Sochiatrist: Inferring the Relationship Between Emotion and Private Social Messages

SACHIN R. PENDSE, Brown University Department of Computer Science

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science, Brown University

This work was done in collaboration with the Sochiatrist team in the Brown Human-Computer Interaction Lab, including Grant Fong, Jessica Fu, Palak Goel, Polina Tamarina, and Jeff Huang.

Abstract

To demonstrate the viability of a platform to extract private message metadata from different sources and consolidate that data into a consistent format, we present Sochiatrist, an automated system that downloads private message data from a group of social media applications, anonymizes the data when possible, and consolidates this data into a consistent format. To demonstrate the utility of this platform, through using Sochiatrist with 25 participants over the course of two weeks, we demonstrate a weak to moderate correlation between private message features and negative affect, demonstrating the efficacy of private message metadata in better understanding the intersections between affect and the use of private social media. A linear model solely using features derived from private social media results in an overall fit (R^2) of 0.16. Through our analysis, we also demonstrate that features such as high amounts of affective language have a moderate level of correlation with negative affect, which may show that people are more likely to express negative sentiment via private messages. To contextualize our findings on the relationship between emotional state and the use of private social messages, we also conduct a time series analysis to isolate the impact of past emotional states on current emotional states from our analysis of private social media and negative affect. Overall, this study shows valuable correlations between the metadata based features of messaging data and emotional state, and the potential for impactful future work.

1 INTRODUCTION

While the most commonly used social media websites [48] were initially created with the intention of objectifying others (Facebook) [24], to share photos with friends (Instagram) [46], and to form communities around shared interests (Pinterest) [21], online social media networks have now become an ubiquitous method for those who have access to connect, communicate, and relate with others. With the global rise of common mental health disorders [40] and prevalent barriers to mental health care [10], strong connections of social support are more crucial now than ever before, and online social media networks enable much of this strong social support for those who might not have access to it otherwise. It has been demonstrated that people with ordinary mental distress, common mental disorders [38] and severe mental illness [37] benefit from being able to voice their experiences and find solidarity with others online. Through a form of empowerment by expression, social media has become a space where self-disclosure of mental distress has become commonplace [2], so much so that even large corporations like The Walt Disney Company have started to produce memes centered around the disclosure of mental distress on social media [16]. The authentic expressions of mental distress and health that social media facilitates make social media a useful space to gain a better understanding of the nuances of *in vivo* mood, affect, and behavior. Some mental health professionals have begun to use information gathering from the social media accounts of their clients to inform how they create treatment strategies for their patients. Though uncommon, this practice, known as the gathering of collateral information, has included monitoring patients' Facebook pages for potential self-harm or for more insight into a potential suicide attempt [19]. The utility of gathering collateral information in better understanding the realities and nuances of patient experience is well recognized by clinicians [19], as is the importance of understanding a client's changing affective states outside of a session [51]. However, situating the practice of gathering collateral information from social media in the context of legacies of the surveillance and institutionalization of those with mental illness [20, 25], it is clear that innovative methods to infer emotional state that *also* empower individual agency over that data are crucial. In the past, much collateral information gathering has been focused around public social media data, but there are significant differences between how people use social media to describe their emotional experience in public and in private [6]. Giving clinicians the power to infer changes in a client's emotional state based on the metadata of private messages would empower clinicians to be able to better understand the day to day affective experiences of their client without the potential for excessive monitoring or unnecessary hospitalization.

In the context of the clinical desire for non-invasive tools to assess the temporal and momentary emotional states of a client, as well as the gap between public presentation and private emotional experience, we present Sochiatrist, an automated system that downloads private social media data from a diverse group of social media applications, anonymizes the data when possible, and consolidates this data into a consistent format for analysis. Using this system with a population of 25 participants over the course of two weeks, we demonstrate the potential utility of metadata inferred from private social media by showing a weak to moderate correlation between metadata based features of private social media and measurements of negative affect. To put these findings into context, we also perform a time series analysis of changes in emotional state over time without use of any private social media data.

Our work has two contributions. First, we demonstrate the viability of the extraction and use of private messages exchanged on social media to infer valuable insights about emotional state. Second, we examine the potential reasons for why certain metadata based features may be more indicative of emotional state than others, and make recommendations for the future use of private messages to infer momentary emotional state.

2 RELATED WORK

In this section, we describe related research on the efficacy of traditional methods of inferring affect and mood, as well as work done to infer and predict affect and mood through the use of mobile sensing data and social media data.

2.1 Clinical Psychology

From the early beginnings of clinical psychology research, the appraisal and evaluation of a participant's affect has often been predicated on self-reported data. These evaluations are often facilitated through questions that force the participant to reflect on past experiences and look retrospectively at how their mood has changed over a period of time in the past. Several studies have demonstrated that individuals experiencing disorders with intense emotional distress as a symptom, such as major depressive disorder or borderline personality disorder, are more likely to have a bias towards negative experiences and feelings when recalling thoughts retrospectively, and have some level of difficulty in remembering specific thoughts [8, 18, 56]. To mitigate the potential for retrospective negative recall bias, the use of Ecological Momentary Assessments (EMA) have become commonly used in clinical psychology research. EMA, also called experience sampling, is a method of collecting data that allows participants to record their thoughts and feelings in a structured and real-time format [12, 49]. An example of a commonly used EMA technique is the use of mobile technology to prompt a research participant to fill out a mood assessment survey or engage in an unstructured thought record at a randomly chosen time [49].

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. For example, in the case of psychotherapy homework, participant attitudes and beliefs towards the task to be completed (such as a strong sense of perfectionism) can affect whether the participant actually completes the task [27]. Other factors, such as the number of questions in an EMA can make participants uncomfortable and less willing to complete an EMA or likely to falsify data [30].

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 and ubiquitous behavioral health data, including GPS based spatial trajectory data, accelerometer based mobility data, and social network data [39, 50]. 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, or what Onnela et al. and Insel dub "digital phenotyping" [26, 39].

Another increasingly popular approach to better understanding participant affect is the gathering of client-specific information through the use of the Internet, such as by patient-targeted Googling [11] or looking at the public social media profiles of clients [19] to get a better understanding of patient experience, particularly when there is the possibility of self-harm or suicide. While this practice is still fairly uncommon except in situations with drastic consequences, discussions on norms for mental health professionals to collect collateral information from private social data are only just beginning to happen [11, 19], and as Fisher et al. note, if social media becomes a bigger part of therapeutic processes, the power differential between a mental health professional and their client could potentially make clients feel uncomfortable providing direct social media data to a clinician [19].

It is clear that there is much interest from clinicians in the potential for non-intrusive methods of understanding changes in client affect and behavior without the use of methods that burden their patients in some way, with many clinicians placing hope in new technologies for a more nuanced understanding of their client's day-to-day experiences.

2.2 Personal Informatics

As a result of the widespread use of mobile phones, a significant amount of research done in mood inference has been done through the use of mobile phone based data. Mohr et al. detail the the work that has been done in this space in their 2017 review paper [36] on mobile sensing and mental health.

In the first major study on mobile sensing and mood, Madan et al., found that there was a significant decrease in total communication ($p < 0.05$), proximity to contacts ($p < 0.05$), and location entropy ($p < 0.001$) when comparing self-reported sad-lonely-depressed participants to other participants [33]. Building upon this initial research, several different experiments were done with the hope of inferring mood via mobile sensing data.

