In your wildest dreams: the language and psychological features of dreams

Kate G. Niederhoffer
Circadia Labs
kate@circadialabs.com

Jonathan Schler
Circadia Labs
schlerj@cs.biu.ac.il

Patrick Crutchley
Qntfy
patrick@qntfy.com

Kate Loveys
Qntfy
kate@qntfy.com

Glen Coppersmith
Qntfy
glen@qntfy.com

Abstract

In this paper, we provide the first quantified exploration of the structure of the language of dreams, their linguistic style and emotional content. We present a collection of digital dream logs as a viable corpus for the growing study of mental health through the lens of language, complementary to the work done examining more traditional social media. This paper is largely exploratory in nature to lay the groundwork for subsequent research in mental health, rather than optimizing a particular text classification task.

1 Introduction

Despite a prominent role in the origin of psychology (Freud, 2013; Jung, 2002), scientific research about the meaning and value of dreams has waned in the 21st century. Cartwright (2008), for one, has argued that dreams lost their prominence in the latter half of the 20th century as psychology attempted to become a more empirical science focused on observable behavior and mental activity and less reliant on memory. In the last decade, the distinctive brain patterns of dreaming have become more identifiable (Siclari et al., 2017) and research has amassed on the impact of dreams on waking life with links to mood (Cartwright, 2013), relationship health (Selterman et al., 2012) and decision-making (Morewedge and Norton, 2009). While scientists debate the purpose of dreams (Barrett, 2007; Cartwright et al., 2006), dreams continue to be a universal and time intensive experience across humanity.

Until recently, dreams remained an offline phenomena, qualitatively separate from other forms of social interaction via social media. Online platforms such as Facebook and Twitter are fertile grounds for research in social science (Wilson et al., 2012; boyd and Ellison, 2007) and more recently, in mental health via computational approaches in text analysis (Pennebaker et al., 2015; De Choudhury et al., 2013; Coppersmith et al., 2014) and network structure (Christakis and Fowler, 2014). However, dreams have remained as private, albeit important conversational currency (Wax, 2004). When dreams are studied, they are gathered from sleep labs, psychotherapeutic and inpatient settings, personal dream journals and occasionally classroom settings where "most recent dreams" and "most vivid dreams" are collected (Domhoff, 2000). The recent development of a social network dedicated to dreams offers scientists unprecedented access to the language of dreams at scale, collected with consistent methodology. Understanding the structure of this large corpus of dreams gives us access to previously unobservable mental activity and enables future research to identify abnormal patterns in themes, emotional tone, and styles associated with mental health diagnoses and therapeutic outcomes.

We begin with a brief overview of the impetus for this work and a discussion of related work in the intersection of dreams and text analysis. We then provide details on the corpus of dreams and discuss our results organized around three research questions. The paper concludes with implications for subsequent research on dreams, both to better understand nuances in the medium, and for mental health purposes.

1.1 Previous research on dream content and text analysis

Dreams are challenging to understand. Dreams are a diverse medium that vary from being perceptual or cognitive, from involving simple settings to complicated narratives, which may be similar or dissimilar to waking life (Siclari et al.,

Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology, pages 13–25,
Vancouver, Canada, August 3, 2017. © 2017 Association for Computational Linguistics

2017). Analyzing them is similarly complex; researchers have put extensive effort into the development of systems to score their global content, specific themes, psychological intensity, and theoretical underpinnings (Schredl, 2010). Different researchers, research goals, collection vehicles and analytic techniques present issues in replication, reliability and the validity of standardized methods for the content analysis of dreams. The Hall-Van de Castle coding system is the most comprehensive protocol for content analysis of dreams, with eight main categories and over 300 sub scales in the dream manual (Hall and Castle, 1966). Categories include: Physical surroundings (e.g. indoor, outdoor), Characters (e.g. persons, animals), Social interactions (e.g. friendly vs. aggressive), Activities (e.g. communication, thinking), Achievement outcomes (e.g. success, failure), Environmental press (e.g. fortune, misfortune), Emotions (e.g. anger, happiness), Descriptive elements (e.g. size, age, color), and Theoretical scales (e.g. castration anxiety, regression).

A handful of studies have used automated text analysis to explore dreams, specifically to discern differences from waking narratives and identify the relationship between dream language and personality (Hawkins and Boyd, in press), for automated sentiment detection (Nadeau et al., 2006) and to distinguish linguistic features from personal narratives (Hendrickx et al., 2016). To our knowledge, no study has examined as large a sample of dreams from a naturalistic setting (neurotypical research participants, online social context) across methodologies for psychological purposes (i.e. non classification/ non hypothesis driven).

Hawkins and Boyd (in press) analyze dreams across three samples of recent dream reports, two undergraduate and one sample from Amazon's Mechanical Turk[1]. Using Linguistic Inquiry and Word Count (Pennebaker et al., 2007), they find a distinctive pattern for recent dreams that differs from the base rate norms for waking narratives, specifically characterized by more function words, common words, pronouns, personal pronouns, first person pronouns, past tense verbs, and more use of words describing leisure activities; less use of present tense and future tense verbs, causation words, second person pronouns, numbers, swear words, and assent words. They did not find consistent relationships between dream language features and personality. Hawkins & Boyd's research paves the way for understanding how and why a dream narrative differs from a waking narrative and what these differences mean from a psychological perspective. For example, what does it mean for a dream to have more function words than a waking narrative? What is the relationship between the content of dreams and the more "invisible" word differences (pronouns, prepositions, articles)?

