing bullying, or receiving low social support without analyzing the content of a patient's messages.

When comparing the results between a feature engineering approach and approaches that utilize machine learning methods, it is important to consider the interpretability and comfort of clinicians in trusting the results from black box algorithms. While ELMo and other deep learning methods may provide slightly improved performance, the engineered features are tied to accepted literature within the clinical sphere. This additionally holds in the methodology and complexity of models. In this case, with the relatively close performance of each embedding method, a more conservative approach may be better in a clinical setting. While using ELMo for the regression task resulted in slightly higher performance (CV RMSE of 5.3 ± 3.6 and 4.2 ± 3.1, clinical validation RMSE 6.5 and 7.0 for the engineered features and ELMo features respectively), there was no similar improvement when used on the classification task. As a result, in future work the benefits of using older but more interpretable methods over newer machine learning methods must be evaluated on a case by case basis.

While public posts were not collected for the college group and were not used in the affect prediction task, public social media data from Facebook, Instagram, and Twitter was collected from the clinical group. Only 9 out of the 12 participants had public data, and this data only made up 2.8 percent of all data. Public social media usage was also incredibly right-skewed, with the median posts being 13, with a maximum of 492. Given the larger amount of data in private messages and the more constant messaging behavior, private messages provide a more granular input for time series data compared to public posts. With both private and public messages, there are 433 EMA periods with content, but when solely looking at public data there only 104 of the EMA periods are non-empty. Additionally, these non-empty blocks only have 5.3 pieces of content in them on average. With the higher usage at both a participant and time period level, messaging data has significant advantages when collecting social media data when having data for a given participant is important.

As noted in the results section, using features derived from ELMo resulted in an increase of performance in the regression task when compared to the engineered features. However, it is important to recognize that these improvements are relatively small in the task, and even smaller in a practical sense. This is supported by a similar performance in the classification task, given that the difference between positive and negative affect at inversions points is typically larger than 5. As a result, the interpretability of engineered features could be a worth-while tradeoff with outright in a clinical setting. In future work, the trade-off between interpretability and performance must be considered.

These results highlight the potential impact and usefulness of analyzing the content and features of messaging data in future work. Even with relatively simple models and features, the SVR was able to predict mood with a CV RMSE = 5.3 on a 50 point scale. With the off the shelf nature of the data extractor, such models could be deployed in a clinical setting where messaging features are used to retroactively determine affect.

4.7 Study Limitations

Given that emotion is both subjective and noisy, it is difficult to tell the upper bounds for accuracy for these particular models. Several participants noted that they occasionally found it difficult to assess their feelings in the moment and may have under-reported their affect as a result. With additional participants, it would be easier to understand the tolerances of affect. Future work with larger samples in different, diverse populations will better understand the impact of these biases.

The length of the study was shorter than previous studies on mood and emotional state, which had data collection periods that ranged from 30 days [41] to 3 years [55]. Additionally, to examine the specific utility of direct message data as a space for people to express their most intense and private emotions [5], we purposely chose to test the extreme case where no public data is used at all, but in practice the system can extract both public and private data for analysis. That being said, having more data, from both the perspective of time and the number of types of data used, could have increased the correlations we found. There is a need for additional work further analyzing and comparing analyses of both direct messaging data and public social media data.

4.8 Summary

Direct social media messages have become a platform for expression of intense, private emotions [5], but current research has left this data space relatively unexplored until the development of the multi-platform extraction and anonymization research tool. To understand the relationship between direct messages and emotional affect, we built and tested models that used the message and EMA data from 25 participants and predicted their individual affect. We have also developed an alternative for predicting emotional affect that places a lower burden on participants and has a higher compliance rate than the current clinical method of Ecological Momentary Assessments [31, 39]. Of the models tested, the HLM had the best performance originally, at the risk of overfitting towards the study sample, but the SVR proved most effective at being a generalizable model during validation using a clinical population. Given the importance of understanding patient's emotional affect outside of sessions, clinicians and researchers will be able to use the interpretation of features along with the correlations with participants' emotional state to inform treatment strategies for patients and research participants.

