Sochiatrist: Signals of Affect in Messaging Data

Talie Massachi Grant Fong Varun Mathur Sachin R. Pendse
Gabriela Hoefer Jessica J. Fu Chong Wang Nikita Ramoji
Nicole R. Nugent Megan L. Ranney Daniel P. Dickstein
Michael F. Armey Ellie Pavlick Jeff Huang

May 4, 2021

Messaging is a common mode of communication, with conversations written informally between individuals. Interpreting emotional affect from messaging data can lead to a powerful form of reflection or act as a support for clinical therapy. Existing analysis techniques for social media commonly use LIWC and VADER for automated sentiment estimation. We correlate LIWC, VADER, and ratings from human reviewers with affect scores from 25 participants. We explore differences in how and when each technique is successful. Results show that human review does better than VADER, the best automated technique, when humans are judging positive affect ($r_s = 0.45$ correlation when confident, $r_s = 0.30$ overall). Surprisingly, human reviewers only do slightly better than VADER when judging negative affect ($r_s = 0.38$ correlation when confident, $r_s = 0.29$ overall). Compared to prior literature, VADER correlates more closely with PANAS scores for private messaging than public social media. We propose, implement, and evaluate some potential improvements to VADER, and find some possible areas of improvement in correlation to PANAS scores. Our results indicate that while any technique that serves as a proxy for PANAS scores has moderate correlation at best, there are some areas to improve the automated techniques by better considering context and timing in conversations.

Contents

1 Introduction

A major part of daily communication in today's world is digital, sent through messaging applications. These applications include popular but disparate platforms: regular texting (SMS and iMessage),

Facebook Messenger, Kik, Twitter, Instagram, WhatsApp, Snapchat, and others. The long-term records of conversations made through these applications may provide a window into a person's past. Here we explore the potential for these messages to serve as a proxy for changes in a person's affect. We raise the question: how much of someone's emotional state can be interpreted from these messages? Doing so may be useful as backstory for a therapy session, helping a user create a timeline of emotional patterns that can be given to their clinical therapist [17], within which trends as well as extreme emotional events are clearly visible.

Another type of online communication, social media, is often considered very similar to direct messaging, but generally broadcasted to a wider range of people. Social media use is often tied to emotional well-being, as social media and online social support have become essential ways for people with ordinary mental distress, common mental disorders [33], and severe mental illness [32] to voice their experiences and find solidarity with others. Additionally, some research suggests that public social media posts reflect affect [19]. However, it is important to note that private messages and public posts on social media have a number of fundamentally different features that affect the way users interact with the platform. While public social media posts are visible to a diverse audience, in many cases completely unknown to the poster, private messages are intended only for people included in the conversation [5]. Studies have also shown that users often consider self-presentation more when posting publicly, or to a wider audience [5, 22], as opposed to a private message or smaller network. Thus private messages and their context may require different approaches in studying an individual's affect.

However, two major obstacles complicate useful programmatic analysis of personal messaging data. First, extracting this data from multiple platforms in a privacy-preserving manner is difficult without specialized technical expertise and preparation. Second, it has not yet been studied whether messages sent (and potentially their metadata) actually correlate with an individual's positive and negative affect. Even the language used during messaging may be different from public posts, given that many users write informally when messaging [36], reflecting that private messages are usually intended only for one recipient who typically has additional context about the sender.

We address the first of these hurdles in this paper by describing the implementation and maintenance of a cross-platform messaging extractor. This extractor has been active for three years and has extracted data from over 350 individuals from a patient population with clinical collaborators; the data from those extractions are part of clinical research which is currently unpublished. A separate study using the same extractor was conducted with a population of college students, which is presented in this paper.

The second of these hurdles is addressed by a comparison between existing sentiment analysis techniques (known techniques called LIWC and VADER), human review, and users' self-reported affect rated on a scale of 10–50 (using the well-known PANAS scale [53]) when prompted throughout the day by their mobile device. Each of these techniques are compared and the differences are explained in the findings.

We find that LIWC (the most common technique used in literature) and VADER (a more recent technique) have high agreement with each other. When it comes to identifying a person's emotions from messaging data, they perform comparably or better than what has been reported with other textual sources in literature (e.g., diary entries [50]). The set of scores found through review by a human panel performs better than VADER over all cases of positive affect and anytime when the panel is confident. The human reviewers are aware of their own limitations in judging the messages, as is evident by them achieving a closer match to actual affect scores as their confidence in their predictions increases. Reviewers found contextual cues to be particularly salient when rating text for affect.

The main contribution is to report how popular automated sentiment analysis techniques, like LIWC and VADER, may predict positive and negative affect scores reported by participants, especially in comparison to human review. Qualitative explanations are provided for the discrepancies between the techniques, and the extraction system that enables this type of evaluation is presented and described.

2 Related Work

2.1 Self-Report Measurements of Affect

Previous works on sentiment analysis, as well as other work from psychology [53, 49] commonly express emotional state in terms of two separate quantities: positive affect and negative affect. Positive affect refers to the extent to which an individual subjectively experiences positive moods and emotions, such as joy, interest, and alertness [30], while negative affect refers to the experiences of negative moods and emotions, such as anxiety, sadness, fear, anger, guilt, and shame [47]. A commonly held belief is that these two metrics are independent: a high negative affect does not necessarily imply a low positive affect and vice versa [53]. Common techniques used to measure positive and negative affect include the Positive And Negative Affect Schedule (PANAS) [53].

The widely used 20-item version of the PANAS has been shown to be an internally consistent self-report assessment of affect [53, 51, 3]. It contains a series of 20 mood-descriptive terms (e.g., interested, determined, upset, ashamed) where participants are asked to rate their level of agreement with each on a 5 point Likert scale ranging from 1 = *very slightly or not at all* to 5 = *very much*. Of these, 10 terms are representative of negative affect, while the remaining 10 represent positive affect. Upon completion, the totals of all negative affect and positive affect questions are summed, producing two scores (one for positive affect and one for negative) with a minimum value of 10 and a maximum of 50 [53].

Given its wide usage and overall reliability, we use PANAS as a ground truth measure of affect in this study. We correlate the results of different sentiment analysis techniques with positive and negative affect scores found through PANAS.

2.2 Sentiment Analysis Techniques for Social Media Text

In this study, we compare ground truth PANAS scores to two automated sentiment analysis techniques known as LIWC and VADER. Linguistic Inquiry and Word Count (LIWC) is one of the most widely used automated sentiment analysis techniques [37]. Additionally, Valence Aware Dictionary for sEntiment Reasoning (VADER) is a sentiment analysis model specifically built for use in contexts such as social media posts [20].

Unlike PANAS, LIWC and VADER are indirect measures used to automatically estimate affect. LIWC is text analysis system developed in 1993 [38] popularly used for sentiment analysis. It is a commonly held standard in both fields of computer science and psychology. After its initial creation, LIWC has been updated a number of times to follow linguistic trends and improve its accuracy, most recently in 2015. LIWC contains a weighted dictionary of words, word stems, and a selection of emoticons. Each of these is labeled with a list of categories and sub-categories that identify, among other things, the affect implied by each word. When given a written text, LIWC compiles scores for it by summing the weighted values assigned to each word in the passage. The original set of words in LIWC's dictionary was based on common emotional and affect rating scales, including PANAS [38].

Updates to the scale since then have been based on cycles of expert analysis and further methods of relevant term discovery [48].

VADER is a freely available rule-based sentiment analysis tool built to improve upon existing techniques, including LIWC and human review [20]. VADER utilizes a larger dictionary of terms, but otherwise works very similarly to LIWC, with a particular focus on analyzing text from social media sources. While not as popular as LIWC, VADER has been used in previous research as a sentiment analysis tool [6, 7].

2.3 Predicting PANAS using Automated Sentiment Analysis

Previous work has found significant but weak correlations between PANAS and automated methods of affect prediction discussed above (ie. LIWC and VADER). This is consistent across texts collected from a variety of textual sources such as social media posts and diary entries [7, 6, 8, 50]. Although all of these previous studies used PANAS scores as a more general measure of affect rather than an in-the-moment measure, they still provide useful comparison points for our purposes.

2.3.1 In Personal Writing and Speech

Tov et al. [50] investigated LIWC predictions of PANAS scores over 21 days. However, instead of using a single PANAS score to describe that period, participants completed the PANAS each night, reflected on their affect that day, and wrote two short diary entries about a good and a bad event from that day. At the end of the study, all of a participant's diary entries were combined and used to generate a LIWC score. This score was then compared in a correlative analysis with the average of all PANAS scores obtained by the participant over the same period, producing a moderate-to-weak, but significant, correlation (positive: $r = 0.21$, $p < 0.01$ and negative: $r = 0.22$, $p < 0.01$).

Cohen et al. [8] wanted to investigate methods of personality detection from autobiographical text. Participants were asked to speak for three minutes on the topic of their choice, though they were subtly encouraged to talk about themselves. After participants finished speaking, they were given a series of psychological tests, including PANAS-X (a version of the PANAS with 60 items instead of 20). Each speech was transcribed and used to generate a LIWC score. Further analyses were run from there, including a correlative analysis between LIWC and PANAS scores. LIWC was found to be moderately and significantly correlated with PANAS positive and negative affect scores ($r = 0.29$, $p < 0.05$ and $r = 0.24$, $p < 0.05$, respectively).

2.3.2 In Social Media

Beasley & Mason [6] investigated the ability of LIWC to predict general affect from public social media posts on Facebook and Twitter. Participants were asked to fill out the PANAS with regard to their general affect. Researchers then collected as many of the participant's posts from Facebook and Twitter as the platforms allowed. PANAS scores were then compared to a VADER analysis over the text of all collected Twitter and Facebook posts. PANAS was also compared to LIWC analyses over all collected posts (going as far back as the platform would allow), and posts only in the month, 6 months, and year preceding the study [6]. For the most part, correlations between PANAS and LIWC and PANAS and VADER were weak, with a highest correlation of $r = 0.13, p < 0.01$.

Following that study, Beasley et al. [7] investigated whether pre-filtering Facebook and Twitter data would improve VADER's accuracy in predicting PANAS scores. Methods for PANAS score generation and Facebook and Twitter data extraction were the same as those in Beasley & Mason [6].

After posts were collected, they were filtered for posts containing pre-selected "patterns of expression," for example,

> "am" (e.g., "I am") + optional up to two words (e.g., "not very") + [affect-related words] (e.g., "happy")

For each participant, only posts that met these filtering criteria were included in the text given to VADER for analysis. A correlative analysis was then performed for PANAS as compared to both VADER scores for Facebook posts, and VADER scores for Twitter posts of participants containing greater than 36 Twitter posts after filtering. Despite this, all correlations were still weak ($r < 0.15$). See Table 9 for exact values.

In this study, we similarly correlate LIWC and VADER scores with PANAS. However, unlike previous studies, we use private messaging data as input for LIWC and VADER as opposed to public social media, diary entries, or transcribed speech. We cross-compare our results with those of previous studies to find patterns in input data that detect affect more successfully.

