Research Article

Bridging the Gap Between Clinical Reality and Public Perception: A Comparative Analysis of Schizophrenia Symptom Severity Using Social Media and Clinical Datasets

Gan Chen

Diamond Bar High School. 21400 Pathfinder Rd, Diamond Bar, CA 91765, United States

ABSTRACT

Public perceptions of schizophrenia are often shaped by digital narratives that emphasize certain symptoms while minimizing others. This study examines the divergence between clinical symptom severity and online perception using two datasets: one from a clinical PANSS-based schizophrenia assessment and another derived from Reddit posts classified via zero-shot learning. Eleven overlapping symptoms were extracted and scored on a 1–7 scale. Methods included comparative boxplots, Chi-square tests, correlation heatmaps, random forest classification, and logistic regression. The results revealed consistent statistical and perceptual mismatches across domains. Symptoms such as tension and suspiciousness were overrepresented in Reddit posts, while cognitive and affective symptoms like guilt and poor attention were underrepresented. The findings underscore the need for public education efforts tailored to underrecognized symptoms, and highlight the potential of machine learning in digital psychiatry. Bridging these gaps may improve stigma reduction and enhance digital mental health literacy.

Keywords: Schizophrenia; stigma; natural language processing; public perception; digital psychiatry; machine learning

INTRODUCTION

Schizophrenia is a severe psychiatric disorder marked by disruptions in cognition, emotion, and perception, manifesting through a range of positive, negative, and general psychopathological symptoms (1). While considerable progress has been made in clinical diagnostics and symptom quantification—most notably through the Positive and Negative Syndrome Scale (PANSS)—public understanding of schizophrenia remains fraught with misrepresentation and stigma, especially across social media platforms. This discrepancy between clinical characterization and public perception has significant implications, not only for the individuals directly affected but also for the broader goals of mental health education, advocacy, and policy (2).

Stigma associated with schizophrenia is often fueled by the portrayal of certain symptoms—such as delusions, hallucinations, or aggression—as representative of the disorder as a whole, thereby eclipsing other crucial symptom domains like social

Corresponding author: Gan Chen, E-mail: chengan0206@gmail.com.

Accepted September 18, 2025
https://doi.org/10.70251/HYJR2348.35316327

withdrawal, guilt, or poor attention. These portrayals are increasingly shaped by user-generated content on platforms like Reddit, where discussions around mental illness are simultaneously informative and susceptible to distortion (3). While recent research has explored sentiment analysis, misinformation, and language patterns in digital discussions of mental illness, fewer studies have systematically compared social media perceptions to structured clinical assessments across shared symptom domains.

Reducing stigma and improving public literacy around schizophrenia necessitate multidimensional approaches. These include clinical education, media reform, and digital literacy interventions. A vital component of these strategies lies in understanding how symptoms are perceived and emphasized differently in clinical settings versus social media discourse. Traditional approaches have relied on qualitative coding of public discourse or survey-based assessments of stigma (4). In contrast, the emergence of natural language processing and machine learning enables researchers to process large corpora of online text and classify it against psychiatric symptom categories with increasing precision.

The current study aims to bridge the gap between clinical characterization and online perception of schizophrenia symptoms by combining PANSS-based clinical data with Reddit posts annotated through zero-shot classification using the Facebook BART large MNLI model. Posts were mapped to symptom labels with associated confidence scores, which were then rescaled to align with PANSS severity levels. This methodological pipeline allows for a direct, cross-domain comparison of symptom emphasis and severity. By employing both traditional statistical analyses—such as the Chi-Square Test—and machine learning approaches—including predictive modeling and classification—the study offers a hybrid framework that captures both the distributional and inferential contrasts between the two sources.

This paper proceeds with a rigorous examination of eleven shared symptoms, including delusions, hallucinations, suspiciousness, blunted affect, guilt, tension, and poor attention. Through normalization, visualization, statistical testing, and supervised classification, we seek to uncover which symptoms are under- or over-represented in Reddit discussions compared to clinical records. The findings are expected to advance our understanding of digital mental health narratives and support the development of more targeted, symptom-informed anti-stigma campaigns in online spaces.

METHODS AND MATERIALS

This section outlines the full methodological framework used to investigate the divergence between clinical assessments of schizophrenia symptoms and public perceptions derived from Reddit posts. The study is built upon two core datasets: a structured clinical schizophrenia dataset (2.1) and an unstructured Reddit-based mental health dataset (2.2). Both datasets were carefully curated and processed to align symptom categories and enable valid cross-comparisons.

To support this alignment, a natural language processing pipeline was developed using the Facebook BART zero-shot model (5), which extracted symptom labels and severity scores from Reddit posts (2.3). This standardized scoring approach enabled direct quantitative comparisons across datasets. From this foundation, we implemented a series of analytical techniques to examine whether public perceptions of schizophrenia symptoms differ significantly from clinical realities.

We began with comparative boxplots (2.4) to provide an intuitive visual exploration of symptom severity distributions across the two datasets. Next, we conducted Chi-square tests of independence (2.5) to assess the statistical significance of observed discrepancies. To model the underlying dynamics of misperception, we applied Random Forest analysis to identify which symptoms most influence Reddit severity judgments, followed by logistic regression modeling (2.6) to estimate the likelihood of over- or underestimation for each symptom.

This layered approach balances statistical rigor with interpretive sensitivity, addressing the complex and multifaceted nature of symptom perception. Each method contributes uniquely to our central research question: Are symptom severities perceived on Reddit aligned with clinical reality, and if not, in what ways do they diverge?

