Original Paper

CROSS-PLATFORM DETECTION OF PSYCHIATRIC HOSPITALIZATION VIA SOCIAL MEDIA DATA: A COMPARISON STUDY

Viet Cuong Nguyen, Nathaniel Lu, John M. Kane, Michael L. Birnbaum, Munmun De Choudhury

Abstract

Background: Previous research has shown the feasibility of utilizing machine learning models trained on social media data from a singular platform (e.g., Facebook or Twitter) in distinguishing individuals with either a diagnosis of mental illness or experiencing an adverse outcome from healthy controls. However, the performance of such models on data from novel social media platforms unseen in the training data (e.g., Instagram, TikTok) have not been investigated by previous literature.

Objective: Our study examines whether it is feasible to build machine learning classifiers that can effectively predict an upcoming psychiatric hospitalization given social media data from platforms unseen in the classifiers' training data, despite the preliminary evidence on identity fragmentation on the investigated social media platforms. It also aims to explain any discrepancies in performances that are found during analysis between intra- and inter-platform classification.

Methods: Windowed timeline data from three platforms among patients with a diagnosis of Schizophrenia Spectrum Disorder (SSD) before a known hospitalization event and healthy controls was gathered: Facebook (N = 254), Twitter (N = 54), and Instagram (N = 134). Then, we utilized a 3 x 3 combinatorial binary classification design to train machine learning classifiers and evaluate their performance on testing data from all available platforms. We further compared results from models within intra-platform experiments (i.e., training and testing data belongs to the same platform) to models within inter-platform experiments (i.e., training and testing data belongs to the different platforms). Finally, we utilized SHapley Additive exPlanations (SHAP) values to extract top predictive features to explain and compare the underlying constructs that predict hospitalization on each platform.

Results: We found that models within intra-platform experiments on average achieved an F1-score of 0.72 in predicting a psychiatric hospitalization due to SSD, which is 68% higher compared to the average of models within inter-platform experiments at an F1-score of

0.428. When investigating the key drivers for divergence in construct validities between models, an analysis of top features for the intra-platform models shows both low predictive feature overlap between the platforms and low pairwise rank correlation (< 0.1) between the platforms' top feature rankings. Furthermore, low average cosine similarity of data between-platforms within-participants in comparison to the same measurement on data within-platforms between-participants points to evidence of identity fragmentation of participants between platforms.

Conclusions: We demonstrated models built on one platform's data to predict critical mental health treatment outcomes, such as a hospitalization, do not generalize to another. In our case, this is due to different social media platforms consistently reflecting different segments of participants' identities**.** With the changing ecosystem of social media use among different demographic groups, and as online identities continue to get fragmented across platforms, further research on holistic approaches to harnessing these diverse data sources is required.

Keywords: schizophrenia; mental health; machine learning; clinical informatics; social media

Introduction & Literature Review

Despite its relatively low prevalence compared to other mental health disorders, the burden of schizophrenia spectrum disorder (SSD) on patients, families, and society is substantial [1]. To mitigate the burden of SSD, early diagnosis and treatment are crucial. However, psychotic disorders including SSD often receive delayed attention and care, resulting in worse health outcomes [2, 3]. At the same time, the use of social media is high amongst patients afflicted by serious psychotic disorders such as SSD, especially among adolescents and young adults when SSD typically emerges [4, 5]. For instance, Birnbaum et al. (2017) studied social media usage among adolescents and young adults with psychotic and mood disorders and found that 97.5% of participants (mean age = 18.3 years) regularly used social media, spending approximately 2.6 ± 2.5 h per day online [4]. Similarly, Miller et al. (2015) studied usage of digital technologies among patients diagnosed with SSD, and found that for participants with access to the Internet, 98% of them reported using at least one social media service and 57% of them use social media daily [5].

Given this information, there has been an established body of research on utilizing social media data to identify and predict psychiatric outcomes of social media users with SSD using machine learning classifiers [6-8]. The most robust data sources available to train these classifiers consist of textual content posted online. Prior work in speech and text analysis among SSD patients has identified reliable linguistic markers associated with SSD, which have been successfully used as features for the classifiers above [7, 9, 10]. These include certain word frequencies, word categories, and self-referential pronouns [11, 12]. Given that usage of image and video-based social media platforms such as Instagram, Snapchat, and TikTok is associated with youths, there has also been prior work in the analysis of images between SSD patients and healthy controls [13, 14]. Hansel et al. (2021) has identified additional image markers associated with SSD such as the image's colorfulness, saturation, and the average number of faces per image [14]. By exploiting these markers, previous research conducted by Birnbaum et al. (2020) and Ernala et al. (2019) built classifiers to distinguish between users with a confirmed diagnosis of SSD versus healthy controls on Facebook and Twitter with Area under the Receiver Operating Characteristic (AUROC) scores of 0.75 and 0.82 respectively [8, 15].

While such results demonstrate the potential of automated techniques in predicting mental health outcomes of individuals with SSD via social media data, many research gaps remain that need to be addressed before psychiatrists can reliably deploy such techniques for clinical purposes. Most prior work in this area primarily focused on a single source of social media data, either exclusively from Twitter or Facebook, for downstream classification and analysis tasks [16]. However, previous research has also shown that many social media users, especially youths, utilize different social media platforms for different purposes due to their variety in affordances and culture. Among youths, Facebook usage is associated with keeping up with close and distant friends, whereas Instagram and Snapchat usage is associated with self-expression uses and gratification [17, 18]. In addition, researchers have argued that social media users have fragmented identities across platforms [19, 20]. Therefore, utilizing only a single source of social media data to build psychiatric hospitalization prediction models might potentially lead to low-sensitivity prediction models, making them unsuitable for clinical purposes. However, few research have quantified the extent to which classifiers trained on data from one social media platform are generalizable to other platforms. To this end, our paper aims to measure the generalizability of social media-based classifiers aimed at predicting upcoming psychiatric hospitalizations to data from unseen social media platforms. In addition, we also aim to surface any evidence of the differing fragmented identities that are reflected on three popular social media platforms: Twitter, Facebook, and Instagram that might affect models' generalizability.

The research question we attempt to answer is as follows: Given the preliminary evidence of fragmented identities that are reflected on the investigated social media platforms, can we build classifiers that can effectively detect users at risk of an upcoming psychiatric hospitalization using social media data from platforms unseen in the training data?

To answer our research question, we collated textual and image content (if available) from consenting participants' social media data from Facebook, Twitter, and Instagram. We then trained platform-specific classifiers to distinguish between social media data belonging to healthy controls and data belonging to SSD patients with an upcoming psychiatric hospitalization. We compare performance of classifiers on testing data between seen and unseen social media platforms from the training data. We also compare and analyze the top predictive features and the feature importance distributions between the three platform-specific classifiers, with a view towards finding potential empirical evidence for fragmented identities between the various social media platforms.

Methods

Recruitment

We recruited participants clinically diagnosed with Schizophrenia Spectrum Disorder (SSD) and that of clinically verified healthy controls between the ages of 15 – 35 years. This data was collected as part of a broader research initiative involving the paper's authors to identify technology-based health information to provide early identification, intervention, and treatment to young adults with SSD [6]. The study was approved by the Institutional Review Board (IRB) of Northwell Health (the coordinating institution) and the IRBs of participating partners. Participants were recruited from 6/23/2016 through 12/4/2020. Written informed consent was obtained for adult participants and legal guardians of participants under 18 years of age. Assent was obtained for participating minors.

