Automatic Detection and Prediction of Psychiatric Hospitalizations From Social Media Posts

Zhengping Jiang
Computer Science Dept.
Columbia University
zj2265@columbia.edu

Jonathan Zomick
Psychology Dept.
Hofstra University
jzomick1@pride.hofstra.edu

Sarah Ita Levitan
Computer Science Dept.
Hunter College, CUNY
sarah.levitan@hunter.cuny.edu

Mark Serper
Psychology Dept.
Hofstra University
Mark.R.Serper@hofstra.edu

Julia Hirschberg
Computer Science Dept.
Columbia University
julia@cs.columbia.edu

Abstract

We address the problem of predicting psychiatric hospitalizations using linguistic features drawn from social media posts. We formulate this novel task and develop an approach to automatically extract time spans of self-reported psychiatric hospitalizations. Using this dataset, we build predictive models of psychiatric hospitalization, comparing feature sets, user vs. post classification, and comparing model performance using a varying time window of posts. Our best model achieves an F1 of .718 using 7 days of posts. Our results suggest that this is a useful framework for collecting hospitalization data, and that social media data can be leveraged to predict acute psychiatric crises before they occur, potentially saving lives and improving outcomes for individuals with mental illness.

1 Introduction

Every year, approximately 1% of adults in the United States are hospitalized for psychiatric reasons, including increased suicidality and psychosis (Elflein, 2020). With the global COVID-19 pandemic, hospitalizations due to suicidality are projected to increase substantially (John et al., 2020), and there is already evidence of the adverse impact of the pandemic on the mental health of individuals around the world (Cullen et al., 2020). Psychiatric hospitalizations typically result from crises among individuals struggling with suicidality and mental illness. The present study aims to predict psychiatric hospitalization due to increased suicidality or a psychotic break before it occurs.

There are several motivations for this research goal. Improving our ability to better predict psychiatric hospitalization helps enable the identification of early warning signs of these crises before they fully develop. Early detection and prediction of acute psychiatric crises is essential for lowering mortality rates and improving overall outcomes for individuals suffering with mental illness. Further, psychiatric hospitalizations place a tremendous burden on limited hospital resources, and involve steep costs for patients as well as taxpayers (Stensland et al., 2012; Owens et al., 2019).

Typically, prediction of psychiatric hospitalization has relied on rich and personalized clinical information for a particular patient. This requirement has limited the size of available datasets, and has also limited the possibility of reaching and helping potential patients who do not have a well-documented psychiatric medical history. In this work we circumvent this limitation by leveraging social media data to train and evaluate predictive models of psychiatric hospitalization. This is a necessary step towards the ultimate goal of predicting behavioral and cognitive changes that often lead to hospitalization. There is a rich literature of computer scientists, psychologists, and psychiatrists taking advantage of the vast amount of social media data – which includes language data of posts and comments, as well as meta-information such as preferences, engagement patterns, and group membership – to gain insights about mental states and behaviors of people with psychiatric disorders.

Building on this successful line of research, we detect engagement patterns combined with self-disclosures to identify potential periods of psychiatric hospitalization. We compile a dataset of these periods, or time spans, along with the posts preceding those periods, and conduct machine learning experiments to automatically predict whether a post precedes a hospitalization or not. Our results suggest that this is a potentially useful approach for predicting psychiatric hospitalizations before they occur. This can enable clinicians to mitigate and hopefully prevent a psychotic break or suicide attempt, helping to save patients' lives and improve outcomes.

The rest of this paper is organized as follows: Section 2 reviews related work and Section 3 describes our novel data collection approach. Section 4 presents our experiments to predict psychiatric hospitalizations, and Section 5 provides analyses of the data and the learned models to gain further insights about the dataset and our results. We conclude in Section 6 and discuss ideas for future work.

2 Related Work

Research over the past decade has supported and validated the use of computational linguistics techniques applied to social media data for predicting and detecting mental illness across a broad range of psychiatric conditions (Guntuku et al., 2017; Wongkoblap et al., 2017). To date, linguistic indicators of psychopathology have been identified for a wide range of psychiatric conditions (Zomick et al., 2019; Coppersmith et al., 2015; Birnbaum et al., 2017; Huang et al., 2017; De Choudhury et al., 2013; Shen and Rudzicz, 2017). Recent work has also looked at detecting and predicting suicidality using linguistic features from social media posts (Du et al., 2018; Coppersmith et al., 2018; Zirikly et al., 2019).