Through the use of location, accelerometer, light, and ambient sound data, Ma et al. created a mood prediction tool called MoodMiner [31]. MoodMiner was able to exactly predict daily moods with 50 percent accuracy on a discrete 1–5 scale for 15 participants over a length of 30 days. Moodscope, a mood inference tool developed by LiKamWa et al., attempted to infer the daily mood average of 32 participants, as measured by EMAs through the use of communication-based metrics, application use patterns, web browser history, and location data [29]. While initial results only had a 66 percent accuracy, over the course of 2 months of training, accuracy was increased to 93 percent. A subsequent attempt by Asselbergs et al. to replicate the results of the Moodscope study with 27 students failed to perform better than chance. In all three of these studies, ground truth metrics to determine mood evaluated mood from a daily or daily average level, rather than a granular level. While impactful, this work does not take into account how rapidly mood and emotion can often change for individuals, especially in the case of individuals with some level of emotional dysregulation [42]. To attempt to predict some of these changes in overall mood, Servia-Rodriguez et al. attempted to use location data, microphone data, accelerometer data, and call/SMS logs to predict self-reported mood state through the use of 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 percent on weekends, but varied between 50 and 65 percent on other days [47].

Research has also been done into using mobile sensing to predict disordered mood states, such as depressed mood. The first study done to relate persistent depressed mood with mobile sensor and audio data, done with 8 members of a continuing care retirement community, was able to relate mobile sensor data (including audio and accelerometer data) with overall mental well-being and depressed mood [41]. Additionally, higher levels of accelerometer-based physical activity were shown by Vallance et al. to be strongly associated with lower rates of depressed mood [52]. In a similar study, through the use of ambient light, GPS, accelerometer, and call log data, Burns et al. attempted to predict depressive mood in 8 participants over the course of 8 weeks [9]. However, while participants were satisfied with the interface of the phone application to evaluate mood, the predictive accuracy of depressive mood in this study was no better than chance. Continuing along this line of research, Wang et al. analyzed audio-based conversation data and the amount of colocation with other students for 48 students over 10 weeks, and found a significant relationship with both stress and depressive affect [53].

2.3 Social Computing

While several studies in mobile sensing and mood inference used call and SMS logs, none used an analysis of the actual content of the messages of participants to infer mood. In parallel to work done in mobile sensing and mood, significant research has been done in the field of social computing that demonstrates that emotional state and mood (as well as the social factors that influence affect and mood) can be inferred and extracted from social media data.

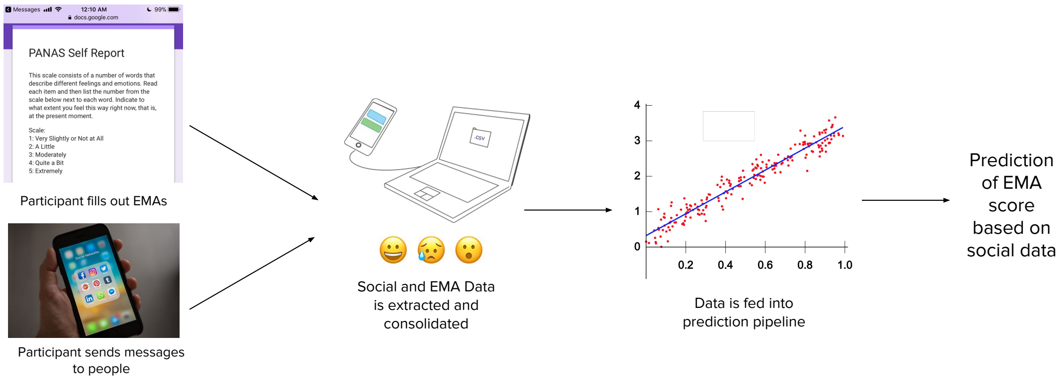
It is well established that the strength and level of social support can lead to improved psychological outcomes, more positive mood, and better overall mental health [45]. As a result, research done to use social media network data to predict the strength of ties between individuals has important implications for mood inference. As demonstrated by Gilbert et al., features drawn from social media posts and messages can be used to predict tie strength with a high level of accuracy [22]. In this study, message and content-based features such as the number of days since the first communication (β = .755), wall words exchanged (β = .299), and wall intimacy words (β = .111) were significant at a $p < .0001$ level in predicting tie strength. Additionally, SMS and call logs, a form of private social data, can also be used to predict the type of relationship an individual has with others [35], but as demonstrated by Wiese et al., the use of this form of metadata can be complicated and fraught with errors due to the nuances of private communication and emotional experience. For example, individuals may feel close to people that they do not communicate with at any significant level due to events from the past or historical rates of communication [55].

Analyses of the metadata and content of social media data have also been used to evaluate and predict mood disorders at a population level [14], with a trained classifier predicting the onset of depression from social media posts with 70% accuracy and 0.74 precision. However, these analyses might not be extensible to smaller and more homogenous populations, as culture and gender have a significant impact on how individuals express symptoms of mental illness and mood on social media [15]. Since expressions of mood can significantly vary between individuals, some work has been done to predict mood on an individual and personalized level. Studies of mood contagion on social media have demonstrated that people often express personal expressions of mood on social media very deeply [2], and that these expressions can be influenced by the content and interactions that individuals have with social media [28]. However, the kind of direct communication that happens via private social messages and strong tie strength with online social connections can moderate the effect of contagion with regards to depressed mood [58]. Data from public social media posts have also been used to predict individual mood instability. Saha et al. used Twitter posts to infer mood instability with a 96% accuracy and F-1 score [43].

Much of the research that has been done on affect and social media has solely looked at the metadata of private and public social media posts, the content of public social media posts, or some hybrid of these two areas. Most work has also focused on mood, or a state that is less temporal and differing in intensity in comparison to immediate affective state [7]. The content of private social messages, or what we define as messages that are sent from one person to a limited number of people via the use of a mobile messaging platform, is a ripe space for analyses of individual and temporal emotional states. Looking specifically at direct communication in a social media setting, Bazarova et al. demonstrated that Facebook users are more likely to share intense and private emotions in direct messages as opposed to network-wide status updates [6]. As demonstrated by O'Leary et al., private direct communication with one other untrained individual in a formal and mental health focused setting online improved clinically significant levels of anxiety and depression [38], lending credence to the potential for private message data to be a valuable signifier of changes in emotional state. It is yet to be seen whether mood can be inferred based on both the metadata *and* the content of informal private social media based messages, and the work done in this paper hopes to bridge that gap.

3 THE SOCHIATRIST SYSTEM

With the potential utility of private messaging metadata in better understanding emotion, and the diverse variety of private social messaging platforms by which people communicate their emotional



Messages 12:10 AM 99%
docs.google.com

PANAS Self Report

This scale consists of a number of words that describe different feelings and emotions. Read each item and then list the number from the scale below next to each word. Indicate to what extent you feel this way right now, that is, at the present moment.

