PreeLinguistic Patterns of Suicide Disclosure on Social Media: A time series clustering approach

David Gomez[1] and Munmun De Choudhury[1]

[1]School of Interactive Computing, Georgia Institute of Technology

June 26, 2024

Abstract

This work contributes to our understanding of broadcasting self-disclosures on social media—specifically surrounding the highly stigmatized topic of suicidality. In particular, we assess (1) whether there are any psycholinguistic patterns post-disclosure, (2) if they reflect therapeutic benefits, and (3) if we can preempt those who would benefit from such disclosures. We analyze public Twitter data of (N=1060) users who have disclosed some form of suicidality. We use Linguistic Inquiry and Word Count (LIWC) along with timeseries clustering to identify temporal-psycholinguistic patterns post-disclosure. We identify two clusters that are differentiated by their use of `filler` words. The majority group (73% of users) appears to experience therapeutic benefit in the form of significantly lower usage of filler words (i.e., higher coherence) than the other group post-disclosure. We then develop a range of machine learning and deep learning classifiers that utilize only pre-disclosure information to predict whether a user would benefit from such disclosures. We achieve modest but positive results, with our best model achieving an AUC score of 0.66 over a baseline of 0.50 and a macro F1 score of 0.64 over a baseline of 0.50—indicating that there is some predictive information in the language pre-disclosure that can preempt whether someone would receive therapeutic benefit from broadcasting self-disclosures. We discuss the implications of our findings for designing new intervention strategies that can improve support provisions for those who disclose suicidality on social media.

1 Introduction

Suicidality continues to represent a complex and significant public health concern in the United States. Suicide rates have steadily increased over the last two decades—approximately 2% each year from 2000-2010 and 3-4% from 2010-2020— except for 2020, which saw a slight decrease which many attribute to the COVID-19 pandemic. In 2021, an estimated 12.3 million American adults seriously thought about suicide, 3.5 million planned a suicide attempt, 1.7 million attempted suicide, and 48,183 were successful [CDC 2021]. Provisional data for the year 2022 show that 49,449 people died by suicide—pending confirmation, this would be the highest figure ever recorded. Sadly, these figures do not capture the physical or psychological pain that led these individuals to take their own lives.

Researchers and practitioners combating suicide do so on several fronts. With the widespread adoption of social media and the advances made in the fields of machine and deep learning, one rich venue for research is the interaction between suicidality, or more broadly mental well-being, and social media. On the one hand, social media data can be used as a lens through which to understand one's mental state [26]. For example, psychological studies have shown that our state of mind can manifest itself in the linguistic features we use to communicate [13, 22], and several important studies have taken this perspective [10, 14, 16, 30]. On the other hand, it has been well-established that social media can affect our mental well-being [4, 21, 23, 31]. While these studies primarily focus on the adversarial effects of social media on mental health, there is also evidence that it can be a source of emotional and social support [1–3]—including for those suffering from suicidality [15].

Outside the context of social media, it is understood that self-disclosure, a process of "making the self known to others" [12], is an important therapeutic element in the achievement of physical and mental well-being [19], as it is a widely adopted mechanism for coping. More recently, research has found evidence that these therapeutic benefits of mental health disclosures *may also extend to social media context* [15, 17]. Important to our study, prior work has established that broadcasting self-disclosures of highly stigmatized mental disorders (e.g., schizophrenia) shows evidence of therapeutic benefit. In [17], Ernala et al. examined 146 clinician-verified disclosures of schizophrenia on Twitter and found evidence of therapeutic benefit across several metrics including improved readability

and coherence in language, future orientation, lower self-preoccupation, and reduced discussion of symptoms and stigma perceptions. Our work fits within this context and contributes toward our understanding of broadcasting self-disclosures on social media—specifically surrounding the highly stigmatized topic of suicide.