Nadeau et al. (2006) also used LIWC on dreams to gauge the efficacy of automated sentiment analysis to bypass human judges or dreamer estimates of emotion. Comparing the performance of LIWC, the General Inquirer, a weighted lexicon (HM) and standard bag of words approach, they find machine learning outperforms human judgments - and specifically demonstrate that LIWC and the GI have the best features for sentiment classification. While a step in a promising direction, Nadeau et al.'s sample was small (100 dreams from 29 individuals) and sentiment was classified on a limited negative scale (4-class, from neutral to highly negative) omitting nuance in the purported emotional content of dreams, c.f. Cartwright (2013).

Hendrickx et al. (2016) looked at the distinguishing features from dreams as compared to personal narratives (diary entries from Reddit and personal stories from Prosebox) via text classification, topic modeling and text coherence. The authors find dreams can be classified with near perfect precision based on the presence of uncertainty markers (somebody, remember, somewhere, recall) and descriptions of scenes (setting, riding, building, swimming, table, room), with lower discourse coherence. Personal narrative markers (non-dream) include time (2014, today, tonight, yesterday, day, months) and conversational expressions (please, :), ?, thanks). Hendrickx et al. also applied LDA topic modeling to explore the main themes in dreams as compared to personal narratives validating the classification results. Dream topics span everyday activities, setting descriptions, and uncertainty expressions. The Hendrickx et al. research is notable in its exploration of male vs. female topic distributions within dreams in addition to comparisons across corpus type (dream vs. personal narrative) though does not explore the relationship between topic

[1] Mechanical Turk users do short human intelligence tasks for small payments. For more see `http://www.mturk.com`.

and emotion and excludes the analysis of function words, which we believe is a critical piece in understanding the psychological value of dreams and dreamers, given previous findings (Chung and Pennebaker, 2007).

1.2 Relevant research on mental health and text analysis

Computational text analysis allows for assessment of larger samples and proactive identification of mental illness. Language in social media can indicate the likelihood a user self-reports a particular mental disorder (Coppersmith et al., 2015), or has received a mental health diagnosis (De Choudhury et al., 2013). The language of online dreams has yet to be analyzed relative to mental health conditions, however prior laboratory research suggests that dream content may differ between clinical conditions. We refer the reader to Skancke et al.'s comprehensive review of dream content grouped by clinical disorder (Skancke et al., 2014). In brief, patterns in emotional tone, themes, and actor focus have been associated with diagnoses of mood and anxiety disorders, schizophrenia, personality, and eating disorders. Though, it remains unclear whether dream content can distinguish between clinical disorders.

Nightmares are especially relevant to mental health, featuring as a diagnostic symptom for post-traumatic stress disorder (Campbell and Germain, 2016), and a common correlate with schizophrenia (Okorome Mume, 2009), depression and anxiety (Swart et al., 2013), and personality disorders (Schredl et al., 2012). Nightmare frequency and intensity have been positively correlated with incidence of suicidal thoughts and behaviors (Bernert et al., 2005), suggesting nightmares could be a near-term risk factor to assess during crisis. In sum, analysis of dream topics and emotional tone may provide some insight to the mental health of the dreamer.

2 Data

Dreams were collected from DreamsCloud, a social network for sharing dreams. DreamsCloud is available to the public; those who register for the site are informed that their data can be used for research purposes. DreamsCloud is moderated by professional dream reflectors who comment on dreams, in addition to the broader community of registered users who can also "like" and comment on dreams.

DreamsCloud has the largest available digital collection of dreams with over 119k dreams from 73k users and an overall community of over 300k registered users. Visitors to the site come from 234 countries (according to Google Analytics) and have shared dreams in 8 languages. DreamsCloud differs from online dream banks in that dreams are voluntarily shared for social purposes rather than collections from research studies.

A random sample of 10k English dreams over 100 words from September 1, 2013 through December 31, 2016 was used in this study. Data cleansing removed 322 dreams due to incorrectly classified language (Spanish), lyrics or news content copied from the Internet by the user, and duplicated data. The remaining sample included 9,678 dreams. No additional data about the gender, age, name, or ethnicity of the participants are included in our study. Only the original dream texts are analyzed. While DreamsCloud has comments and conversations around many of these dreams, we put off analysis of commentary for subsequent research and focus directly on the first-person accounts of dreams. The average length of dreams in the sample is 208 words (SD = 116.7). Data is organized by an encrypted alphanumeric Dreamer ID and a unique, encrypted alphanumeric Dream ID for each dream logged.

2.1 Ethical considerations

While community members agree to Terms of Service that explicitly state their content is owned by the company and will be used for research purposes, the nature of the content is very intimate. Because of the unknowns about the science behind why we dream, what our dreams mean, how dreams are related to life events, there is less of a stigma about sharing otherwise private or bizarre information. The site refers to dream-sharing as an "anonymous-as-you-want" activity. Although the analyses in this paper are structural and aggregate in nature, deeper analysis of this data could raise privacy concerns as well as questions about appropriate intervention. Our hope is that additional research in this area will shed light on the relationship between dreaming and waking life to help address these questions.