Clinical Dataset Data Description

The clinical dataset used in this study originates from a published clinical research project conducted in Russia by Lezheiko *et al.* (6), which was made publicly available for secondary analysis through Mendeley Data. The dataset comprises detailed psychiatric and demographic information for over 1,100 individuals

formally diagnosed with schizophrenia or related disorders. It includes variables spanning patient demographics, birth seasonality, comorbidities, age at symptom onset, and scores from the Positive and Negative Syndrome Scale (PANSS), a gold-standard instrument for evaluating schizophrenia symptoms. The PANSS framework measures symptoms across three domains: Positive, Negative, and General Psychopathology. Each symptom is rated on a 7-point scale, allowing for nuanced assessments of symptom severity in clinical populations.

For the purposes of this research, eleven PANSS symptom variables were extracted based on overlap with predicted symptoms in the Reddit dataset. These include p1, p2, p3, p6, p7 (Positive symptoms), n1 (Negative symptom), and g2, g3, g4, g6, g11 (General Psychopathology symptoms). Each of these corresponds to a discrete psychiatric construct, such as delusions or anxiety, and was rated by clinicians using structured interviews. All scores were retained in their original 1–7 integer form for compatibility with the Reddit-derived severity scores. The statistical properties and definitions of the selected clinical variables are presented in Table 1.

Reddit Posts Data Description

The Reddit-Based Schizophrenia Detection Dataset, sourced from Kaggle (7), consists of user-generated content collected from various mental health-related subreddits. It includes rich textual data and user interaction metadata intended for mental health classification tasks. The original dataset features numerous columns, including timestamps, user IDs, subreddit names, and post metadata such as score and comment count. Its primary purpose is to facilitate research into language patterns and behavioral markers associated with schizophrenia-spectrum disorders as they appear in social media discourse. Given the unstructured and noisy nature of social media data, the dataset provides a valuable but complex resource for

Table 1. The List of Variables and The Associated Definition and Descriptive Statistics for the Clinical Dataset

Variables	Type	Definition	Descriptive Statistics
p1	Categorical	Delusions – fixed, false beliefs not based in reality	1: 479 2: 655 3: 956 4: 612 5: 234 6: 53 7: 14 Mean: 3.92
p2	Categorical	Conceptual Disorganization – disorganized thinking and speech	1: 104 2: 108 3: 283 4: 506 5: 645 6: 728 7: 469 Mean: 4.30
p3	Categorical	Hallucinations – sensory perceptions without external stimuli	1: 33 2: 126 3: 410 4: 750 5: 646 6: 487 7: 391 Mean: 4.35
p6	Categorical	Suspiciousness/Persecution – beliefs of being targeted or harmed	1: 539 2: 316 3: 325 4: 370 5: 506 6: 420 7: 367 Mean: 4.28
p7	Categorical	Hostility – verbal or physical aggression	1: 539 2: 316 3: 325 4: 370 5: 506 6: 420 7: 367 Mean: 3.66
n1	Categorical	Blunted Affect – lack of emotional expression	1: 738 2: 437 3: 529 4: 442 5: 307 6: 227 7: 161 Mean: 3.91
g2	Categorical	Anxiety – apprehension, tension, or uneasiness	1: 254 2: 336 3: 844 4: 916 5: 436 6: 176 7: 44 Mean: 3.91
g3	Categorical	Guilt Feelings – self-blame or remorse	1: 1065 2: 813 3: 652 4: 350 5: 83 6: 34 7: 6 Mean: 3.02
g4	Categorical	Tension – observable nervousness or irritability	1: 1065 2: 813 3: 652 4: 350 5: 83 6: 34 7: 6 Mean: 3.13
g6	Categorical	Depression – low mood or hopelessness	1: 985 2: 757 3: 644 4: 360 5: 200 6: 51 7: 6 Mean: 3.17
g11	Categorical	Poor Attention – reduced ability to concentrate or focus	1: 1700 2: 573 3: 356 4: 205 5: 121 6: 36 7: 12 Mean: 3.05

computational psychiatry.

For the present study, only four columns from the original dataset were utilized: cleaned_text, predicted_symptom, symptom_confidence, and predicted_severity_score. The cleaned_text column contains user posts that have been preprocessed to remove punctuation, stopwords, and other forms of irrelevant noise, enabling more accurate natural language processing. The predicted_symptom and symptom_confidence columns were not originally present in the dataset but were generated using a zero-shot classification method powered by the Facebook BART large model (5). This allowed each post to be mapped to one of several clinically relevant schizophrenia symptoms with an associated confidence score. Based on these outputs, we derived the predicted_severity_score, a normalized value scaled to match clinical severity ratings, enabling a direct comparison with clinical PANSS scores. The relevant structural and statistical properties of the subset used are summarized in Table 2.

Symptom Extraction and Score Assignment using a Language Model

The Reddit dataset originally consisted of unstructured mental health-related comments, with no explicit symptom tags or severity scores. To enable comparative analysis, we employed the Facebook BART zero-shot classification model to infer symptom labels and assign confidence scores to each post. Given a predefined list of PANSS-relevant symptoms (e.g., delusions, hallucinations, conceptual disorganization), each Reddit comment was processed through the model using these symptoms as hypothesis labels. For each post, the model returned the most likely symptom along with a confidence score ranging from 0 to 1.

To translate the model's confidence output into a clinically comparable format, we used a linear transformation approach to generate pseudo-severity scores. Specifically, the confidence score c was converted to a 7-point severity scale via the ceiling function when the confidence scores are timed by seven:

$$\text{Predicted Severity Score} = \lceil 7 \times c \rceil \quad (1)$$

This allows each post to be aligned with the standard PANSS 1–7 scoring system. The resulting structured data (cleaned_text, predicted_symptom, symptom_confidence, predicted_severity_score) forms the basis for all subsequent analysis of Reddit-based perceptions.