In this paper, we assess whether any discernible psycholinguistic patterns differentiate those who self-disclose their suicidality on social media and whether these patterns reflect therapeutic benefits. Furthermore, we assess whether we can preemptively identify those who would benefit from such disclosures. More specifically, we address the following questions: *(a) Are there any psycholinguistic patterns that differentiate those who self-disclose their suicidality on social media? (b) If so, do they show evidence of therapeutic benefits to such disclosures? (c) Is it possible to anticipate who and who would not receive such a benefit?*

To address these questions, we examine publicly shared Twitter posts from users who have disclosed some form of suicidality. We use the widely adopted and vetted linguistic lexicon Linguistic Inquiry and Word Count (LIWC) to extract temporal psycholinguistic signals post-disclosure, we conduct a timeseries clustering analysis to identify distinct response behaviors of those who publicly disclose their suicidality. We find that a majority of users maintain lower usage of filler words, which may be interpreted as higher coherence, in their language post disclosure—providing further evidence of the therapeutic benefits to self-disclosures of suicidality. Additionally, we developed classifiers that utilize only pre-disclosure language to preempt whether a user *would benefit* from such disclosures. These classifiers achieve modest but positive results indicating that pre-disclosure language alone contains some predictive signals that can preempt whether someone would benefit from disclosing suicidality on social media—however, we note that further study is warranted to determine other possible predictors of therapeutic benefit.

These findings have important implications for designing new intervention strategies that can improve support provisions for those who disclose suicidality on social media. For example, our findings suggest that it may be possible to preemptively identify those who would benefit from disclosing their suicidality on social media and provide them with additional opportunities for richer forms of expression. Similarly, those who are predicted not to benefit from such disclosures could be provided with alternative intervention strategies and/or support provisions.

The remainder of this article is as follows. In section 2, we describe our data collection process

and the methods used to address our research questions. In section 3, we present our results. In section 4, we discuss the implications of our findings and their limitations. Finally, in section 5, we conclude with a summary of our findings and directions for future study.

2 Data and Methods

2.1 Data

We utilized public Twitter data via the Twitter API v2 for this research[1]. We take great care to protect the privacy of the users; we do not present direct quotes from any data, or any identifying information.

Our data collection proceeded in two phases. First, we identified a set of suicidal tweets via a case-insensitive query for tweets that contained some form of suicidality. Building on prior literature, we collected tweets for 38 suicide-relevant keywords [6, 9]; the complete list is reproduced in Table 1. Observe that these keywords capture a wide range of suicidality. Some phrases can capture those explicitly considering harming themselves (e.g., stab, shoot, hang myself), others capture the desire for death (e.g., sleep forever, never wake up, asleep and never wake), or the lack of desire to live (e.g., don't want to exist, don't want to go on, don't want to wake up, not want to be alive). These keywords even capture more subtle forms of suicidality, like the miscalculation of perceived burdensomeness [34] (e.g., better off dead, my death would). Also, note that these keywords were originally used to query tweets from the 2015-2017 time frame, thus we adopt the same time frame for our data collection process, as expressions of suicidality may have changed.

From this initial query of suicidal tweets, we manually filtered benign or overtly non-suicidal tweets. We decided to include suicide "jokes" in our dataset (e.g., brb killing myself). From a psychological perspective, humor plays an interesting role in suicidal disclosures. By framing one's suicide self-disclosure as a joke, the discloser gets to express themselves without (or with less risk of) social stigma or rejection [X]. These phrases were thus included in the analysis (see Table 1). We then collected the timelines of each user six months before and after the date of disclosure. We removed users with multiple suicide disclosures. This means that for those users in our dataset whose disclosures resemble "jokes," these jokes only occurred once in a year—perhaps adding more

[1]The data collection for this project was carried out before Twitter's acquisition by X Inc.

significance to their occurrence. We also removed those users with excessively few (not conducive to statistical analyses) and frequent (likely bots) tweets, and those whose primary language was not English. This resulted in a final dataset of 1.4M tweets from N=1060 users. Table 1 tabulates the keywords used to collect tweets along with examples of included/excluded self-disclosure tweets. Table 2 provides summary statistics for our dataset. Figure 1 depicts the top 10 keywords with the highest frequencies within the set of disclosure tweets. Note that some disclosure tweets contained multiple keywords.

Suicide-Relevant Keywords
hang myself, stab myself, drug myself, ready to die, take my life, shoot myself, end this pain, ending it all, stop the pain, never wake up, sleep forever, poison myself, killing myself, to hurt myself, my death would, want to end it, cutting myself, die in my sleep, to live anymore, want to be gone, take it anymore, better off dead, tired of living, take my own life, not worth living, feeling hopeless, dont want to live, isnt worth living, dont want to exist, dont want to go on, want it to be over, my life isnt worth, put an end to this, nothing to live for, my life is pointless, dont want to wake up, not want to be alive, asleep and never wake

Examples of Included Tweets
I'm going to **take my life** because I'm fucking over this shit.
I really wish I could go to **sleep and never wake up**.
brb **killing myself**.