3 Results

Three approaches are used to examine the dream narratives: content analysis using an LDA topic model (Blei et al., 2003), analysis of linguistic style via function words using LIWC (Pennebaker et al., 2015), and categorization of emotions using an emotion classification model (Coppersmith et al., 2016).

3.1 The topical structure of dreams

Topic models are statistical models which discover topics in a corpus. Topic modeling is especially useful in large data, where it is too cumbersome to extract the topics manually. Due to the large volume of dreams in our corpus and the lack of prior knowledge about their subjects, we follow other content-based studies in employing topic modeling to understand the content of the dreams (Kireyev et al., 2009; Yin et al., 2011; Chae et al., 2012; Mitchell et al., 2015; Hendrickx et al., 2016). We analyzed the topical structure of the dream corpus using a popular topic modeling algorithm, latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is an algorithm for the automated discovery of topics. LDA treats documents as a mixture of topics, and topics as a mixture of words. Each topic discovered by LDA is represented by a probability distribution which conveys the affinity for a given word to that particular topic.

We used the LDA implementation available in the Mallet package (McCallum, 2002). We converted the text to lower case and, because the topic analysis is focused on content of dream narratives, excluded all function words and punctuation marks. (Function and style will be considered in the following section.) No reduction in inflection (i.e. stemming, lemmatization) was performed to satisfy the goals of exploring the nuance of dream narratives as a medium and subsequently make inferences about the psychological orientation of the authors (see section 3.2). Further, in order to make more valid comparisons to the existing literature based on human coding, it is important to understand how distributions of singular vs. plural nouns and present vs. past tense verbs, for example are distributed topically. We selected 25 topics for LDA to infer and used 2000 iterations of Gibbs sampling to fit the model. The number of topics was informed by maximizing the computed information gain of the resulting feature sets, while maintaining a reasonable training time.

LDA provides insightful information about the topics in the corpus. However, interpreting the 'aboutness' of a topic based on a list of words requires human judgment based on term frequency, exclusivity, meaning, and subjective inference. Interestingly, we found 23 of 25 topics to be interpretable based on semantic meaning and 2 (Topics 17 and 22) which appeared more syntactically related. Most heavily weighted topic words are quoted in results tables, and the full 25-topic distribution with manual labeling is included in Appendix A. Note that the topic number is randomly assigned by LDA and does not indicate anything meaningful like rank, weight, or importance.

Although we utilize a 25-topic solution as compared to Hendrickx et al.'s 50-topic solution, we see some consistency in the topics identified as characteristic of dream narratives. Specifically, we see similar support for the continuity hypothesis of dreams - that dreams are a continuation of waking life activities - in topics such as Topic 19 about School, Topic 12 about food and eating, and Topic 15 about driving and cars. Similar to their research, we also see clustering of present tense verbs in Topic 0, a water topic (11), and home settings topic (5). We see an almost exact replication of their "dreaming in general," in our Topic 18. Comprehensive comparisons in distributions or characteristic words are not possible with the data their published research makes available.

In inspecting the topical distribution and noting the support for the continuity hypothesis, what also stands out is the lack of support for the 'dreams-as-psychotic-state' hypothesis. Beginning with Freud and Jung, researchers have drawn similarities between dreaming and psychosis. These similarities range from phenomenological to neurobiological, qualitatively manifested as a loosening of associations, incongruity and bizarreness of personal experience, and distortion of time and space parameters (Scarone et al., 2008). Reviewing the content of our 25-topic solution, we see no reason to interpret the clustering of words within any given topic as incongruous nor do we detect support for the content to be evaluated as "bizarre" (Hobson et al., 1987). The topics instead appear closely aligned with reality, reflective or overt (actions) and covert (thoughts) behaviors and demonstrate semantic congruity within topic. However, an automated approach to coding as subjective a construct as bizarreness demands inspection beyond content words alone.

LDA is an effective means to understand the distribution of content words in a given corpus. Importantly, it was developed for the purpose of dimensionality reduction - document summarization and information retrieval (Blei et al., 2003). Some of the assumptions that enable the algorithms behind topic models, such as the exclusion of words that have no content relevance (e.g. function words), leave room for additional methods to explore the psychological meaning of a given document, the author's mindset, and emotions.

3.2 The linguistic style of dreams

Recent research on language from a psychological perspective demonstrates that function word use reflects and is a reliable marker of personality and a range of social and psychological processes, cognitive thinking styles and psychological states (Pennebaker, 2011). Pennebaker proposes that function words are the infrastructure for thought and perspective: they connect (e.g. conjunctions, auxiliary verbs), shape (e.g. pronouns) and organize (e.g. articles, prepositions) content. Content is important in dreams, and often metaphorical (Lakoff, 1993). The style in which we remember and share our dreams can give important clues to how we make sense of our dreams, and in turn, ourselves. Said another way, our goals in this paper are not just to explore the stuff that dreams are made of but the style of dreams as a reflection of the dreamers' psychological states. With multiple lenses on the data, we can obtain an enhanced picture of the psychological value of the corpus.

LIWC categorizes the words in a given text into approximately 80 variables. Variables represent the proportion of words in a given document (i.e. dream) that correspond to a lexicon composed of different categories of words, including function words (pronouns, prepositions), affect words (positive emotion, anxiety), and content words (money, religion, leisure activities). We reduced the window of interest in LIWC categories to function words, affect, and cognitive processes, as justified by what remains from the LDA analysis (e.g. functions words) and comparisons to results from the empirical literature described thus far (Hawkins and Boyd, in press; Nadeau et al., 2006). Table 1 shows the means and SDs for all LIWC categories within the Linguistic Processes dictionaries with Cognitive, Social and Affective Processes added. Unweighted means from the aggregated sample of expressive writing in Pennebaker et al. (2015) are provided for context.