For participants with SSD between the ages of 15 and 35 (n = 141), diagnoses were based on clinical assessment of the most recent episode and were extracted from participant's medical records at the time of their consent. Participants in this group were recruited from Northwell Health's Zucker Hillside Hospital and from collaborating institutions located in East Lansing, Michigan. Participants were excluded if they have an IQ of below 70 (per clinical assessment), autism spectrum disorder and substance-induced psychotic disorder.

Additionally, healthy volunteers between the ages of 15 and 35 (n = 127) were approached and recruited from an existing database of eligible individuals who had already been screened for prior research projects at Zucker Hillside Hospital and had agreed to be recontacted for additional research opportunities. Healthy status was determined either by the Structured Clinical Interview for DSM Disorders conducted within the past two years or by the Psychiatric Diagnostic Screening Questionnaire [21, 22]. Participants were excluded if clinically significant psychiatric symptoms were identified during the screening process. Additional healthy volunteers were recruited from a southeastern university via an online student community research recruitment site. Finally, healthy volunteers were also recruited from the collaborating institutions located in East Lansing, Michigan.

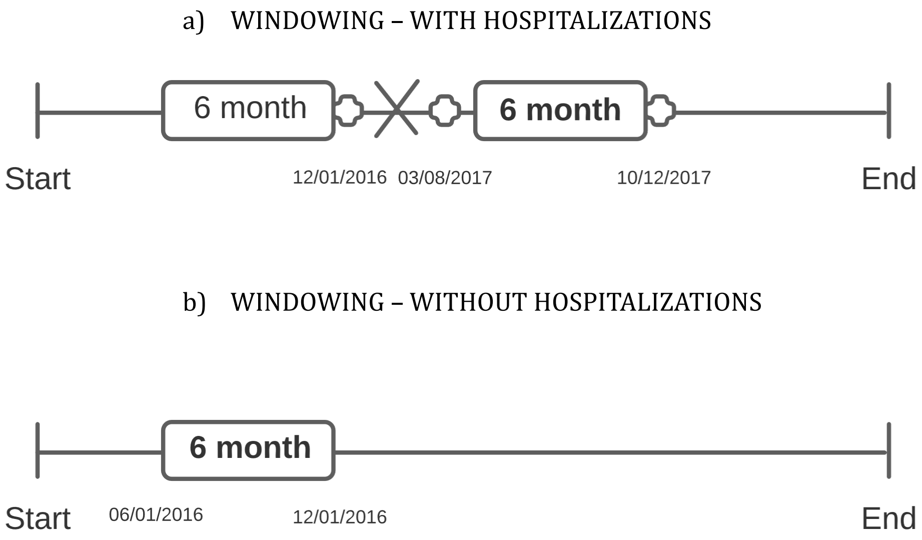
Data Collection

All consenting participants were asked to download and share their Facebook, Twitter, and Instagram data archives. We collected all linguistic content from participants' Facebook and Twitter archives, i.e., status updates and comments on Facebook and posts shared on Twitter. In addition, we collected image content from participants' Facebook and Instagram archives, including profile pictures and Story photos.

Finally, we also collected medical history for each participant (following consent and adoption of HIPAA compliant policies). This included primary and secondary diagnosis codes, the total number of hospitalizations, and admission and discharge dates per each hospitalization event. Hospitalization data was collected from the medical records at the time of consent. As all consented patient participants in the study had also received care at the Zucker Hillside Hospital, the medical records at the hospital were accurate and up to date to the best of the hospital's efforts. We only counted psychiatric hospitalizations (not any hospitalizations for other non-psychiatric reasons). Thereafter, the study team accessed the corresponding consented patient's medical record to extract all of their recorded hospitalization events in a similar manner to previous studies utilizing this source of data [6, 23].

Then, we collected social media data within all available platforms for each participant with at least one known hospitalization event within a 6-month window before the latest hospitalization event, ensuring that there were no hospitalization events within these six months. This was done to ensure that the data gathered was representative of the subject's healthy mental status before symptomatic exacerbation and subsequent hospitalization. A six-month period, which we refer to as the "windowed data," was selected because it represents an interval of time long enough to identify changes signaling symptomatic exacerbation while also containing sufficient data required to train machine learning models. For healthy control participants without any hospitalizations, we randomly sampled a non-empty six-month window of social media data for each available social media platform; non-empty meaning there was at least some social media activity.

Figure 1: Diagram representing the windowing process used to gather participant's social media data before hospitalization events. Bold text represents the selected data windows. Crosses represent hospitalization events. X represents invalid data windows.



a) WINDOWING – WITH HOSPITALIZATIONS

b) WINDOWING – WITHOUT HOSPITALIZATIONS

Table 1: Demographic and clinical characteristics for participants.

Characteristic	SSD	Control	Full Sample
Age (in years), mean (SD)	24.86 (5.49)	24.57 (5.82)	24.73 (5.64)
Sex, n (%)			
Male	89 (63%)	38 (30%)	127 (47%)
Female	52 (37%)	89 (70%)	141 (53%)
Race/ethnicity, n (%)			
African American/Black	64 (45%)	19 (15%)	83 (31%)
Asian	20 (14%)	23 (18%)	43 (16%)
Caucasian	37 (26%)	75 (59%)	112 (42%)
Mixed Race/Other	15 (11%)	5 (4%)	20 (7%)
Hispanic	5 (4%)	4 (3%)	9 (3%)

Pacific Islander	0 (0%)	1 (1%)	1 (1%)
Primary Diagnosis, n (%)			
Schizophrenia	67 (48%)	N/A	67 (25%)
Schizophreniform	26 (18%)	N/A	26 (10%)
Schizoaffective	25 (18%)	N/A	25 (9%)
Unspecified schizophrenia spectrum disorders	23 (16%)	N/A	23 (9%)
No Diagnosis	N/A	127 (100%)	127 (47%)

Feature Engineering

To encode participant's social media data for downstream classification and analysis tasks outlined in our research objectives, we identified and extracted the following categories of features from these data for all three of the investigated social platforms. The specific feature categories were chosen based on relevant prior literature, particularly relating to the use of social media data to infer mental health attributes and psychiatric outcomes [7, 8]. Note that all features were computed at the individual participant level. More details about this process can be found in the Supplementary Information (Multimedia Appendix 1)

Feature Selection

Using the features described above, for each of the three examined social media platforms, we encode available participants' textual and image data on Facebook and Instagram into 613-dimensional feature vectors, and textual data on Twitter into 590-dimensional feature vectors. This yields us a Facebook dataset of dimension 254 * 613, a Twitter dataset of dimension 51 * 590, and an Instagram dataset of dimension 134 * 613. We shall refer to these datasets respectively as F, T, and I respectively for Facebook, Twitter, and Instagram.