Examples of Excluded Tweets
This movie is so good I don't **want it to end**.
I **don't want to wake up** at 4am tomorrow.

Table 1: Suicide-relevant keywords and phrases used for Twitter data collection. (Source [6] citing [9])

Number of Users	1,060
Number of Tweets	1,399,148
Mean Number of Tweets per User	1,320
Median Number of Tweets per User	583

Table 2: Summary statistics for suicide-disclosure dataset.

2.2 Methods

Our analyses proceeded in two phases. In Phase I, we explore temporal-psycholinguistic clusters *post-disclosure* within the set of Twitter users who have disclosed suicidality. We convert tweets

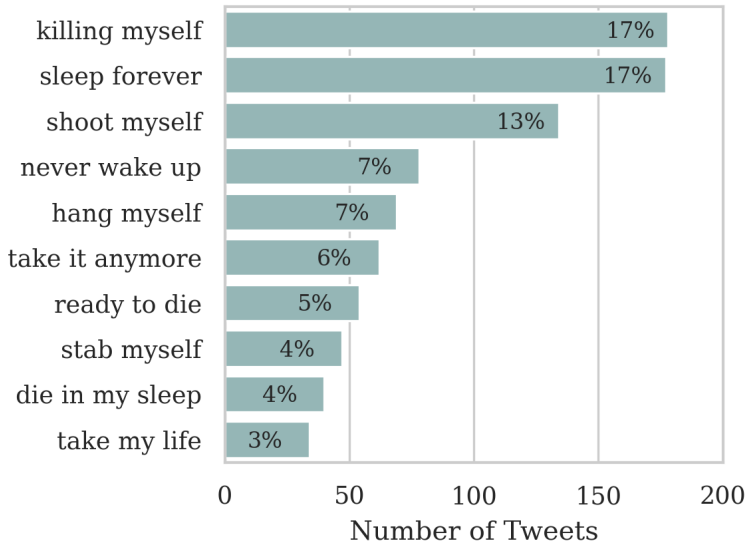


Figure 1: Top 10 suicidal keywords within the set of suicidal disclosures. (N=1060) *I'd like to include a sample of tweets for the top 2-3 keywords exploded out to the side.*

to timeseries via LIWC, then leverage univariate and multivariate timeseries clustering algorithms for this purpose. This phase results in each user being assigned a label according to the cluster it was assigned. In Phase II, we use these labels in a subsequent classification task, whereby we use only *pre-disclosure* data to predict to which cluster a user belongs—i.e., to predict how a user *would respond* to disclosing suicidality. These phases can be summarized as the timeseries clustering and textual classification phases, respectively. In the following subsections, we discuss the methodology for each.

2.3 Phase I. Timeseries Clustering

To construct psycholinguistic timeseries for each user, we use the Linguistic Inquiry and Word Count (LIWC) version 2015 [25]. This is a widely used text analysis software that counts the percentage of words in a given text that fall into a set of predefined categories or dimensions. These categories capture various aspects of language including linguistic/grammatical dimensions such as function words, pronouns, articles, verbs etc. as well as psychological aspects such as affective, cognitive, and social processes. In version 2015, there are 74 such categories—a complete description of which can be found in [25]. As with prior versions, these categories were developed hierarchically with some dimensions subsumed under others. For example, the category `anger` is a subcategory of the

broader category `negative emotion` which is a subcategory of the broader category `affect`. To more specifically connect LIWC categories to psychological constructs, we considered only the 61 categories with no subcategories. *I would like to explore higher-level categories as well. Perhaps these would better capture 'multidimensional' constructs.*