As compared to the base rates from expressive writing (Pennebaker et al., 2015), a dream narrative comes across as a first person (1st person pronouns) account of a past event (past tense) with particular attention to people (family, friends, women, and men), objects (articles), locations (prepositions) and what is seen, heard, and felt (perceptual processes) more than known or understood (cognitive processes).

Low cognitive processes (M = 9.29; SD = 3.48) would suggest dreamers are not on a search for meaning in sharing their dreams, however it is unclear if this is a case of displaced cognitive processing due to the more dominant perceptual experience of dreams. Previous research indicates that narrative coherence has an inverse relationship with cognitive processing words (Klein and Boals, 2010; Boals et al., 2011). Boals et al. (2011) show that cognitive process words are related to sense making as a process which occurs prior to the development of a narrative (sense making as an outcome). This might suggest that dreamers do not tend to be caught up in why they had a given dream as much as explaining what happened. In other words, dreams are shared as complete stories. A dream narrative's low proportion of emotion words (Mean Affect = 3.42, SD= 1.90) are unexpected given recent research on the emotion regulatory function of dreams and call for additional investigation, which we address below. One possibility is the sensitivity of a lexicon-based instrument to the way in which emotions are expressed in dream narratives. In general, our findings are consistent with Hawkins and Boyd (in press), despite differences in the collection vehicle (recall: Hawkins and Boyd use the 'most recent dream' and 'most vivid dream' paradigm) and previous version of LIWC (2007 vs. 2015).

3.3 How is language style related to the content of dreams?

To explore the relationship between dream topic and language style, we focus on function words only: pronouns, prepositions, articles, auxiliary verbs, and negations. In particular, we use an index composed of the proportions of these classes of words called the Categorical Dynamic Index (CDI; Pennebaker et al. 2014) that measures the

	Dreams (n=9,678)		Expressive Writing (n=6,179)	
	Mean	SD	Mean	SD
Word Count	208.85	116.61	408.94	248.23
Words per sentence	30.34	40.49	18.42	14.89
Words < 6 letters	11.66	3.41	13.62	4.12
Dictionary words	91.87	4.06	91.93	5.03
Total Function Words	60.04	4.32	58.27	6.26
Total Pronouns	19.72	4.31	18.03	5.36
Personal Pronoun	14.87	4.17	12.74	4.28
1st person sing.	9.54	3.36	8.66	4.25
1st person plur.	1.24	1.54	0.81	1.22
2nd person	0.27	0.65	0.68	2.14
3rd person sing.	3.06	2.71	2.01	2.95
3rd person plur.	0.77	1.05	0.57	0.82
Impersonal Pronoun	4.82	2.13	5.28	2.36
Articles	6.99	2.62	5.7	2.56
Prepositions	13.99	2.67	14.27	2.82
Auxiliary verbs	8.08	2.38	9.25	3.06
Adverbs	5.03	2	6.02	2.3
Conjunctions	8.52	2.62	7.46	2.06
Negations	1.4	1	1.69	1.25
Cognitive Processes	9.29	3.48	12.52	5.11
Social Processes	11.18	5.07	8.69	5.46
Affective Processes	3.42	1.9	4.77	2.59
Positive emotion	1.64	1.4	2.57	1.63
Negative emotion	1.75	1.37	2.12	1.74

Table 1: Linguistic Processes Categories in LIWC2015

extent to which thinking is Categorical (high prepositions, articles) versus Dynamic (pronouns, auxiliary verbs).

The CDI is a simple unit-weighted computation which adds the proportions of articles and prepositions and subtracts personal pronouns, impersonal pronouns, auxiliary verbs, conjunctions, adverbs and negations. It has been shown to be a reliable marker of cognitive style which we use to understand differences in the experience of various topics in dreams. Being categorical versus dynamic are different ways of sense-making. One of the goals of our research is to understand how people use "the dream" as a medium on the path to self insight and social connection. In the most basic sense, do people share dreams about certain topics as a narrative personal experiences indicating changes over time? Do certain topics lend themselves to a more distant style- stories of what happened to whom with precise descriptions of events and goals?

The top five Categorical dream topics and top five Dynamic topics are depicted in Table 2. Topics that are the most categorical are primarily marked by physical environments: trees, sky, house, beach, road. Dynamic dream narratives are characterized by intimate relationships (baby, mom, boyfriend, sister) and experiences (remember, time). The CDI acts a shortcut to identify those dreams that are experienced as a narrative, potentially offering cues to the role of the dreamer as the main character, a distinguishing factor in dreams of healthy controls as compared to psychiatric patient samples (Skancke et al., 2014). Additionally, this shortcut points to a style of dream that would be difficult to discern with a topical lens only; that is, interpersonal situations with multiple characters and complex relationships. Interestingly, Cartwright et al. (1984) find that complex dreams containing multiple characters and shifts of scenes were one marker of depression remission in their five month longitudinal REM tracking study. Appendix B includes two samples of dreams with high and low CDI scores.