2.4 Personal Disclosure across Multiple Social Media Platforms

In order to effectively predict affect from social media data, it is first important to understand how people interact with and communicate over these platforms.

Most previous work exploring affect detection in social media focuses on a single platform during analysis. The majority of these works specifically analyze Twitter data [11, 14, 16, 21, 35, 41], but some studies examined platforms such as Instagram [40] and Facebook [10]. Still others focus on two or more social media platforms [6, 7], a decision that seems prudent given that as of 2018, the Pew Research Center reported that 73% of social media users use multiple social media platforms (Median = 3) [46]. Furthermore, the social media landscape makes rapid advances, and platforms that are popular now may fade out over time. For example, in 2015, 71% of teens ages 13–17 reported using Facebook, but by 2018, the percent of teens 13–17 dropped to 51% [2].

Previous research on sentiment analysis also tends to focus on public social media data. This assumes that people primarily post about their emotions publicly. However, studies have shown contradictory results on whether people are more likely to share emotions publicly or privately [4, 27, 35]. User preference for sharing emotions more openly in public posts or private messages seems to be influenced by platform [35, 4]. When investigating trends in self-disclosure on Twitter, Park et al. found some individuals prefer to share their emotions in public Tweets, rather than in Twitter private messaging [35]. They attribute this to individuals using public spaces on Twitter as an area for emotional reassurance and self-expression. In contrast, Bazarova et al. found that Facebook users disclosed more sensitive information in private channels than in public posts [4]. They suggest that when people cannot control the target of their disclosure, such as in public posts, then users are less likely to disclose. Lottridge & Bentley similarly found that users were more likely to share news links with a goal to start conversations in private chats rather than publicly [27]. Furthermore, studies have shown that people often present themselves differently on their public social media in order to self-promote (eg. [26]) indicating that public social media posts may not accurately reflect a user's emotional state.

In this study, we sought to bridge a gap in the literature by investigating the performance of automated methods of affect prediction over private messaging data. We believe that private messages extracted from a variety of platforms may more closely predict affect than previous social media studies [7, 6]. In many ways, private messages are more similar to direct speech [8] than to public social media posts. So based on the success by Cohen et al., private messages may be a

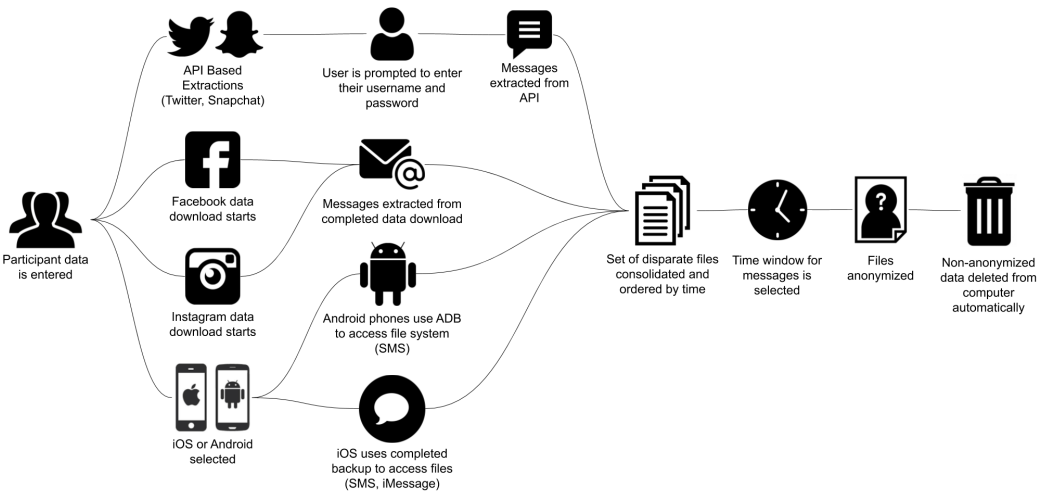

Figure 1: Sochiatrist extracts, consolidates, and pseudonymizes the data to develop models predicting affect based on messaging data.

viable textual source for affect detection. Furthermore, by extracting data from a variety of different platforms we can protect against the rapid changes in social media platform popularity and have a more holistic view of a user's communication, especially for the many users who do not use one platform exclusively.

3 Sochiatrist System

We developed Sochiatrist (a portmanteau of the words "social" and "psychiatrist"), an application that uses data-scraping methods to automate the retroactive extraction of a participant's messages and public posts from popular social media platforms (Figure 1). The system is retroactive in that it can collect social media data from within any specified past date range in a single run. It is non-invasive and does not require the participant to install any long-term software or tracking system. It is also built to be privacy-and-consent-first: participants must be physically present to input their login information for each messaging platform during an extraction and all extracted data is pseudo-anonymized. The system is unique because it allows efficient extraction of *both public posts and private messaging data* from many social media platforms. Code and instructions on using the Sochiatrist system will be released publicly at the completion of the clinical studies (currently funded by the National Institute of Health) that make use of the system.

3.1 Data Extraction Methods

The Sochiatrist system enables researchers to collect public posts and private messages from multiple web and mobile based platforms. Supported platforms have changed over time due to changes in data availability and social network popularity. At the time of writing this paper, Sochiatrist supports Facebook, Twitter, Instagram, Snapchat, and SMS/iOS Message extraction on both Android and iOS phones.

For platforms whose data is stored locally (e.g. SMS, WhatsApp), the Sochiatrist system extracts data directly from the phone, which must be connected to the computer being used for extraction. The system creates a backup file from the phone and then reads this backup file to extract messages.

For iOS devices, this backup file contains a `SQLite` database where text messages are stored. The location of this database is known from previous work in computer forensics to contain messaging data [34, 9]. For Android devices, the backup file is a CSV file that is read from the phone over ADB (Android Debug Bridge). In all cases, the participant must provide the password to their phone to allow this process to take place. Under no circumstance does extracting messages require rooting or jail-breaking, hard-to-reverse procedures that may damage a participant's phone, or unauthorized access to encrypted data.

For web-based platforms (e.g. Facebook, Twitter, Instagram, Snapchat), the Sochiatrist system downloads messages directly from the web, as a data download either provided by the website (Facebook, similar to Saha et al. [42], and Instagram), from an application programming interface (Snapchat), or directly from the website through the use of a web scraping script (Twitter). For all of these methods, the participant enters their username and password into the system and these credentials are used to provide the legitimate authentication required to gain access to the wanted data. There is no use of unauthorized access to data from any platform.

After extraction, messages from all platforms are compiled in a consistent format and sorted by timestamp. The user is prompted to specify a desired time range and only data within that time range is saved. To ensure privacy, the data is pseudo-anonymized (see Section 3.2.1) and all intermediary files created during the process are automatically deleted, leaving the participant's phone and the user's laptop untouched. The final pseudo-anonymized data is written to disk in CSV format on the computer running the application.

The Sochiatrist system has a graphical user interface. It displays clear, step-by-step instructions and allows for non-technical research assistants to run extractions with minimal training. It is also resilient to many types of user errors such as incorrect input formats or incorrect passwords, and attempts to provide helpful correcting instructions upon failure. Over the course of the past 3 years, the system has been used by clinical research assistants without computational backgrounds to successfully extract data from over 350 study participants, e.g. Ranney et al. [39].

Currently, the Sochiatrist system is used in five different clinical studies. The system is fully supported by our team: we release regular updates with patches and offer technical support to resolve issues that we discover through dedicated automated bug reporting mechanisms. With the current timeline of studies that use the extractor, this support will continue for at least the next 5 years.

3.2 Privacy and Consent

The Sochiatrist system is specifically designed to respect participants' consent and privacy. During data extraction, participants must be physically present to input their login information for each messaging platform used or the password to their phone, which builds participant consent into the core of the system. Special care is taken to ensure that there is no data persisted to the computer running the system that would ever allow a future unauthorized extraction. All data extractions are also legal and make use of legitimate authorization methods. There is a system that pseudo-anonymizes the final data, and all non-anonymized intermediary files are irrecoverably deleted. Table 1 shows an example output of the Sochiatrist system. For each message, the dataset includes the timestamp, the text of the message, whether the message was sent or received, an ID representing the sender's name, and which social media platform the message was exchanged on.

Timestamp	Status	Message	To/From	Platform
7/1/17 12:32	sent	I haven't been feeling 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	sent	Are you free at #:##?	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.

3.2.1 Pseudo-Anonymization

Private messages are by definition personal and it is natural that participants may not want their identity or the identity of any of their conversation partners linked to extracted data. To respect this, for every message extracted, the "to/from" field is replaced by a randomly generated alphanumeric string. To maintain consistency, the same conversation partner is always replaced by the same string, although the same person messaging across different platforms will receive different identifiers for each platform.

However, there is also textual content in messages that may be identifying. As a simple example, if either conversational participant sends "Hi [name]", then identity is likely compromised. The free text in messages is therefore processed to remove names and numbers. The system has a constant database of common names that it augments with the participant's Facebook friends (obtained through the same data dump that contains Facebook messages) during anonymization. It uses this list to detect and remove names within the free text. As above, the names are replaced with a random alphanumeric string which is consistent across the use of the same name. Simple regular expression searches were used to detect different forms of the name such as possessives and different capitalization. All numbers in the free text are also replaced with the '#' character, which anonymizes shared phone numbers, account numbers, addresses, meeting times, and other such identifiers.

This pseudo-anonymization removes some sensitive information that messages can contain. However, this is of course an imperfect system and cannot be guaranteed to anonymize all identifying information in the extracted data, which is why we refer to it as pseudo-anonymization. This technique misses cases where people use nicknames for people that they message (e.g. Auntie, Honey), when people exchange messages with someone who has a common English word in their name (e.g. April, Hope), or cases when a name happens to not be in the custom dictionary of names generated. Locations are also not deidentified. However, the topic of how to completely anonymize free text is an extremely complex one, and it is our hope that as modern anonymization techniques improve so will the anonymization capabilities of our system.

4 Methods

We ran a study to investigate the relationship between affect and private messaging data, collecting private messaging data from a sample of undergraduate students at Brown University. Undergraduate students were studied due to the heavy usage of messaging platforms in this population [18]. Sentiment analysis and human review methods were used to estimate self-reported PANAS scores from their messages. The performance of these estimations were compared later during analysis. All procedures were reviewed by our institution's human subjects review office, (the IRB), and passed through a full board review (the highest level of review) on July 6, 2017.

Negative Words	Distressed	Guilty	Upset	Scared	Irritable
Positive Words	Alert	Excited	Strong	Inspired	Proud

Table 2: Examples of negative and positive words used in the PANAS survey. Participants are asked to fill out a Likert scale from 1–5 for each word to answer the question "to what extent do you feel this way right now"

4.1 Study Procedure

Participants were recruited using posters and Facebook posts in groups created for undergraduate students of various class levels. Participants were required to have an Android or iOS phone to join the study, and use at least one messaging application supported by the Sochiatrist system. Of 28 students who agreed to participate, 3 decided against participation during the study, 2 for privacy reasons and 1 for undisclosed reasons. Their data was not collected and thus excluded from the study. The final set of 25 participants included 20 females and 5 males, with an age range of 17–22 years of age (Mean = 19 years). 68% of our participants ($N = 19$) used an iOS device for the study while the other 6 used an Android device.