Scale:
1: Very Slightly or Not at All
2: A Little
3: Moderately
4: Quite a Bit
5: Extremely

Participant fills out EMAs
Participant sends messages to people
csv
Social and EMA Data is extracted and consolidated
4
3
2
1
0
0.2
0.4
0.6
0.8
1.0
Data is fed into prediction pipeline
Prediction of EMA score based on social data

Fig. 1. Using the Sochiatrist System to extract data and predict EMA score

state, we set out to create a system that extracts data from different platforms and consolidates it into a consistent format for analysis.

This integrated system, Sochiatrist (a portmanteau of the words "social" and "psychiatrist"), is comprised of three main parts:

(1) A data extractor that collects and anonymizes private messaging data from a variety of social platforms
(2) A survey prompting system that collects survey data from participants via text message and email
(3) A prediction pipeline that predicts the results of the survey data based on extracted private messaging data.

The process of using Sochiatrist is illustrated in Figure 1.

3.1 Social Data Extraction

Previous studies that made use of private messages used a variety of different methods to collect enough messages to build a representation of a user's messaging habits, style, and content. These methods include having the user personally request the individual pages that contained messages from the social media website [22], or using some form of software to scrape data as the participants chat with their social connections in real-time and creating a recording of all actions they take on their monitor during this live chat session [34]. These past methods required an individual to physically be present in the lab for a significant amount of time while any conversational data was being collected, and restricted the amount and variety of data that could have been collected.

Past social computing research that has made use of private social media data has generally focused around one social media platform [6, 22], but individuals tend to use more than one social media platform. In a 2018 Pew Research Center survey of American adults, it was found that 73 percent of the public uses more than one of the eight social media platforms assessed in the survey [48]. The multiplatform use of social media that is commonplace today demonstrates the necessity of a tool that enables social media and user experience researchers to collect many different forms of social media data from a variety of platforms in a consistent format for analysis.

The Sochiatrist system fills this need by enabling researchers to collect data from different social media platforms through the use of an automated script. Drawing from knowledge in computer forensics, for phone based private social messaging platforms (i.e. SMS, WhatsApp, Kik), we identify the location where the database that contains all messages sent and received by the phone are stored. Once the phone is backed up to a lab computer, we then use a series of database queries to

Timestamp	Sent or Received	Message	Conversation Participant	Platform
7/1/17 12:32	sent	Thinking about skateboarding a bit at #:## when it gets cooler.	6af224	facebookmessage
7/1/17 12:32	received	Have a good time 7ac988	6af224	facebookmessage
7/2/17 18:01	sent	Eating some Cajun food now!	72c7e0	instagram_DM

Table 1. A sample dataset from the Sochiatrist Social Data Extractor

get the messages from each database, and store them in individual Comma Separated Values (CSV) files. In parallel, for web-based social messages (i.e. Facebook, Instagram), we scrape private social messaging data directly from the browser-based website, either from a data download provided by the website, from an API, or directly from the website through the use of a web scraping script. After all messages have been extracted, the messages are then compiled into one CSV file, and the user is prompted to specify the time range of messages for their final CSV file. Specific messages from that time range are filtered out, and identifying names in the data are replaced with then anonymized identifier, using an anonymization method as specified in the section below. Only data within the specified time range is collected.

An example of data extracted using the Sochiatrist Social Data Extractor appears in Figure 1. For each message, the generated dataset includes the timestamp of when the message was sent or received, the text of the message, whether the message was sent or received, a hashed (and thus anonymized) version of the participant's name and any numbers that might be inside of the messages (such as phone or credit card numbers, and addresses), and what social media platform the message was exchanged on.

The Sochiatrist Data Extractor can be used to extract Facebook direct messages, Instagram direct messages, Twitter direct messages, Kik messages, WhatsApp messages, and SMS messages (including iMessages on iOS). The Data Extractor is compatible with the listed platforms on both Android and iOS.

SMS/iMessage, Kik, and WhatsApp are mobile instant messaging platforms primarily used by individuals to send direct and group messages to other people. While primarily an image sharing application, 71 percent of people between the ages of 18–24 use Instagram [48], with direct messaging being a commonly used feature in the application among young people, as noticed by our clinical collaborators. While Facebook and Twitter have been characterized as primarily public social media platforms in previous literature [48], significant work suggests that direct messaging via public social media platforms is used to discuss private disclosures and intimate information, such as in the case of Facebook messaging [5].

3.2 Anonymizer

To anonymize the names of the sender and receiver, the Sochiatrist Anonymizer takes each name as a string and uses a SHA1 hashing function with a random salt to create a random alphanumeric string to represent the name of each individual. This can be seen in the fourth column of Figure 1. To anonymize names within messages, the Sochiatrist Anonymizer creates a personalized list of the names of people in the individual's social circles by taking the first name of all of the participant's Facebook friends. The Anonymizer then looks for instances of the names from the individual's Facebook friend list (including special cases, such as all capital letters or with punctuation after), and then replaces those instances with the hashed version of the name. If the participant did not provide Facebook data (n = 3 in our study), then only the sender and receiver of each message was anonymized..

	Average (Per Participant)	**Max**	**Min**	**Median**	σ
Message Statistics					
All Messages	3,068	16,697	358	1,897	3,531
Sent Messages	1,797	9,826	296	1,123	1,973
Received Messages	1,269	7,141	25	603	1,572
Applications Used	2.24	3	1	2	0.71
EMA Statistics					
Rate of Compliance	96.10	104.76	80.96	97.62	6.33
Positive Affect Score	23.95	50	7	24	8.20
Negative Affect Score	16.87	49	2	14	7.26

Table 2. Message and EMA Statistics. With a median of 2 and a average of 2.24 apps per participant, it is clear that private social communication happens via several different mediums. EMA compliance was above 100 percent due to participants occasionally completing assessments unprompted.

3.3 Survey Prompter

A survey prompting service was developed to support the random sampling involved with sending out EMAs. Participants would be directed to a customized Google Form containing the questionnaire associated with the EMA. The URL sent to each participant includes their unique identifier which could then be used to correspond to their responses. Through the use of the system, each EMA is sent out at random times, at least an hour apart, during each of their pre-defined available periods each day.

3.4 Prediction Pipeline

To process, generate and test different models with the collected data, an extensible and modular prediction pipeline was developed. This Python-based implementation uses the scikit-learn library and provides simple interfaces to add and switch-out features and parsing logic for any type of participant data.

4 METHOD

To test the viability of inferring emotional state based on private social media use, we designed an experiment to test whether we can use retroactively collected private social media to model and predict the scores of participants on an EMA. Over the course of two weeks, participants would complete an EMA three times each day during their waking hours. At the end of the two weeks, we collected their social media data, and consolidated it with their EMA data to perform our analysis. As part of a related study, participants also tracked their fitness data through the use of a Microsoft Band. The study was reviewed by our institution's Human Subjects Board, and data collection happened between November 14th 2017 and December 17th 2017.

4.1 Study Procedure

Recruitment for the study happened via posters and Facebook posts in Facebook groups created for undergraduate students at different class levels at [location anonymized for submission]. From these recruitment methods, 129 students signed up to participate in the study, and of the 129 students that signed up for the study, 25 participants participated for the full two week length of the study. Specific demographic information on these participants can be found in the Data Analysis section below. To be able to participate in the study, participants were required to have a cell phone running either the Android or iOS operating system, and use some type of private social messaging application. As part of the onboarding process, participants were asked what their sleep and wake times were, in order to only send participants EMAs during their waking hours. If participants asked why their data was being collected, they were told that it was part of a personal informatics research project in order to avoid the potential response bias of participants elaborating more or less on their emotional state in their messages as a result of being in the study.