We carefully construct the timeseries as follows. For each tweet in a user's timeline (+/- 6 months from disclosure), we compute the LIWC scores for each category. (For example, the tweet "brb killing myself" would receive a score of 0.33 for `death`, 0.33 for `first-person-pronoun`, and scores of 0 for all remaining categories. [*VERIFY*]) Then, we aggregate scores by day. If there were multiple tweets in a given day, we take the average of the scores. For those days with no tweets, we impute the previous day's scores (i.e., forward-fill missing values). We smooth the data using LOESS smoothing [8] with a 7-day parameter (to account for weekly patterns) and normalize the entire timeseries such that each has a zero-mean and unitary standard deviation. Then we crop the timeseries from the day of disclosure to w weeks post-disclosure, where w is a hyperparameter. We consider four timeseries lengths after disclosure: 1, 2, 3, and 4 weeks, corresponding to 7, 14, 21, and 28 days, respectively. Thus, we have a 61-dimensional timeseries for each user with one data point per day with lengths of 7, 14, 21, and 28 days post-disclosure. *There is some information leak via smoothing—I think you acknowledged this. Also, I think we should reconsider normalizing AD data with BD data. We originally settled on this being an information leak, but on the contrary, it's really 'factoring out' BD data from AD data—decoupling the two even more.*

We then perform timeseries clustering on these timeseries. We use the `tslearn` Python package [33] to perform the timeseries clustering, and we use the `kmeans` algorithm with Dynamic Time Warping (DTW) distance measure, which allows for some amount of temporal distortion between the two sequences [28]. Note that for the `kmeans` clustering algorithm, we need to specify the number of clusters k as a hyperparameter. We consider 3 values for k: 2, 3, and 4. To evaluate the quality of the clustering solutions, we used the `silhouette` score [29]. We discuss this metric in more detail in the next section.

2.3.1 Silhouette Score

The silhouette score is a metric used to evaluate the quality of a clustering solution; it measures how similar an object is to its own cluster compared to other clusters [29]. It ranges from -1

to 1, with higher values indicating that the object is well matched to its own cluster and poorly matched to neighboring clusters We also considered other clustering evaluation metrics–namely, the Calinski-Harabasz [5] and the Davies-Bouldin [35] indices—but we choose the silhouette score as it is bounded and thus more easily interpretable. Note that a silhouette score of 0 indicates that the object is on the boundary between two clusters, and negative values indicate that the object may be assigned to the wrong cluster. Mathematically, the silhouette score is defined as follows. Let a_i be the average distance between i and all other objects in the same cluster $\mathcal{C}_I$, and let b_i be the average distance between i and all objects in the nearest cluster $\mathcal{C}_J$. Then, for a given object i, the silhouette score is defined as follows.

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{1}$$

with

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, j \neq i} d(i,j) \tag{2}$$

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \tag{3}$$

$$\tag{4}$$

where $d(i,j)$ is the some distance metric between objects i and j. In the context of timeseries clustering, $d(i,j)$ is the DTW distance measure. The silhouette score is illustrated graphically in Figure 2 with a notional Euclidean dataset.

Note that the silhouette score is computed for each object in the sample. Thus, we can average the silhouette scores by cluster (5) or take a global average (6) to get a single score for the entire clustering solution.

$$\bar{s}_{C_I} = \frac{1}{|C_I|} \sum_{i \in C_I} s_i \tag{5}$$

$$\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{6}$$

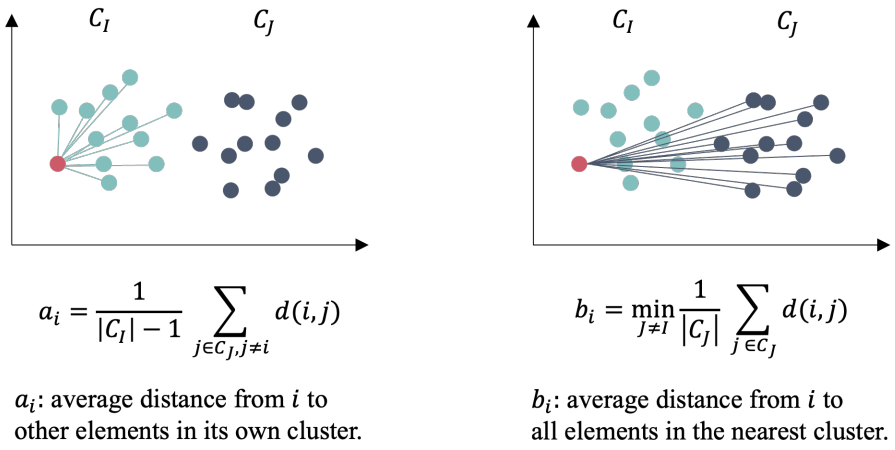


Figure 2: Illustration of the silhouette score components. a_i denotes the average distance between i and all other objects in the same cluster $\mathcal{C}_I$. b_i denotes the average distance between i and all objects in the nearest cluster $\mathcal{C}_J$.