While the majority of past research has compared specific psychiatric conditions with healthy control groups, more recent work has begun analyzing and identifying unique differences and discriminators among psychiatric conditions (Jiang et al., 2020; Cohan et al., 2018a; Coppersmith et al., 2015). As this area progresses, we have begun to investigate whether this technology can be used beyond detection of mental illness for detecting severity of symptomatology and prediction of acute psychiatric episodes that result in hospitalization. This would benefit patients by alerting clinicians to worsening symptoms, allowing for early intervention care and potential mitigation. Relatedly, advancements in machine learning techniques have led to the development of advanced models for predicting psychiatric crises such as increased suicidality and psychotic episodes using a multimodal approach based on clinical data (Koutsouleris et al., 2021). However, to date, these studies have relied exclusively on clinical data and medical data. To our knowledge, this is the first study to leverage a large dataset of publicly available social media posts for predicting psychiatric hospitalization.

3 Data Collection

In this section we describe the pipeline components of our dataset construction process, in the order in which they are applied.[1] Table 1 presents the overall statistics of our dataset.

Candidates	TI	SC	#Posts
95,904	318	128	7,077

Table 1: Overall dataset statistics, where **Candidates** are the total number of users we examined, **TI** corresponds to number of users from which we extracted hospitalization identification with time-span information and **SC** corresponds to the number of users having posts collected for the 21 days directly before the refined hospitalization span. **#Posts** are number of posts from these spans in total.

3.1 Candidate Collection

We begin data collection by identifying candidate Reddit users who may be at risk for a psychiatric hospitalization. We focus on two user groups: those that self-identify with a psychiatric disorder, and those that self-identify with suicidal ideation or attempted suicide. To identify such users, we leverage subreddits, or forums on Reddit dedicated to specific topics. Following Shing et al. (2018), we collect posts from the r/SuicideWatch (SW) subreddit, and following (Cohan et al., 2018b; Jiang et al., 2020) we collect posts from subreddits related to 8 different mental health conditions: obsessive compulsive disorder (OCD), schizophrenia (SZ), borderline personality disorder (BPD), post-traumatic stress disorder (PTSD), eating disorder (ED), major depression dis-order (MDD), general

[1]This study received IRB approval and all human subjects protection guidelines were followed.

anxiety disorder (GAD) and bipolar disorder. We then use regular expression matching to extract self-identification statements from these posts to form our candidate user pool. Our data collection methods yield 69,682 candidates for suicidal risk and 35,606 candidates for mental health conditions.

3.2 Hospitalization Time Span Identification

After identifying nearly 100k candidate Reddit users at risk for psychiatric hospitalization, we designed an approach to identify users from that pool that have been hospitalized for psychiatric reasons. While previous work has shown that regular expression matching alone is able to create high precision mental health datasets (Coppersmith et al., 2014; Cohan et al., 2018b; Jiang et al., 2020), it is far more difficult to automatically construct a dataset with more fine-grained information. MacAvaney et al. (2018) created a dataset of self-disclosures of depression on Reddit, which includes manually annotated temporal information about the diagnosis date. In our case, it is important to not only identify users that self-disclose psychiatric hospitalizations, but also to pinpoint the time span of the hospital stay. There are several challenges associated with this task: First, we need to ensure that the correct time span is identified when a user mentions multiple events in a single post, and avoid identifying a time span that is not associated with the identified hospitalization instance. Second, there are various ways an adverbial phrase of time could be attached to a predicate, making regular expression design difficult. A third challenge is that some time-related words having other common synsets (e.g. "May").

We address the above mentioned problems by (1) sentence-tokenizing the posts and performing all our matching at sentence-level; and (2) running a state-of-the-art semantic role labeling model first to identify the likely span for regular expression matching. Specifically, we only parse the [ARG-TMP] temporal field related to the hospitalization event, identified by the pre-trained SRL model (Shi and Lin, 2019) provided by AllenNLP (Gardner et al., 2018). When the identification is precise to date level we allow ± 7 days of flexibility. In total, we extracted 72 hospitalization time spans from the SuicideWatch user group, and 349 time-spans from the psychiatric disorders user group. A clinical psychologist trainee manually reviewed all 421 spans and found that 69.12% of them were clearly correctly identified and relevant hospitalizations, while the other time-spans were not incorrect but simply lacked enough context in the post for confident labeling. This validates our proposed time-span identification approach, and suggests that further context (e.g. other posts in the same thread) may be useful to improve time-span identification.