Each day, participants were asked to complete the Positive and Negative Affect Schedule (PANAS). The PANAS is a self-reported questionnaire that measures positive and negative affect [54]. The version of the PANAS used for this study was 20 questions, with 10 questions measuring positive affect and 10 questions measuring negative affect via a Likert scale, for a potential maximum score of 50 each for positive and negative affect, and a potential minimum score of 0 (which would correspond with a user choosing not to answer any questions and then submitting the EMA) for each. We made the decision to give non-answers a score of 0 instead of using complete case analysis or some form of imputation for the missing data. In making this decision, we decided to treat the lack of an answer as a reflection of the participant's uncertainty about whether they had any affect that related to the PANAS question being asked, based on qualitative interviews done after the experiment. Based on the Likert scale provided, this uncertainty would most closely relate to a score of 0 (or no sense of relation to the dimension of affect being assessed). We used the PANAS to assess participant's momentary emotional states due the fact that the PANAS has been externally validated and is known as a highly internally consistent [54] assessment of affect as a part of EMA in a variety of clinical psychology settings [3, 51]. Participants were prompted to complete the PANAS via a private Google form sent by email or text message three times each day, with each prompt being sent at a random time during each third of the participant's waking hours. After receiving a prompt, participants had an hour to complete the survey, and would receive a reminder prompt after 30 minutes if the survey was not submitted by the halftime mark during that hour. If participants did not complete a survey within the hour after they received the survey, it did not count towards their compensation.

After completing the two weeks, participants were asked qualitative questions about their experience participating in the study. The Sochiatrist Social Data Extractor was then used to gather and consolidate the data they produced over the course of the two weeks of the study. Participants were then compensated a maximum of 60 dollars for their participation in the study, with amount of compensation based on the amount of compliance with completing EMAs, wearing the Microsoft Band, and the provision of some amount of social data.

During the course of the study, three participants discontinued their participation in the study. Two participants discontinued their participation for reasons related to privacy, and one participant discontinued their participation in the study for undisclosed reasons. During the data extraction process, one participant was unable to extract their text message data, as they were unable to remember their iPhone encrypted backup password, but were able to get message data from their other social media accounts. Another was not able to extract their text message data, as their phone was not compatible with the lab computer from which we performed extractions, but were able to get social private messaging data from their other accounts. In total, from November 14th 2017 to December 17th 2017, 25 participants participated in the study for the full 2 week length of the study.

4.2 Data Analysis

Data was collected via the Sochiatrist Social Data Extractor, as well as through the use of custom made Google Forms and Google Sheets to collect EMA data, as shown in Figure 1.

Of the 25 participants we collected data from, 20 were women, and 5 were men. With regards to type of operating system, 6 were Android users, and 19 were iOS users. The age of participants

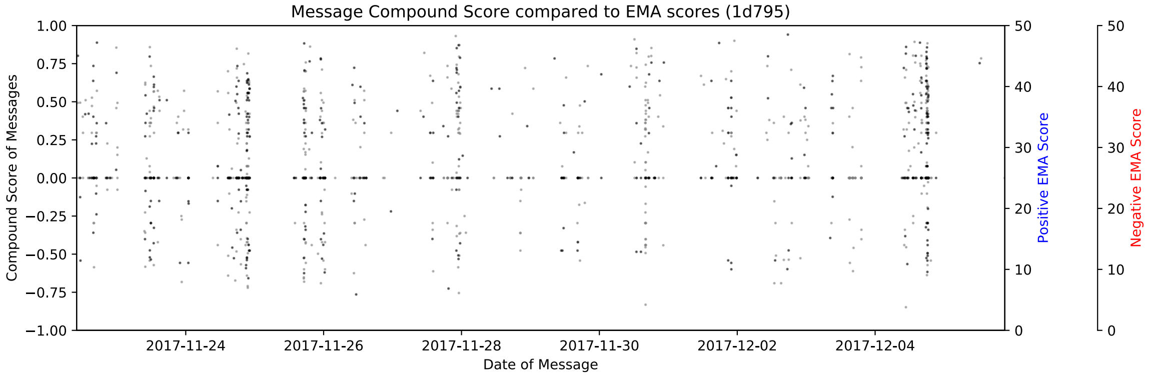


Fig. 2. The initial iteration of message visualizations used to determine features, shown for a single example participant. This displays positive EMA scores, negative EMA scores, and message sentiment. Black dots represent sent messages, and grey represents received. It was also interesting to see that positive and negative EMA scores tend to have an inverse relationship.

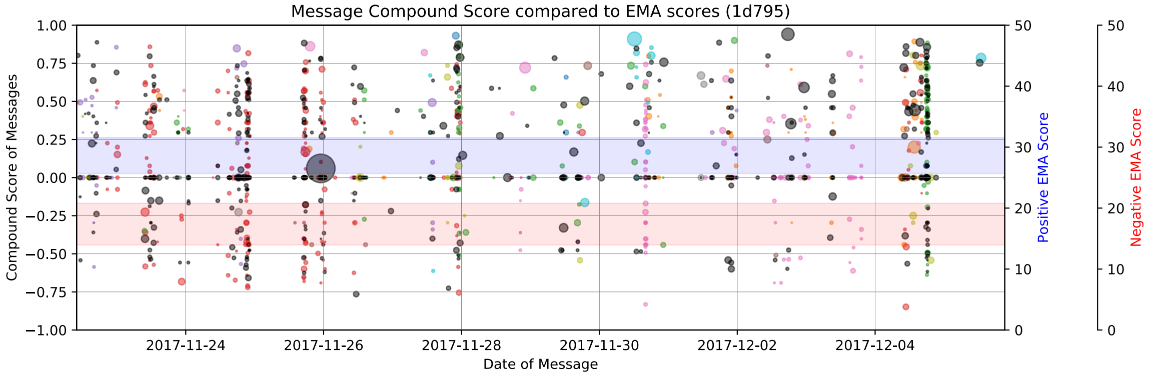


Fig. 3. The final iteration of message visualizations. In addition to the attributes mentioned in Figure 2, the EMA scores have a zone of usual scores (mean ± one standard deviation). The message points are scaled to message length in words, and color coded to individual users. We noticed that the sentiment of messages during participant's daily waking hours varied quite a bit, such that graphing the sentiment of these messages formed a clearly visible "stripe" on the graph, with messages scoring across the entire spectrum of compound sentiment (from 1.00 to -1.00) each day.

ranged from 17–22 years old, with the average age of the participants being 19 years old. Over the course of the study, we collected a total of 76,693 messages. Specific data about the messages we collected, as well as the rates of compliance with EMA prompts can be seen in Table 2 The maximum EMA compliance rate was above 100 percent, as some actually did extra EMAs without being prompted or compensated for the additional EMAs.

The most used messaging application we were able to collect data from was text messaging/iMessaging (n = 23) and the least used application was Twitter Direct Messages (n = 1).