5 Drug Detection

5.1 Introduction

One of the challenges of using messaging data in a clinical psychology context is analyzing the data in a scalable, qualitative way. Current best practices require hand labeling of all data by multiple people, which is time consuming and expensive. In addition, given the given sensitive nature of data, crowd sourcing labels through services like Amazon Mechanical Turk are not possible. Thus, the iteration time of answering questions about the content of the data is slow, limiting the ability to discover qualitative insights about the data. One such "coding" task is to determine of messages contain mentions of substance use.

This task is particularly challenging, given that terms are often overloaded with multiple meanings that can change irrespective of the structure of the sentence. For example, "coke" could refer to the popular soft drink or an illicit substance. The class imbalance of this data is also difficult, as only a small number of messages contain drug references, which make machine learning approaches impractical.

Using the Sochiatrist Data Extractor to gather data on 92 adolescents, we propose using a semi-supervised approach to gather and query messages related in topic to questions asked in clinical surveys. This approach uses keywords gathered from existing clinical psychology surveys in conjunction with anchored correlation

explanation [22] to flag messages with drug mentions. These selected messages can then be passed to a research assistant to verify results and remove false positives. In this scenario, recall is much more valuable than precision, as flagging false positives is relatively easy for a human, but finding false negatives would require a reading of all messages. From a clinical perspective, this additionally provides better understanding to the clinician of events happening out of session. While drug use was focused on due to concrete labels, these methods have been applied to detecting positive social support, discussions of trauma, and PTSD symptoms.

5.2 Related Work

5.2.1 Clinical Psychology "Coding"

Within clinical psychology, the process for determining labels for individual messages is refereed to "coding." A typical approach is to use a piece of software, such as NVivo to manually highlight and annotate text. This process can be open or follow a template, with both methods being widely accepted but applicable in different scenarios [6]. Once these "codes" are generated, they can be exported from the program into statistical analysis software [57].

With the struggles of coding well documented, some attempts have been made to partially automate coding. Marathe et al. found that in order for semi-automated coding to be helpful, clinicians needed to retain control over their labels, as well as have easy to understand reasoning behind the suggestions. They tested simple keyword matching, augmented keyword matching, and logical combinations of keywords. The logical filtering of keywords achieved 88% precision and 82% recall [43]. Yan et al. was able to use an active learning approach to achieve 70% precision but only 8% recall for coding tasks [37].

5.2.2 Seeded Topic Modeling and Classification

Much work has gone into developing topic models and document classification, with both unsupervised and semi-supervised methods being explored. Latent Dirichlet Allocation (LDA) remains a popular choice for generating topics in an unsupervised manner. However, past work has shown that LDA struggles with short texts due to the sparcity of data within documents [25]. While some methods have attempted to improve the accuracy of LDA on short texts such as messages through pooling[44] and using Gaussian mixture models[58], these unsupervised methods are ill suited to the drug detection problem given the low number of messages that contain mentions.

Semi-supervised approaches to topic modeling have also been explored, typically with an emphasis on document classification. Jagarlamud et al. and Chen et al. have explored altering the LDA algorithm to incorporate seed words when calculating topic distributions[29, 10]. Seeded topic modeling has also been explored with non generative models, which have been been show to be more effective at generating topics form short text documents[22]. Seed words have also been used for document classification, where word co-occurrences have been used to generate both general and document level topics[34, 35].

5.2.3 Drug Use Detection

The issue of drug detection in social media has also been explored, but only focusing on public social media and with the goal of preventing illicit substances from being advertised and sold on cerain platforms. Cameron et al. developed PREDOSE, a tool which scraped the web searching for drug mentions. However, this approach required a highly organized and refined ontology in order to detect drug mentions [8, 32]. Other tools have been created to passively monitor sources for drug mentions, but required manual filtering and analysis of data [65]. Other approaches have used structured heterogeneous information networks derived from Twitter user data to detect opioid addiction using social network graphs [17]. However, little work has been done in detection of drug mentions at an individual level.