2.3.2 Feature Subset Selection

In the context of multivariate clustering, we recognize that some dimensions may be more conducive to clustering than others. Consider again the notional Euclidean dataset now depicted in Figure 3. We see that the data is separable along the horizontal axis but not along the vertical axis. Thus, the vertical dimension does not contribute to the clustering solution. As the dimensionality of the problem increases, each dimension that does not contribute to the clustering solution can only add sparsity to the dataset, which can degrade the quality of the clustering solution. Given our modest sample size (N=1060) and relatively high dimensionality ($d = 61$), we need a method for evaluating the utility of a given dimension, such that we can select the optimal subset for clustering.

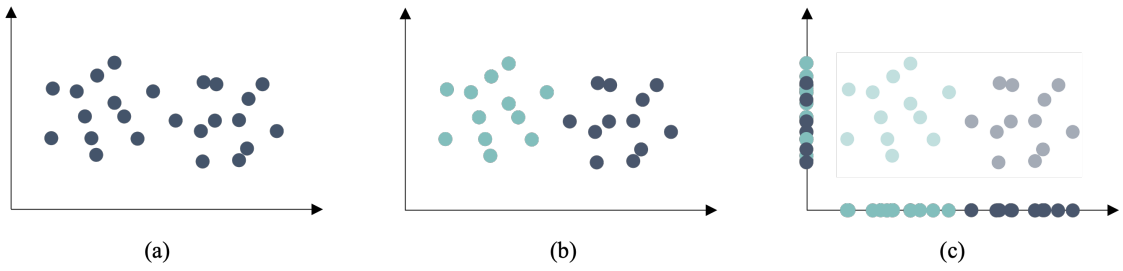


Figure 3: **Some dimensions may be more conducive to clustering than others.** (a) Raw dataset, (b) clustering solution, and (c) clustering solution projected onto constituent dimensions. The horizontal dimension is more conducive to clustering than the vertical dimension.

We explored conventional dimensionality reduction methods, such as Principal Component Analysis (PCA), but we found that these methods do not preserve the interpretability of the dimensions, which is essential if we are to connect our psycholinguistic clusters back to psychological constructs. Thus, we developed a method to evaluate the utility of a given dimension, which we then used to select an optimal subset of dimensions for further analysis. We refer to this method as *projected silhouette scores*.

2.3.3 Feature Subset Selection via Projected Silhouette Scores

A graphical depiction of the projected silhouette score calculation is shown in Figure 4. In short, the process involves first obtaining a multivariate clustering solution and projecting the data onto the constituent dimensions. Then, the silhouette scores are computed for each sample where the distances used to compute a_i and b_i are as measured on the projected dimensions. If there are d dimensions in the data, this method returns d silhouette scores for every sample in the dataset.