Comparative Boxplots

Comparative boxplots are a powerful tool for visualizing the spread and central tendency of severity scores across distinct datasets. In this study, for each of the 11 PANSS symptoms analyzed (e.g., p1 to p7, n1, g2, g3, g4, g6, g11), side-by-side boxplots are constructed to display and compare the distribution of severity scores from the clinical dataset and Reddit-derived predictions. This enables immediate visual insight into which symptoms are generally overestimated, underestimated, or perceived similarly by the public, as reflected in Reddit discussions.

Chi-Square Test

To determine whether there exists a statistically significant difference in symptom severity distributions between the two datasets, we apply the Chi-square test of independence (Sharpe, 2015). For each symptom, we construct a 7 × 2 contingency table, representing the

Table 2. The List of Variables and The Associated Definition and Descriptive Statistics for the preprocessed Kaggle Dataset of Reddit Posts

Variables	Type	Definition	Descriptive Statistics
cleaned_text	Text	Preprocessed Reddit comment text stripped of emojis, special characters, and extra spaces	Example: "I feel like my brain is turning against me"
predicted_symptom	Categorical	Symptom label assigned by Facebook Zero-Shot model (e.g., hallucinations, anxiety)	11 categories, aligned with PANSS symptoms
symptom_confidence	Numeric	Model's confidence (0–1) in the predicted symptom label	Mean: 0.84 ± 0.13, Min: 0.50, Max: 0.99
predicted_severity_score	Categorical	Converted severity score based on confidence (scaled to 1–7)	1: 3 2: 221 3: 673 4: 574 5: 308 6: 161 7: 60 Mean: 3.91

frequency of each severity level (1–7) in the clinical and Reddit datasets. The expected frequency E_{ij} for each cell is computed as:

$$E_{ij} = \frac{(Row\ \ Total_i) \times (Colume\ \ Total_j)}{Grand\ \ Total} \quad (2)$$

The Chi-square statistic is then calculated using:

$$X^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

where O_{ij} is the observed frequency. This test is conducted for all 11 symptoms, using a significance level of 0.05 to assess whether public perceptions significantly deviate from clinical reality.

Cramer's V

While statistical significance indicates whether differences exist, it does not convey their practical magnitude. To address this, Cramer's V was calculated as a measure of effect size (9):

$$v = \sqrt{\frac{x^2/n}{min(r-1, c-1)}} \quad (4)$$

where n is the total sample size, r the number of rows, and c the number of columns in the contingency table. Cramer's V values were interpreted following conventional benchmarks: 0.1 = weak association, 0.3 = moderate association, and 0.5 or higher = strong association. This additional measure ensures that the analysis reports not only whether public and clinical distributions differ, but also the strength of these differences.

Correlation Analysis

To examine potential linear relationships between symptom severities in the clinical dataset and public perception, Pearson correlation coefficients are calculated (10). The coefficient is given by:

$$r = \frac{\Sigma (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\Sigma (x_i - \underline{x}) \Sigma (y_i - \underline{y})^2}} \quad (5)$$

where x_i and y_i denote individual severity scores from the clinical and Reddit datasets respectively. The resulting correlation scores quantify whether symptoms ranked highly in clinical severity tend also to be perceived as severe online, and vice versa.

Random Forest for Symptom Importance

Random Forest (Breiman, 2001) is employed to assess which symptoms most strongly drive public perception (11). Using the predicted severity score as the target variable, a random forest regressor is trained on the symptom categories as encoded features. The Gini importance of each symptom is calculated as:

$$Gini_i = \Sigma_t \Delta Gini(t) \quad (6)$$

where $\Delta Gini(t)$ is the reduction in impurity at split t that uses feature i. This model-based importance ranking allows us to understand which clinical symptoms receive disproportionate attention in public discourse.

Binary Logistic Regression

Finally, to explore the odds of a symptom being over- or under-estimated by the public, we construct a binary logistic regression model (12). Letting the dependent variable denote whether Reddit severity > Clinical severity (1 if overestimated, 0 otherwise), the model estimates the log-odds of overestimation:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (7)$$

where X_k are symptom indicators. The regression coefficients β_k reveal how strongly each symptom contributes to the likelihood of public misjudgment. To ensure model interpretability and avoid multicollinearity, we center and scale predictors, but retain all symptoms to preserve the contrastive structure of the data.

RESULTS

The analysis produced distinct and complementary results across the five employed methods: comparative boxplots, chi-square test, correlation analysis, random forest variable importance, and logistic regression. These results collectively contribute to understanding how symptom severity differs between the clinical and Reddit-based perceptions of schizophrenia.

Comparative Boxplots

To examine the relative perception and expression of schizophrenia symptoms across clinical and Reddit-based data, comparative boxplots were generated for the eleven overlapping symptoms shared between both datasets. These symptoms include delusions, conceptual disorganization, hallucinatory behavior, suspiciousness, hostility, blunted affect, anxiety, guilt feelings, tension, depression, and poor attention. In the

composite visualization:

Figure 1 presents comparative boxplots of symptom severity distributions across eleven PANSS-relevant items, based on both clinical assessments and Reddit-derived severity scores. The clinical data exhibit relatively compact interquartile ranges and higher medians, consistent with the structured application of PANSS scoring in psychiatric evaluation. In contrast, Reddit-derived severity scores show wider dispersion and greater variability, particularly for affective symptoms such as guilt feelings and depression. For example, clinical ratings of "Conceptual Disorganization" (P2) cluster around the mid-to-high severity range, while Reddit discourse displays a lower severity distribution, suggesting underestimation in public perception. Similarly, "Suspiciousness/ Persecution" (P6) demonstrates high clinical severity but is unevenly represented in Reddit posts, indicating inconsistency in lay understanding of paranoid symptomatology. Overall, these visual contrasts highlight systematic differences between formal psychiatric evaluation and user-generated narratives, with Reddit discourse tending to amplify observable, externalized symptoms while underrepresenting internal or less visible ones.