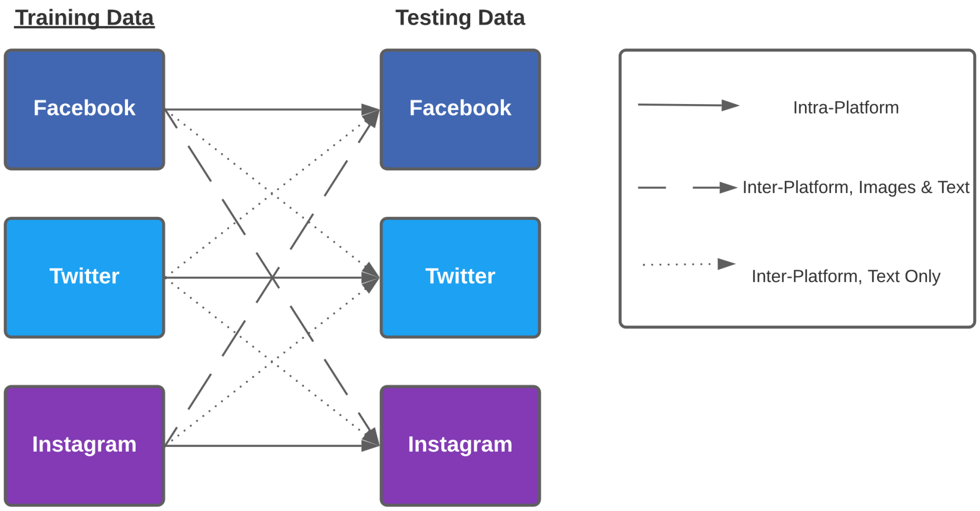
As the feature set might contain features that are noisy and irrelevant, the classification models may be unstable and produce suboptimal results [24]. To maximize the predictive power of the models while also reducing redundancy and computational resources needed to train them, feature selection methods were employed [24]. More specifically, we adopted the ANOVA F-test to rank the features based on their F-statistic under the test, which has been shown to produce optimal feature sets in previous research surrounding classification of social media data belonging to SSD patients [8, 11].

We trained a random-forest model, with 5-fold stratified cross-validation to finetune hyperparameters, on datasets F, T, I with an 80:20 train-test split, only using the top-k percent of features based on the ranking given by the ANOVA F-test on the classification where k is between 10 and 100 in increments of 10. Via an examination of the evaluation metrics on the test sets (described below in the Classification Algorithms and Metrics section), we determined that using only top-20% of the features (based on their F-statistic under the ANOVA F-test) yielded the best results on unseen data across all three platforms. We will be using this subset of features moving forward.

Combinatorial Classification Methods

To answer the research question laid out in the Introduction, we adopted a 3x3 combinatorial classification design, where we trained and tested machine learning models on the psychiatric hospitalization prediction task using all possible pairs of training and testing dataset. Figure 2 provides a visual description of our experimental design. For intra-platform experiments (where the training data and testing data came from the same platform, e.g., train on Facebook data, test on Facebook data), we trained and tested the models on an 80-20 train-test label-stratified split based on scikit-learn's train_test_split() function [25]. For inter-platform experiments (where the training data and testing data came from the different platforms, e.g., train on Facebook data, test on Instagram data), we trained the model on the entirety of the training dataset and evaluated it on the entirety of the testing dataset.

Figure 2: Diagram representing the classification experiments performed and their nature within the 3 x 3 combinatorial design.



Classification Algorithms and Metrics

For both intra- and inter-platform experiments, training data represented by the top 20% features (as described in the Feature Selection section) was fed into a model to learn the classification task. We tried training the model over several algorithms including the random forest, logistic regression, and support vector machine and multi-layered perceptron [26]. We selected these algorithms as they represented a variety of different types of learning algorithms [26]. This ensures that our analysis of performance differences between intra- and inter-platform experiments will hold irrespective of the learning algorithm selection. We utilized scikit-learn's implementation (version 0.24.1) for all the aforementioned algorithms [25]. For each algorithm, we fine-tuned its hyperparameters using 5-fold stratified cross-validation via scikit-learn's GridSearchCV() pipeline, retaining the best hyperparameters per algorithm for analysis [25]. For each classification algorithm, the chosen hyperparameters are given below (all other hyperparameters are left default according to scikit-learn's specification):

- Random Forest: max_depth: 15, n_estimators: 100, max_features: None
- Logistic Regression: penalty: l2, C: 0.1
- Support Vector Machine: kernel: rbf, C: 0.01, gamma: scale
- Multi-layered Perceptron: alpha: 0.0001, hidden_layer_sizes = (512, 256, 128)

We measured the performance of models using the following metrics, all of which are commonly used in binary classification models. Here, we abbreviate the number of true positives, true negatives, false positives and false negatives as TP, TN, FP, FN respectively [27]:

- Accuracy: Also known as Rand Accuracy, it is the ratio between correct predictions and all predictions.
  - Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
- Precision: The ratio between correct positive predictions and the total number of positive predictions.
  - Precision = $\frac{TP}{TP + FP}$
- Recall: The ratio between correct positive predictions and the total number of true positive instances.
  - Precision = $\frac{TP}{TP + FN}$
- F1: The harmonic mean between the precision and recall.
  - F1 = $2 * \frac{Precision * Recall}{Precision + Recall}$ = $\frac{TP}{TP + ½(FN + FP)}$
- Area under Receiver Operating Characteristic (ROC) Curve (AUROC): The area under the ROC curve, which plots the False Positive Rate (FPR) against the True Positive Rate (TPR). In practice, it is often estimated using the trapezoidal rule with the formula
  - AUROC = $1 - \frac{FPR + FNR}{2}$ = $1 - \frac{\frac{FP}{FP + TN} + \frac{FN}{FN + TP}}{2}$

Feature Importance Selection

We used Shapley Additive exPlanations (SHAP) to examine how certain features affect our model's decision to predict the user with potential psychiatric hospitalization due to SSD, given their social media data from the three inspected social media platforms. Our decision to use SHAP rather than other explainability methods stems from the fact that SHAP is not only model-agnostic but also the most theoretically-sound explainability framework out of the available options. This is because SHAP feature scores can be calculated for localized samples and the entire global dataset [28]. SHAP is based on top of Shapley values, a game-theoretic concept that intuitively describes each feature's contribution to the outcome after considering all possible combinations of features [29].

For each of the intra-platform experiments within the 3x3 combinatorial design and each machine learning model, we calculated the average SHAP values for each of the features (that is, their importance to the prediction) across all the instances within the testing set. Then, we recorded the list of features sorted by descending order according to the average SHAP values measured by each model. For models with native support for feature importance extraction, including Random Forest (Gini importance) and Logistic Regression (feature coefficients), we also calculated and recorded them in an equivalent manner to SHAP values.

Robustness Checks

To ensure our findings regarding differences of model performance between models and between intra- and inter-platform experiments still hold when certain aspects about training and testing datasets are made more ideal, we performed several robustness checks which are described in the Supplementary Information (Multimedia Appendix 1)