4.3 Feature Selection

To generate and evaluate features for our model of emotional states and private social messaging data, we began by generating visualizations of the compound sentiment of individual messages juxtaposed against the EMA scores recorded by a participant at given times. To generate the sentiment of each message, we used Python's Natural Language Tookit with the VADER sentiment lexicon [23], designed specifically for a social media context, to annotate each message with its compound sentiment score. Each message point was gray if received, and black if sent. This private social messaging visualization can be seen in Figure 2. From this graph, we noticed that the sentiment of the messages exchanged within a participant's daily waking hours tended to vary quite a bit, enough so that graphing the sentiment of these messages formed a "stripe" on the graph, with messages ranging from a 1.00 to -1.00 compound sentiment each day. It was also interesting to see that positive and negative EMA scores tend to have an inverse relationship, as seen in Figure 2.

Based on work done showing that a high exchange rate of messages is indicative of a more intimate relationship [22], and that private messages tend to have more intense and negative content [6], we suspected that the size of messages and the relationship with the other participant in the conversation may also have some interaction between emotional state. We visualized this idea in Figure 3, with the size of each message point scaled to the word count of the individual message text. Sent messages were colored black, while received messages were colored in a different color for each particular sender, to better differentiate conversations within a stripe, as well as the people the participant consistently conversed with. This can be seen in Figure 3. From these graphs, we were able to infer that long messages that also have an exceptionally high or low sentiment seem to indicate a change in EMA score, regardless of whether they are sent or received. We were also able to infer that the number of people messaged and the density of messages sent during a "striped" period of waking hours seemed to have some relation to EMA score.

We organized insights we saw in the generated visualizations into features that fit four different general categories:

- **Sentiment Based Features**: These features are derived from a compound score of sentiment, ranging from +1.00 to -1.00, as calculated via the VADER sentiment lexicon [23].
- **Content Independent Features**: These features are not dependent on the content of the messages, and could be calculated solely based on private messaging metadata, and do not require the actual text of a message or any data on 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**: Inspired by the trend we noticed for people to consistently talk to certain people each day, these features are counts that are stratified based on whether the underlying messages are being exchanged with people that a participant sends messages to frequently.

A full list of features that were used in our linear regression model can be seen in Table 3.

In designing our features, we made some specific design choices:

- While looking for messages that were outliers in terms of sentiment and length, we found the number of messages that were in the 99th percentile of message sentiment magnitude for an individual's set of private social messages. The “sentilength score" of outliers was then calculated by multiplying the length of a message in words with the magnitude of the sentiment of that message. This served to weight long, highly sentimental messages as the most influential. For a certain EMA period, the overall sentilength score of each message was summed to provide the feature.
- The social network based features were limited to the top five most messaged people over the course of the two weeks based on Dunbar's [17] work and MacCarron et al.'s supporting research [32] theorizing that there are approximately 5 people that an individual is cognitively able to maintain a close relationship with at a given time.

- Features we designed that were not inspired by our analysis of the visualization included the count of emotion related language, and the count of mental health support phrases. To create a corpus of emotion related language, we scraped a list of emotion words from a popular website used by therapists to help individuals better identify their emotions [1], in both singular noun (e.g. "anxiety") and adjective (e.g. "anxious") form. Drawing on De Choudhury et al.'s use of data from Reddit mental health communities to study online mental health language [13], we also utilized a corpus of mental health support language generated from some of the same Reddit support groups. To create this corpus, via an online dataset of all Reddit posts from 2015 [4], we scraped all text posted in 2015 from from the Depression, Anxiety, and Suicide Watch subreddits, to isolate support language used around the kinds of negative affective patterns that are common to common mental disorders. We then compiled a list of the 250 most used trigrams and four-grams on each website to parse out mental health support *phrases*, filtering out any stopwords that were not first or second person nouns (i.e. "I" or "you") since these are commonly used when providing mental health support. Top phrases included "I don't know" and "I feel like I." Based on the most common unigrams from this dataset, we also included counts of the words "sorry," "I," and the word stem "feel."

4.4 Affective Modeling via Multiple Linear Regression

We attempt to model the participant's negative affect score using a multiple linear regression model, with features being calculated via messages sent and received from the beginning of the previous assessment to the end of next assessment according to Table 3. In constructing our model, our ground truth was the sum of the negative affect based PANAS questions, consistent with how the PANAS is used to assess overall negative affect. We chose to predict negative affect EMA scores to follow from Bazarova et al.'s finding that individuals tend to describe more negative and intense emotions in private messages [6], and as a result, our features (such as the count of instances of mental health support language, or counts of the word "sorry") are drawn from places where people express negative affect. In addition, we chose to use data from messages exchanged both in the time between the previous EMA and a given EMA, and the time between a given EMA and the next EMA. We made this decision since messages exchanged during the time period preceding the EMA could serve as a predictor to emotional state, while messages exchanged after the EMA could serve as as a reaction to the emotional state that had been assessed at the time of EMA.

As a result of the fact that all individual features were count-based, the aggregated private social messaging dataset occasionally contains extreme counts in some EMA periods. In order to reduce the positive skew of these count-based features, following Gilbert et al. [22], we added one to each count, and then log transformed the count before calculating correlation coefficients and inputting them our features into our linear regression model.

After doing this analysis, to contextualize our results on the predictability of emotional states via private message data, we also performed an Autoregressive integrated moving average (ARIMA) analysis on the isolated EMA data without taking into account any private social media data, and compared the subsequent affective forecasting models.

5 RESULTS

5.1 Feature Analysis

In order to relate these features to EMA scores, the Pearson correlation was calculated for each feature. Feature data for each EMA period was combined across all participants into a single dataset to calculate the overall correlation to the features. In addition, the correlations for each feature were individually calculated for each participant, and the subsequent maximum, minimum, mean,

Features	Correlation Coefficient	Maximum	Minimum	Average	Median	σ
Sentiment Based Features						
Number of sent messages with an exceptional sentiment multiplied by message length (sentilength, sent)	0.25	0.42	-0.29	0.03	-0.01	0.20
Number of received messages with an exceptional sentiment multiplied by message length (sentilength, received)	0.18	0.43	-0.27	0.01	0.00	0.16
Content Independent Features						
Number of messages sent by the participant	0.25	0.27	-0.23	0.03	0.05	0.15
Number of unique users communicated with	0.23	0.44	-0.37	0.01	0.01	0.20
Number of messages received by the participant	0.22	0.33	-0.29	0.01	0.00	0.17
Content Dependent Features						
Number of emotion words used in messages sent by the participant	0.29	0.31	-0.28	0.03	0.04	0.16
Number of times "sorry" was used in messages sent or received by the participant	0.27	0.38	-0.39	0.02	0.01	0.16
Number of times "I" was used in messages sent or received by the participant	0.27	0.37	-0.39	0.02	0.01	0.16
Number of times the word stem "feel" was used in messages sent or received by the participant	0.26	0.41	-0.32	0.06	0.05	0.18
Number of emotion words used in messages received by the participant	0.24	0.40	-0.35	0.00	-0.01	0.17
Number of mental health support phrases used in messages sent and received by the participant	0.22	0.36	-0.34	0.04	0.07	0.17
Number of characters in all messages during an EMA period	0.21	0.36	-0.33	0.04	0.06	0.17
Network Aware Features						
Number of messages sent by the participant that were not to the top 5 people they communicate with	0.28	0.36	-0.28	0.01	0.00	0.18
Number of emotion words used in messages sent by the participant to the top 5 people they communicate with	0.27	0.40	-0.35	0.04	0.04	0.20
Number of messages received by the participant that were not from the top 5 people they communicate with	0.26	0.42	-0.27	0.02	0.00	0.18
Number of messages sent by the participant to the top 5 people they communicate with	0.23	0.28	-0.32	0.03	0.03	0.16
Number of messages received by the participant from the top 5 people they communicate with	0.16	0.29	-0.44	0.00	0.01	0.18
Number of emotion words used in messages received by the participant to the top 5 people they communicate with	0.15	0.45	-0.41	0.00	0.00	0.21

Table 3. List of Features (All features resulted in $p < .001$)
The correlation coefficient was calculated using feature and EMA values from every single participant. The Maximum and Minimum columns refer to the maximum and minimum Pearson correlation coefficient found for any individual user. As seen in the table, there was significant variability in the level of correlation between emotional state and private messaging data for individual participants.

median, and standard deviation were calculated for each feature. These correlations can be found in Table 3.