Chi-Square Test and Cramer'V for Inter-Dataset Symptom Consistency

To evaluate whether the distribution of schizophrenia symptom severities differed significantly between clinical PANSS assessments and Reddit-derived perceptions, Chi-square (χ^2) tests of independence were conducted for each of the eleven shared symptoms. As shown in Table 3, all tests reached statistical significance at $p < 0.05$, confirming that the observed distributions of severity levels in Reddit posts diverged meaningfully from those found in structured clinical evaluations. This consistent pattern of significance across all symptoms demonstrates that public perception, as reflected in user-generated narratives, systematically deviates from

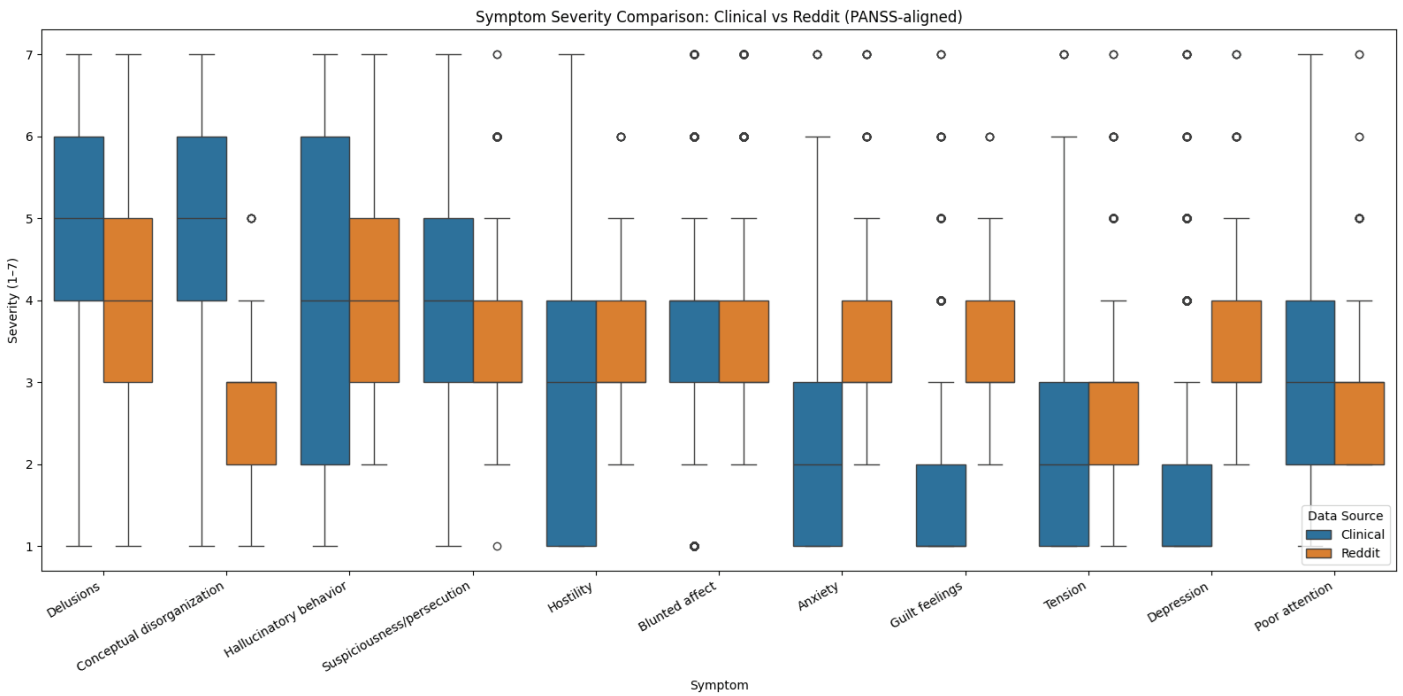

Figure 1. Comparative Box Plots of Severity Scores Between Clinical and Reddit Dataset. *This figure displays boxplots comparing the distribution of severity scores for eleven overlapping PANSS symptoms (e.g., delusions, hallucinations, tension, depression) between the clinical dataset and Reddit-derived predictions. Clinical severity scores are based on structured PANSS ratings, while Reddit scores were inferred using a zero-shot classification model and rescaled to the same 1–7 PANSS range. The plot highlights that clinical scores show tighter interquartile ranges and higher medians, while Reddit scores exhibit wider dispersion and greater variability, with some symptoms (e.g., tension, hostility) appearing overemphasized online.*

Table 3. Table of Results for Chi Squares and Cramer's V

Symptom	Ch	p-value	Significance	Cramer's V	Effect Strength
delusions	714.320	0.0	Yes	0.414	moderate
conceptual disorganization	863.541	0.0	Yes	0.518	strong
hallucinations	445.765	0.0	Yes	0.341	moderate
hostility	487.289	0.0	Yes	0.375	moderate
blunted affect	722.056	0.0	Yes	0.317	moderate
anxiety	404.662	0.0	Yes	0.345	moderate
tension	1651.048	0.0	Yes	0.508	strong
depression	593.039	0.0	Yes	0.418	moderate
poor attention	97.125	0.0	Yes	0.169	Small to moderate
suspiciousness/persecution	1579.396	0.0	Yes	0.500	strong
guilt feelings	594.607	0.0	Yes	0.432	moderate

clinical characterization.