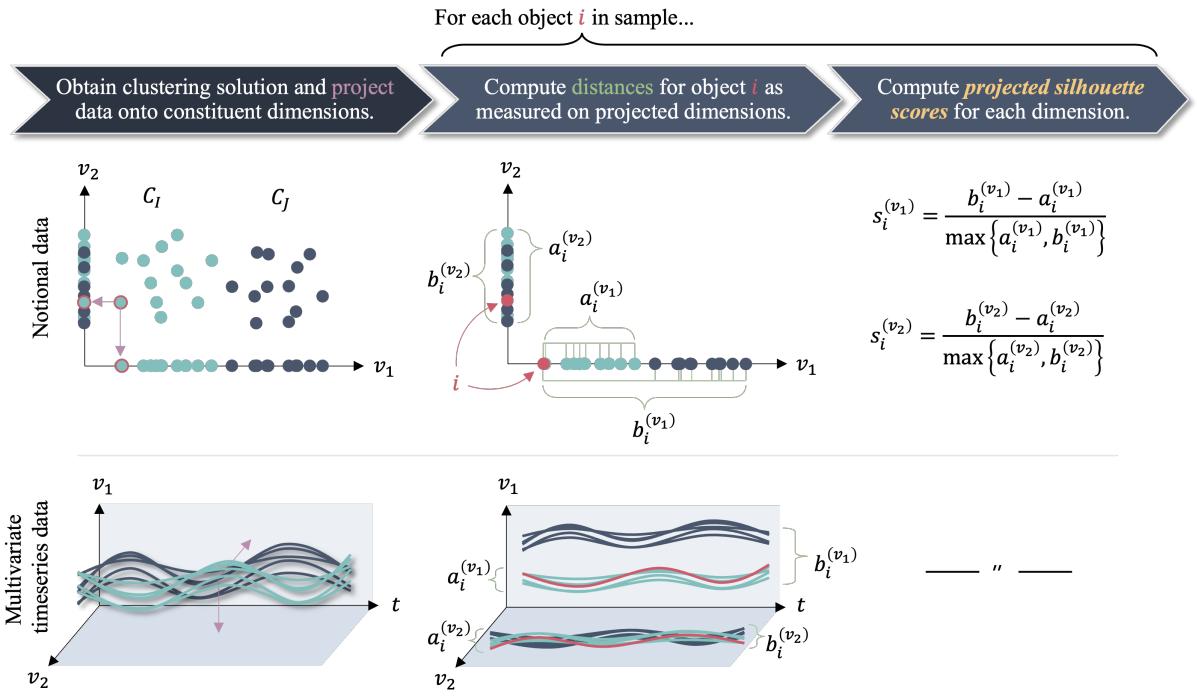


Figure 4: **Projected silhouette score calculation.** From left to right: (1) Obtain clustering solution and project data onto constituent dimensions. Then for each sample: (2) Compute silhouette distances for each object as measured on the projected dimensions. (3) Compute the *projected silhouette scores* for each dimension.

Once we have the projected silhouette scores for each dimension *for every sample*, we can ag-

gregate by cluster to get average silhouette scores by *both dimension and cluster*. With this $d \times k$ matrix of silhouette scores—where k is the number of clusters and d is the number of dimensions—we can average across the d-axis to obtain scores by cluster, or we can average across the k-axis to obtain scores by dimension. It's this latter approach that we used to compute a single score for each dimension to assess the quality of the clustering solution on that dimension/projection. Finally, note that we can also take a global average of all samples to obtain a *macro* score for the entire clustering solution.

Using these projected silhouette scores to evaluate the utility of a given dimension, we conducted various feature subset selection (FSS) experiments to determine the optimal subset of LIWC features (dimensions) to use for clustering. First, we simply took each LIWC dimension in isolation and performed univariate timeseries clustering (UTSC). Then, we computed the projected silhouette scores for each dimension and selected the dimension with the highest score. Second, we also conducted multivariate timeseries clustering (MTSC) experiments. We quickly found that using all 61 dimensions yielded invalid solutions (i.e., negative silhouette scores), thus we reduced the dimensionality via a backward feature selection (BFS) algorithm. In the backward selection approach, we started with all $d = 61$ dimensions and iteratively removed the worst performing dimension, as measured by the projected silhouette score, until we reached a terminating condition. We employed two such conditions: terminating with the *highest scoring model* (as measured by the macro silhouette score) or terminating with the *first valid model* (i.e., with all positive silhouette scores). Note that with either terminating condition, the final solution may be univariate.

For each experiment, we varied both the number of clusters k as well as the timeseries length w post-disclosure. We considered a test matrix comprised of $k = 2, 3, 4$ clusters and $w = 1, 2, 3, 4$ weeks post-disclosure. We typically found increasing the number of clusters and/or the timeseries length post-disclosure degrades cluster quality. That said, a longer timeseries or a model with more clusters and/or dimensions may be more robust than the alternative. Thus in some cases, we considered clustering solutions which may not have the highest macro silhouette score but may be more robust and generalizable.

From these analyses, we found four candidate clustering solutions, see Table 3: two from the univariate experiments and two from the multivariate experiments. (These will be presented more thoroughly in Section 3.) We explored all these candidate clustering solutions as the labels for our

classification task, though we found the best classification performance with Univariate Solution #1, which incidentally was also the clustering solution with the highest macro silhouette score. In Section 3, we discuss this solution in more detail.

Candidate Solution	(clusters, weeks)	LIWC Dim.(s)	Silhouette
Univariate Solution #1	$(k = 2, w = 2)$	`filler`	0.56
Univariate Solution #2	$(k = 2, w = 1)$	`anxiety`	0.53
BFS Highest Scoring	$(k = 2, w = 1)$	`focuspresent`	0.43
BFS First Valid	$(k = 2, w = 1)$	`verb, focuspresent, auxverb`	0.25

Table 3: Candidate clustering solutions considered for use in the classification task. Univariate Solution #1 was selected as the best candidate.