In general, features based off sent messages were better correlated with the negative EMA scores than received messages. This suggests that user's emotional states are correlated with the messages that they send, and less about messages are received. The highest correlation observed on the aggregated dataset was a correlation of .29 for the count of emotion words sent by a participant. This might follow research done that shows that people use social media platforms to seek social support in the wake of an event associated with strong negative affect [44], showing that this is the case in private messages as well.

To our surprise, when solely isolating the sentiment of messages exchanged during an EMA period, and using that sentiment as a feature, the sentiment itself was not as predictive as other features on our aggregated dataset. For example, the feature corresponding to the average compound sentiment score of messages during an EMA block had a correlation coefficient of -0.071 with a p < 0.05. This suggests that there is little to to no correlation between the average sentiment of messages sent and received during a time period, and the negative affect of the participant. This may speak to variability between the specific wording and style of language that participants use when speaking to others in times of negative affect, but less variabilty with regards to general behavioral patterns regarding private social messaging during times of distress.

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, with a positive correlation of .45. However, it was also observed that (for the same feature), the lowest correlation was -.45, showing that emotion words in messages from closely tied connections might actually be indicative of an improvement in negative affect for some. This variability speaks to the known variability in predisposition for particular affective states based on external personality-based factors [57]. However, while there is significant variability on an individual level, the correlations of the data in aggregate suggest that certain affective patterns exhibit themselves on a population level, such as sending messages with strongly emotional language 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 our time-series analysis below.

5.2 Affective Forecasting

These features listed in Table 3 were then incorporated into a multiple linear regression to generate a basic predictive model. This resulted in a R^2 of 0.164 and an F-statistic of 11.99. While all used features resulted in a weak to moderate correlation to negative EMA scores, only a select number of features resulted in significant Standardized β values, as shown in Table 4. Features that had a high level of correlation with the aggregated dataset, such as the number of emotion words sent, resulted in an insignificant p-value in the linear regression model. This suggests that while correlated, on a small-scale level of analysis, they have little predictive power for specific negative EMA scores.

Other features did result in a p-value under .05, such as the number of messages sent by the participant that were not to the top 5 people they communicate with and the number of unique users communicated with. In addition, the number of times the word stem "feel" was used in messages resulted in a p-value of < 0.01. From these p-values we can infer that these features provide a significant role in predicting negative EMA scores. This can be seen in Table 4, but not a strongly predictive role.

The count of sent messages with an exceptional sentiment (sentilength, sent) and the number of messages sent by the participant to the top 5 people they communicate with both resulted in a p-value of less than 0.001, suggesting a high predictive power to negative EMA scores. The highly significant p-value of sent messages with an exceptional sentiment suggests that looking at the outlying messages serves as a strong predictor of negative mood, a confirmation of our inference from looking at the visualized data in Figure 3.

Surprisingly, the high p-value of number of messages sent by the participant to the top 5 people they communicate with as well as the corresponding positive β value (0.484) suggests that the rate a client converses with their friends is highly predictive of negative mood. This may be explained by clients being more likely to seek social support from close friends during times of distress. This finding may be of interest to clinicians, as both of these features may be useful in providing a non-intrusive way to estimate negative emotion during a certain time period.

As expected, there were stronger correlations to negative affect with our features than positive affect, as seen in Table 5. This may demonstrate that participants were more likely to talk about negative things in private messages, as supported by past research where private social media content was intense and negative than public social media [6]. Building on Bazarova et al.'s work, our findings show that there is a correlation between specific private messaging patterns and negative affect.

Features	β Value
Sentiment Based Features	
Number of sent messages with an exceptional sentiment multiplied by message length (Sentilength, Sent) ***	0.15
Number of received messages with an exceptional sentiment multiplied by message length (Sentilength, Received)	0.03
Content Independent Features	
Number of messages sent by the participant *	-0.38
Number of unique users communicated with *	0.14
Number of messages received by the participant *	-0.29
Content Dependent Features	
Number of emotion words used in messages sent by the participant	0.04
Number of times "sorry" was used in messages sent or received by the participant	1.47
Number of times "I" was used in messages sent or received by the participant	-1.28
Number of times the word stem "feel" was used in messages sent or received by the participant **	0.12
Number of emotion words used in messages received by the participant	0.07
Number of mental health support phrases used in messages sent and received by the participant	-0.30
Number of characters in all messages during an EMA period	0.42
Network Aware Features	
Number of messages sent by the participant that were not to the top 5 people they communicate with *	0.21
Number of emotion words used in messages sent by the participant to the top 5 people they communicate with	0.043
Number of messages received by the participant that were not from the top 5 people they communicate with	-0.04
Number of messages sent by the participant to the top 5 people they communicate with ***	0.48
Number of messages received by the participant from the top 5 people they communicate with *	-0.25
Number of emotion words used in messages received by the participant to the top 5 people they communicate with	-0.02

Table 4. List of Features and their Corresponding Standardized β Values (p-value level of significance: 0.001:***, 0.01:**, 0.05:*)

5.3 Time Series Analysis

To isolate the effect of previous emotional state on future emotional, and put our findings on the relationship between private social messaging data and negative affect into context, we also performed a time-series analysis on the isolated EMA data. An autoregressive moving average (ARIMA) model was used to see if past EMA values could accurately predict future negative affect. This method effectively decomposes the EMA scores into a seasonal component and into an overall trend in order to better predict future points. As part of our analysis, each individual had their

Features	**Negative Affect**	**Positive Affect**
Sentiment Based Features		
Number of sent messages with an exceptional sentiment (sentilength, sent)	0.2593	0.1551
Number of received messages with an exceptional sentiment (sentilength, received)	0.1783	0.0904
Content Independent Features		
Number of messages sent by the participant	0.2584	0.1277
Number of unique users communicated with	0.2384	0.2095
Number of messages received by the participant	0.2277	0.1277
Content Dependent Features		
Number of emotion words used in messages sent by the participant	0.2963	0.1872
Number of times the word "sorry" was used in messages sent or received by the participant	0.2742	0.1075
Number of times the word "I" was used in messages sent or received by the participant	0.273	0.1328
Number of times the word stem "feel" was used in messages sent or received by the participant	0.2665	0.043
Number of emotion words used in messages received by the participant	0.2381	0.1075
Number of mental health support phrases used in messages sent or received by the participant	0.2197	0.1235
Number of characters in all messages in the time in an EMA period	0.2179	0.1181
Network Aware Features		
Number of messages sent by the participant that were not to the top 5 people they communicate with	0.287	0.1388
Number of emotion words used in messages sent by the participant to the top 5 people they communicate with	0.2763	0.1732
Number of messages received by the participant that were not from the top 5 people they communicate with	0.2654	0.1902
Number of messages sent by the participant to the top 5 people they communicate with	0.2352	0.0993
Number of messages received by the participant from the top 5 people they communicate with	0.1615	0.0651
Number of emotion words used in messages received by the participant to the top 5 people they communicate with	0.1536	0.062