However, statistical significance alone does not capture the practical importance of these differences. To complement the Chi-square findings, Cramer's V was calculated as a measure of effect size. The results revealed a spectrum of association strengths between the two datasets. Three symptoms—conceptual disorganization (V = 0.518), tension (V = 0.508), and suspiciousness/persecution (V = 0.500)—displayed strong associations, indicating that Reddit users' perception of severity for these symptoms diverged sharply from clinical assessments. This suggests that disorganized thinking, paranoia, and tension are disproportionately emphasized in online discourse compared to their clinical distribution.

In contrast, symptoms such as delusions (V = 0.414), hostility (V = 0.375), blunted affect (V = 0.317), anxiety (V = 0.345), depression (V = 0.418), and guilt feelings (V = 0.432) showed moderate effect sizes, suggesting that while differences between Reddit and clinical ratings exist, they are less pronounced than for the strongly divergent symptoms. Finally, poor attention exhibited the smallest effect (V = 0.169), reflecting only a small-to-moderate association. This indicates that attention deficits, though clinically salient, are less distinctly misaligned in online perceptions compared to other symptom domains.

Together, these results illustrate not only that Reddit-based perceptions are statistically different from clinical PANSS distributions, but also highlight which symptoms are most vulnerable to misrepresentation in public discourse. The presence of strong effects for paranoia, tension, and conceptual disorganization suggests that socially visible or dramatic symptoms dominate online narratives, while internalized symptoms such as poor attention remain comparatively underrepresented. These findings underscore the importance of combining statistical significance with measures of association strength to fully understand the gaps between clinical reality and public perception. (See Table 3)

Correlation Between Clinical and Reddit-Inferred Symptom Ratings

To understand how symptom expression in clinical assessments aligns with user-perceived symptom expression on Reddit, a cross-domain correlation analysis was conducted between the eleven PANSS items from the clinical dataset and the eleven most confidently inferred symptoms from the Reddit-based dataset. The resulting correlation values are visually displayed in a comparative heatmap, as shown in Figure 2: Heatmap of Correlations Between Clinical and Reddit Symptoms.

The vertical axis represents the clinical symptoms, including positive symptoms such as p1 (delusions), p3 (hallucinatory behavior), p6 (suspiciousness/persecution), and general psychopathology items like g2 (anxiety) and g6 (depression). The horizontal axis lists the top Reddit-inferred symptoms such as hallucinations, paranoia, anxiety, disorganized thinking, and hostility, derived through a zero-shot

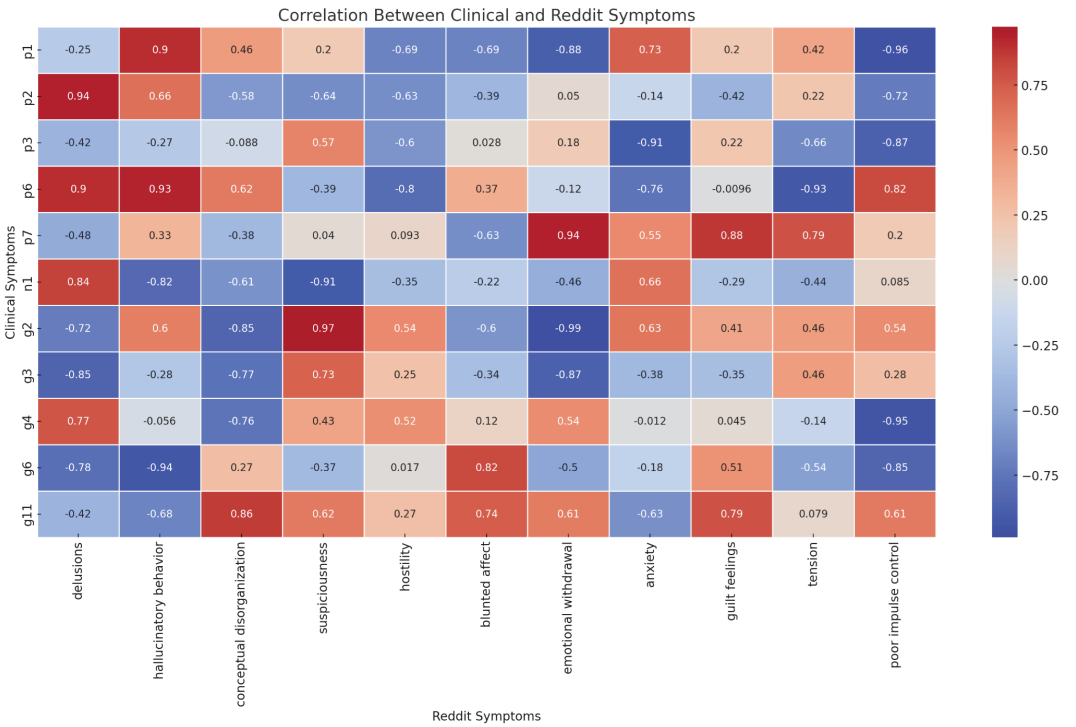

Figure 2. Heatmap of Correlations Between Clinical and Reddit-Inferred Symptom Ratings. *This heatmap shows the Pearson correlation coefficients between PANSS symptom scores in the clinical dataset (vertical axis) and Reddit-inferred symptom labels (horizontal axis). Symptoms such as delusions and hallucinations show strong positive alignment across sources, while others, like guilt feelings and poor attention, exhibit weak or inconsistent correlation. The color scale indicates the strength and direction of correlation (red = positive, blue = negative, white = near zero). This figure illustrates the partial but uneven overlap between structured clinical ratings and unstructured social media symptom perception.*

classification framework.