Platform:	**SMS**	**Facebook**		**Instagram**		**Twitter**		**WhatsApp**
	Messages	Messages	Posts	Messages	Posts	Messages	Tweets	Messages
Number of Messages	133,890	27,389	684	2,850	455	57	850	1,828

Table 7: With the drug use group, SMS was by far the most popular messaging platform. As seen with the other clinical population, the number of messages (166,136) greatly exceed the number of public posts (1,866)

	Nicotine	**Alcohol**	**Cannabis**	**Stimulants**	**Sedatives**	**Cocaine**	**Opioids**	**PCP**	**Hallucinogens**
Number of Participants	13	14	19	0	1	3	1	0	0

Table 8: Out of all the measured drug categories, cannabis was discussed by the most participants, followed by nicotine and alcohol. This uneven distribution also meant that the number of messages were biased towards these more popular substances.

5.3 Study Procedure

5.3.1 Data Collection

Using the Sochiatrist Data extractor, private and public data was gathered from adolescents between the age of 13-18 who had experienced a traumatic event that had placed them in the ER. This data was collected by clinical collaborators at the Bradley Hasbro Children's Research Hospital for an ongoing study. Private and public social media data was retroactively collected from two weeks before to two weeks after the incident on Facebook, Instagram, Twitter, Kik, and WhatsApp.

During the course of the study, the participants completed a series of surveys that recorded their drug use. The Adolescent Alcohol and Drug Involvement Scale (AADIS) and Modified Version of the Fagerstrom Tolerance Questionnaire (mFTQ) were self reported and The Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS) was facilitated by a research assistant. These surveys asked about use of nicotine, alcohol, cannabis, stimulants, sedatives, cocaine, opioids, PCP, and hallucinogens. The keywords for search were derived from the categories and examples within these surveys, which can be seen in Appendix A.

5.3.2 Data Analysis

In total, 168,003 pieces of social media data were gathered from the 93 participants between January 4th, 2017 and February 8th 2019. Demographic and phone data was not provided. Out of these, 1,866 were public posts, 72,979 was sent content, 94,545 pieces of content were received.

Drug mention labels were generated manually by reading messages and validated between two research assistants. 28 of the participants mentioned drugs as classified by the drug use survey with 515 messages containing explicit substance mentions. A breakdown of drug specific mentions can be seen in Table 8.

5.4 Topic Searching

In order to search for drug mentions in the data, keyword search and a semi-supervised topic model were explored. For all tests, the keywords remain the same. In order to asses the performance of each of these methods, the number of messages correctly flagged by the algorithm and the number of false positives were recorded. This was then compared to the performance of picking these messages by chance.

For the keyword search approach, a few different methods were tried. First, a basic search was performed over the messages using the keywords derived from the clinical surveys. In this method, a regular expression was used to detect keywords regardless of capitalization. A fuzzy search was then performed across all tokens, allowing for slight misspellings to be captured. A match was determined by dividing the Levenshtein edit distance by the alignment length and was considered a match when the distance ratio was greater than 0.9. This cutoff was chosen after doing a parameter search to find the point when the number of correct messages stopped increasing. The messages were tagged with their part of speech using the `nlkt` part of speech (POS) tagger and tokens were compared with the keywords as segmented by part of speech.

One downside of this searching approach is that it does not capture the latent correlations between words. In order to search for underlying distributions of related words, Correlation Explanation (CoreEx) was used

Model	**Precision**	**Recall**
Basic Search	0.07	0.34
Fuzzy Search	0.06	0.43
POS Fuzzy Search	0.35	0.33
CoreEx with 1 Class	0.14	0.48
CoreEx with 15 Classes	0.11	0.91

Table 9: The best model performance as measured by recall used CoreEx with 15 individual classes for each type of substance. The high accuracy of this approach is likely attributed to the capturing of new types of e-cigarettes in the topics.

with anchor words to identify if there were any additional underlying distributions of words that would be helpful in determining topics. CoreEx was chosen due to it's higher performance in modeling topics in short texts compared to LDA [22]. Before being fed into the model, English stop words were filtered out. Two approaches were used with CoreEx, with one topic seeded with all drug use phrases, and one with 15 topics seeded for each individual category of drug.

5.5 Results

Out of the models that used search based approaches, the unrestricted fuzzy search had the best performance as measured by recall (0.43). However, when the fuzzy search was forced to consider parts of speech when tagging, the precision increased dramatically from 0.06 to 0.35 at the expense of recall dropping from 0.43 to 0.33. The decrease in performance could be explained by the informal nature of messaging conversations, which often times do not follow formal writing conventions. Thus, part of speech taggers assuming these formal conventions could struggle with the correctly parsing messaging data. Additionally, this demonstrates that placing restrictions on search parameters is a effective way to increase precision, but often at the expense of recall. A full list of results can be seen in Table 9.