3.3 Span Refinement

We observe that the most common duration of the span identified is one month, and it is desirable to have hospitalization time identified on a more fine-grained scale. For example, a user might mention that they were hospitalized "last June," without providing specific start and end dates of their hospital stay. Coppersmith et al. (2017); Coppersmith et al. (2018) shows that social media provides information in the "clinical whitespace." Inspired by them, we further identify rare media blackout periods in the previously found plausible hospitalization span, and use them as a proxy to a ground truth hospitalization period. To do this, we fit an exponential distribution on users' social media posting activity, and define a rare media blackout period as the time span of inactivity where the occurrence probability is less than a certain threshold r. This process also provides us with other benefits, as we are able to characterize irregularities like throw-away accounts. Figure 1 is an example of such irregularities, where the user became significantly more active after the identified span; therefore we hypothesize that most of their posts would be related to their mental health condition and perhaps their hospitalization experience. In contrast, Figure 2 is an example of users who actively use their social media before and after the hospitalization blackout. We believe these users and their posts are potentially more useful for research, because they include posts on a wide range of topics over long periods of time, both before and after a psychiatric hospitalization. However, in this paper we make no further use of the features other than to select posts that directly precede a blackout period. When multiple rare media blackout periods are found for an identified span, we empirically select the one with the longest overlap with the span.

4 Prediction of Psychiatric Hospitalization

Having collected a dataset of proposed hospitalization spans and preceding posts, we use our col-

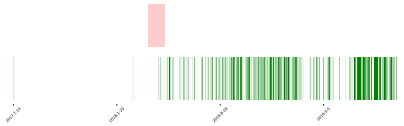

Figure 1: Irregular Reddit activity plots (green), where the user is significantly more active (darker) after the plausible hospitalization span (red).

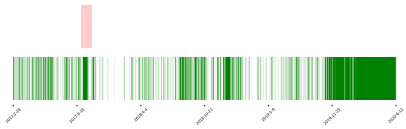

Figure 2: Regular Reddit activity plots, where the user generally post smoothly with some minor irregularity around the plausible hospitalization span (red).

lected dataset to build predictive models of psychiatric hospitalizations. We experiment with two different task formulations: post-level prediction and user-level prediction. Post-level prediction involves a binary classification for each post, determining whether the post is followed by valid hospitalization span or not. User-level prediction classifies a group of posts from a user in a given time window to predict whether the user will be hospitalized. In order to train classification models, we first need to select negative samples as a control group for our experiments. We describe our methods of pairing negative samples in subsection 4.1. We experiment with three set of features: unigram, bigram[2] and LIWC (Pennebaker et al., 2007, 2015) features. We perform hyper-parameter grid search to optimize performance. For all features we use the Naive-Bayes classifier, as it has been found to perform well on small datasets (NG and Jordan, 2002). We pre-process the text by lower-casing all input posts and, following the guidelines of (Benton et al., 2017), we de-identify posts by anonymizing URLs and replacing usernames with randomly generated strings.

4.1 Pairing Negative Samples

To form a challenging prediction task, we compile negative samples for classification by selecting control users from the same candidate pool that the target hospitalization group was selected from. The control users are those who do not have associated hospitalization time spans, but did have similar media blackout periods (described in subsection 3.3).

[2]Due to the size of our dataset, we set a minimum document count of 5 for bigram features.

We group spans by number of post before the span in a prescribed time window of length d days. For each positive span we randomly sample a span from the negative span pool that has a similar number of posts, creating a balanced classification task. Note that we expect this task to be difficult because the control users either self-identified with mental health conditions or posted in the SW subreddit. For post-level classification, we use the same set of posts sampled on the user-level.

4.2 Classification

Table 2 shows mean F-1 scores from cross-validation on both user-level and post-level tasks. In all experiments, we set the span selection probability threshold $t = 0.1$. For user-level and post-level performance comparison, we set the inclusion number of days to $d = 21$.

	1-gram	1,2-gram	LIWC
user-level	0.687	0.698	0.655
post-level	0.601	0.622	0.584

Table 2: Experiment result in F-1, with different features on both tasks.

The best performance of 0.698 F1 is obtained using bigrams for the user-level task. In general, user-level classification results in better F-1 scores, indicating that more context is likely crucial to success in psychiatric hospitalization prediction. N-gram features outperform LIWC features for both tasks, and adding bigram features perform better than unigrams alone. Overall, the model performance with a small amount of data is promising, well above a 50% random baseline.