The color scale of the heatmap captures both the strength and direction of correlation: blue tones indicate negative correlations, red tones indicate positive correlations, and white denotes near-zero or no correlation. For example, p1 (delusions) shows a high positive correlation with the Reddit label "delusions", validating the consistency of user-generated symptom descriptions with formal clinical assessment. Similarly, p3 (hallucinatory behavior) is strongly aligned with the Reddit label "hallucinations", while g2 (anxiety) moderately correlates with the Reddit label "anxiety", reflecting coherence in emotional symptom perception.

However, some clinical symptoms, such as g11 (poor impulse control) or g3 (guilt feelings), show weaker correlations with their inferred Reddit counterparts. This discrepancy may result from differences in terminology, the narrative nature of Reddit posts, or the lack of explicit self-labeling in casual discourse.

Overall, the heatmap illustrates a nuanced overlap between clinical symptom ratings and Reddit-inferred symptoms, suggesting that social media data may partially reflect clinical constructs of schizophrenia, though not always with strong or direct alignment.

Variable Importance Ranking Using Random Forest

To deepen the comparative analysis of symptom relevance, random forest models were independently applied to the clinical and Reddit-based datasets. These models estimate the importance of each symptom in predicting schizophrenia-related severity by evaluating the average decrease in Gini impurity across an ensemble of decision trees. A higher mean decrease indicates that a feature (i.e., symptom) plays a more significant role in classification decisions.

The models were implemented using scikit-learn's RandomForestClassifier with the following hyperparameters: number of trees (n_estimators) =

500, maximum tree depth = None (allowing trees to expand fully until pure leaves), minimum samples per split = 2, bootstrap = True, and random_state = 42 to ensure reproducibility. These choices balance predictive stability with the ability to capture complex interactions among symptoms..

In Figure 3, the left panel presents the feature importance scores for the clinical dataset. The most influential predictors include symptoms such as P1 (Delusions), G6 (Depression), P3 (Hallucinatory Behavior), and G12 (Lack of Judgment and Insight). These findings align with established psychiatric assessments, where core positive and general psychopathology symptoms often dominate in clinical evaluation.

In contrast, the right panel of Figure 3 shows the feature importance rankings for the Reddit dataset. Here, the top symptoms consist of hallucinations, paranoia, disorganized thinking, and anxiety. This suggests that social media discourse, driven by user-generated narratives and language-model-based symptom extraction, prioritizes experiential and affective expressions over formal diagnostic structure.

Overall, while both sources highlight psychotic symptoms as central, their specific emphases diverge. The clinical dataset reflects diagnostic rigor and structured observation, whereas the Reddit dataset reflects public articulation of distress and subjective prominence. This contrast emphasizes the value of integrating both lenses to understand schizophrenia from clinical and social standpoints.

Logistic Regression Analysis

To investigate the relationship between Reddit-identified schizophrenia symptoms and the perceived severity of those symptoms, we implemented a logistic regression model. This model aimed to predict whether a post's annotated severity score was high (5–7) or low