RESULTS

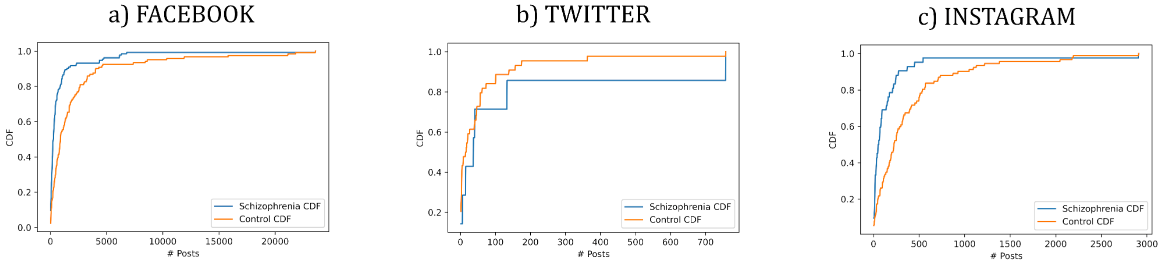
Data Characteristics

In total, 268 participants (mean age 24.73 years; male: 47.4%; Schizophrenia Spectrum Disorder (SSD): 52.6%) with non-empty windowed data for at least one platform were included. Out of those participants, 254 (254/268, 94.8%; SSD: 52.4%) had valid windowed Facebook data, 51 (51/268, 19%; SSD: 13.7%) had valid windowed Twitter data and 134 (134/268, 50%; SSD: 31.3%) had valid windowed Instagram data. For participants with valid data for more than one platform, 47 (47/268, 17.5%; SSD: 10.6%) had valid data for both Facebook and Twitter, 38 (38/268, 14.2%; SSD: 10.5%) had valid data for both Twitter and Instagram, and 119 (119 / 268, 44.4%; SSD: 28.6%) had valid data for both Facebook and Instagram. Finally, 38 (38/268, 14.2%; SSD: 10.5%) participants had valid data for all three platforms. Table 1 describes the demographic and clinical characteristics of these 268 participants. Table 2 describes summary statistics, including mean, median, etc. for these windowed data for each of the three social media platforms, grouped by clinical status (SSD vs Control). Figure 3 describes the distribution of available posts for participants within each of the three investigated platforms

Table 2: Summary statistics for windowed data for both the control class and the SSD class (i.e., participants hospitalized with SSD). Here, we consider data from Facebook, Twitter, and Instagram as mentioned above.

	Facebook		**Twitter**		**Instagram**	
	SSD Class	Control Class	SSD Class	Control Class	SSD Class	Control Class
Total #Users, n (%)	133 (52%)	121 (48%)	7 (14%)	44 (86%)	42 (31%)	92 (69%)
Total #Posts, n (%)	114793 (68%)	54632 (32%)	991 (4%)	22786 (96%)	7111 (30%)	16440 (70%)
Avg #Posts	863.1	451.5	141.6	519.9	169.3	178.7
Median #Posts	260	184	37	138	54.5	103
Max #Posts	23589	4852	758	7056	2909	1328
Min #Posts	2	1	1	1	1	1

Figure 3: Cumulative Distribution Function (CDF) curves of users and their #posts for the SSD and control classes per dataset: (a) Facebook (left), b) Twitter (center), c) Instagram (right).



a) FACEBOOK

b) TWITTER

c) INSTAGRAM

Results of Combinatorial Classification

We report the full results of the intra-platform experiments in Table 3. We also report the full results of inter-platform experiments in Table 4. Finally, we report the Receiver Operating Characteristic (ROC) curves for the best-performing logistic regression model for the experiments in Tables 3 and 4 in Figure 4.

Elaborating on the results from Table 3, we found that among the four classification algorithms that we utilized, the logistic regression model performed the best across the three intra-platform experiments, with the best performances for all of them. More elaborately, for the intraplatform experiments, performance reaches its peak with the logistic regression model with an average F1 score of 0.72, Accuracy of 0.81, and AUROC of 0.749. In contrast, the worst-performing model (in this case, multi-layered perceptron) achieved an average F1-score of 0.521, Accuracy of 0.714, and AUROC of 0.623 for the intra-platform experiments. We will be thus using the logistic regression model for further analysis regarding feature importance between platforms. These results align with prior research and thus could be considered a soft replication of those findings [8, 15].

In contrast, by aggregating metrics for the inter-platform experiments presented in Table 4, the average F1 score was lowered to 0.428 (Accuracy = 0.559, AUROC = 0.533) for the logistic regression model. This constitutes, on average, a drop of 40%, 31.4%, and 28.8% in F1 score, Accuracy, and AUROC score from the intra-platform experiments, respectively. As just demonstrated, when comparing the effectiveness of models between intra-platform and inter-platform experiments, we found a consistent drop in performance for all the investigated social media platforms. The drop in test F1, given the best-performing logistic regression model, is the most drastic for Facebook at 0.364 F1 (46%) and least drastic for Twitter at 0.08 F1 (14%), averaging a drop of 0.285 F1 (40%), going from 0.713 for intra-platform experiments to 0.428 for inter-platform experiments). Such trends hold even when disparities in dataset size and dual-platform data availability (as described in the Methods section under Robustness Checks) are applied to the training and testing data. (Multimedia Appendix 1)

Table 3: Classification results for all intra-platform classification experiments. Here, for instance, Facebook indicates the Facebook-Facebook experiment. The following metrics are utilized: Accuracy (Acc), Precision (P), Recall (R), F1, and Area under the Receiver Operating Characteristic Curve (AUROC).

	Facebook					**Twitter**					**Instagram**				
Models	Acc	P	R	F1	AUROC	Acc	P	R	F1	AUROC	Acc	P	R	F1	AUROC
Random Forest	0.739	0.739	0.738	0.738	0.709	0.745	0.150	0.116	0.116	0.494	0.7	0.648	0.637	0.637	0.681
SVM	0.72	0.74	0.69	0.71	0.72	0.85	0.54	0.45	0.46	0.69	0.74	0.73	0.75	0.74	0.80
MLP	0.50	0.40	0.50	0.36	0.51	0.84	0.45	0.45	0.42	0.69	0.79	0.77	0.79	0.77	0.84
Logistic Regressi	0.759	0.767	0.758	0.756	0.727	0.881	0.742	0.6	0.63	0.772	0.792	0.771	0.801	0.773	0.848

Table 4: Classification results for inter-platform classification experiments. The abbreviations for metrics are identical to that used in Table 3.