To maintain anonymity, we did not collect any personally identifying information beyond gender and age. Recruitment materials were distributed in four Facebook groups: one group for each undergraduate class. All students in each year received an invitation to their respective group upon university admittance, and these groups were commonly used for events, announcements, and general communication between students at the university. We therefore assume each group to be representative of the diversity of the student body as a whole, and we assume our sample to be similarly representative of the students.

We measured participants' affect through self-reported PANAS surveys collected via a method known as Ecological Momentary Assessment (EMA), commonly used in field studies pertaining to mood or affect [1, 54]. During EMA, participants are notified over text to complete a survey or task. This notification appears a set number of times per day, as chosen by the experimenter [44]. Notification time (and whether that stays consistent day-to-day) is also at the discretion of the experimenter. EMA aims to minimize recall bias, maximize ecological validity, and allow study of microprocesses that influence behavior in real-world contexts [45].

In line with typical EMA protocols, 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. Links to surveys were sent via email or text, depending on the participant's preference, and were administered by an online Google form pre-filled with the participant's unique identifier. The survey presented each of the 20 PANAS attributes with a 5-point Likert scale, and asked participants how they felt at that specific moment in time. 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. Extra PANAS surveys that were completed without a prompt were discarded from analysis.

The PANAS survey used for this study had 20 words, 10 measuring positive affect "PANAS(+)" and 10 measuring negative affect "PANAS(–)". Participants are asked to fill out a Likert scale from 1–5 for each word included in the PANAS in answer to the question "to what extent do you feel this way right now". The responses are simply summed up for a potential maximum score of 50 on each affect scale, and a minimum score of 10. Some examples of the words included in the survey are in Table 2.

We then collected participant's private messaging data from these two weeks using the Sochiatrist system, which consolidated the private messaging data they produced over the two weeks of the study.

Messaging Platform	Messages Sent	Messages Received	Total Messages	Participants
Facebook Message	16,506	24,348	40,854	22
Text Message	14,878	19,183	34,061	23
WhatsApp Message	329	1,206	1,535	4
Twitter DM	0	196	196	1
Instagram DM	15	0	15	5
Total	31,728	44,933	76,661	25

Table 3: Summary statistics of the messages send and received analyzed in this study, and the number of participants using each platform. Snapchat messages were unavailable at the time.

The ability to retroactively extract messages both simplifies the study procedure and reduces the risk of the study influencing naturalistic conversation behavior that would normally take place. This data includes messages extracted from the participant's online Facebook, Instagram, and Twitter services and messages extracted from WhatsApp, Kik and SMS (including iMessage) applications on the participant's Android or iOS phone. *Third-party messages (messages received by the participant) were not used during the analysis.* In other words, the conversations were only analyzed from the participant's sent messages.

Participants did not report challenges or objections with the Sochiatrist system, but rather, reported the study in general to be straightforward and understandable. That being said, there were two potential participants that decided not to continue with the study due to privacy reasons, and we must consider that participants in the study had already self-selected to be comfortable with sharing their data through the Sochiatrist system.

Upon completion of the study, participants were debriefed in the lab. They answered general questions about their opinions of mood tracking, issues they faced, and the process overall. Participants appreciated tracking their mood through EMAs and reported interest in continuing to track their emotional state, even after study termination. Finally, 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 (which was not used for this paper's analysis due to the focus on messaging data).

The study took place from November 14, 2017 to December 17, 2017. Over the course of the study, 1,009 PANAS surveys were collected. However, out of these, 55 were incomplete and were discarded. Survey compliance ranged from 81–100% with a median compliance rate of 98%. PANAS(+) and PANAS(–) scores were calculated from each of the 954 complete surveys. The mean PANAS(+) score was 23.9 and the mean PANAS(–) score was 16.9. See Figure 2 for a visualization of the distribution. These values are similar to the positive and negative affect population means estimated by Watson et al. [53] of 29.7 and 14.8 respectively. During the Sochiatrist download, $76,661$ messages (Mean = $3,068$ messages per participant, Median = $1,897$, Range = 358–16,697) were collected across all participants. The messaging platform used by the most participants was SMS text messaging/iMessage ($N = 23$), while only one user received Twitter direct messages. Further messaging statistics can be seen in Table 3.

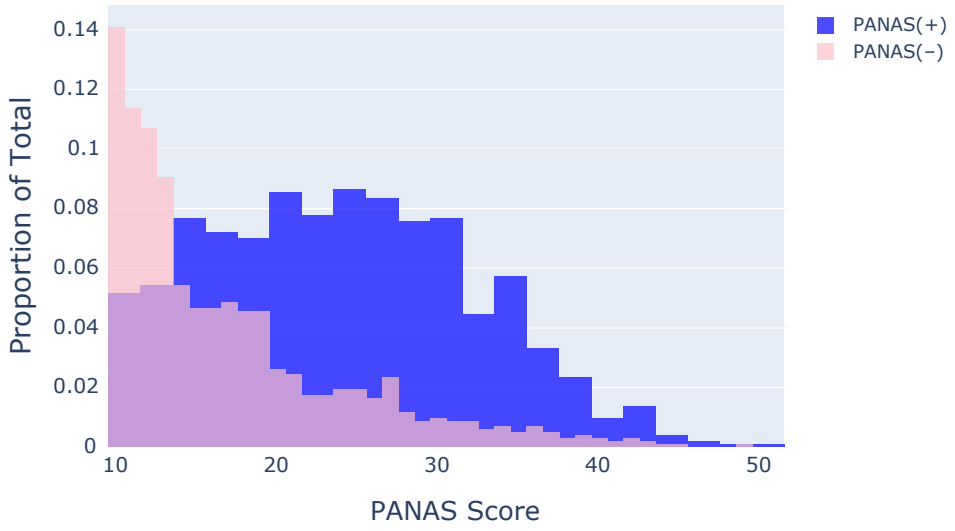

Figure 2: Histogram for PANAS(−) (pink) compared to the histogram for PANAS(+) (blue), with the overlap in purple. Note that PANAS(−) has a lower mean than PANAS(+) and is a more skewed distribution.

4.2 Data Processing and Session Generation

Third-party messages (messages received by the participant) were removed from all data collected. We matched each PANAS score with all the messages sent in a 24 hour window surrounding it ($\pm$12 hours of when the PANAS was completed); messages in one 24 hour window corresponded to a single PANAS score as its "session." We discarded all sessions that did not have at least one message sent on any tracked platform in the two hours surrounding its actual PANAS survey time, or where reported PANAS scores were both 10, as this likely indicated that the participant did not fill out the survey in accordance with their true affect. It is important to note that although we consider 24 hours of data in a session, the actual PANAS survey asked people how they feel *at that moment* in accordance with EMA.

We chose to use a 24 hour window ($\pm$12 hours), as opposed to three intervals based on PANAS score report times for two reasons. First, we wanted the most accurate results from our human reviewers for a solid best case comparison. Human reviewers requested to see all messages occurring within one day of a PANAS score for more context, and we wanted to keep the time window for LIWC/VADER analysis consistent with human review for more accurate comparisons. Second, intervals between two PANAS score measurementsvaried due to the randomized prompting of the EMA (some as short as two hours) so some sessions would include a disproportionate number of messages due to the uneven windows, especially if a session included a period of sleep.

We also note that during investigation using shorter time periods for LIWC and VADER analysis ($\pm 2, 4, 8$ hours) we found that the longer interval ($\pm$12 hours) correlated most strongly with the ground truth PANAS scores. Of the shorter intervals, PANAS(−) correlations were comparable (within 0.05 for LIWC, within 0.1 for VADER) to when sessions used a $\pm$12 hour time period. PANAS(+) correlations for the $\pm$8 hour set were also comparable to $\pm$12 hours (within 0.01), but $\pm$2 and $\pm$4 hours correlated much less closely to PANAS(+) (a difference of more than 0.12).

4.3 Automated Sentiment Analysis

Two methods of sentiment analysis, LIWC and VADER, were used and their accuracy was compared. These are the most common methods used for identifying sentiment in social media messages (see Related Work).

4.3.1 LIWC

The authors of LIWC describe it as "...the gold standard in computerized text analysis. Learn how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations." Of the 19 studies we reviewed pertaining to online communication and emotional state, 13 of them used LIWC as a measure of affect [35, 14, 6, 25, 24, 23, 50, 16, 41, 12, 13, 10, 31]. It is a commonly held standard in fields of both computer science and psychology as a sentiment analysis tool. Due to its widespread usage in research as a sentiment analysis tool, we included LIWC as a key estimator for affect from private messaging data.

The LIWC2015 tool was used to analyze all the messages in each entire session. For each session, the tool outputs many metrics. We took the `posemo` score as the PANAS(+) estimate and the `negemo` score as the PANAS(–) estimate. These scores are not on the same 10–50 scale that PANAS uses and do not necessarily scale linearly with PANAS scores. For this reason, all correlations computed later are non-parametric (i.e. Spearman's correlation).

4.3.2 VADER

Developed after LIWC, VADER was an attempt to produce a rule-based sentiment analysis method that was optimized for social media text, and is also popular in the field [20]. The authors of VADER release it as, "a gold-standard sentiment lexicon that is especially attuned to microblog-like contexts."

The NLTK distribution of VADER from the `nltk.sentiment.vader` package was used to analyze all the messages in each entire session. For each session, the tool produces a `pos`, `neg`, `neu` and `compound` score. We took the `pos` score as the PANAS(+) estimate and the `neg` score as the PANAS(–) estimate. As in LIWC, these scores were not scaled in any way.

4.4 Human Review Process

Human Review is important for two reasons. Firstly, analysis of the differences between human review and sentiment analysis methods may provide insight into what may be lacking from these methods in PANAS predictions so that we may build better estimators. Automated sentiment analysis can reveal how existing methods perform on textual message content, but does not project what is possible with additional context from the conversation. Human reviewers reading the conversation can get a broader understanding of the pace of conversation from timestamps, tone, and changing topics, and can even infer the relationship of the participant and conversational partner. Secondly, it allows us to calibrate the task. With access to just a snippet of a conversation—no context about the person, the scenario, etc.—it may be extremely difficult to predict a PANAS score. Human labels may allow us to understand better what is perceived as good performance on this task and what we may hope to achieve in the future.

A group of three reviewers estimated PANAS(+) and PANAS(–) from a sample of sessions. These three reviewers are authors on this paper, but they did not participate in running the study nor did they interact with any of the participants. To select the sessions, a stratified sample of

sessions over participants was taken by randomly selecting between 4–6 eligible sessions per study participant (for a total of 110 sessions). There were a few restrictions to prevent bias in the reviewers labels. Any session previously rated by any reviewer (in tests or pilots for reviewing) was removed from the set of eligible sessions. Any session with fewer than two messages within two hours of the PANAS rating time were also removed from the set of eligible sessions. Lastly, in order to prevent large of amounts of overlap in the text between rated sessions, we only included one session per participant per day in the set of sessions to be labeled. We selected this set of sessions to get as broad a range of different participants as possible without over representing any single participant, and subsequently over representing one particular texting style.