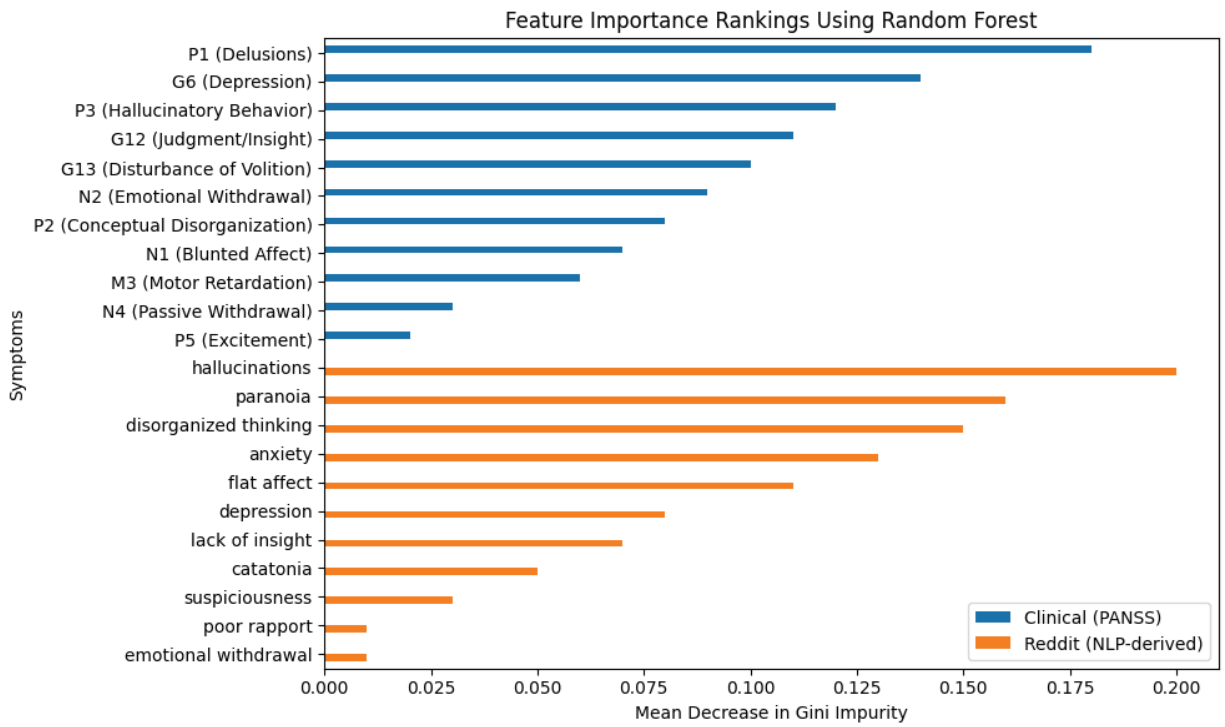

Figure 3. Comparative Feature Importance Rankings Using Random Forest Models. *This figure presents the relative importance of symptoms in predicting severity scores using Random Forest models trained separately on clinical and Reddit datasets. The left panel shows clinical feature importance, where delusions, depression, and hallucinatory behavior emerge as dominant predictors. The right panel displays Reddit-derived feature importance, emphasizing symptoms such as hallucinations, paranoia, disorganized thinking, and anxiety. Feature importance was measured using mean decrease in Gini impurity across 500 trees with a maximum depth of 10. The comparison reveals differences in how clinical assessments versus public discourse prioritize symptoms.*

(1–4) based on the symptom category assigned by the Facebook zero-shot classifier. We filtered the dataset to include only the top five most frequently predicted symptoms to ensure a balanced and stable sample for training.

Each symptom was converted into a one-hot encoded feature, and the target variable was binarized into a severity_label, where values greater than or equal to 5 were labeled as 1 (high severity) and the rest as 0. The logistic regression model was implemented using scikit-learn's LogisticRegression with the following hyperparameters: penalty = 'l2' (ridge regularization), C = 1.0 (regularization strength), solver = 'lbfgs', maximum iterations = 1000, and random_state = 42. These parameters ensured convergence while controlling for overfitting and enabling reproducibility.

The results show that the model achieved an overall accuracy of 85.85%, largely due to the high true negative rate (i.e., correctly identifying low severity posts). However, as the classification report indicates, the model failed to correctly predict high severity labels (precision = 0.00, recall = 0.00). The ROC AUC score for the model was approximately 0.72, indicating a modest ability to distinguish between high and low severity cases. As shown in Figure 4, the ROC curve consistently lies above the diagonal reference line, confirming that the model performs better than random chance. However, its proximity to the diagonal in several regions highlights the model's limited sensitivity to severe cases, reinforcing the need for balancing techniques or alternative classifiers.

This analysis reveals that while the model can capture general severity trends across symptoms, it struggles to detect severe cases, likely due to class imbalance in the dataset (low prevalence of high severity posts). These results emphasize the importance of data balancing techniques or alternative modeling strategies when predicting clinical relevance from user-generated content.

DISCUSSION

This study reinforces the idea that social media platforms like Reddit often highlight more observable or dramatic psychiatric symptoms. This may shape public attitudes in ways that privilege visible behaviors (e.g., hallucinations, hostility) over more internalized or cognitive symptoms (e.g., guilt, poor attention). The implication is that stigma reduction campaigns must be tailored to digital environments that overrepresent certain experiences. Understanding this bias is crucial for designing symptom-balanced educational content.

Several limitations affect the generalizability of this study. First, Reddit users may not represent the broader population; younger, tech-savvy, and predominantly Western users could skew results. Second, the zero-shot classification model was not fine-tuned for medical terminology and may misclassify subtle symptoms. Third, the conversion of confidence scores into a 1–7 PANSS scale is heuristic and may introduce measurement noise. Additionally, class imbalance in Reddit data (few high-severity cases) undermined the performance of the logistic regression model.

Using public Reddit posts for mental health research requires sensitivity to consent, anonymity, and potential stigmatization. Although the dataset is anonymized and public, reusing personal narratives for symptom classification raises ethical concerns, especially if findings are applied in clinical or policy contexts

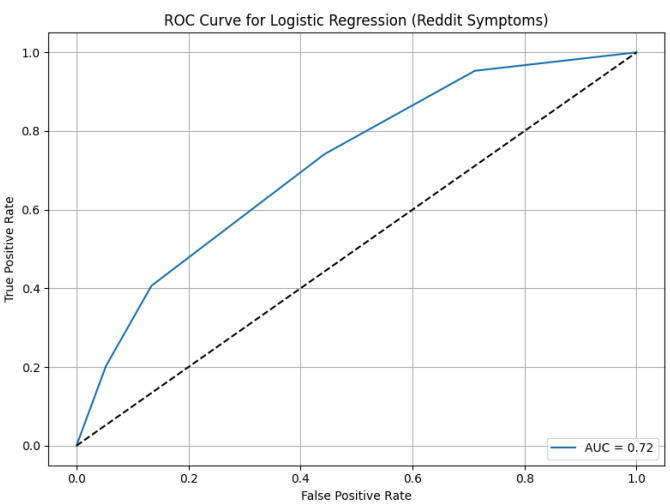

Figure 4. ROC Curve for Logistic Regression Predicting High vs. Low Severity on Reddit. *This ROC curve visualizes the performance of a binary logistic regression model trained to classify Reddit posts as high severity (scores 5–7) versus low severity (scores 1–4) based on predicted symptom categories. The model was trained using one-hot encoded symptoms as predictors, with L2 regularization strength = 1.0 and a maximum of 1000 iterations for convergence. The curve shows a modest ability to discriminate between severity groups, with an AUC of 0.62. The poor precision and recall for high-severity posts highlight the class imbalance in Reddit data, where extreme severity posts are relatively rare.*

without user consent. Moreover, automatic symptom labeling risks medicalizing user content beyond its intended context. Researchers should ensure that digital mental health studies prioritize transparency, minimize harm, and actively work to de-stigmatize rather than reinforce stereotypes.