Out of the semi-supervised approaches, the CoreEx model with 15 classes performed the best, with a recall of 0.91, compared to a recall of 0.48 for simply using one class. Precision for both was similarly low. Much of the improvement with the 15 class topic model was caused by the capturing of the word "Juul" and "pod," a new type of electronic cigarette and its refill. With just a single class, the algorithm struggled to identify these substance specific actions and relations. This highlights an issue with using clinical surveys as keyword sources as they often are not often representative of the current phrases that are being used for certain topics. However it is a good demonstration of how semi-supervised methods can overcome these issues.

5.5.1 Public vs. Private posts

When comparing the public to private data in this dataset, the importance of private messaging when detecting rare events is evident. In total, only 53 out of the 92 participants had any form of public data, which only made up 1.1% of all data gathered. In addition, given the already small number of messages with drug mentions, only 8 public posts contained any drug mentions from 4 participants. This shows when searching for uncommon and specific mentions within social media, the additional volume provided by private messages allows us to build a more in depth representation for a greater number of individuals. This is particularly important if such a tool were used in a clinical setting, as a tool that could only be run on 60% of patients would be largely ineffective.

5.6 Discussion

When searching for messages, the semi-supervised topic modeling approaches outperformed the basic search tasks. This is largely be explained by those model's ability to detect and use the latent relationships between words, such that messages that don't have keywords in them can still be detected. Splitting the topics into 15 individual categories also improved performance, which was likely due to the increased granularity of the

topics. For example, the word "smoking" would be related to nicotine use but not alcohol use. When lumped together, these individual relationships can be lost.

One area where all these methods struggled was with drug use phrases that have other, more common uses. For example, "speed" is listed within the clinical surveys as a slang word for Amphetamines. However, in most cases the word "speed" is associated with the rate of movement and thus the semi-supervised topics converged around transportation items such as "car", "train", and "limit". These categories with overloaded phrases made up a mast majority of false positives in the CoreEx approach, and have the opportunity for future refinement.

None of these approaches was able to capture all messages that were labeled as containing drug mentions. As a result, hand labeling is still necessary when high accuracy is required. However, these methods take multiple orders of magnitudes to run (seconds versus weeks). This enables RA's to answer questions about the content of messages at a much higher rate. These methods can ultimately be used to query messages based on topics in a fraction of the time required to hand label sensitive social media data. In addition to saving time and energy, filtering messages using this message also means that fewer messages need to be read by research assistants. Such privacy improvements are necessary to have messaging and social media data make an impact in a clinical sphere.

The recall of this approach are comparable to the results of Marathe et al. and Yan et al. but failed to match the precision of Marathe et al.'s logic based keyword search [43, 37]. However, both of these methods required continuous human feedback. Whether the additional time spent fine-tuning the models is more efficient than checking for a number of false positives still needs to be evaluated with user studies.

5.7 Limitations and Future Work

While these approaches provide a solution to improve the ability to query messages from specific topics, there is still much room for improvement. One area that still needs to be explored is the ability of these techniques to capture more abstract concepts. For example, depressive symptoms or instances of positive social support. In order to better search for these messages, it could be beneficial to include sentiment as classified through VADER into the detection process. Another possibility is to create a more rule-based querying process. For example, examples of positive social support should have largely positive sentiment, and thus items that have large negative sentiment could be discarded automatically. For topics that are more common, active learning approaches could also be explored, where a machine learning model is tuned through human input already required for filtering false positives.

These methods are also ultimately still techniques that need a level of coding expertise to use. In order to make an impact in clinical research, tools and systems need to be developed using these methods such that they are accessible to everyone. When user interviews were performed with clinical psychologists, all expressed dissatisfaction with their current "coding" tools. New tools are necessary to make messaging data usable and accessible to a level where it could provide clinically impactful information to a nontechnical audience.