Table 5. A comparison between correlation coefficients for negative affect and positive affect. Correlations are consistently higher for negative affect than positive affect. Interestingly, the feature with the highest correlation to positive affect was the number of unique users communicated with (at a correlation of .2095), which had a similar level of correlation for negative affect. This may show that people are likely to reach out to many people during both times of intense positive *and* negative affect. All correlations had a p-value less than .05.

own ARIMA fit generated for them with optimized parameters for that person. The R^2 value for this fit was then calculated by checking the fit's predicted values against the actual values that the fit was trained on and taking the square of the correlation coefficient between the fit's predicted values and the actual values. This analysis resulted in an R^2 of 0.37. This suggests that only 37% of the variation in EMA score can be predicted by previous EMA values, showing the complexity and difficulty of predicting variation in EMA scores. Since emotional state is subject to sudden changes, and does not seem to have an identifiable pattern over the course of two weeks of data collection, using retrospective passive data alone to predict the exact scores of positive and negative affect in future emotional states may not be the best method of predicting emotional state in a way that is meaningful. The correlations associated with features that lend themselves to particularly predictive models may be more useful in understanding the causes and reactions to negative affect in the context of widespread private social media message use, especially when interpreted by a human. This idea is further expanded upon in the Discussion below.

5.4 Participant Reflection and Feasibility

On a qualitative level, when assessing their own experience as a part of the study, several participants noted feeling more aware of their emotional state due to the daily prompting and needed to fill out the EMA. While some noted that they felt like they were able to approach their mood with more equanimity due to the surveys, none felt like they changed their messaging habits as a result of being part of the study.

Several participants questioned the validity of the PANAS scale at accurately measuring affect, noting that some terms in the PANAS (such as assessing the level of "Scared" and the level of "Nervous") were so similar that participants consistently rated them at the same level. Other participants noted that some of the assessments in the PANAS seemed too strong (e.g. Determined), so these were ranked lower throughout the course of the study. While participants found filling out the survey quick and simple to do, the time constraint of needing to fill out the EMA within the course of an hour from the initial prompting was mentioned by participants as being somewhat burdensome, demonstrating the need and potential for non-invasive ways of measuring affect continuously throughout a day.

Several also noted that they occasionally found it difficult to assess their feelings in the moment, and thus may have underreported their affect. Some participants also claimed to never have strong feelings over the course of the study, and also claimed that their mood stayed relatively and stably low throughout the course of the study, and thus their emotional state did not vary much during the two weeks that they participated in the study.

6 DISCUSSION

6.1 Implications

In conducting this study, we solely used private social media data as a method of testing the efficacy of private message metadata in inferring emotional state. Our research demonstrates that inferring the specific scores quantifying the emotional states that individuals experience daily is a difficult task. From the results of correlation testing, linear regression and a time series analysis, it is clear that assessing a participant's momentary emotional affect is difficult solely using private social media data. The large variation in personality as well as the varying usage of social media platforms means that the exact EMA score corresponding to specific emotional states are difficult to accurately predict in any general model.

However, from the demonstrated moderate correlations over a short period of time with a relatively small sample, it is likely that one can infer the overall trajectory of a client's emotional state based on metadata extracted from private social media data, which demonstrates the utility and value of using private messaging metadata in using technology to better understand the emotional experience of individuals in a non-invasive way. As elaborated on in our Related Works section, most previous work done in the past did not comprehensively utilize private messaging metadata from a *in vivo* setting, choosing instead to focus on public social media posts [43] or private messaging data collected in a lab [22, 34]. The Sochiatrist System and our associated experimental use of it demonstrates the feasibility and value of doing work with private messaging metadata. Though using private social messaging data along with other forms of data used in prior work, it may be possible to make significant and accurate predictions of individual emotional state at a given point in time.

On an individual and qualitative level, the limited correlation we found between relatively innocuous features (like the number of messages sent or the number of times the word "sorry" is used) and negative affect shows the value in potentially reporting this data to clinicians. Features that have moderate to high correlations could be reported to clinicians along with traditional EMA based methods of self-reported affect, and instead of an automated algorithm, a clinician could use their interpretation of this metadata along with the corresponding correlations with an individual's emotional state to better inform treatment strategies and planning for future treatment pathways and outcomes for their client. Metadata based statistics, such as the number of unique users messaged since the last psychotherapy session, could be used by clinicians as non-threatening mechanisms to explore a patient's past experience and serve as a method of reducing the impact of difficulties in memory recall among clients experiencing intense emotional distress.

The value of private message metadata that we have demonstrated could also put more agency in the hands of those experiencing mental distress. The value of metadata with the significant correlations demonstrated in this study is twofold for both clients and therapists. As negative and intense emotion is more often discussed in private messages than public ones, clinicians could use metadata from private social messages to develop a better understanding of their patients' day-to-day inner experiences. From the client's side, deciding what metadata from private messages to disclose could give patients more agency over how their data is being used by their clinician, if at all, and reduce the cognitive burden associated with formulating and retelling changes in their emotional state when in session with a clinician.

6.2 Correlation vs Causation

From our work, it is unclear whether sending more private social messages is indicative of a negative event happening (and thus the corresponding increase in negative affect), whether the act of sending messages to a higher number of people causes an increase in negative affect, or whether there is some external cause independent of messaging patterns or specific events that is causing an increase in negative affect. It is likely that all three cases can be seen at different points in our dataset, but it is impossible to infer causation simply from analyzing metadata.

This is why a clinician-augmented approach to the use of metadata to infer emotional state is critical. The use of private social messaging data could be dangerous if used without the guidance of a licensed clinician to help interpret increases in negative affect and their corresponding significantly correlated features, and to parse out the behavioral patterns that might be causing the increase in negative affect. As observed in our study, one participant wondered whether their unchanging emotional state was atypical, and felt some level of stress in pondering that question.

6.3 Limitations and Future Work

Our study had some limitations. To examine the specific utility of private social message data, we did not use public social media data, which may have been useful in analyzing how the presentation of public posts and the features of private social messaging interact and relate to the emotional state of an individual.

The length of our study was also significantly shorter than other previous studies on mood and emotional state, which had data collection periods that ranged from 30 days [31] to 3 years [47]. It is very possible that having more data, from both the perspective of time and the number of types of data used, could have increased the correlations we found. For example, we used the top five people messaged during the period of two weeks as a measure of intimate relationships, but it is possible that there may have been an event (e.g. a group project, a sporting event) that directly affected who the top five people messaged were over the course of the two weeks. Having data from a longer amount of time might be useful in better teasing out social networking patterns within direct communication habits.