	LDA Topic	Words Characterizing Topic	Correlation with CDI (Pearson's r)
Categorical	13	walking tree trees small forest	0.25
	8	see sky plave flying building	0.21
	5	room door house floor stairs	0.2
	11	water pool beach boat swimming	0.17
	15	car driving road bus drive truck	0.11
Dynamic	21	baby hospital boy pregnant girl	-0.12
	4	mom dad house brother sister	-0.13
	18	remember know time think	-0.16
	17	guy phone told boyfriend	-0.22
	9	friend guy boyfriend friends	-0.34

Table 2: Top and Bottom Five dream Topics on CDI continuum

3.4 The emotional landscape of dreams

One of the goals of this paper is to investigate how emotions are revealed in dreams, which emotions, and how they vary with the topics that emerge. One prominent hypothesis in dream research posits that the function of dreams is to help regulate negative emotion by "intervening" between waking emotional concerns and post sleep mood (Cartwright, 2008). Much of the literature points to a central role for emotions in dreams, yet there are inconsistencies in the frequencies of the emotional array detected and their valance. The inconsistencies are dependent on a similar variety of reasons to those cited above which make standardized dream content analysis challenging, with the added challenge that make emotions difficult to detect and discern in the broader computer science literature (Sikka et al., 2014; Schredl and Doll, 1998). For example, Merritt et al. (1994) tested a small student population (n=20) and found that there are an average of 3.6 emotions per dream with 95% of dreams having at least one emotion, with fear being the most pervasive. This is directionally consistent with Hall and Castle (1966) who find negative emotions to be more prominent, however the frequencies vary. Sikka et al. (2014) find consistent differences in the external judgments of emotions in dreams as compared to self ratings. The predicted labels of each dream narrative should not be taken as a definitive representation of the overall emotion of that narrative (a difficult task for even human annotators to accomplish consistently; see Purver and Battersby 2012). Instead, these results should be viewed as an additional feature of each narrative, able to be evaluated automatically and quickly to gain insight and explore broader trends.

In our exploration of language style with a lexicon-based approach, LIWC detected a low proportion of affect (Mean Affect = 3.42, SD= 1.90). To assess the emotional content of dreams in an unsupervised manner (i.e., without annotating each narrative manually), we turn to a model for classifying emotional content from text. (We briefly summarize here, but for complete details, see Coppersmith et al. 2016.) A series of character language models (one for each of anger, fear, joy, sadness, surprise, and no emotion) are trained on a large corpus of Twitter data with an included emotional hashtag, e.g., "#anger". Tweets containing indications of sarcasm were removed. Tweets were labeled by the emotional hashtag contained, and then that hashtag was removed for training the model, thus learning what words might contribute to something being tagged "#anger". A two-step semi-supervised process is used to produce the no-emotion model, since most tweets with emotional content are not labeled with #[emotion]. (We also scored each narrative using the Mohammad and Turney 2013 NRC Emotional Lexicon and opted for the character language models for greater vocabulary coverage and possible explicit "no emotion" label.)

We apply each of the emotion character language models (CLM) to each of the dream narratives, producing a probability that each narrative's content results from each emotion's CLM. We then label that narrative with the maximum-probability emotion. Concretely, we expect dreams to have a mixture of emotions, and this technique is likely to surface the dominant emotion in the dream (as measured by the number of words used that indicate that emotion). Percent breakdown of predicted emotion labels were as follows: *sadness*, 31.6%; *fear*, 21.0%; *surprise*, 19.9%; *joy*, 18.7%; *anger*, 8.7%; *no emotion*, 0.0%. Only two narratives out of almost 10,000 were labeled *no-emotion*, and only 6 had the *no-emotion* label above 10% of the estimated emotional content within a dream; see caveats of this approach below.

To continue to deepen our understanding of the psychological value of the corpus and gain insight on the relationship between dream content and emotion, we correlate each emotion's CLM probability with each of the 25 LDA topics. Table 3 shows the most positively-correlated topic and most negatively-correlated topic for each emotion. Consistent with previous research (Merritt et al., 1994; Hall and Castle, 1966), we demonstrate emotions present in all dreams, with more negative than positive emotion: 61.3% negative emotions (sadness, fear, anger), and sadness as the dominant emotion. Drawbacks of this approach of relying on self-stated emotional content tags are outlined in Coppersmith et al. (2016). In short, even given the two-step semi-supervised method of obtaining the most emotionally neutral tweets possible to use as *no-emotion* exemplars, it is likely that some nontrivial percentage of the tweets contain significant emotional content. In addition, even in a single tweet, emotional content is often mixed, and the training method employed allows for only one label that may not be sufficiently descriptive. Perhaps the largest caveat of these results comes from the mismatch between the Twitter data the model was trained on and the dream data it is applied to here. The featurization and parameters of the model are optimized for Twitter messages that are constrained to 140 characters, while the dream narratives are 1,047 characters on average (SD 716). Content varies as well; the dream narratives, at least in theory, have a consistent purpose and theme: recounting the content of a dream. Content of tweets is incredibly varied, from a segment of a story, meant to be read in the context of additional tweets; to a single hyperlink, perhaps with a few words of commentary; to a single emoji repeated 140 times. Future research directions include training a semi-supervised emotion classifier that includes the dream narratives to generalize better across domains.