When labeling a session, human reviewers were shown the timestamp of the PANAS score along with the corresponding 24 hours (±12 hours) of messages. For each message, reviewers could see the text, its timestamp, the anonymized numeric participant ID, and the anonymized alphanumeric ID for the conversation the message was a part of. In addition, reviewers knew the population means for positive and negative affect of 29.7 and 14.8, respectively, as estimated by Watson et al. [53]. Note that this is not the same as the sample mean of PANAS(+) or PANAS(–) from the sample of sessions labeled, or even the sample mean from all the sessions present in our data.

To start, all reviewers familiarized themselves with the PANAS survey. As mentioned above, they were given the mean PANAS(+) and PANAS(–) scores estimated by Watson et al. [53]. Before any discussion, reviewers also each individually rated a set of messages with a shorter time window (±2 hours) as a test. Reviewers were shown the scores of the other reviewers, as well as the true PANAS scores, for the sessions in this test set. Following the test, reviewers requested that a full day of messages be included in a session for context, and that they would prefer to be able to discuss scores with another before deciding on a rating. The set of sessions in this test set was fully disjoint from the sessions used in final analysis.

For the final human review process whose data was included in this study, all three reviewers were placed together in a video call. With the data described above available to them, each reviewer individually proposed overall PANAS(+) and PANAS(–) scores (each between 10 and 50) for a given session. When reviewers disagreed about the values they proposed (which happened in the majority of cases), they were given the opportunity to simply accept another reviewer's proposed score. If a unanimous decision was still not reached at this point, then each reviewer explained their reasoning for the scores they proposed. One reviewer would then propose a new set of PANAS predictions based on the given explanations. The other reviewers could then accept the proposed values or propose an alternate set of scores, with explanation if appropriate. Reviewers repeated this last step until unanimous consensus on positive and negative affect scores were agreed upon, resulting in two final scores, one for PANAS(+) and one for PANAS(–) . Time to reach consensus averaged about four minutes. Scores were predicted on a scale of 10 to 50 to be consistent with the PANAS scale. Reviewers looked for cues in content and tone, as well as time of day the survey was taken, time of year, and their own previous experience with the situations described in the text. When reviewers struggled to identify any signal in a session, they rated it as "neutral", falling back on population means mentioned above. Reviewers also each took individual notes on each EMA, which included the group's reasoning, as well as any particularly important aspects of the text. Finally, reviewers came to a unanimous decision as to whether they felt confident in the rating or not. Confidence was a binary outcome (yes/no). It was recognized that confidence would be a subjective decision and there were no strict heuristics used by reviewers. Some things they expected to use to determine confidence were the strength of a particular emotion and how clearly the participant expressed their feelings in their messages.

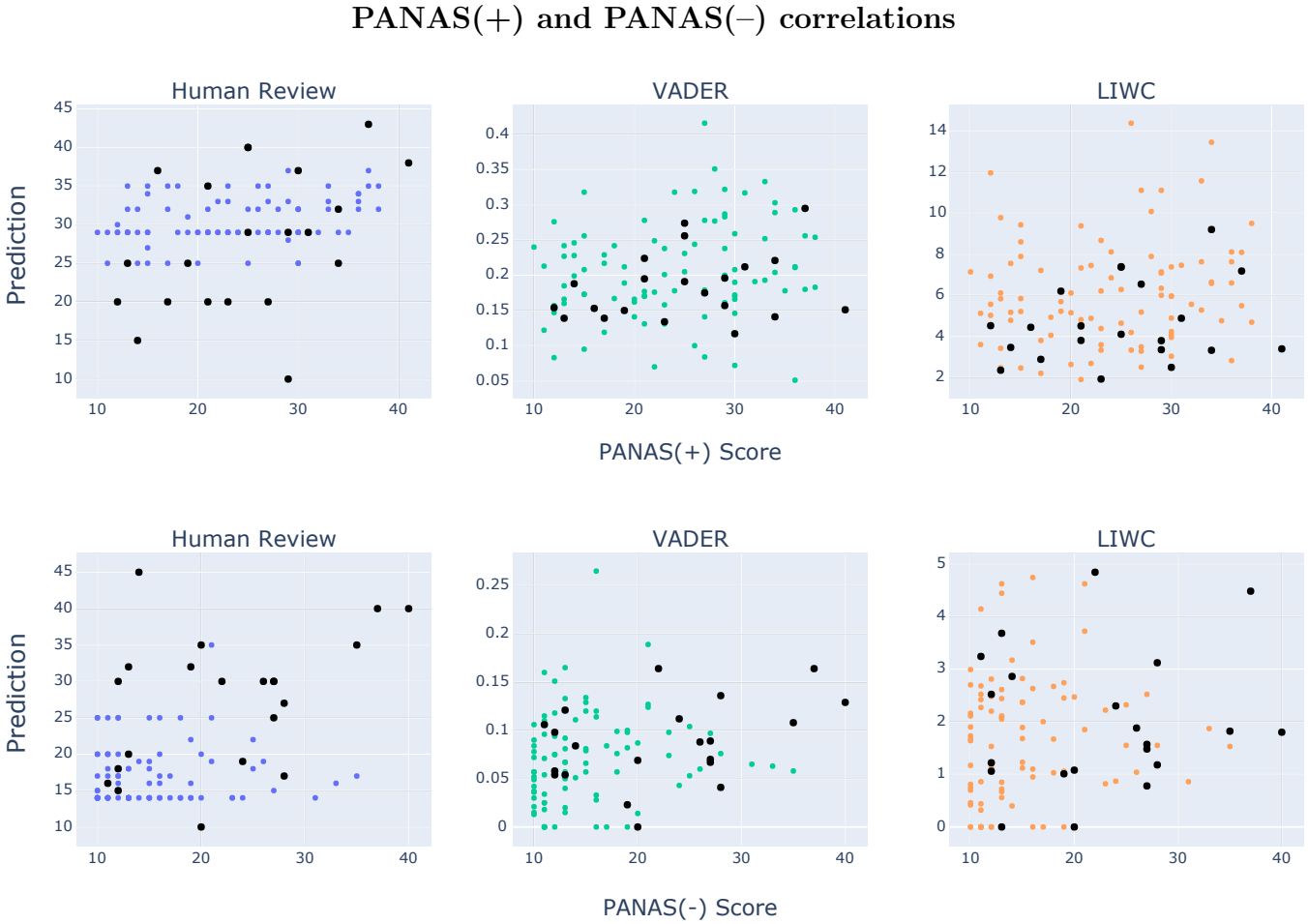

Figure 3: Self-reported PANAS(+) and PANAS(–) scores plotted against predicted scores. The x-axis is PANAS(+) in the top plots and PANAS(–) in the bottom plots while the y-axis is the algorithm prediction (each technique uses a different scale). The dots in black are the points labeled confident by human reviewers.

5 Results

We analyzed three sentiment analysis techniques—LIWC, VADER, and human review—to investigate their ability to predict affect from private messaging data. Correlations were calculated using the non-parametric Spearman's correlation due to the lack of a linear relationship between the sentiment analysis scores and PANAS (thus, a Pearson's correlation is not used). Predicted versus actual PANAS score comparisons for each sentiment analysis technique can be seen in Figure 3.

5.1 Individual Performance of Techniques

Tables 4 and 5 show both individual performance measures (how well a sentiment analysis technique correlates with PANAS affect scores), as well as the correlation between techniques.

Self-reported positive and negative PANAS scores, PANAS(+) and PANAS(–), were treated as the ground truth for the participant's affect. In absolute terms, PANAS(+) scores tended to be higher ($Mean = 23.95$, $SD = 8.20$) than PANAS(–) scores ($Mean = 16.87$, $SD = 7.26$). Between the two scores, PANAS(+) and PANAS(–) scores had a weak, insignificant correlation (Pearson's

	VADER	LIWC	Human Review		
			All	Confident	Not Confident
PANAS(+)	*0.20	0.14	**0.30	*0.45	**0.28
VADER	-	**0.75	**0.27	0.31	*0.22
LIWC	-	-	**0.29	0.36	*0.24

Table 4: Spearman's correlations between various techniques for estimating *positive* affect scores from PANAS, i.e. PANAS(+). Human review performs best when the reviewers feel confident about their estimate. LIWC and VADER are highly correlated, but less correlated with human review. They are less correlated with PANAS(+) compared to human review with confidence ignored ("All" column). *$p < 0.05$, **$p < 0.01$

coefficient $r = 0.16$, $p = 0.09$). This reflects the expected independence of these measures, where high positive affect does not necessarily mean having low negative affect and vice versa.

For the stratified sample of 110 sessions chosen for analysis, PANAS(+) ($Mean = 24.04$, $SD = 7.87$) was higher than PANAS(–) ($Mean = 16.51$, $SD = 6.88$). For the subset of sessions that human reviewers reported confident, PANAS(+) was slightly higher than the 110 analyzed ($Mean = 27.70$, $SD = 9.04$) and PANAS(–) was substantially higher ($Mean = 27.75$, $SD = 9.25$). Other techniques were used to predict the PANAS affect scores with the messaging data, and are reported in the following subsections.

5.1.1 Human Review

The three human reviewers examined the timestamped private messages for the sampled 110 sessions and attempted to predict their associated PANAS(+) and PANAS(–) scores, marking 21 of these sessions as confident. As expected, these ratings moderately predicted the PANAS(+) and PANAS(–) from the user-reported EMAs. For PANAS(+), human review was more accurate compared to automated techniques, with a correlation ($r_s = 0.31$, $p < 0.01$). For PANAS(–), human review still correlated moderately ($r_s = 0.28$, $p < 0.01$).

When only the sessions where reviewers were confident of their scores ($N = 21$) are considered, human review outperformed all other analyzed techniques when estimating PANAS(+). Among these sessions, the human review estimates correlated more strongly with PANAS(+) than when not confident ($r_s = 0.45$, $p = 0.04$). PANAS(–) estimates showed similar behavior ($r_s = 0.38$, $p = 0.09$), though this correlation was not significant. The human reviewers performed much better when confident than when not confident, meaning they are able to judge whether there was enough signal in the conversations to accurately predict affect. When reviewers were less confident, their accuracy was relatively inconsistent. Reviewers' ability to predict PANAS(+) remained moderately accurate ($r_s = 0.28$, $p < 0.01$), while ability to predict PANAS(–) dropped sharply ($r_s = 0.11$, $p = 0.29$). These findings indicate that although the task of generally predicting affect is difficult, there may be identifiable scenarios where the task is more tractable.