5.8 Summary

In this section, the usage of seeded topic modeling for message querying was explored using data from adolescents after traumatic incidents. Keyword searching and semi-supervised approaches were tried, with a CoreEx model achieving 91% recall. However, precision remained poor across all tested models. Such methods provide ways to quickly query large amounts of data to unearth messages that are related to clinically important surveys and topics without the need for human input, both decreasing the effort required to flag messages as well as decreasing the number of messages that need to be read, increasing the privacy of participants.

6 Ethical Considerations

As demonstrated by these tasks, private messaging data is a powerful, yet underutilized data source for clinical psychology and behavioral work. However, given the sensitive nature of messaging data, one must always consider the ethical implications of using such a data source. In this next section, we will explore

some of the ethical concerns with utilizing this data, and how the Sochiatrist Social Data Extractor is built with a privacy and consent first approach.

The first way that consent is built into the system is through the process required for each platform to be extracted. Participants must be physically present throughout the whole process in order to enter their own user information after consenting to extraction for each platform. No login information, including authentication tokens, are stored on disc. Additionally, all data is anonymized immediately after extraction to further ensure patient privacy. This data enables researchers to computationally compare sent and received messages and look for hints of contagion. Clinicians can gain a deeper understanding of patients' emotional states without the need for invasive procedures like forcing them to explain and relive traumatic experiences or through monitoring patients' social media profiles without prior consent. When using messaging data, it is important to balance comprehensiveness and privacy. These procedures were additionally reviewed by the NIMH Human Subjects Department as well as Brown's institutional IRB.

Given the structure of the system, we secure user privacy by automating the analysis of the messages so there is no direct need for researchers to read the message contents. Patients are in full control of what platforms they will allow to be analyzed and must give full consent in order for their messages to be extracted and their emotional state to be successfully inferred. This emphasis on participant agency is why the ability to blacklist conversations was introduced. While it may remove data that could be useful for analysis, it ensures agency is placed on the participant to share what they are comfortable with.

For the clinical groups, the researchers are not part of the treatment team. Though the treatment team is consulted to confirm that it is acceptable to approach a family about the research, a key part of the consent and assent procedures involves a process of making clear to families that the research is distinct from the clinical care. Further reinforcing the separation between the data collection of this research and clinical care, families are approached and consented while inpatient but procedures related to ecological momentary assessment (EMA) and assessment of social media use all occur after discharge. As with any research involving informed consent, it is possible that participants may have changed any number of behaviors given their awareness of being monitored and a key part of the consent procedures involves making clear to participants that we are only interested in information that teens feel willing and comfortable to share.

We also want to consider concerns about analyzing third party data, the messages written by participants' contacts instead of by the participants themselves. While separate from ethical considerations, U.S. courts have established that the sender of text messages [54] and other electronic communications [47] loses a reasonable expectation of privacy after the message has been sent to the recipient of the message. In United States v. Nosal and Facebook Inc. v. Power Ventures Inc., the courts have opined that consenting to an external analysis of one's personal data from a social media website does not violate the Terms of Service of the website until the company explicitly revokes permission. We have followed these guidelines to the best of our ability, but issues of user agency over their data require larger public discourse.

Additionally, since we don't have any third party messages or social media content beyond what has been sent to participants, it would not be possible to analyze third party senders using the Sochiatrist system. All data is anonymized to further mitigate user privacy concerns. Nonetheless, as we showed in the analysis, predictive models performed only marginally worse when limited to features derived from sent messages. This presents itself as a way to provide similar clinical impact without the need for analyzing third party messages. These findings open the door for additional work using message content without the mentioned privacy concerns.

The trade-off between continued research and privacy must be also considered if the extractor is used in clinical tools. Given the current research goals, all content is downloaded such that new analysis techniques can be developed and tested. However, in order to preserve the privacy of participants within a clinical context, one could only harvest feature vectors and never export raw text messages. If deployed into a clinical setting, the balance between privacy and continued research work must be carefully weighed.

While messaging data holds much promise in its use in a clinical sphere, it is important to consider a cost benefit analysis when deploying such systems. While the Sochiatrist Data Extractor is built from a participant consent focused framework, it does not mean that data can not be misused. One way that these risks are minimized is through the continued involvement of clinical psychologists in the study and design process who can provide a better perspective on the benefits that certain use-cases could bring to the table.