4.3 Performance Over Time

We again run experiments for user-level classification with another more strictly paired control group that satisfies the pairing constraints mentioned in subsection 4.1 for $d \in \{1, 7, 14, 21\}$. Table 3 shows the performance change as the window length increases. The results suggest that using a wider context is useful in predicting hospitalization blackouts, and the best performance was obtained using unigrams extracted from 7 days of posts.

5 Lexical Analysis

Figure 3 shows the list of most predictive words for the unigram model. We see that many words correspond to time duration (e.g. “week”, “month”),

d (days)	1-gram	1,2-gram
1	0.678	0.676
7	0.718	0.695
14	0.697	0.692
21	0.708	0.706

Table 3: F-1 performance with different features on different window lengths

medical professions (e.g., "med", "doctor", "hospital") and conversation (e.g., "sorry", "thanks"). We hypothesize that these may correspond to users' frequent online posts seeking advice and describing conditions. Indeed we observe some posts conforming to this pattern through manual examinations.

care, come, person, taken, stuff, able, hear, weeks, &, definitely, bit, let, doctor, does, makes, point, home, tell, times, sorry, family, months, hope, little, use, yeah, sleep, maybe, best, new, post, told, night, probably, voices, went, great, isn, meds, bot, moderator, school, days, thought, week, doesn, trying, started, working, used, mom, message, thank, long, doing, hospital, having, try, hard, love, year, thanks, bad, getting, actually, pretty, sure, thing, help, better, years, life, ll, need, said, right, say, didn, work, way, did, make, lot, day, got, things, url, want, going, feel, good, think, people, time, know, ve, really, don, like, just

Figure 3: The top 100 most predictive words for the hospitalized group by the uni-gram model.

6 Conclusion and Future Work

We present a novel social media data collection method for identifying hospitalization time spans and design a novel classification task for predicting psychiatric hospitalizations. We experiment with multiple linguistic feature sets and task formulations, including user-level and post-level classification, as well as varying the time window of posts used. Our results suggest that this is a useful framework for collecting data related to psychiatric hospitalization, and that social media data can be leveraged to predict psychiatric crises before they occur. In our ongoing and future work, we plan to conduct further analysis of the language of pre-hospitalization posts to gain insights about linguistic patterns and changes that occur as the user experiences a psychiatric crisis. We also plan to improve the data collection process to achieve better precision and to expand to a larger scale. We hope that an improved understanding of the linguistic cues that precede psychiatric hospitalizations, as well as improvements in automatic prediction of hospitalizations, will enable interventions that can potentially save lives and improve outcomes for individuals with mental illness.

References

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, 19(8):e289.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018a. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018b. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

G. Coppersmith, C. Hilland, O. Frieder, and R. Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 393–396.

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 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

W Cullen, G Gulati, and BD Kelly. 2020. Mental health in the covid-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312.

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

Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 18(2):77–87.

John Elflein. 2020. Mental health service use in the past year among u.s. adults from 2002 to 2019, by type of care. https://www.statista.com/statistics/252316/type-of-mental-health-service-used-by-us-adults-since-2002/. Accessed: 2021-03-15.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Yen-Hao Huang, Lin-Hung Wei, and Yi-Shin Chen. 2017. Detection of the prodromal phase of bipolar disorder from psychological and phonological aspects in social media. *arXiv preprint arXiv:1712.09183*.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.

Ann John, Jane Pirkis, David Gunnell, Louis Appleby, and Jacqui Morrissey. 2020. Trends in suicide during the covid-19 pandemic.

Nikolaos Koutsouleris, Dominic B Dwyer, Franziska Degenhardt, Carlo Maj, Maria Fernanda Urquijo-Castro, Rachele Sanfelici, David Popovic, Oemer Oeztuerk, Shalaila S Haas, Johanna Weiske, et al. 2021. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA psychiatry*, 78(2):195–209.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. *arXiv preprint arXiv:1806.07916*.

Andrew NG and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

Pamela L Owens, Kathryn R Fingar, Kimberly W McDermott, Pradip K Muhuri, and Kevin C Heslin. 2019. Inpatient stays involving mental and substance use disorders, 2016: Statistical brief# 249. *Healthcare cost and utilization project (HCUP) statistical briefs*.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report, TX: University of Texas at Austin.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Michael Stensland, Peter R Watson, and Kyle L Grazier. 2012. An examination of costs, charges, and payments for inpatient psychiatric treatment in community hospitals. *Psychiatric Services*, 63(7):666–671.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6):e228.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83.