Reviewers tend to be more confident when PANAS(–) was higher. The mean PANAS(–) for confident predictions is 27.0 (SD=9.45) , as compared to the mean for not confident predictions of 17.1 (SD=3.97) . However, this trend does not hold true across PANAS(+) scores. This may be due to the fact that all human language has a known positivity bias [15], meaning that negative speech is relatively rare. Therefore, when negative emotions are expressed, it may become a more clear indicator of a high PANAS(–) that raters are able to detect and feel confident in their detection.

	VADER	LIWC	Human Review		
			All	Confident	Not Confident
PANAS(–)	**0.28	0.17	**0.29	0.38	0.11
VADER	-	**0.77	**0.31	0.32	*0.26
LIWC	-	-	*0.22	0.31	0.19

Table 5: Spearman correlations between various techniques for estimating *negative* affect scores from PANAS, i.e. PANAS(–). Human review performs better when the reviewers feel confident about their estimate, although sometimes not statistically significant. LIWC and VADER are highly correlated, but less correlated with human review. Even so, VADER performs similarly to human review with confidence ignored ("All" column) $^*p < 0.05$, $^{**}p < 0.01$

5.1.2 LIWC

When trying to predict affect, LIWC weakly correlated with both PANAS(+) and PANAS(–) scores (not statistically significant in both cases). Over the 110 sessions in our randomly sampled testing set, LIWC received a Spearman's $r_s = 0.14$, $p = 0.14$ with regard to PANAS(+) scores, indicating a non-statistically significant but weak correlation. For PANAS(–) scores, LIWC received a Spearman's $r_s = 0.17$, $p = 0.08$, which is a similarly weak, not statistically significant correlation.

This may indicate that words used at a specific time have some correlation to the affect of a person at that same time. There seems to be some relationship between the messages an individual sends and their affect, but it is hard to draw conclusions in this case given the lack of significance.

5.1.3 VADER

Across both PANAS(+) and PANAS(–) scores, VADER performed moderately well. In particular when PANAS(+) scores were treated as ground truth, VADER outperformed LIWC, though both were outstripped by human review, earning a Spearman's correlation of $r_s = 0.20$, $p = 0.03$. VADER exceeded our expectations across ground truth PANAS(–) scores, obtaining $r_s = 0.28$, $p < 0.01$ and performing better than LIWC and nearly as well as human review.

This indicates that VADER is a clearly better choice than LIWC for identifying affect from private messages. It suggests that VADER approaches a potential upper bound for how well an automated technique can perform given that VADER analysis approaches human accuracy, especially in the case of negative affect detection. Even when selecting only the most confident of predictions for negative affect, human reviewers only reach $r_s = 0.38$ correlation, which is only slightly better than VADER's prediction. Though VADER performs comparatively less well for positive affect prediction, these findings indicate a promising trend for automated affect detection.

5.2 Comparative Performance between Techniques

Aside from their correlation with PANAS scores, each technique can be compared with each other to identify discrepancies. We can use these discrepancies to detect how we might be able to improve the techniques in the future.

LIWC and VADER were found to be significantly strongly correlated across both PANAS(+) ($r_s = 0.75$, $p < 0.01$) and PANAS(–) ($r_s = 0.77$, $p < 0.01$) scores. This is as expected, since VADER was originally partially based on LIWC's lexical dictionary. Furthermore, since both techniques were developed with human predictions of textual indications of emotion [38, 20], we would assume

both to have high correlations to human review scores. In accordance, both LIWC and VADER correlated more strongly with human review than with PANAS scores when taking into account all samples, especially for PANAS(+) scores (Tables 4 and 5).

LIWC's correlation with human review was found to be statistically significant in PANAS(+) ($r_s = 0.29$, $p < 0.01$) and PANAS(–) ($r_s = 0.22$, $p = 0.02$). VADER followed this trend, obtaining a statistically significant correlation in both PANAS(+) and PANAS(–) ($r_s = 0.27$, $p < 0.01$, and $r_s = 0.31$, $p < 0.01$, respectively), as well as outperforming LIWC in the PANAS(–) case.

Similar trends of LIWC and VADER prediction of PANAS scores were seen when comparing sessions where human raters were confident in their scores. Although some were moderate, none of these correlations were found to be statistically significant. This lack of significance is likely due to the low statistical power due to the lack sparsity of confident ratings ($N = 21$). In these sessions, human reviewers predicted PANAS scores more accurately than both LIWC and VADER. VADER correlated more strongly with PANAS(–) ($r_s = 0.35$, $p = 0.11$) than PANAS(+) ($r_s = 0.25$, $p = 0.27$). VADER also outperformed LIWC in these cases.

Interestingly, unlike VADER, LIWC correlated more strongly with confident scores in PANAS(+) than in PANAS(–) ($r_s = 0.16$, $p = 0.48$ and $r_s = 0.08$, $p = 0.71$, respectively). Despite its similarity to VADER, it is notable that LIWC severely under-performed comparatively.

In comparison, across sessions when reviewers were not confident, the accuracy of reviewer scores and LIWC and VADER's predictions dropped for the most part (LIWC has a high correlation to PANAS(–) on the not-confident sessions). As before, for PANAS(+) VADER ($r_s = 0.21$, $p = 0.05$) outperformed LIWC ($r_s = 0.14$, $p = 0.19$), but both performed worse than human reviewers ($r_s = 0.28$, $p < 0.01$). Note that both human review and VADER correlated significantly, but LIWC did not. However, for these sessions where human reviewers were not confident, VADER actually predicted PANAS(–) ($r_s = 0.26$, $p = 0.01$) better than both LIWC ($r_s = 0.18$, $p = 0.08$) and human review ($r_s = 0.11$, $p = 0.29$). This means that human reviewers' relatively successful performance for PANAS(–) prediction was heavily dependent on sessions with confident ratings (21 of 110 sessions). Also notable here is that, despite the relatively high number of sessions included in the set ($N = 89$), human review correlation with PANAS(–) was not significant (Tables 4 and 5).

VADER's overall higher performance in comparison to LIWC over all PANAS score predictions raises several questions. One might expect LIWC and VADER to perform comparatively given their similar structure and the high correlation between their scores. However, because VADER is specifically optimized for social media [20], VADER is likely better than LIWC at interpreting patterns of speech common to social media platforms, and therefore more closely mirror patterns of human review.

5.3 Mispredictions Across All Techniques

Differences in LIWC, VADER, and human review ratings can occur for a variety of reasons. We have identified several cases where we observed common mispredictions for each of the three techniques when predicting self-reported affect. To pinpoint mispredictions, we manually reviewed the performance of each metric. Particularly, we selected sessions where sentiment analysis methods were highly unsuccessful in emotion prediction, or where different methods arrived at substantially different scores for the same session. We then conducted a qualitative analysis of message content for these selected sessions and found several common patterns at points of inaccurate prediction.

Session	PANAS		Human		VADER		Select example illustrating different types of mispredictions
	(+)	(–)	(+)	(–)	(+)	(–)	
410	34	35	25	35	0.14	0.11	"...i really wish i could spend less time venting but i dont know how to not breakdown... i am just emotionally mentally and physically exhausted..."
707	29	37	10	40	0.20	0.16	"i cried myself to sleep for like 30 minutes once u left ... yesterday", "during the day i'm just feeling hopeless"
258	14	11	32	14	0.20	0.10	"HAPPY BIRTHDAY", "Amazing! I think it's great so I hope they will too!! Yayyy!"
668	18	25	35	22	0.16	0.10	"I'm sad and tired just laying in bed" and "I almost started sobbing"
749	13	28	25	27	0.14	0.04	"I tried chatting with my professor today about getting my paper deadline pushed", "I'm asking my chiropractor... she can probably get permission quicker than my physical therapist"
948	13	15	35	14	0.17	0.06	"Guess who might have just won ##$??", "bro thats rough", "don't worry I'm going home"

Table 6: Perturbed examples where the PANAS scores by participants differed from those decided during human review. In these sessions, the PANAS(–) scores were similar, but the human labeled scores were lower than PANAS for sessions 410 and 707, higher in sessions 258, 749, and 948. All examples have been perturbed for privacy. Note that the average PANAS(+) score is 29.7, the average PANAS(–) score is 14.8, and both have a range of 10–50. Values above and below these averages can be treated as "relatively high" and "relatively low" values for both PANAS and human review scores. Though VADER is on a different scale (from 0 to 1), values above or below its means (0.21 for PANAS(+) prediction and 0.07 for PANAS(–)) can also be treated as "relatively high" or "relatively low".

5.3.1 Over-reliance on tone or context

LIWC and VADER function fundamentally by making predictions based on the vocabulary and tone used in a session. Human reviewers, however, rely much more heavily on context and situational content when trying to predict. Both of these focuses can lead to major errors in affect prediction, depending on the session.

We find exclusive reliance on tone, as LIWC and VADER do, to be an inconsistent indicator of true affect values, likely because it does not scale for situational context. We see many cases where human reviewers note using contextual clues to help predict affect, and then proceed to predict much more accurately than LIWC and VADER on the same session. One example of this is session 749, where a participant is trying to reschedule a paper deadline, and requires a dean's permission and a doctor's note, as shown in Table 6. Since the language used is not angry and contains few traditionally negative words, LIWC and VADER predict a low PANAS(–) score while the human review accurately predicts a high PANAS(–) score, based on the shared knowledge that situations such as these are often very frustrating.

In contrast, human reviewers on occasion focused too much on message content, especially when confident. For instance, in session 948 (shown in Table 6) one participant mentioned winning money from a study they had taken part in, and using it to buy gifts for their parents. The reviewers rated this as high PANAS(+), and noted that they had imagined the participant would be excited over winning money. However, the participant had an unusually low PANAS(+), a fact that was identified by the automated techniques. While the human reviewers made an assumption about the participant's mood based on the content of the conversation they read, LIWC and VADER's focus on sentence tone was able to more accurately predict ground truth scores.

Further investigation into additional latent clues for affect may be warranted. As a result, it is unclear which technique is better in this sense, as both have clear shortcomings.

5.3.2 Message weight based on temporal distance

Human reviewers reported placing very different weight on messages based on their temporal distance from PANAS completion, giving messages closer to PANAS completion time greater weight than those that are further. LIWC and VADER, on the other hand, place equal weight on all messages in a given time frame.

We would intuitively assume that messages sent very long before or after a PANAS measurement would have little bearing on the measurement itself. And indeed, there are instances where we see that discounting very distant messages seems to improve prediction accuracy. However, it is important to also consider situations where very distant messages can provide evidence of the actual affect of a participant. For instance, in one session reviewers noted that the message, "I just wanted to make myself feel a little better" was noted as indicative of negative affect, but was ultimately discounted when making their prediction because it was sent the previous day, and the tone had become less negative in recent messages. The true PANAS(–) was higher than reviewers predicted. It may be that this message was indicative of a broader negative mood state, and human reviewers missed the signs of this. LIWC and VADER, however, would perform better in these situations as they do not discount statements based on temporal location.