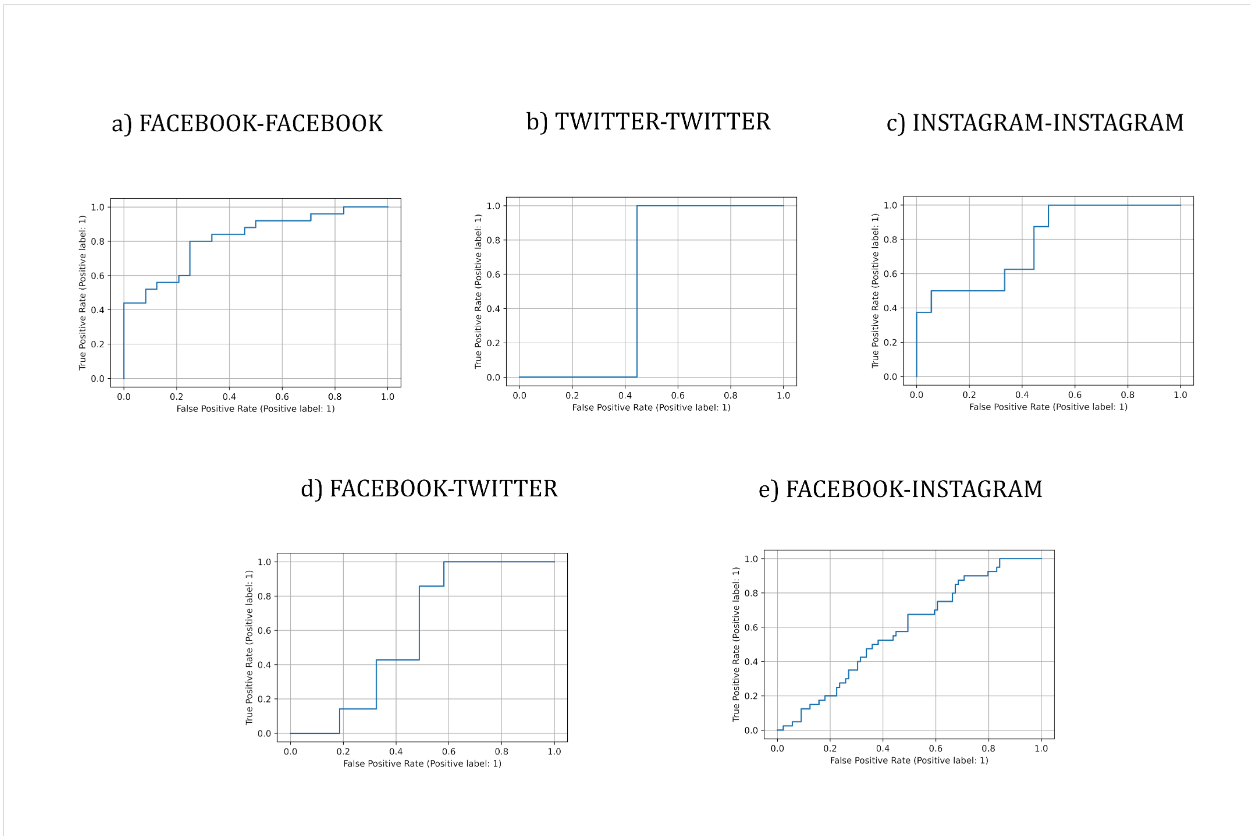
Training Data: Facebook										
	Twitter					**Instagram**				
Models	Acc	P	R	F1	AUROC	Acc	P	R	F1	AUROC
Random Forest	0.392	0.221	0.88	0.354	0.579	0.379	0.328	0.952	0.488	0.537
SVM	0.545	0.253	0.72	0.373	0.612	0.432	0.337	0.860	0.483	0.550
MLP	0.587	0.240	0.55	0.334	0.573	0.435	0.332	0.812	0.471	0.539
Logistic Regression	0.628	0.246	0.47	0.323	0.567	0.472	0.344	0.775	0.476	0.555

Training Data: Twitter										
	Facebook					**Instagram**				
Models	Acc	P	R	F1	AUROC	Acc	P	R	F1	AUROC
Random Forest	0.531	0.569	0.378	0.452	0.536	0.628	0.331	0.207	0.252	0.512
SVM	0.514	0.53	0.537	0.530	0.513	0.563	0.340	0.42	0.373	0.523
MLP	0.533	0.561	0.440	0.492	0.536	0.557	0.325	0.395	0.356	0.512
Logistic Regression	0.534	0.552	0.522	0.535	0.535	0.578	0.362	0.47	0.408	0.548

Training Data: Instagram										
	Facebook					**Twitter**				
Models	Acc	P	R	F1	AUROC	Acc	P	R	F1	AUROC
Random Forest	0.51	0.523	0.612	0.563	0.507	0.751	0.369	0.42	0.386	0.624
SVM	0.524	0.544	0.51	0.524	0.525	0.691	0.213	0.25	0.229	0.521
MLP	0.554	0.584	0.48	0.526	0.557	0.683	0.201	0.23	0.214	0.51

Logistic Regression	0.516	0.524	0.689	0.595	0.51	0.628	0.256	0.52	0.342	0.587

Figure 4: Receiving Operator Characteristic (ROC) curves for the classification experiments, given the best logistic regression model. a), b), c) are curves for Facebook, Twitter, and Instagram intra-platform results respectively from Table 3. d), e) are the ROC curves for inter-platform experiments from Table 4 where Facebook is utilized as the training data.



a) FACEBOOK-FACEBOOK

b) TWITTER-TWITTER

c) INSTAGRAM-INSTAGRAM

d) FACEBOOK-TWITTER

e) FACEBOOK-INSTAGRAM

Feature Importance Analysis

We hypothesize that the decrease in performance going from intra-platform experiments to inter-platform experiments, as presented above, is driven by differences in feature importance learned by models when trained on data from different social media platforms (even when they shared the same feature set). By extracting the list of SHAP features from the models per the method described above, we found support for this hypothesis. Specifically, we observed little overlap between them across platforms among the top 25 features for each model and platform (when holding the model constant). On average, there were only 4.66 overlapping features for the same logistic regression classification model across platforms (the best performing model, based on the above discussions). In addition, we also found that list of feature importance for each of the platforms, based on the logistic regression model, have very weak rank correlation pairwise. Fully enumerating on the statistical results for Kendall's rank correlation coefficient, we found very weak rank correlation between the ranked lists of feature importance for Facebook and Twitter ($\tau_b = 0.0{,}81$ $p = 0.003$), Facebook and Instagram ($\tau_b = 0.041, p = 0.013$), and Twitter and Instagram ($\tau_b = 0.055, p = 0.05$). We report the average SHAP values and logistic regression coefficients of the top 10 features based on their SHAP values, along with their average value in the SSD class and the control class in Table 5.

Table 5: Top 10 features for the Logistic Regression (LR) model for each of the platforms (Linguistic Inquiry and Word Count (LIWC) features are italicized) based on their Shapley Additive exPlanations (SHAP) values.

Platform: Facebook					
Feature acronym	**Feature**	**SHAP value**	**LR**	**Avg in SSD**	**Avg in**
Avg_post_readability	Average post readability, as measured using the Simple Measure of Gobbledygook	0.761	-0.268	5.6341	6.8048
Quant	Ratio of words within the "quantifiers"	0.4195	-0.189	0.0012	0.0016
Negemo	Ratio of words within the "negative	0.0953	0.244	0.0043	0.0031
Money	Ratio of words within the	0.0739	-0.216	0.0007	0.0011
Swear	Ratio of words within the "swear"	0.0628	0.236	0.0017	0.0007
Ratio_octile8	Ratio of activities made from 21:00 to 00:00	0.0443	0.077	0.1443	0.1241
Ratio_octile7	Ratio of activities made from 18:00 to 21:00	0.0409	0.177	0.1561	0.1054
Anger	Ratio of words within the "anger"	0.0095	0.191	0.0018	0.0009

Dream	Ratio of “dream” within the overall bag of words	0.0077	0.224	0.2028	0.0746
Fun	Ratio of “fun” within the overall bag of words	0.0043	-0.209	0.5722	1.1315

Platform: Twitter					
Feature acronym	**Feature description**	**SHAP value**	**LR Coefficient**	**Avg in SSD**	**Avg in Control**
Conj	Ratio of words within the “conjunctions” category	0.2319	-0.063	0.0001	0.0003
Adj	Ratio of words within the “adjectives” category	0.1825	-0.05	0.0057	0.0080
Avg_post_negativity	Average post negativity, as calculated using the Valence Aware Dictionary for sEntiment Reasoning (VADER) library	0.1509	0.082	0.071	0.0519
Male	Ratio of words within the “male” category	0.1355	0.039	0.0011	0.0007
Ratio_octile_8	Ratio of activities made from 21:00 to 00:00	0.1265	0.045	0.0231	0.1227
Ingest	Ratio of words within the “ingest” category	0.0627	-0.056	0.0003	0.0014
Insight	Ratio of words within the “insight” category	0.0516	0.053	0.0044	0.0035
Power	Ratio of words within the “power” category	0.0308	-0.058	0.0024	0.0042
We	Ratio of words within the “we” category	0.0196	-0.056	0.0001	0.0002
Prep	Ratio of words within the “prepositions” category	0.0117	0.063	0.0028	0.0017