7 Future work

While the public and private social media data extracted using Sochiatrist have shown promise in a variety of tasks, there is still much work to be done before such a tool would be ready to launch in a clinical setting. While accessibility towards non-technical audiences was a focus when building the tool, the use of a script as the interface is only accessible to trained clinical researchers. In order to be adopted on a wider scale, the experience of using the extractor is required. Work is currently being done to create a GUI for the tool such that minimal training for the tool is required. There are also current plans for the extractor to support additional platforms such as Snapchat. In addition to making the extractor easier to use, more work must be done to translate the insights provided by the features and models into the hands of psychiatrists.

As discussed in the previous section, ethics and privacy play an important role within this work. In order to make a clinical impact, future tools and methods need to take a proactive approach in addressing these issues. While privacy is integrated into the extractor, privacy needs to play a larger role within the analysis of the data and presentation of results to clinicians. Specifically, the balance between preserving privacy by not making messages viewable with the explainability of showing important messages to clinicians needs to be explored. These challenges must be addressed both at a modeling level as well as a user interaction level in order to ensure that these two goals are weighted properly.

While some comparisons between public and private data was performed, there still is more work needed to understand how these two forms of content relate to each other. In addition, the differences in messaging behavior across platforms is still unknown. Such insights could be helpful for tasks such as affect prediction and weighting public and private posts differently. This weighting of different messages is also important, since much of messaging data can be considered noise for downstream tasks. New methods of filtering this noise and weighting important messages also needs to be explored.

Given the dense nature of this sort of data, there is more work to be done in developing novel models specifically targeted at messaging data. Specifically, the temporal nature of talking with others, the social graph, and content of messages were all explored separately with the engineered features in the affective modeling section. However, these unique aspects of the data should be integrated with models at a more basic level to capture the unique structure of the data.

While the use of private messages as a data source is promising as demonstrated by our tasks, there are many unanswered questions at both a basic level as well at an applied level that will required more work to understand. However, with the Sochiatirst Data Extractor, the overhead in collecting the data to answer these questions has never been as straightforward.

8 Conclusion

The Sochiatrist Social Media Extractor creates a robust, scalable way for clinical psychologists to collect public and private social media in a research setting, enabling access to data that would otherwise be inaccessible. Built from the ground up with patient agency and consent in mind, the tool provides additional protections to research participants through anonymization and name blacklisting. The usefulness of this data has also been demonstrated through the prediction of both positive and negative EMA through engineered features and features derived from pre-trained language models. New methodologies were also explored in raising messages that contain drug use in order to limit the need for hand labeling in a clinical setting. Given the viability of using messaging data as a data source, the Sochiatirst tool opens an avenue for clinicians to more fully interpret patients' lives utilizing passively generated data, as well as providing insights in how mental states are reflected and affected by messaging.

9 Acknowledgements

First and foremost, I would like to thank my advisor Jeff Huang for helping guide me over the years. Without his support, this project would not have been possible. I am incredibly grateful to have had the chance to work on a project that combined my passions of computer science and mental health. I also want to thank my reader Carsten Eickhoff for spending his time on this project.

A huge thank you to our clinical collaborators who have used the data extractor in their studies. I particularly want to thank Nicole Nugent who has provided countless hours of her own time to discuss this project and the extractor and how it fits into a clinical context. The current deployment of the data extractor in clinical settings would not be possible without the incredible RA's who use the system. Thank you for all the feedback that you have all provided that has made the system what it is today. Special thanks to Sara Schulwolf, who came up with the original idea that inspired the drug detection portion of this work.

This project has also allowed me to work with fantastic people here at Brown. Thank you to Jessica Fu, Sachin Pendse, Valintina Cano, Nikita Ramoji, Chong Wang, and Varun Mathur for being amazing! I am also incredibly grateful for all the friends and family who supported me over the years.

Finally I would like to thank Angela Cheng for providing feedback and supporting me through this entire process while putting up with my constant shenanigans.