5.3.3 Inability to dissociate positive and negative affect

As mentioned earlier, PANAS(+) and PANAS(–) are considered independent dimensions of affect [53]. While intuitively we would assume a high PANAS(–) to indicate a low PANAS(+) and vice versa

(i.e. a negative correlation), this is not found to be the case with actual PANAS scores. PANAS scores in our data actually showed the opposite with a positive correlation ($r = 0.20$, $p < 0.01$).

Despite this, we found a negative correlation between predicted PANAS(+) and PANAS(–) scores for all sentiment analysis techniques. This correlation was particularly prevalent in human review ($r = -0.51$, $p < 0.01$), though LIWC and VADER both had weak, non-statistically significant, negative correlations between PANAS(+) and PANAS(–) ($r = -0.06$, $p > 0.05$, and $r = -0.16$, $p > 0.05$, respectively). The general failure to dissociate positive and negative affect is clearly exemplified in situations where participants report high scores for both PANAS(+) and PANAS(–). For example, in session 410 shown in Table 6, the participant messaged, "i am just emotionally mentally and physically exhausted." Despite this they reported both a high PANAS(+) and a high PANAS(–). Both human reviewers and VADER underestimated PANAS(+) scores for this session, likely due to its explicit account of negativity. Situations such as this one explain many of the mispredictions in human review ratings. The fact that human reviewers were the most susceptible to this type of error may have interesting implications in the clinical setting: would a licensed psychiatrist make the same "mistake"?

6 Improvements on Automated Systems

Based on the mispredictions we identified, we tried to improve the accuracy of automated sentiment analysis systems when applied to private messages. VADER performed comparatively well when predicting affect up to this point, even correlating comparably to human reviewers for ground truth negative affect, so we used VADER as a basis for improvements. As the inability to dissociate positive and negative affect was more relevant to human review, we focused our improvements primarily on incorporating context and message weight based on temporal distance to the model.

6.1 Analysis of Contextual Elements

We first isolated the set of sessions where humans perform much better than VADER. For each of the 110 sessions that were rated by human reviewers, we found the z-score for the positive and negative PANAS scores, hand ratings, and VADER scores. We then selected the cases where, for either positive or negative affect, human ratings had the same sign as the ground truth scores and VADER had the opposite sign, or human ratings were at least one standard deviation closer to the ground truth scores than VADER scores were.

These sessions and the human reviewer comments and reasoning for their ratings were analyzed for themes that might be contextual indicators of affect. These contextual themes can be split further into two categories, those based on content, and those based on metadata.

6.2 Improvements on VADER Using Content-Based Contextual Themes

VADER makes its predictions based on the given text string passed into the program. Any improvements based on content would use only the same information as that fed to VADER, therefore we tested our content-based improvements on top of the existing VADER system. As a basis for this, we use an improved version of VADER, currently available on Github, a different version to the one used in previous experiments. The performance of this version can be seen in Table 7.

For these changes, we focused on a further constrained set of sessions (the "sample set"). We pulled all sessions where humans performed much better than VADER for *both* positive and negative

	PANAS(+)	PANAS(–)
VADER (applied to full text of sent messages)	*0.23	**0.28
Infinite Punctuation	*0.23	**0.27
All Caps	*0.23	**0.28
Add Swears to Booster Dict	*0.23	**0.24
Add Swears to Booster Dict And Count Regular Weight	*0.23	**0.27
Add Swears to Booster, Count Regular Weight, Extend Booster Influence	***0.24**	**0.28
VADER (averages)	0.13	**0.25
VADER Weighted (averages)	0.09	****0.26**

Table 7:

affect. That is, where the z-score of human reviewer scoring was at least one standard deviation closer to the ground truth z-score, or was the same sign as ground truth while VADERs was not for both the positive and negative PANAS scores.

We built and tested modifications on VADER on this subset and analyzed how these changes were reflected in VADER scores. For each session in the subset, we calculated VADER scores *per sent message*, first with unmodified VADER as a base case, and then with our potential improvements. We then compared the message by message changes in VADER's scoring, to see what the exact effects of our changes were.

After investigation of the VADER code, we found that all of our suggested content-based improvements were already implemented in VADER in some form, though not the exact implementation we wanted. This indicates that the patterns we identified were likely also found by the developers of VADER in their own experiments. However, while our proposals were in line with patterns already covered in VADER, we proposed potential changes or expansions to the current implementation.

6.2.1 Repeated Punctuation as a Marker of Emphasis

Our qualitative analysis found that repeated punctuation was often a good indicator of emphasized emotion. For example, the message "I'm so excited" likely indicates positive affect, but the similar message "I'm so excited!!!!!" is a much stronger indicator of positive affect. Similarly, "I'm so mad!!!" indicates a stronger level of negative affect than "I'm so mad" without the repeated exclamation marks.

The number of times a punctuation mark is repeated also has an effect on the emphasis. Two question marks indicates a higher level of emotionality than one, three indicates a higher level than two, etc.

In its current form, VADER only applies increased emphasis to a message for between 2 and 4 exclamation points or question marks. However, this doesn't account for cases where people use many punctuation marks to convey even stronger emotional responses. We found many examples where participants used more than 4 punctuation marks in their messages, thus we allowed for infinitely many punctuation marks to apply emphasis.

Following this change we saw no changes to scores in the sample set. We find multiple instances where we would expect a result, but because the message originally had no positive or negative weight, an emphasis does not cause any change. For example, session 17 in Table 11 is, to humans, a clearly negative message, further emphasized by the frustrated statement "brown funds me!!!!!!!!!!!", however it lacks any clearly negative individual terms so VADER marks it as neutral and the

repeated exclamation points have no effect. The system did seem to be impacted correctly by mostly simple, positive statements, for example session 410 in Table 11 is correctly categorized as a more positive statement than before. When we applied this modification overall, we saw little change, with a slight decline in the correlation between VADER and ground truth negative affect scores, as shown in Table 7.

6.2.2 Capitalization as a Marker of Emphasis

Similarly to how repeated punctuation is used, we found that capitalization is also used as a marker for emphasis. For example, the message "WOW AMAZING" indicates a higher level of positive affect than uncapitalized "wow amazing".

VADER in its current form does apply emphasis to capitalized words, but only when the entire message is not capitalized. The above example messages would therefore return the same VADER scores. This may be because in most cases positive and negative VADER scores compared to each other, so if both positive and negative values are emphasized equally (ie. all given text is capitalized) then the ratio is unchanged and emphasis needn't be applied.

However, in the context of our system, the positive and negative scores for a single message aren't compared with each other, but to those of *other messages*. Taken in the context of a conversation, equal increases for positive and negative values in a message are still useful to record.

Therefore, we try removing this restriction, and allow emphasis to be applied to all capitalized words, regardless of the capitalization of other words in the given string.

Following this change we saw minor improvements in the sample set, for example the message from session 1027 shown in Table 11. However, this was mostly restricted to cases where we expected positive and negative affect to increase relative to the original VADER scores. Sessions where scores were meant to decrease performed poorly (ex. session 515 as shown in Table 11). Furthermore, we again encountered the problem where no terms in the given message have a weight, so emphasis causes no change in the final score, for example session 17, under Capitalization to Mark Emphasis, in Table 11. Given the half and half nature of effects on the sample set, it perhaps makes sense that overall scores saw little change, with a slight decline in the correlation between VADER and ground truth negative affect scores, as shown in Table 7.

6.2.3 Swears as Markers of Emphasis

Intuitively, we would assume that swear words, abbreviations of swears, or replacements for them, are indicators of negative affect. We see this reflected in the VADER lexicon, where most of these terms are given strong negative values.

However, our analysis showed that there are also cases where swear words are not used as a negative exclamation, but instead are used for emphasis. For example, in the message "that's so fking funny", the term "fking" here is being used to emphasize the term "funny", so treating it simply as a negative term would lose the contextual significance of how it was used.

VADER does have an implementation for terms used for emphasis. It includes a dictionary of "booster" terms that strengthen the positive or negative association (depending on whether the affected term is positive or negative) to any terms that occur within three words of the booster term. These terms are also ignored when it comes to the weight they would normally have on the VADER score. The terms in the dictionary include mostly adverbs, but also a small set of other words and phrases that would be used for emphasis (eg. "extremely", "enormous", "friggin").

Based on our analysis, we expect a broader set of swear words to be used for emphasis than are currently represented in the booster dictionary. We therefore added an additional set of swear words

	PANAS(+)	PANAS(–)
VADER (applied to full text of sent messages)	*0.23	**0.28
Time of PANAS recording	0.01	0.01
Average Time Between Sent Messages	*-0.07	**-0.32
Average Time Between Received Messages	**-0.16	**-0.25
Average Sent Message Length	0.05	0.00
Average Received Message Length	**-0.13	-0.06
Number of Sent Messages	**0.11	**0.20
Number of Received Messages	**0.13	**0.20
Number of Sent Facebook Messages	0.01	**0.29
Number of Received Facebook Messages	0.03	**0.24
Number of Sent Text Messages	**0.09	*-0.06
Number of Received Text Messages	**0.12	**-0.11

Table 8: Spearman's correlations between various metadata-based contextual measures and ground truth PANAS scores. While no methods perform comparably or better than VADER overall (for both positive *and* negative estimation), two measures (average time between sent messages and number of sent facebook messages, shown bolded) correlate comparably or better than VADER for *negative* PANAS scores. *$p < 0.05$, **$p < 0.01$

as booster terms, listed in Appendix A, Table 10.

This change showed some improvements, but primarily worsened results in the sample set. Closer inspection shows that this seemed to be primarily due to the fact that booster terms are not checked for their own positive or negative scores. Because swear words were added to the booster dictionary, the negative scores that would normally be ascribed to them are no longer being calculated. Analysis of the sample set shows that the decrease in negative scoring from those terms is likely the cause of a decrease in negative score accuracy that we see when we test this version overall (see Table 7).

Based on this finding, we modified VADER so that booster terms both applied their normal weight as well as a boost to terms within the three words following the booster term.

While this did mitigate the problem and swear words were no longer being treated as neutral terms, we found that in the sample set there was a miniscule effect to VADER scores. This was due to the fact that most of the booster terms we added were further than three words away from another term with a weight, so the boost was not being applied to anything. We then increased the range at which a boost was applied, so that terms up to 10 words away would be affected by the boost, with the weight of the boost decreased by 5% for each additional step of distance between the booster and the boosted term.

This showed great results in the sample set. While there were still some instances where VADER scores changed in an unfavorable direction, they were heavily outweighed by desired changes to results. Beyond that, many of the cases where the algorithm performed poorly were actually as a result of other terms in the booster set, for example, in Appendix B, Table 11, session 1027 has the message "and ill be out in the world so thatll be fun :) If I decide to drop ill come see you", in this case, the term "so" is acting as a booster and affecting scores, it was not part of the additions we made. Despite the positive results, overall this modification has an extremely small effect on results, with a .01 increase to correlation for PANAS(+) compared to original VADER.