	Topic number	Correlation with topic (Spearman ρ)	Words characterizing topic
Anger	16	0.187	people kill man trying guy gun shot killed
	9	-0.08	friend guy boyfriend friends love girl
Fear	18	0.17	remember know time think felt life feeling
	19	-0.139	school class teacher high game friend friends
Joy	0	0.151	see says look know comes walk run looks
	9	-0.13	friend guy boyfriend friends love girl
Sadness	9	0.237	friend guy boyfriend friends love girl
	0	-0.101	see says look know comes walk run looks

Table 3: Most positively and negatively-correlated topics for each emotion

4 Conclusion

Our paper presents three types of analyses on an innovative corpus. First we explored the content of dreams with LDA topic modeling. The results demonstrate topics easily interpreted by a human including everyday activity, dreaming itself, and themes common in the dream literature (teeth, animals, flying). These results are consistent with the limited amount of existing research in this area. Our second lens on the data using LIWC portrays dreams, in general, as first person accounts of past events with disproportionate social references and abstract descriptions of settings. Dreams tend to focus on perceptual processes more than cognitive processes. However, there are qualitative distinctions in the content of dreams such that certain topics are experienced as dynamic and others, more categorical. Lastly, we further explored the emotional content in dreams with an unsupervised approach. Our results indicate that emotion is present in dreams and is disproportionately negative, with the most common emotion being sadness. With a sensitive tool, emotion can help disambiguate content in dreams that would otherwise be lumped together, for example dreams about friends, romance, and love which show a complex configuration of emotion.

One major question that underlies this paper is whether we are investigating how we dream or how we story and share our dreams. In future research, we hope to compare dream data to other corpora to better understand how this way of knowing a person, through their dreams, is related to other forms of self expression. Identifying a reasonable comparative dataset for dreams collected from a social network is challenging. This data set is unique in its length (e.g. 140 character Tweets vs. 210 word dreams), content (intimate and quotidian content), and purpose (these dreams are shared for social connection and interaction) making most social media, which would otherwise present the appropriate scale and date range, a poor fit.

Interpreting topics in dreams is extra challenging because there is no ground truth. Language style and emotional classification enhance our understanding of topics and the mindset of a given dreamer, but it is as of yet unclear whether there are individual differences in the way dreams are experienced, or whether dreams are 'victims' of our memories and are yet another corpus to explore the same individual differences we might see in conscious thought. Continued research on dreams over time, dreamers across media and a variety of facets within dream data as compared to different outcome measures (personality, etc.)

will help address this concern.

Another limitation in our research is lack of information about potential skew in the data. For example, there may be biases in who shares dreams and why; who knows about and has access to the social network. We also did not have access to ground truth of user mental health information, so we did not analyze dream content relative to clinical disorders. At this time, site behavior is unreliable at the level of dream reporting to tell us whether there is any systematic bias in who provides dreams. Future studies will certainly explore demographic variables including age, sex, race, socioeconomic status, education level, in addition to variables related to belief in dreams, dream frequency and other psychological attributes which would make people more or less likely to share their dreams. Additionally, future research could investigate associations between mental disorder diagnoses and the content of dreams. This is a preliminary investigation into a vast data set with many additional variables to explore.

Much like this field has used social media data as a lens to study the conscious waking perceptions, emotions, and thought processes of individuals with mental health conditions, we see this as a complementary set of quantifiable signals related to the person's unconscious processes. While more traditional social media data is a convolution of the person's internal state and the world they inhabit, we see this dream data as a convolution of their dreaming self, as recalled and recorded by their waking self. Considered in context of the Fluid Vulnerability Theory, dream content could serve as one of many dynamic, near-term risk factors for detecting transitions into psychological crisis (Rudd, 2006). Given the richness of social media data for uncovering unknown signals related to mental health, we strongly suspect this data may hold similar and complementary power.

In sum, our paper offers preliminary evidence that the language of dreams can be an insightful contribution to human-centric big data, as a means for an enhanced understanding of human behavior and cognition alongside standard psychological means and modern neuroimaging. Paired with large scale analysis of social media language, Internet behavior, and wearable sensor information that predict mental health, the language of dreams could serve as an additional data source from which to evaluate mental health by digital life traces.

References

Deirdre Barrett. 2007. An evolutionary theory of dreams and problem-solving. In *The New Science of Dreaming: Content, Recall, and Personality Correlates*, Praeger Publishers, volume 2, pages 133–154.

Rebecca A. Bernert, Thomas E. Joiner, Kelly C. Cukrowicz, Norman B. Schmidt, and Barry Krakow. 2005. Suicidality and sleep disturbances. *Sleep* 28(9):1135–1141.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Adriel Boals, Jonathan B. Banks, Lisa M. Hathaway, and Darnell Schuettler. 2011. Coping with Stressful Events: Use of Cognitive Words in Stressful Narratives and the Meaning-Making Process. *Journal of Social and Clinical Psychology* 30(4):378–403. https://doi.org/10.1521/jscp.2011.30.4.378.

danah m. boyd and Nicole B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13(1):210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Rebecca L. Campbell and Anne Germain. 2016. Nightmares and Posttraumatic Stress Disorder (PTSD). *Current Sleep Medicine Reports* 2(2):74–80. https://doi.org/10.1007/s40675-016-0037-0.

R. Cartwright. 2013. History of the Study of Dreams. In Clete A. Kushida, editor, *Encyclopedia of Sleep*, Academic Press, Waltham, pages 124–128. DOI: 10.1016/B978-0-12-378610-4.00028-0.