Platform: Instagram					
Feature acronym	**Feature description**	**SHAP value**	**LR Coefficient**	**Avg in SSD**	**Avg in Control**
Avg_post_readability	Average post readability, as measured using the Simple Measure of Gobbledygook (SMOG)	0.761	-0.203	5.1018	6.2564
Space	Ratio of words within the “space” category	0.733	-0.147	0.0031	0.0042
Affiliation	Ratio of words within the “affiliation” category	0.6839	-0.181	0.0032	0.0056
Friend	Ratio of words within the “friend” category	0.5336	-0.159	0.0009	0.0018
Female	Ratio of words within the “female” category	0.4576	-0.168	0.0008	0.0019
Sad	Ratio of words within the “sad” category	0.4554	0.113	0.0011	0.0007
Quant	Ratio of words within the “quantifier” category	0.4195	-0.118	0.0012	0.0019
Away	Ratio of “away” within the overall bag of words	0.4064	-0.105	0.0768	0.2505
Assent	Ratio of words within the “assent” category	0.3913	-0.102	0.0008	0.0013
Next	Ratio of “next” within the overall bag of words	0.3854	-0.12	0.0957	0.6466

Table 6: Example (paraphrased and deidentified) posts representative of example top features to distinguish between SSD and control classes. Words indicative of the features are highlighted in light blue.

Feature	**Example Paraphrased Post**
NegEmo	I fear to try and fail, because i don’t want to be part of the STATISTIC of people that failed. It hurts when the opportunity passes by though.’

Swear	Omfg the Damn mf #struggle to stay the fking sleep I'm like wtf this isn't fair I hate my Damn neck hurting like this shit isn't cool this pain waking me up every Damn hr
Sad	Im a useless sorry sob
Anger (Anger)	Yo stay tf out my room unless we fucking cause I'm tired too tired for this shit 😑 and all my shit better be where i left it

Attributing Divergent Construct Validity of Models to Divergent Identities Online

What could explain the observed differences in construct validities of the intra-platform models? Early in the paper we posited that these differences might stem from people's identities being fragmented across different platforms. To situate that these divergent identities are indeed the drivers behind differential cross-platform model construct validities, and by extension, performance, we adopted a strategy to measure the differences within the extracted feature space between the investigated platforms for a given participant. As social media data for participants on all platforms are encoded via feature vectors in this work, we calculated the pairwise similarity between platform-specific data using cosine similarity [30]. More specifically, we calculated the average cosine similarity within-subjects between-platforms and compared it with the average cosine similarity between-subjects within-platforms for SSD participants with data on all three platforms. Given that even within the same social media platform, different people can have unique modes of expressing their unique identities, we used the latter as a baseline for assessing whether fragments of identities representing an individual across platforms diverge more or less than the divergence of identities between individuals.

We found that the average between-platforms, within-participants cosine similarity is 0.3093 for Facebook-Twitter, 0.2304 for Facebook-Instagram and 0.3905 for Twitter-Instagram. This is either lower than or similar to the average within-platforms, between-participants cosine similarity for the investigated platforms: 0.5072 for Facebook, 0.5427 for Twitter and 0.373 for Instagram respectively. The same trend holds even when calculating the averages utilizing data from both SSD and HC participants with data from all three platforms.

DISCUSSION

Principal Findings

Our paper aims to measure the ability (or inability) of mental health classifiers to generalize across platforms and surface evidence for fragmented identities on social media among patients suffering from SSD. Overall, we found that, across the board, models trained on data from social media platform have poor generalizability when evaluated on data from other social media platforms, even when holding the feature set constant across training and testing data. This trend holds true even when the two robustness tests where the same participants and dataset size are used in the training and testing data (as described in the Methods section). This trend is also true even when the training data comes from a platform with high data availability and the testing data comes from a platform with low data availability. For instance, the best F1-score of the intra-platform models for Twitter (F1 = 0.63) is 0.257 (69%) higher compared to the best F1-score of the inter-platform models for Twitter where the training data comes from Facebook (F1 = 0.373).

Next, we discuss findings regarding feature importance in more details. Firstly, looking at the theoretical validity of the top 10 features per platform and interpretation of the sign of features' logistic regression coefficient, we find alignment with prior literature and evidence of clinical meaningfulness [7, 8, 11]. For instance, given the positive coefficient from the trained logistic regression model presented in Table 5, higher levels of usage of lexicon indicative of Negative Emotions are highly predictive of SSD for Facebook (see example post in Table 6 highlighting words like "fear," "fail," and "hurts"). This confirms literature noting that a reduced ability to feel or express pleasure (anhedonia) is common in patients with SSD [31]. Similarly, prior research has found anger related terms commonly appearing in social media postings before the onset of early psychosis as well as preceding a psychiatric hospitalization [32]. This may explain why higher levels of usage of lexicon indicative of the LIWC category Anger is also highly predictive of SSD for Facebook (example post in Table 6 containing Anger words like "shit" and "fucking"). Finally, words and phrases like those in the LIWC Sad category (e.g., "useless," "sorry," "sob") point to typical negative symptoms of SSD [33]. They can be indicative of a decreased sense of purpose and a seeming lack of interest in the world [33]. Models trained on Instagram successfully pick up such cues from the postings, where higher usage of such vocabulary is indicative of an impending psychiatric hospitalization due to SSD.

That said, each model corresponding to each platform seems to pick up contrasting signals from their respective training data, which is why we note the low overlap in top SHAP features above. Among the few that overlap in the top 10 features reported above, we find "avg_post_readability" to be picked up as a highly predictive feature by both Facebook and Instagram models, while "ratio_octile8" by both Facebook and Twitter. Poor readability of the written text is a known negative symptom of schizophrenia and related psychotic disorders as observed in prior work [34]. In addition, higher levels of late-night activity such as web or social media use, captured in the "ratio_octile8" feature, have been known to be associated with deteriorated mental health [35]. Finally, we found significant divergence in the distribution of feature importance between the platforms, as indicated by the low pairwise Kendall's tau (< 0.1) for the platforms' feature importance rankings. These qualitative and quantitative results, broadly imply that the models are getting trained on considerably different data sources with differing content and context of use, which is likely contributing to poor cross-platform model generalization.

At the crux of these differences, we find the models have inherently different construct validity across platforms. Data on each platform reflects only a segment of an individual's identity – a segment that may be absent in another platform. The fragmentation of one's identity on social media can be most clearly seen within participants with data on all three platforms. In the analysis presented at the end of the Results section, we found low average pairwise cosine similarities within-participants between-platforms, especially when comparing to cosine similarities of different participants within the same platform. This indicates that even within the same feature space for the same participant, social media data between platforms is likely to diverge into multiple distinct directions mapping to these fragments of identities. This divergence at least equal to, if not even more so than the divergence in identity presentation between different individuals within the same social media platform. Therefore, when models trained on data from one platform learn this specific fragment of identity, they are less effective on testing data that capture a different identity.