Appendices

A Drug Use Labels

The following are the seed words used for the drug deteciton task. These were taken from AADIS, K-SADS, and mFTQ:

cannabis, smoking, THC, liquor, stimulants, pot, mescaline, xanax, uppers, blunts, dexedrine, sedatives, blow, percodan, hypnotics, shrooms, oxycontin, hallucinogens, rush, morphine, ecstasy, codeine, barbiturates, LSD, librium, demerol, solvents, glue, cocaine, crystal, paint, alcohol, hashish, wine, diet, freebase, blues, mushrooms, beer, hash, ludes, ectasy, powder, gasoline, PCP, horse, ether, anxiolytics, valium, cigars, chloroform, benzodiazepine, prozac, amphetamines, cigarettes, peyote, weed, heroin, spraycans, meth, grass, methadone, MDA, acid, tobacco, ritalin, smack, coke, downers, speed, opium, whiteout, quaalude, crack, marijuana, opioids, quaaludes, inhalants, pills, psychedelics, rock

References

[1] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, pages 538–539, 2018.

[2] Mawulolo K Ameko, Lihua Cai, Mehdi Boukhechba, Alexander Daros, Philip I Chow, Bethany A Teachman, Matthew S Gerber, and Laura E Barnes. Cluster-based approach to improve affect recognition from passively sensed data. *arXiv preprint arXiv:1802.00029*, 2018.

[3] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte. Sensitive self-disclosures, responses, and social support on instagram: The case of #depression. In *Proceedings of CSCW*, pages 1485–1500, 2017.

[4] Michael F Armey, Janis H Crowther, and Ivan W Miller. Changes in ecological momentary assessment reported affect associated with episodes of nonsuicidal self-injury. *Behavior Therapy*, 42(4):579–588, 2011.

[5] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 154–164, 2015.

[6] Erik Blair. A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences*, 6(1):14–29, 2015.

[7] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, sep 2016.

[8] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. Predose: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics*, 46(6):985–997, 2013.

[9] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005, 2013.

[10] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Leveraging multi-domain prior knowledge in topic models. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[11] Andrew Coles. iPhone backup database hashes: which filenames do they use? *iPhone Backup Extractor: Recover Your Lost Data*, 2012.

[12] Mihaly Csikszentmihalyi and Reed Larson. Validity and reliability of the experience-sampling method. In *Flow and the Foundations of Positive Psychology*, pages 35–54. Springer, 2014.

[13] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, pages 71–80, 2014.

[14] Amy Dryman and William W Eaton. Affective symptoms associated with the onset of major depression in the community: findings from the us national institute of mental health epidemiologic catchment area program. *Acta Psychiatrica Scandinavica*, 84(1):1–5, 1991.

[15] Robin IM Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 6(5):178–190, 1998.

[16] Ulrich W Ebner-Priemer, Janice Kuo, Stacy Shaw Welch, Tanja Thielgen, Steffen Witte, Martin Bohus, and Marsha M Linehan. A valence-dependent group-specific recall bias of retrospective self-reports: A study of borderline personality disorder in everyday life. *The Journal of Nervous and Mental Disease*, 194(10):774–779, 2006.

[17] Yujie Fan, Yiming Zhang, Yanfang Ye, Wanhong Zheng, et al. Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1259–1267. ACM, 2017.

[18] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[19] Carl E Fisher and Paul S Appelbaum. Beyond googling: The ethics of using patients' electronic footprints in psychiatric practice. *Harvard Review of Psychiatry*, 25(4):170–179, 2017.

[20] Andrew J. Flanagin. IM online: Instant messaging use among college students. *Communication Research Reports*, 22(3):175–187, 2005.

[21] Michel Foucault. Madness and civilization (r. howard, trans.). *New York: Pantheon*, pages 265–268, 1965.

[22] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017.

[23] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

[24] Dave Holmes. From iron gaze to nursing care: mental health nursing in the era of panopticism. *Journal of Psychiatric and Mental Health Nursing*, 8(1):7–15, 2001.

[25] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. acm, 2010.

[26] Justin Hunt and Daniel Eisenberg. Mental health problems and help-seeking behavior among college students. *Journal of Adolescent Health*, 46(1):3–10, jan 2010.

[27] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, pages 216–225, 2014.

[28] Thomas R Insel. Digital phenotyping: technology for a new science of behavior. *Journal of the American Medical Association*, 318(13):1215–1216, 2017.

[29] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[30] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 17–33, 2017.

[31] Nikolaos Kazantzis, Frank P Deane, and Kevin R Ronan. Assessing compliance with homework assignments: Review and recommendations for clinical practice. *Journal of Clinical Psychology*, 60(6):627–641, 2004.