6.3 Improvements Using Metadata-Based Themes

While VADER and the potential improvements discussed previously use message content as the main indicator of affect, there are also more general contextual elements that may be helpful in predictions. Here we look at contextual elements using message metadata only as a predictor of affect. Metadata as a predictor of affect is further backed by previous work, for example studies have shown that day of the week [52] and time of day [16] can effect population affect.

We primarily investigate 4 types of metadata, time of affect recording, message frequency, message length, and message platform, with modifiers regarding whether messages were sent or received. Correlations for each of these measures with ground truth PANAS scores are shown in Table 8.

With the exception of two measurements, the average time between sent messages and the number of sent Facebook messages, no measurement performed even comparably with VADER for either positive or negative affect. Both of these exception correlate comparably or better than VADER to *negative* PANAS scores, and very poorly with positive PANAS scores. They are also likely covariates, since the time period over which we look at messages is constant ($\pm$12 hours from the time the PANAS score is recorded) and Facebook is the most common messaging platform, as the number of Facebook messages sent of the 24 hour period increases, we would expect the average time between sent messages to decrease. Which is consistent with our findings here.

Interestingly, with those same two cases as exceptions, across the board otherwise metrics using received messages performed better than equivalent metrics using sent messages. This may indicate that *the reactions of others* to messages may be a better indicator of current affective state than ones own messages. Perhaps because a conversation partner theoretically knows the sender better and can more correctly interpret and then react to messages sent.

6.4 Addition of Message Weight by Temporal Distance

To this point we investigated improvements based on the misprediction regarding lack of context considered by automated systems. However, here we address the misprediction regarding message weight based on temporal distance to affect recording.

As weight must be applied on a message by message basis, for these calculations we must compare to the correlation between PANAS and baseline VADER scores calculated by *averaging per-message VADER scores for the whole session.* These alternative baseline can be seen in Table 7. Final weighted scores are also calculated as an average of the final weighted scores of each message in the session.

We use a the exponential equation $weight = 1.11 * (0.79^{|timeDelta|})$ to calculate the weight for each message. This equation was chosen based on the following metrics 1) it allows for a slightly higher weight to be applied for messages sent at exactly the time at which the affect score was recorded, 2) because it is exponential, the result will always be greater than zero, 3) "halfway" to maximum temporal distance we allow from the time of recording (6 hours out) has a weight close to 25%, this chosen based on human reviewer comments.

When we apply this to all sessions and look at how the z-scores for the sample set of sessions have changed, we see improvement across the board. When compared to z-scores using the averaging method for VADER score calculation, all 5 of 5 sessions in the sample set perform better, for both positive and negative affect. Despite this, when we look holistically at the correlations between the weighted and unweighted VADER versions, weighting seems to have decreased accuracy for positive affect (neither version was significant), and only very marginally improved correlations for negative affect (by 0.01, see Table 7).

7 Discussion

7.1 Reflection on Affect Predictions

7.1.1 Features of private messaging shared with other successful textual sources

When compared with previous research on affect detection, private messaging is found to be a promising textual source for analysis. As shown in Table 9, previous research comparing LIWC or VADER to PANAS scores has used public social media data [6, 7], personal diary entries [50], or spoken responses [8]. Across multiple studies examined, social media posts were consistently outperformed by all other textual sources. This may indicate a level of emotional distancing from public social media posts as opposed to something more private like a diary, or something relatively unfiltered like speech. Overall, direct speech and our VADER analysis over private messaging data were the most accurate predictors of affect, aside from human review.

Across all studies discussed, LIWC analysis over natural speech performed with the highest accuracy [8]. Private messaging data and direct speech share several key components, which likely contributes to their increased reflection of affect. For instance, both encourage rapid communication and have an assumed lack of statement revision. A study conducted by Lyddy et al. revealed texts to serve as a method of fast, unedited communication, with the average word count being 14.3 words and 17% containing misspellings (SD=12.0) [28]. Thousands of initialisms (e.g. LOL, brb, and ttyl) have been developed to aid in communicating quickly and further limiting a need for revision.

Because private messages are less likely to be revised to fit a wider audience, messaging data is likely to be more reflective of true affect than public data. When users revise statements, original emotional tone may be edited out. We speculate that one of the main reasons LIWC and VADER were able to detect affect from direct speech was because speech affords little time for editing and influencing tone. In contrast, public posts on social media may be edited heavily before posting. This calls for further investigation into the influence of statement revisions on the accuracy of sentiment analysis techniques, as well as further explanations for the strong reflection of affect found in private messages.

The second most successful technique found in previous studies was LIWC analysis over short daily diary entries [50]. The shared relative success of our method and this one may be due to the similarities between diary entries and private messages. Both are comparatively private methods of communication and thus have a level of perceived control when it comes to who has access to the text. While messages can be forwarded or shared with others outside of the conversation, users in a private messaging conversations still decide who sees the message. This is especially apparent when compared with public social media posts, which are visible to anyone, even people entirely outside of the poster's network.

In addition to the previous points about rapid communication and lack of revision, this idea of privacy and control can also be extended to the Cohen et al. speech study based on how it was conducted [8]. Participants were speaking directly to a single group (i.e. the researchers), and could reasonably assume that no one but the researchers would ever have access to that information. It may be due to the combination of all of these factors that LIWC analysis over natural speech by Cohen et al. was so successful.

7.1.2 Comparing LIWC, VADER, and Human Review

While both human review and VADER showed significant correlations in this study, it should be noted that LIWC analysis using private messaging data was not more accurate than VADER analysis

	Beasley & Mason				Tov et al.	Cohen et al.	Beasley et al.†		This Paper	
	LIWC		VADER†							
	Facebook	Twitter	Facebook	Twitter	Diary Study	Speech	Facebook	Twitter	LIWC	VADER
PANAS(+)	-0.07	0.05	0.05	0.07	**0.21	*0.24	**0.14	*0.14	0.14	*0.20
PANAS(–)	-0.11	0.04	0.02	0.09	**0.22	*0.29	**0.13	**0.16	0.17	**0.28

Table 9: Different sources of text data from various prior literature report low correlations between PANAS and LIWC; † indicates a correlation between PANAS and VADER rather than PANAS and LIWC (Beasley & Mason, Beasley et al.); Note that Beasley & Mason achieve slightly higher correlations when using a wider time range than the one presented here. We only include the results for the time range shown as it is the time range most similar to ours. Compared to other studies, messaging data analyzed in our paper achieves a similar accuracy range with LIWC, but notably better with VADER, especially for PANAS(–). $^*p < 0.05$, $^{**}p < 0.01$

over public posts done in previous studies [7, 6]. It seems, then, that when looking at social media data (i.e. public posts or private messages over social media platforms) VADER is truly a more accurate predictor of affect, likely because it was specifically optimized for use with social media.

Even after analysis and implementation of potential improvements to VADER, no improvement was able to perform better than human review. However, the marginal improvements reveal that improvements to automated systems such as VADER do have the potential for improvement. Particularly, incorporation of contextual metadata shows promise. The only tested metric that performed better than human review for either positive or negative affect was average time between sent messages ($r = -0.32, p < .01$). That same metric performed extremely poorly as a correlation to positive affect, however.

With this being said, the most accurate affect detection technique tested (i.e. human review) still had only moderate strength. Although developers of LIWC and VADER sought to incorporate aspects of human review, the primary intent of both systems is to identify tone, without taking context into account. As human reviewers reported primarily focusing on content over sentence tone, this contrast likely accounts for the disparity between human review and LIWC/VADER's ability to identify affect. Because human review detected affect better than LIWC/VADER, it is logical to conclude that sentence tone may carry less weight on true affect scores. It has become apparent that, while these measures may be useful in conjunction with other work, they should not be treated as a replacement for direct measures of affect (such as PANAS).

7.1.3 Suggestions for future work

Qualitative results indicate that LIWC and VADER, systems relying heavily on word tone, perform poorly when affect manifests itself "contextually" in the text (e.g. a participant describes a situation that may not have negative language but was in context a negative experience).

Although human reviewers had better performance than all other sentiment analysis techniques, their correlations with true PANAS scores were still moderate at best. This implies that human interpretation may not be the "gold standard" for understanding human emotion. Although it is intuitive to use humans as ground truth for sentiment analysis, our findings reveal this may not be reflective of true affect.

As mentioned previously, studies have found that users may prefer to share emotional posts differently on different platforms [4, 35]. We have postulated that when editing posted content, users may remove emotional tone and dampen tone indicators used by LIWC and VADER. However, it is important also to consider why users may do this. We know that users are likely to not only

intentionally project a certain persona when posting online, but also tend to do something similar among social groups in person [5]. It makes sense, then, that users are likely to do the same over direct messaging, and emotional expression will not necessarily match actual emotional state, though the discrepancy is likely to a lesser degree than in public posts due to previously discussed features of private messaging. It is therefore likely that predictions by human reviewers will be influenced by participant's attempts to obscure their feelings in messages. Perhaps we need to look beyond message content in order to more accurately identify emotional state.

We argue that a more optimal method would combine insights from: sentence tone, used primarily in sentiment analysis; insights from situational context, noted in human review; and also more obscure data outside of message content. We tested a number of different representations of these insights, finding promise especially in data outside of message content. This warrants further investigation into alternative parameters for use in affect detection in order to develop a more refined sentiment analysis technique. These parameters could include: who does a participant talk to, how often do they message, and when do they message? Previous studies have shown that there is fluctuation in population affect depending on the day of the week [52] and possibly even the time of day [16]. If participants have a fixed routine (something that does not hold for our college population data), these trends may be even more significant. While contextual cues contribute to stronger affect detection, as seen in human review, they alone are insufficient indicators.

Until such a technique is developed, we recommend solely using self-report scales, such as PANAS, as ground truth for affect.

7.2 Ethical Considerations

A key requirement for the functionality of Sochiatrist is informed consent and a benefit to the user themselves. Participants are physically present for all data extraction, and they enter their own login information after consenting to extraction for each platform. The participants have an opportunity to review the messages that will be extracted, and to remove specific lines from the dataset. These procedures are consistent with the human subjects full board review that occurred before the start of the study.

Studies in the past have addressed ethical concerns in a few ways. Most other studies collected publicly available posts only (e.g. [6, 7, 11, 14, 16]), others only collect data participants specifically record for the study (e.g. diary entries [50] or recorded speech [8]), while yet others ask participants to copy over messages by hand in the lab [5]. All of these methods aim to only collect data that participants are actively comfortable sharing with the researchers or the general public (in the case of public posts). However, we believe that by requiring participants to specifically opt-in to each individual platform, and also providing them the option to additionally remove any individual messages or messaging threads after extraction, achieves the same goal. This is supported by the fact that none of our participants mentioned objections or discomfort with sharing their data during post-study followup interviews.

It is important to note that only messages sent from a participant are included in this study. Therefore messages from other individuals to the participant (third parties) are not collected in this study, as those senders were unable to consent to message extraction. Additionally, pseudo-anonymization reduces the risk that participants expose third parties' names or identifying numbers in their own messages.