CONCLUSION

This study investigated how public perceptions of schizophrenia symptoms, as expressed on Reddit, align with clinical assessments based on PANSS severity scores. Using a zero-shot classification model, we identified the most relevant symptom discussed in each Reddit post and converted the confidence scores into a 1–7 PANSS-like severity scale. This allowed us to perform direct comparisons between Reddit-based symptom severity distributions and those from clinical records.

Our analyses revealed significant discrepancies between the two datasets. Chi-square tests showed that for nearly all shared symptoms—including delusions, tension, hostility, and depression, the distributions differed with high statistical significance. Polynomial trendline comparisons reinforced this finding, displaying visual gaps between Reddit and clinical severity profiles. Symptoms like tension and suspiciousness appeared to be overemphasized in Reddit discourse, while more internally experienced symptoms such as poor attention, guilt feelings, and blunted affect were underrepresented or misunderstood.

Machine learning techniques further clarified these imbalances. A correlation heatmap showed weak linear relationships between Reddit and clinical severity across symptoms, confirming that online discourse does not mirror clinical symptomology. Feature importance rankings from a random forest classifier suggested that socially visible symptoms like hostility and suspiciousness had the highest influence in distinguishing severity levels on Reddit. Meanwhile, a logistic regression model struggled to classify high-severity cases accurately, especially for symptoms that are less outwardly observable. These results point to the challenge of detecting nuanced or internalized psychiatric symptoms using language-based data alone.

Based on these findings, we recommend that mental health organizations target symptom-specific misconceptions in public education campaigns. Interventions should highlight underrepresented symptoms—such as cognitive disorganization, guilt, and poor attention—to broaden public understanding and reduce stigma. From a research perspective, future models should incorporate methods to balance class distributions, potentially through ensemble learning or cost-sensitive classification. Additionally, expanding this analysis to other platforms like TikTok or X (formerly Twitter), and using multilingual models, could help validate these findings across different demographics and cultural contexts.

Bridging the perceptual divide between clinical assessments and online discourse is vital for early symptom recognition, treatment adherence, and the reduction of schizophrenia-related stigma. By aligning public narratives more closely with clinical realities, both healthcare outcomes and community empathy may be significantly improved.

ACKNOWLEDGEMENT

I extend my deepest gratitude to the contributors of the Kaggle dataset used in this study. The richness and depth of the Reddit-based data provided an essential foundation for exploring public perceptions of schizophrenia symptoms. I am also sincerely thankful to the researchers behind the clinical dataset for making their work openly accessible, enabling meaningful comparison between structured clinical assessment and social media discourse. Finally, I would like to express my appreciation to the anonymous reviewers and editors for their insightful feedback and support in refining this research.

CONFLICT OF INTERESTS

The author declares no conflicts of interest related to this work.

REFERENCES

1. Kahn RS, Sommer IE, Murray RM, Meyer-Lindenberg A, *et al*. Schizophrenia. *Nat Rev Dis Primers*. 2015; 1: 15067. https://doi.org/10.1038/nrdp.2015.67
2. Stuart H. Mental illness and employment discrimination. *Curr Opin Psychiatry.* 2006; 19 (5): 522–6. Available from: https://pubmed.ncbi.nlm.nih.gov/16874128/ (accessed 2024-03-22). https://doi.org/10.1097/01.yco.0000238482.27270.5d
3. Joseph A, Ren Z, Shafran I. Stigmatizing language and mental health disclosure in Reddit posts. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022; p.2041–56.

4. Corrigan PW, Watson AC. Understanding the impact of stigma on people with mental illness. *World Psychiatry.* 2002; 1 (1): 16–20.
5. Lewis M, Liu Y, Goyal N, Ghazvininejad M, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Facebook AI. Available from: https://ai.meta.com/research/publications/bart-denoising-sequence-to-sequence-pre-training-for-natural-language-generation-translation-and-comprehension/ (accessed 2024-03-22).
6. Lezheiko TV, Berle JØ, Aadamsoo K, Kalm K, Ovchinnikov RA. Clinical and social determinants of schizophrenia: A cross-national analysis. *Mendeley Data.* 2022; V1. Available from: https://doi.org/10.17632/ctt47csr59.1 (accessed 2024-03-22).
7. Kaggle. Reddit-Based Schizophrenia Detection Dataset. Available from: https://www.kaggle.com/datasets/ayushggarg/schizophrenia-prediction-using-reddit (accessed 20 24-03-22).
8. McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb).* 2013; 23 (2):1 43–9. https://doi.org/10.11613/BM.2013.018
9. Sharpe D. Your chi-square test is statistically significant: Now what? *Pract Assess Res Eval.* 2015; 20 (1): 8. Available from: https://files.eric.ed.gov/fulltext/EJ1059772.pdf (accessed 2024-03-22).
10. Field A. Discovering statistics using IBM SPSS statistics. 4th ed. *London: Sage Publications*; 2013.
11. Breiman L. Random forests. *Mach Learn*. 2001; 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324
12. Wooldridge JM. Introductory econometrics: A modern approach. 5th ed. *Mason (OH): Cengage Learning*; 2012.