2.4 Phase II. Textual Classification

2.4.1 Feature Engineering

For the second phase of our analysis, we train several text classifiers to preempt to which temporal cluster a user would belong using only pre-disclosure information. To mitigate concerns of possible information leaks, we utilize different features for the classification task. Instead of LIWC scores, the inputs to our classifiers are the raw text of the last M tweets before disclosure. We experimented with different values of M and found optimal performance with $M = 50$ tweets before disclosure *Perhaps we include a figure for this in the supplementary materials. Then again, there are many small such hyperparameters, and there's no need to be exhaustive*. As is standard practice, we lowercase all text and remove all non-ASCII characters. We experimented with different vectorization schemes, including bag-of-words (BOW) with or without Term Frequency-Inverse Document Frequency (TF-IDF) normalization, but we found better performance without it. We also experimented with different n-grams and found better performance with b-grams.

2.4.2 Model Development

We developed the following models spanning a range of complexity: logistic regression (LR), support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), and DistilBERT (BERT). With the exception of DistilBERT, all models were trained using the `scikit-learn` Python package [24]. DistilBERT was trained via the `keras` Python package [7]. We set up our classifiers to handle

single-label, multi-class classification, although as shown in Table 3, all candidate clustering solutions had $k = 2$ clusters. These models, along with key hyperparameters, are summarized in Table 4. As will be discussed, we found largely similar results across all models with no one model starkly outperforming the others.

Model	Key Hyperparameters
LR	`solver='lbfgs'`
SVM	`kernel='linear'`
RF	`n_estimators=100, criterion='gini'`
MLP	`hidden_layer_sizes=(128,), activation='relu', solver='lbfgs'`
DistilBERT	`epochs=10, batch_size=16, learning_rate=3e-5`

Table 4: Text classification models and key hyperparameters.

2.4.3 Model Training

An 80/20 train/test split was used for all models. For some clustering solutions, the classes (i.e., labels) were imbalanced. To address this, we employed two techniques to improve performance on the minority class: (1) data augmentation via synonym replacement to balance the classes (training data only) and (2) optimal threshold tuning.

Synonym replacement was performed with the `NLPAug` Python library [20]. We did not replace stopwords or words less than 3 characters. For the training set, we augmented samples from the minority class until the classes were balanced. For the test set, we employed test time augmentation (TTA) to afford a more robust evaluation of our models. As a reminder, TTA is a technique where we augment each test sample n times (in contrast to balancing the samples) and average the predictions. We did this to maintain a representative distribution of classes while augmenting the size of our test set. We chose $n = 4$, such that there were 5 total predictions per sample.

The optimal threshold was determined analytically using the g-means algorithm [18]. The g-means algorithm is a method for determining the optimal threshold for a given classification problem. It is based on the geometric mean of the True Positive Rate (TPR) and False Positive Rate (FPR). The geometric mean is a measure of central tendency that is less sensitive to outliers than the arithmetic mean. The optimal threshold that which maximizes the geometric mean of the TPR and FPR. Mathematically, the optimal threshold is defined as follows.

$$\theta^* = \text{argmax}_\theta \sqrt{\text{TPR}(\theta) \times \text{FPR}(\theta)} \tag{7}$$

2.4.4 Model Evaluation

We used standard (imbalanced) classification performance metrics of precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC) to evaluate the models. Both the class-specific and macro scores (average across classes) are reported. Finally, we evaluated our models against three different baselines, one that always predicts the majority class (BL1), one that randomly selects a label with equal probability (BL2), and one that predicts the majority class with probability at parity with the class imbalance (BL3).

3 Results

3.1 Phase I: Timeseries Clustering Results

The results for both our univariate and multivariate timeseries clustering results are shown in Figures 5 and 6, respectively.