Once we move beyond this study and into day to day usage, this kind of data extraction can be used as a personal tracking tool. Users may use message extraction and analysis to track mood events and gain personalized insights into their own emotional triggers. Other mainstream self-tracking

systems store data in remote servers, while the extraction and analysis for our system is local, and therefore can be used without an external party accessing the data. By storing this data locally, extracted data is no more accessible than users' private messages on applications they already have installed on their phones.

Clinicians can also benefit from a more overarching or long-term view of patient mood states. Sochiatrist or a system like it could be used by patients to generate data they are comfortable sharing. This would allow clinicians to gain deeper understanding of patients' affect without the need for invasive procedures like forcing them to explain and relive traumatic experiences or going through and monitoring patients' messaging history by hand. Patients may find it less stressful to present information to their therapists using a data extraction system such as Sochiatrist, rather than explaining a situation verbally.

7.3 Limitations

There were a few areas where this study was limited. First, we were unable to make a direct comparison between public and private messages to prove strongly our hypothesis that private messages may be a richer data source for estimating affect. To examine the specific utility of direct message data as a space for people to express their most intense and private emotions [4], we purposely chose to test the extreme case where no public data is used at all. In practice, Sochiatrist can extract both public and private data for analysis, but for this study, nearly 99.9% of the messages in our collected data from the participants are from private messaging sources and thus there is an insufficient sample of public messages to make any meaningful comparison between the two. However, the fact that public messages are such a rare occurrence in our study demographic is an indicator that further investigation into private messages may be required.

The study population was limited to college undergraduates in the United States, so due to differences in expression across cultures and age groups, the results may not be generalizable to other demographics. Additionally, the timeframe of this study could have introduced several confounds. Participant data was collected from November to December, around the end of the fall semester. During this period of time, many college students were likely to be focused on academics and exams, which could influence their affect and introduce bias into the data. However, we still believe that the methods and systems provided here make it possible to reproduce similar studies across other demographics.

We also may not have had access to all of participants' messaging history. There are common messaging apps that Sochiatrist does not support, such as Telegram and Signal. Our study design ensured that participants must use one of the Sochiatrist-supported applications, but does not guarantee that it is their "main" texting application or that there is not a significant amount of data missing.

The length of our study was shorter than previous studies on mood and affect, which had data collection periods that ranged from 30 days [29] to 3 years [43]. This means that we don't have a long enough timespan per-participant to train participant-specific models. Furthermore, the number of participants ($N = 25$) was also small. It is generally more difficult to collect a large dataset of private messages than a large dataset of public posts, since people are hesitant to share their private data. However, in the future it would be ideal to collect data over a longer period of time with a larger sample of participants.

Due to resource limitations, there was also not as broad a set of human reviewers as we would have liked. The same three reviewers, all from an American collegiate background, reviewed each session. This makes it difficult to conclude that the human reviewed scores can be produced consistently.

A further study with reviewers from more diverse backgrounds would be needed to make broader claims about human review. Crowdsourcing solutions are unfortunately not always a viable method of human review, since the data is only pseudo-anonymized and sharing private data publicly on the internet quickly runs afoul of privacy standards. Future work computing the inter-rater agreement for this task is required to reveal how many human reviewers are required for a low variance estimate; if inter-rater agreement is low, then many reviewers are required, whereas if the inter-rater agreement is high, fewer are needed.

8 Conclusion

In this paper, we investigated the ability for sentiment analysis techniques to find an emotional signal in messaging data. Human review, especially when the reviewers were confident, had the highest correlation with PANAS, while VADER achieved similar correlation for negative affect. VADER, outperforming LIWC, is reasonably accurate for predicting affect, but does so differently than human review, as the correlation between the two is low. Both human review and VADER performed well on messaging data in comparison to past studies, most of which used public social media data. VADER's relatively strong performance when using messaging data supports a more in depth investigation of the difference between private messages and public social media posts. However, it should be noted that all correlations with PANAS, while many were significant, had only moderate strength at best. We would caution against treating any of these methods as a replacement for user-reported affect data gathered through PANAS.

What is perhaps most promising in the future is identifying the parts of human review that the automated techniques perform poorly in. Due to the evidence that human reviewers have the ability to gauge when they are likely more accurate, there must be some explanatory signal that the current automated techniques are not yet catching. Human reviewers are able to interpret both context and timing, and moderate the message based on the social relationship, likely playing a role in this. After identifying and trying to bridge these discrepancies, we were able to show very slight improvements to the automated VADER system, with particular success with metadata-based contextual information. These improvements indicate opportunities for better sentiment analysis in the future, which could feature measures that are currently only noticed by humans.

References

[1] Marije aan het Rot, Koen Hogenelst, and Robert A. Schoevers. Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, 32(6):510 – 523, 2012.

[2] Monica Anderson and Jingjing Jiang. Teens, Social Media & Technology 2018, May 2018.

[3] 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.

[4] Natalya N Bazarova and Yoon Hyung Choi. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657, 2014.

[5] Natalya N Bazarova, Jessie G Taft, Yoon Hyung Choi, and Dan Cosley. Managing impressions and relationships on facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology*, 32(2):121–141, 2013.

[6] Asaf Beasley and Winter Mason. Emotional States vs. Emotional Words in Social Media. In *Proceedings of the ACM Web Science Conference*, pages 1–10, Oxford, United Kingdom, 2015. ACM Press.

[7] Asaf Beasley, Winter Mason, and Eliot Smith. Inferring emotions and self-relevant domains in social media: Challenges and future directions. *Translational Issues in Psychological Science*, 2(3):238–247, 2016.

[8] Alex Cohen, Kyle Minor, Lauren Baillie, and Amanda Dahir. Clarifying the linguistic signature: Measuring personality from natural speech. *Journal of personality assessment*, 90:559–63, 12 2008.

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

[10] Lorenzo Coviello, Yunkyu Sohn, Adam D.I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. Detecting emotional contagion in massive social networks. *PLoS One*, 9(3), 3 2014.

[11] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[12] Munmun De Choudhury and Scott Counts. The nature of emotional expression in social media: measurement, inference and utility. *Human Computer Interaction Consortium (HCIC)*, 2012.

[13] Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, nervous or surprised? classification of human affective states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*, pages 435–438, 2012.

[14] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. pages 128–137. AAAI, July 2013.

[15] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, February 2015.

[16] Fabon Dzogang, Stafford Lightman, and Nello Cristianini. Circadian mood variations in Twitter content. *Brain and Neuroscience Advances*, 1:239821281774450, January 2017.

[17] 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.

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

[19] Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 214–225, 2019.

[20] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. pages 216–225, 2015.

[21] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. pages 151–160, 01 2011.

[22] Funda Kivran-Swaine and Mor Naaman. Network properties and social sharing of emotions in social awareness streams. In *Proceedings of the ACM 2011 Conference on Computer supported cooperative work*, pages 379–382, 2011.

[23] Adam D.I. Kramer. An unobtrusive behavioral model of "gross national happiness". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 287–290, New York, NY, USA, 2010. Association for Computing Machinery.

[24] Adam D.I. Kramer. The spread of emotion via facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 767–770, New York, NY, USA, 2012. Association for Computing Machinery.

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

[26] Nicole Krämer and Stephan Winter. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *Journal of Media Psychology: Theories, Methods, and Applications*, 20:106–, 01 2008.

[27] Danielle Lottridge and Frank R Bentley. Let's hate together: How people share news in messaging, social, and public networks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[28] Fiona Lyddy, Francesca Farina, James Hanney, Lynn Farrell, and Niamh Kelly O'Neill. An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication*, 19(3):546–561, 2014.

[29] 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.

[30] David N. Miller. *Positive Affect*, pages 1121–1122. Springer US, Boston, MA, 2011.

[31] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado, June 5 2015. Association for Computational Linguistics.

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

[33] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. "Suddenly, we got to become therapists for each other": Designing peer support chats for mental health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

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

[35] Myoungouk Park, D.W. McDonald, and M. Cha. Perception differences between the depressed and non-depressed users in twitter. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 476–485, 01 2013.

[36] Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

[37] James Pennebaker, Martha Francis, and Roger Booth. Linguistic inquiry and word count (liwc). 01 1999.

[38] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. 2015.

[39] Megan Ranney, Kyler Lehrbach, Nicholas Scott, Nicole Nugent, Alison Riese, Jeff Huang, Grant Fong, and Rochelle Rosen. Insights into adolescent online conflict through qualitative analysis of online messages. page 9. Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020.

[40] Andrew G. Reece and Christopher M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):1–12, December 2017.

[41] Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, 7(1):13006, December 2017.

[42] 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.

[43] 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.

[44] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4(1):1–32, 2008. PMID: 18509902.

[45] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[46] Aaron Smith and Monica Anderson. Social Media Use 2018: Demographics and Statistics, March 2018.

[47] Deborah M. Stringer. *Negative Affect*, pages 1303–1304. Springer New York, New York, NY, 2013.

[48] Yla R. Tausczik and James W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, March 2010.

[49] Auke Tellegen. Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. page 681–706, 1985.

[50] William Tov, Kok Leong Ng, Han Lin, and Lin Ge Qiu. Detecting well-being via computerized content analysis of brief diary entries. *Psychological assessment*, 25 4:1069–78, 2013.

[51] 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.

[52] Wei Wang, Ivan Hernandez, Daniel Newman, Jibo He, and Jiang Bian. Twitter analysis: Studying us weekly trends in work stress and emotion. *Applied Psychology*, 65:355–378, 04 2016.

[53] 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):1063–1070, 1988.

[54] Yong Sook Yang, Gi Wook Ryu, and Mona Choi. Methodological strategies for ecological momentary assessment to evaluate mood and stress in adult patients using mobile phones: Systematic review. *JMIR mHealth and uHealth*, 7(4), 1 2019.

Appendices

A

fuck	shit	damn	darn	shoot	yikes	
fking	af	bitchin	bitching	hell	frick	omg

Table 10: List of terms added to booster dict in VADER

B

VADER Version	Session	Desired Change To VADER Scores		Actual Change To VADER Scores		Select example illustrating VADER score change
		(+)	(−)	(+)	(−)	
Repeated Punctuation to Mark Emphasis	17	increase	increase	no change	no change	"fine I can ask again although I ALREADY DID brown funds me!!!!!!!!!!! i told you already!"
	410	increase	increase	increase	no change	"happy birthday!!!!!"
Capitalization to Mark Emphasis	1027	increase	increase	no change	increase	"THAT DOESN'T HELP"
	515	decrease	decrease	no change	increase	"WHAT THE HECK HOW DID THEY FIND OUT"
	17	increase	increase	no change	no change	"I'M AN ADULT SO I CAN FEED MYSELF"
Swears As Boosters, Regular Weight Counted, With Influence Distance Extended	315	increase	decrease	increase	decrease	"I saw the email, good shit. here's warm, but im excited to see my fam friends"
	1027	increase	increase	increase	decrease	"and ill be out in the world so thatll be fun :) If I decide to drop ill come see you"

Table 11: Perturbed examples where VADER modifications worked well or poorly.