Rosalind Cartwright. 2008. The Contribution of the Psychology of Sleep and Dreaming to Understanding Sleep-Disordered Patients. *Sleep Medicine Clinics* 3(2):157–166. https://doi.org/10.1016/j.jsmc.2008.01.002.

Rosalind Cartwright, Mehmet Y. Agargun, Jennifer Kirkby, and Julie Kabat Friedman. 2006. Relation of dreams to waking concerns. *Psychiatry Research* 141(3):261–270. https://doi.org/10.1016/j.psychres.2005.05.013.

Rosalind D Cartwright, Stephen Lloyd, Sara Knight, and Irene Trenholme. 1984. Broken dreams: A study of the effects of divorce and depression on dream content. *Psychiatry* 47(3):251–259.

J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. pages 143–152. https://doi.org/10.1109/VAST.2012.6400557.

Nicholas A. Christakis and James H. Fowler. 2014. Friendship and natural selection. *Proceedings of the National Academy of Sciences* 111(Supplement 3):10796–10801. https://doi.org/10.1073/pnas.1400825111.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Social Communication* .

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics, Denver, Colorado, USA.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Tony Wood. 2016. Exploratory data analysis of social media prior to a suicide attempt. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics, San Diego, California, USA.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th ACM International Conference on Web Science*.

G. William Domhoff. 2000. Methods and measures for the study of dream content. In Meir H. Kryger, Thomas Roth, and William C. Dement, editors, *Principles and Practice of Sleep Medicine*, W. B. Saunders, Philadelphia.

Sigmund Freud. 2013. *The Interpretation Of Dreams*. Read Books Ltd. Google-Books-ID: U0t8CgAAQBAJ.

Calvin Springer Hall and Robert L. Van de Castle. 1966. *The content analysis of dreams*. Appleton-Century-Crofts.

R. C. II Hawkins and Ryan L. Boyd. in press. Such stuff as dreams are made on: Dream language, {LIWC} norms, and personality correlates. *Dreaming* .

Iris Hendrickx, Louis Onrust, Florian Kunneman, Ali Hürriyetoğlu, Antal van den Bosch, and Wessel Stoop. 2016. Unraveling reported dreams with text analytics. *arXiv:1612.03659 [cs]* ArXiv: 1612.03659. http://arxiv.org/abs/1612.03659.

J Allan Hobson, Steven A Hoffman, Rita Helfand, and Delia Kostner. 1987. Dream bizarreness and the activation-synthesis hypothesis. *Human neurobiology* .

Carl Gustav Jung. 2002. *Dreams*. Routledge. Google-Books-ID: SWvdQyo_ZX0C.

Kirill Kireyev, Leysia Palen, and Kenneth M. Anderson. 2009. Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*. volume 1.

Kitty Klein and Adriel Boals. 2010. Coherence and Narrative Structure in Personal Accounts of Stressful Experiences. *Journal of Social and Clinical Psychology* 29(3):256–280. https://doi.org/10.1521/jscp.2010.29.3.256.

George Lakoff. 1993. How metaphor structures dreams: The theory of conceptual metaphor applied to dream analysis. *Dreaming* 3(2):77–98. https://doi.org/10.1037/h0094373.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu. [Online; accessed 2015-03-02].

Jane M. Merritt, Robert Stickgold, Edward Pace-Schott, Julie Williams, and J. Allan Hobson. 1994. Emotion Profiles in the Dreams of Men and Women. *Consciousness and Cognition* 3(1):46–60. https://doi.org/10.1006/ccog.1994.1004.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics, Denver, Colorado, USA.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon 29(3):436–465.

Carey K. Morewedge and Michael I. Norton. 2009. When dreaming is believing: the (motivated) interpretation of dreams. *Journal of Personality and Social Psychology* 96(2):249–264. https://doi.org/10.1037/a0013264.

David Nadeau, Catherine Sabourin, Joseph De Koninck, Stan Matwin, and Peter D. Turney. 2006. Automatic dream sentiment analysis. In *Proceedings of the workshop on computational aesthetics at the twenty-first national conference on artificial intelligence (AAAI-06)*. Boston, USA.

Celestine Okorome Mume. 2009. Nightmare in schizophrenic and depressed patients. *The European Journal of Psychiatry* 23(3):177–183.

James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211(2828):42–45.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report. https://repositories.lib.utexas.edu/handle/2152/31333.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE* 9(12):e115844. https://doi.org/10.1371/journal.pone.0115844.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX.

Matthew Purver and Stuart Battersby. 2012. Experimenting with Distant Supervision for Emotion Classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 482–491. http://dl.acm.org/citation.cfm?id=2380816.2380875.

M David Rudd. 2006. Fluid vulnerability theory: A cognitive approach to understanding the process of acute and chronic suicide risk. .

Silvio Scarone, Maria Laura Manzone, Orsola Gambini, Ilde Kantzas, Ivan Limosani, Armando D'agostino, and J Allan Hobson. 2008. The dream as a model for psychosis: an experimental approach using bizarreness as a cognitive marker. *Schizophrenia Bulletin* 34(3):515–522.

Michael Schredl. 2010. Dream content analysis: Basic principles. *International Journal of Dream Research* 3(1):65–73. https://doi.org/10.11588/ijodr.2010.1.474.