Future work could expand on the results we have demonstrated by using more sophisticated machine learning techniques, and making predictions by gathering data in real time through the use of a phone application to record other sources of mobile data, including location and accelerometer data, as seen in studies that attempt to predict mood. Similarly to how we derived features from online mental health communities on Reddit, future papers could expand the scope of where language-based features are derived from, such as supplementing the dataset with Twitter data [43], and replicating our work with a more diverse group of participants with varied cultural expressions of affect on social media [15].

7 CONCLUSION

People experiencing forms of intense emotional distress, such as depression and suicidal ideation, often find it difficult to remember specific events from day-to-day life, especially in evaluating positive memories and experiences [56]. To better assist their clients, it is important for therapists to get a good idea of how the emotional states of their clients change from day to day, and use that as part of the process of identifying and changing maladaptive behavior patterns. Methods of doing this in the past have focused around EMAs, which can often be burdensome for clients and participants in clinical psychology research studies. In this study, we seek to understand the utility of private social media data in inferring temporal and momentary emotional state. To do so, we designed and implemented the Sochiatrist system, an automated system that downloads private social media data from a variety of platforms, and consolidates it into a consistent format. We then used this system to demonstrate the value of using private messaging data to understand and potentially predict affect. While we were unable to demonstrate that it is feasible to predict the exact EMA score associated with a state of negative affect, we find that certain private social media message features have mild to moderate correlations with negative affect, demonstrating the utility of using private social messaging data in inferring attributes that have affect as a component, such as mood or affective state. These correlations, though moderate, may be useful to better understand the factors that influence negative affect, without putting any additional burden of self-assessment on those being assessed or studied.

8 ACKNOWLEDGMENTS

Special thanks go out to members of the Brown Human-Computer Interaction Lab, Project LETS at Brown, the participants in our study, and to both of my advisors, Jeff Huang and Nicole Nugent. I feel very blessed to have gotten the opportunity to work on a project that brought me so much meaning and joy.

REFERENCES

[1] [n. d.]. List of Emotions (Worksheet). *Therapist Aid* ([n. d.]). https://www.therapistaid.com/therapy-worksheet/list-of-emotions

[2] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression.. In *Proceedings of CSCW*. 1485–1500.

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

[4] Jason Baumgartner. 2015. Complete Public Reddit Comments Corpus. (2015). https://archive.org/details/2015_reddit_comments_corpus

[5] Natalya N Bazarova. 2012. Public intimacy: Disclosure interpretation and social judgments on Facebook. *Journal of Communication* 62, 5 (2012), 815–832.

[6] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. 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*. ACM, 154–164.

[7] Christopher Beedie, Peter Terry, and Andrew Lane. 2005. Distinctions between emotion and mood. *Cognition & Emotion* 19, 6 (2005), 847–878.

[8] Dror Ben-Zeev, Michael A Young, and Joshua W Madsen. 2009. Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion* 23, 5 (2009), 1021–1040.

[9] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* 13, 3 (2011).

[10] Sarah Clement, Oliver Schauman, T Graham, F Maggioni, Sara Evans-Lacko, N Bezborodovs, C Morgan, N Rüsch, JSL Brown, and G Thornicroft. 2015. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological medicine* 45, 1 (2015), 11–27.

[11] Brian K Clinton, Benjamin C Silverman, and David H Brendel. 2010. Patient-targeted googling: the ethics of searching online for patient information. *Harvard Review of Psychiatry* 18, 2 (2010), 103–112.

[12] Mihaly Csikszentmihalyi and Reed Larson. 2014. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*. Springer, 35–54.

[13] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.. In *ICWSM*.

[14] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[15] Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 353–369.

[16] The Walt Disney Corporation (@Disney). 2018. "Makes no difference who you are. When someone compliments you, but you're dead inside.". (2018).

[17] RI Dunbar. 1998. The social brain hypothesis. *brain* 9, 10 (1998), 178–190.

[18] Ulrich W Ebner-Priemer, Janice Kuo, Stacy Shaw Welch, Tanja Thielgen, Steffen Witte, Martin Bohus, and Marsha M Linehan. 2006. 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 (2006), 774–779.

[19] Carl E Fisher and Paul S Appelbaum. 2017. Beyond Googling: The Ethics of Using Patients' Electronic Footprints in Psychiatric Practice. *Harvard review of psychiatry* 25, 4 (2017), 170–179.

[20] Michel Foucault. 1965. Madness and civilization (R. Howard, trans.). *New York: Pantheon* (1965), 265–268.

[21] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. 2013. I need to try this?: a statistical overview of pinterest. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2427–2436.

[22] Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 211–220.

[23] Eric CJ Hutto Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

[24] Claire Hoffman. 2010. "The Battle For Facebook". *Rolling Stone* (2010).

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

[26] Thomas R Insel. 2017. Digital phenotyping: technology for a new science of behavior. *Jama* 318, 13 (2017), 1215–1216.

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

[28] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

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

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

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

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

[33] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 291–300.

[34] Joanne Meredith and Elizabeth Stokoe. 2014. Repair: Comparing Facebook âĂŸchatâĂŹwith spoken interaction. *Discourse & Communication* 8, 2 (2014), 181–207.

[35] Jun-Ki Min, Jason Wiese, Jason I Hong, and John Zimmerman. 2013. Mining smartphone data to classify life-facets of social relationships. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 285–294.

[36] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47.

[37] JA Naslund, KA Aschbrenner, LA Marsch, and SJ Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[38] Katie O'Leary and Wobbrock J.O. Pratt W. Schueller, S. 2018. Suddenly, we got to become therapists for each other": Designing Peer Support Chats for Mental Health.. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM.

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

[40] World Health Organization et al. 2017. Depression and other common mental disorders: global health estimates. (2017).

[41] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.

[42] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[43] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 95.

[44] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. *Association for the Advancement of Artificial Intelligence* (2018).

[45] Catherine Schaefer, James C Coyne, and Richard S Lazarus. 1981. The health-related functions of social support. *Journal of behavioral medicine* 4, 4 (1981), 381–406.

[46] Somini Sengupta, Nicole Perlroth, and Jenna Wortham. 2012. Behind Instagram's Success, Networking the Old Way". *The New York Times* (2012).

[47] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. 2017. 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*. International World Wide Web Conferences Steering Committee, 103–112.

[48] Aaron Smith and Monica Anderson. 2018. Social Media Use in 2018. *Pew Research Center* (2018).

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

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

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

[52] Jeff K Vallance, Elisabeth AH Winkler, Paul A Gardiner, Genevieve N Healy, Brigid M Lynch, and Neville Owen. 2011. Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006). *Preventive medicine* 53, 4-5 (2011), 284–288.

[53] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[54] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[55] Jason Wiese, Jun-Ki Min, Jason I Hong, and John Zimmerman. 2015. You never call, you never write: Call and sms logs do not always indicate tie strength. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 765–774.

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

[57] Kay Wilson and Eleonora Gullone. 1999. The relationship between personality and affect over the lifespan. *Personality and Individual Differences* 27, 6 (1999), 1141–1156.

[58] Haifeng Xu, Tuan Quang Phan, and Bernard Tan. 2013. How does online social network change my mood? an empirical study of depression contagion on social network sites using text-mining. (2013).