[32] Francois R Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W Nahhas, Silvia S Martins, Edward W Boyer, and Robert G Carlson. "those edibles hit hard": Exploration of twitter data on cannabis edibles in the us. *Drug and alcohol dependence*, 164:64–70, 2016.

[33] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[34] Chenliang Li, Shiqian Chen, Jian Xing, Aixin Sun, and Zongyang Ma. Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems (TOIS)*, 37(1):9, 2018.

[35] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. Effective document labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 85–94. ACM, 2016.

[36] Jiguang Liang, Bo Jiang, Rongchao Yin, Chonghua Wang, JianLong Tan, and Shuo Bai. Rtpmf: Leveraging user and message embeddings for retweeting behavior prediction. *Procedia Computer Science*, 80:356–365, 2016.

[37] Jasy Liew, McCracken Nancy Yan, Suet, and Kevin Crowston. Semi-automatic content analysis of qualitative data. In *iConference 2014 Proceedings*. iSchools, March 2014.

[38] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[39] Mark D Litt, Ned L Cooney, and Priscilla Morse. Ecological momentary assessment (EMA) with treated alcoholics: methodological problems and potential solutions. *Health Psychology*, 17(1):48, 1998.

[40] Vicki Liu, Carmen Banea, and Rada Mihalcea. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483. IEEE, 2017.

[41] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. Daily mood assessment based on mobile phone sensing. In *Wearable and implantable body sensor networks (BSN), 2012 ninth international conference on*, pages 142–147. IEEE, 2012.

[42] Pádraig MacCarron, Kimmo Kaski, and Robin Dunbar. Calling Dunbar's numbers. *Social Networks*, 47:151–155, 2016.

[43] Megh Marathe and Kentaro Toyama. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 348:1–348:12, New York, NY, USA, 2018. ACM.

[44] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM, 2013.

[45] Joanne Meredith and Elizabeth Stokoe. Repair: Comparing facebook 'chat' with spoken interaction. *Discourse & Communication*, 8(2):181–207, 2014.

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[47] Brian Mund. Social media searches and the reasonable expectation of privacy. *Yale JL & Tech.*, 19:238, 2017.

[48] Jukka-Pekka Onnela and Scott L Rauch. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41(7):1691, 2016.

[49] Kenneth M Ovens and Gordon Morison. Forensic analysis of kik messenger on ios devices. *Digital Investigation*, 17:40–52, 2016.

[50] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[51] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[52] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, pages 5–23, 2016.

[53] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):95, 2017.

[54] Joel Samaha. *Criminal Procedure*. Wadsworth Publishing Company, 2016.

[55] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *Proceedings of the 26th International Conference on World Wide Web*, pages 103–112, 2017.

[56] Aaron Smith and Monica Anderson. Social media use in 2018. *Pew Research Center*, 2018.

[57] Chareen L Snelson. Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods*, 15(1):1609406915624574, 2016.

[58] Vivek Kumar Rangarajan Sridhar. Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 192–200, 2015.

[59] Lindsey B Stone, Benjamin L Hankin, Brandon E Gibb, and John RZ Abela. Co-rumination predicts the onset of depressive disorders during adolescence. *Journal of Abnormal Psychology*, 120(3):752, 2011.

[60] Jessie Sun, H Andrew Schwartz, Youngseo Son, Margaret L Kern, and Simine Vazire. The language of well-being: Tracking fluctuations in emotion experience through everyday speech. 2019.

[61] John Torous, Rohn Friedman, and Matcheri Keshavan. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth and uHealth*, 2(1), 2014.

[62] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, 3(2), 2016.

[63] Timothy J Trull, Marika B Solhan, Sarah L Tragesser, Seungmin Jahng, Phillip K Wood, Thomas M Piasecki, and David Watson. Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, 117(3):647, 2008.

[64] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(6):1063, 1988.

[65] Paul A Watters and Nigel Phair. Detecting illicit drugs on social media using automated social media intelligence analysis (asmia). In *Cyberspace safety and security*, pages 66–76. Springer, 2012.

[66] JMG Williams and J Scott. Autobiographical memory in depression. *Psychological Medicine*, 18(3):689–695, 1988.

[67] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. *CoRR*, abs/1812.04224, 2018.