Michael Schredl and Evelyn Doll. 1998. Emotions in Diary Dreams. *Consciousness and Cognition* 7(4):634–646. https://doi.org/10.1006/ccog.1998.0356.

Michael Schredl, Franc Paul, Iris Reinhard, Ulrich Walter Ebner-Priemer, Christian Schmahl, and Martin Bohus. 2012. Sleep and dreaming in patients with borderline personality disorder: A polysomnographic study. *Psychiatry Research* 200(23):430–436. https://doi.org/10.1016/j.psychres.2012.04.036.

Dylan Selterman, Deirdre Barrett, and Patrick McNamara. 2012. Attachment, sleep and dreams. In *Encyclopedia of Sleep and Dreams*, Greenwood Publishers, Santa Barbara, CA.

Francesca Siclari, Benjamin Baird, Lampros Perogamvros, Giulio Bernardi, Joshua J. LaRocque, Brady Riedner, Melanie Boly, Bradley R. Postle, and Giulio Tononi. 2017. The neural correlates of dreaming. *Nature Neuroscience* advance online publication. https://doi.org/10.1038/nn.4545.

Pilleriin Sikka, Katja Valli, Tiina Virta, and Antti Revonsuo. 2014. I know how you felt last night, or do i? self-and external ratings of emotions in rem sleep dreams. *Consciousness and cognition* 25:51–66.

Joacim Skancke, Ingrid Holsen, and Michael Schredl. 2014. Continuity between waking life and dreams of psychiatric patients: A review and discussion of the implications for dream research. *International Journal of Dream Research* 7(1):39–53. http://journals.ub.uni-heidelberg.de/index.php/IJoDR/article/view/12184.

Marijke L. Swart, Annette M. van Schagen, Jaap Lancee, and Jan van den Bout. 2013. Prevalence of Nightmare Disorder in Psychiatric Outpatients. *Psychotherapy and Psychosomatics* 82(4):267–268. https://doi.org/10.1159/000343590.

Murray L. Wax. 2004. Dream sharing as social practice. *Dreaming* 14(2-3):83–93. https://doi.org/10.1037/1053-0797.14.2-3.83.

Robert E. Wilson, Samuel D. Gosling, and Lindsay T. Graham. 2012. A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science* 7(3):203–220. https://doi.org/10.1177/1745691612442904.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical Topic Discovery and Comparison. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '11, pages 247–256. https://doi.org/10.1145/1963405.1963443.

Appendix A: Full list of LDA topics

Topic	Label	Top words
0	Active first person dreams	see says look know comes walk run looks find wake
1	Sex dreams, some explicit	girl guy sex room bathroom wanted girls shower naked talking
2	Animal dreams	dog house cat snake dogs trying black big came bear
3	Metadreaming	room bed woke sleep night wake asleep time felt see
4	Family presence	mom dad house brother sister told came saw home family
5	Strange homes and settings	room door house floor stairs old open window building doors
6	About family members	house family husband mother old son sister home daughter father
7	Friendship	friend friends party people wedding church best seemed told wanted
8	Flying	see sky plane flying building ground people fire city air fly high huge storm
9	Young love	friend guy boyfriend friends love girl told talking felt life real know
10	Teeth, limbs, body parts	felt face eyes body hand head see looked blood feel
11	Water	water pool beach boat swimming ocean river ship people lake
12	Food and eating	food table sitting people eating eat kitchen restaurant left bathroom
13	Picturesque landscapes	walking tree trees small area forest place beautiful hill little
14	Performance	work people thought asked show working wanted office told music
15	Driving and Cars	car driving road bus drive truck train seat drove home street
16	Violence	people kill man trying guy gun shot killed group dead knife die
17	Friends and Exes	guy phone told boyfriend remember call girl friend asked know
18	Dream sense-making	remember know time think felt life feeling thing real people feel knew
19	School dreams	school class teacher high game friend friends old girl walking time
20	Colorful dreams	white hair black man looked wearing blue dark see red woman light
21	Pregnancy and baby	baby hospital boy pregnant girl know child told little woke
22	Cinematic, sophisticated dreams	woman name life person place words world read help found
23	Shopping and money	store find people money place shop found work left mall
24	Chase dreams	ran saw looked came running house told woke tried door

Appendix B: Sample dreams by CDI

Categorical	This dream appears to me as if it were movie. A crowd of people are running away from a horde of zombies. The crowd of people run up a skyscraper. The zombies are running and still chasing them. At the top of the building, the people are stranded and can hear the dead catching up to them on the stairs. One man in a brown overcoat pulls a leather tome out of his coat and flips through it. "THE PROPHECY IS COMING TRUE!" He yells. The clouds part above them and an angel made entirely out of tiny swords floats down. The people all marvel for a moment. Then the angel disintegrates into a cloud of blades and flies at the zombie horde, decimating them. As this happens, Japanese rock music starts playing. The scene cuts to a montage of zombie people and cows getting disintegrated as credits roll past the "screen" in front of my eyes. I wake up.
Dynamic	So I was going to this thing and my crush was there. It was this hill and it was snowing. So I ran and hugged my crush when I saw him because we're bestfriends. So then I saw one of my old friends. He told me he liked me like 3 years ago. So I hugged him too because I haven't seen him forever. So then I got tired so we sat down at this table and the guy who told me he liked me (this was in real life when he told me) but in my dream he sat next to me and bought me a drink and we kinda just smiled at each other for a while. And that's it.