Comparison with Prior Work

Our findings provide replicative validity to several threads in prior research. Specifically, we found that the performance of models trained on social media data with clinically verified labels (i.e., SSD or control) is consistent with similar models presented in previous research, including those trained on similar patient populations and similar clinical sites [6, 8]. Furthermore, linguistic differences reflecting serious mental health conditions between social media platforms found in our work have also been elucidated upon in previous work. For instance, Guntuku et al. (2019) found that there is little overlap between words indicative of stress on Twitter and Facebook [36]. Additionally, our findings regarding the low performance of models for inter-platform tasks compared to intra-platform tasks follows a similar vein as Ernala et al. (2019) [8]. In this paper, they found that despite the overwhelming advantage in data availability, models trained on social media data with self-reported labels significantly underperforms models trained on social media data with clinically verified labels when evaluated on clinical testing data [8]. Like our experiments, such a difference in performance in Ernala et al. (2019) was also noted to be caused by a mismatch in important features learned by the different models to differentiate between language and activity patterns deployed by SSD patients versus healthy controls [8]. Overall, our analysis combined with previous results suggests that construct validities of predictive models trained on data from different social media platforms are dissimilar, reinforcing the need for continued exploration of novel social media-based early identification strategies, with a special emphasis on uniting distinct fragment of identities for accurate identification and intervention.

Clinical Implications

Our findings provide important implications for mental health research and practice. Hospitalization prediction for psychiatric illnesses by harnessing digital trace data has been of significant interest in recent years. These prior works have explored the utility of smartphone sensor data (i.e., geolocation, physical activity, phone usage, and speech), wearables, and social media activity to predict symptom fluctuations as well as to understand the diagnostic process and hospitalization identification [6, 37-40]. Our work extends this body of research by critically examining how machine learning efforts that harness data from singular sources may not be readily applicable to support hospitalization prediction in contexts where the same source of data is not present. For these models to be usable in the real world, we advocate for a comprehensive approach where clinicians look to patterns gleaned through the integration of different data sources while augmenting their decision-making with objective measures derived from digital trace data. Social media data is also increasingly becoming part of consultations [41, 42]. Therefore, we suggest clinicians consider both acknowledging and incorporating collateral information spanning multiple platforms in the way they monitor symptomatic exacerbation in their patients and modify treatment to prevent further hospitalizations.

Finally, digital interventions that are touted to be powered by social media data should consider this significant aspect of fragmented online identities of patients [43, 44]. To intervene at the right time, at the right place, for the right person, a comprehensive approach to understanding a patient's context for hospitalization prediction would be beneficial. Still, we recognize that in a domain as sensitive as mental health, combining data sources may further complicate the privacy and ethical risks to those who contribute their data – research has shown that information integration can enable the discovery of otherwise latent attributes, some of which may present grave feelings of discomfort and violation to individuals [45, 46]. We therefore urge caution and call for new standards to protect the confidentiality and the rights of this sensitive population and ensure that the enabled technologies are used in the service of positive outcomes for the patients.

Limitations and Future Work

Our work contains some limitations that could be addressed in future research. First, despite the use of data augmentation techniques to rebalance the ratio between SSD data and control data for each dataset, and to make the dataset size for the three examined platforms (i.e., Instagram, Twitter, and Facebook) comparable to each other, we acknowledge that limited quantity of available data may have impacted the observed classification performance. While it is widely recognized that patient social media data is challenging to collect as is the case here, future research may consider the potential of creating benchmarked large datasets that may support better reproducible research in this field [47]. Next, we acknowledge the demographics dissimilarity between participants with SSD and healthy controls, which may be a confounding factor in our study design. Furthermore, our methods did not examine nor extract any features concerning video data, which are available in Facebook and especially Instagram. Given that youths nowadays are increasingly expressing themselves on social media via videos (especially on video-centric platforms such as TikTok), future research should aim to fill in these gaps so that we can ensure the completeness of one's mental health records expressed on social media and other forms of networked communication. Along these lines, future research may also consider data from additional novel social media platforms that are increasingly being used by youth for their social goals, such as Snapchat and TikTok. Finally, it will be worthwhile to examine additional clinical questions, such as suicidal risk, to explore the extent to which identity fragmentation across social media platforms may impact the quality of inferences made from these data.

CONCLUSION

In this work, we showed that it is challenging to build effective models for predicting future psychiatric hospitalizations of SSD patients on new social media data from platforms previously unseen in the models' training data. Specifically, we demonstrated models built on one platform's data do not generalize to another because each platform consistently reflects different segments of participants' identities. This fragmentation of identity is empirically backed up by both significant differences in construct validity of intra-platform classifiers and divergent feature vectors within-participants between the three investigated social media platforms. To ensure effective incorporation of digital technology into early psychosis intervention, especially in the prevention of relapse hospitalizations, further research must explore precisely how symptoms of mental illness manifest online through changing patterns of language and activity on various platforms, as well as how comprehensive, ethical and effective treatment and engagement strategies should be devised that function seamlessly across patients' fragmented online identities.

Acknowledgements

This research was partly funded through NIMH grant R01MH117172 (PI: De Choudhury; Co-PIs: Birnbaum, Kane). The research team acknowledges the assistance of Anna Van Meter and Asra Ali in early phases of the patient data collection. We also thank members of the Social Dynamics and Well-Being (SocWeB) Lab at Georgia Tech for their valuable feedback during various phases of the study.

Conflicts of Interest

None Declared

Abbreviations

SSD: Schizophrenia Spectrum Disorder

SMOTE: Synthetic Minority Oversampling Technique

LIWC: Linguistic Inquiry and Word Count

SHAP: Shapley Additive exPlanations

VADER: Valence Aware Dictionary for sEntiment Reasoning

SMOG: Simple Measure of Gobbledygook

AUROC: Area under Receiver Operating Characteristic

REFERENCES

1. Wolthaus JE, Dingemans PM, Schene AH, Linszen DH, Wiersma D, Van Den Bosch RJ, et al. Caregiver burden in recent-onset schizophrenia and spectrum disorders: the influence of symptoms and personality traits. J Nerv Ment Dis. 2002 Apr;190(4):241-7. PMID: 11960085. doi: 10.1097/00005053-200204000-00005.
2. Birchwood M, Macmillan F. Early intervention in schizophrenia. Aust N Z J Psychiatry. 1993 Sep;27(3):374-8. PMID: 8250779. doi: 10.3109/00048679309075792.
3. Lieberman JA, Fenton WS. Delayed detection of psychosis: causes, consequences, and effect on public health. Am J Psychiatry. 2000 Nov;157(11):1727-30. PMID: 11058464. doi: 10.1176/appi.ajp.157.11.1727.
4. Birnbaum ML, Rizvi AF, Correll CU, Kane JM, Confino J. Role of social media and the I nternet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. Early intervention in psychiatry. 2017;11(4):290-5.
5. Miller BJ, Stewart A, Schrimsher J, Peeples DA, Buckley PF. How connected are people with schizophrenia? Cell phone, computer, email, and social media use. Psychiatry Research. 2015;225:458-63.
6. Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, A RVM, De Choudhury M, et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr. 2019 Oct 7;5(1):17. PMID: 31591400. doi: 10.1038/s41537-019-0085-9.
7. Mitchell M, Hollingshead K, Coppersmith GA, editors. Quantifying the Language of Schizophrenia in Social Media. CLPsych@HLT-NAACL; 2015.
8. Ernala SK, Birnbaum ML, Candan KA, Rizvi AF, Sterling WA, Kane JM, et al., editors. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. Proceedings of the 2019 chi conference on human factors in computing systems; 2019.
9. Rekhi G, Ang MS, Lee J. Clinical determinants of social media use in individuals with schizophrenia. PLoS ONE. 2019;14.
10. Zomick J, Levitan SI, Serper MR. Linguistic Analysis of Schizophrenia in Reddit Posts. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. 2019.
11. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. J Med Internet Res. 2017 Aug 14;19(8):e289. PMID: 28807891. doi: 10.2196/jmir.7956.
12. Ernala SK, Rizvi AF, Birnbaum ML, Kane JM, De Choudhury M. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. Proceedings of the ACM on Human-Computer Interaction. 2017;1(CSCW):1-27.
13. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021;1:1-4.
14. Hansel K, Lin IW, Sobolev M, Muscat W, Yum-Chan S, De Choudhury M, et al. Utilizing Instagram Data to Identify Usage Patterns Associated With Schizophrenia Spectrum Disorders. Front Psychiatry. 2021;12:691327. PMID: 34483987. doi: 10.3389/fpsyt.2021.691327.
15. Birnbaum ML, Norel R, Van Meter A, Ali AF, Arenare E, Eyigoz E, et al. Identifying signals associated with psychiatric illness utilizing language and images posted to

Facebook. NPJ Schizophr. 2020 Dec 3;6(1):38. PMID: 33273468. doi: 10.1038/s41537-020-00125-0.

16. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med. 2020;3(1):43. PMID: 32219184. doi: 10.1038/s41746-020-0233-7.

17. Kircaburun K, Griffiths MD. Instagram addiction and the Big Five of personality: The mediating role of self-liking. J Behav Addict. 2018 Mar 1;7(1):158-70. PMID: 29461086. doi: 10.1556/2006.7.2018.15.

18. Bayer JB, Ellison NB, Schoenebeck SY, Falk EB. Sharing the small moments: ephemeral social interaction on Snapchat. Information, Communication & Society. 2016;19(7):956-77.

19. Purwaningtyas MPF, Alicya DA. The Fragmented Self: Having Multiple Accounts in Instagram Usage Practice among Indonesian Youth. Jurnal Media dan Komunikasi Indonesia. 2020.

20. Gündüz U. The effect of social media on identity construction. Mediterranean Journal of Social Sciences. 2017;8(5):85.

21. First MB. Structured clinical interview for DSM-IV axis I disorders. Biometrics Research Department. 1997.

22. Zimmerman M, Mattia JI. A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. Archives of general psychiatry. 2001;58 8:787-94.

23. Ernala SK, Kashiparekh KH, Bolous A, Ali A, Kane JM, Birnbaum ML, et al. A social media study on mental health status transitions surrounding psychiatric hospitalizations. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1-32.

24. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-82.

25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. ArXiv. 2011;abs/1201.0490.

26. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.

27. Powers DMW, editor. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. 2008.

28. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems (NEURIPS)2017.

29. Winter E. Chapter 53 The shapley value. Handbook of Game Theory With Economic Applications. 2002;3:2025-54.

30. Jurafsky D. Speech & language processing: Pearson Education India; 2000. ISBN: 8131716724.

31. Kwapil TR. Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. Journal of abnormal psychology. 1998;107(4):558.

32. Ringer JM, Lysaker PH. Anger expression styles in schizophrenia spectrum disorders: associations with anxiety, paranoia, emotion recognition, and trauma history. J Nerv Ment Dis. 2014 Dec;202(12):853-8. PMID: 25386763. doi: 10.1097/NMD.0000000000000212.

33. Liu J, Chua JJ, Chong SA, Subramaniam M, Mahendran R. The impact of emotion dysregulation on positive and negative symptoms in schizophrenia spectrum disorders: A

systematic review. J Clin Psychol. 2020 Apr;76(4):612-24. PMID: 31909833. doi: 10.1002/jclp.22915.

34. Kuperberg G, Caplan D. Language dysfunction in schizophrenia. Neuropsychiatry. 2003;2:444-66.

35. Palmese LB, DeGeorge PC, Ratliff JC, Srihari VH, Wexler BE, Krystal AD, et al. Insomnia is frequent in schizophrenia and associated with night eating and obesity. Schizophrenia Research. 2011;133:238-43.

36. Guntuku SC, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH, editors. Understanding and measuring psychological stress using social media. Proceedings of the international AAAI conference on web and social media; 2019.

37. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric rehabilitation journal. 2015;38(3):218.

38. Birnbaum ML, Kulkarni PP, Van Meter A, Chen V, Rizvi AF, Arenare E, et al. Utilizing Machine Learning on Internet Search Activity to Support the Diagnostic Process and Relapse Detection in Young Individuals With Early Psychosis: Feasibility Study. JMIR Ment Health. 2020 Sep 1;7(9):e19348. PMID: 32870161. doi: 10.2196/19348.

39. Eisner E, Bucci S, Berry N, Emsley R, Barrowclough C, Drake RJ. Feasibility of using a smartphone app to assess early signs, basic symptoms and psychotic symptoms over six months: a preliminary report. Schizophrenia research. 2019;208:105-13.

40. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, et al. Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: A BiAffect Digital Phenotyping Study. J Med Internet Res. 2018 Jul 20;20(7):e241. PMID: 30030209. doi: 10.2196/jmir.9775.

41. Fisher CE, Appelbaum PS. Beyond googling: the ethics of using patients' electronic footprints in psychiatric practice. Harvard review of psychiatry. 2017;25(4):170-9.

42. Rieger A, Gaines A, Barnett I, Baldassano CF, Connolly Gibbons MB, Crits-Christoph P. Psychiatry Outpatients' Willingness to Share Social Media Posts and Smartphone Data for Research and Clinical Purposes: Survey Study. JMIR Form Res. 2019 Aug 29;3(3):e14329. PMID: 31493326. doi: 10.2196/14329.

43. Yoo DW, Birnbaum ML, Van Meter AR, Ali AF, Arenare E, Abowd GD, et al. Designing a Clinician-Facing Tool for Using Insights From Patients' Social Media Activity: Iterative Co-Design Approach. JMIR Mental Health. 2020;7.

44. Yoo DW, Ernala SK, Saket B, Weir D, Arenare E, Ali AF, et al. Clinician Perspectives on Using Computational Mental Health Insights From Patients' Social Media Activities: Design and Qualitative Evaluation of a Prototype. JMIR Mental Health. 2021;8.

45. Terrasse M, Gorin M, Sisti D. Social Media, E-Health, and Medical Ethics. Hastings Cent Rep. 2019 Jan;49(1):24-33. PMID: 30790306. doi: 10.1002/hast.975.

46. Thieme A, Belgrave D, Sano A, Doherty G. Machine learning applications. Interactions. 2020;27:6 - 7.

47. Househ M, Grainger R, Petersen C, Bamidis P, Merolli M. Balancing between privacy and patient needs for health information in the age of participatory health and social media: a scoping review. Yearbook of medical informatics. 2018;27(01):029-36.

48. Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing & Management. 2003;39(1):45-65.

49. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 2010;29:24 - 54.

50. Mclaughlin GH. SMOG Grading - A New Readability Formula. The Journal of Reading. 1969.

51. Hutto C, Gilbert E, editors. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media; 2014.

52. Garimella VRK, Alfayad A, Weber I, editors. Social media image analysis for public health. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; 2016.

53. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.