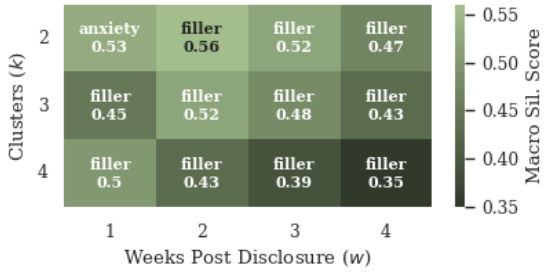
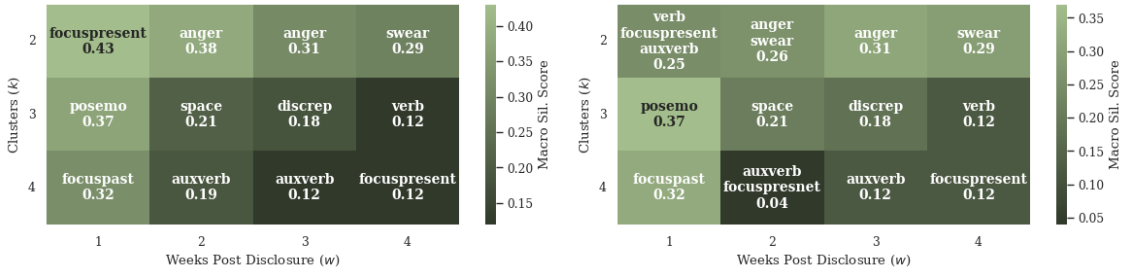


Figure 5: **Univariate timeseries clustering results over the $k \times w$ test matrix.** The reported scores are macro silhouette scores (average across all samples), and the reported dimensions are those with the highest macro silhouette score.

Consider first the univariate results in Figure 5. This figure reads as follows. Each cell in the matrix represents a clustering solution with k clusters and w weeks post-disclosure. The reported scores are the macro silhouette scores (average across all samples), and the reported dimensions are those with the highest macro silhouette scores. We see that for almost all combinations of k and w, the highest scoring LIWC dimension is the `filler` category. The sole exception is the $k = 2$



(a) Highest scoring model.

(b) First valid model.

Figure 6: **Multivariate timeseries clustering results over the $k \times w$ test matrix using feature subset selection (FFS) via projected silhouette scores.** In (a), the FFS algorithm terminates with the *highest scoring model*, and in (b), the FFS algorithm terminates with the *first valid model*.

and $w = 1$ solution, in which `anxiety` is the highest. The highest scoring clustering solution is the `filler` solution with $k = 2$ and $w = 2$ weeks post-disclosure with the `anxiety` solution with $k = 2$ and $w = 1$ weeks post-disclosure coming in a close second. We extract these two solutions as candidate solutions for the classification task. These correspond to Univariate Solutions #1 and #2, respectively, in Table 3.

Now consider the multivariate timeseries clustering (MTSC) results in Figure 6. A reminder that we performed feature subset selection (FSS) via projected silhouette scores as described in Section 2.3.3 with two terminating conditions. Figure 6a depicts the results for the *highest scoring model* terminating condition in which we systematically removed the worst dimension so long as the macro silhouette score increased. Figure 6b depicts the results for the *first valid model* terminating condition in which we removed the worst dimension until we reached a valid solution (i.e., all positive silhouette scores).

Consider Figure 6a corresponding to the MTSC FFS solution using the *highest scoring model* terminating condition. The salient features are three-fold. First, all solutions are 1-dimensional—meaning that the FFS algorithm terminated with a univariate solution—though none of these 1-dimensional solutions surpass the best model from Figure 5. In other words, comparing Figures 5 and 6a element-wise, we see that the univariate solutions outperform the multivariate solutions. This implies that at some point during the FFS algorithm, what would be the best-performing dimension (e.g., `filler`) was removed. Second, there is much more variability in the final LIWC dimension that emerged from the FSS algorithms compared to the purely univariate clustering

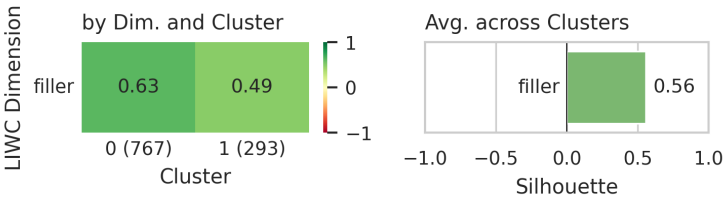
experiments shown in Figure 5. Finally, the highest scoring model is the `focuspresent` solution with $k = 2$ and $w = 1$ weeks post-disclosure. From this experiment, we extract this `focuspresent` clustering model as a candidate solution for the classification task—this corresponds to BFS Highest Scoring candidate in Table 3.

Figure 6b reads similarly. Many solutions are 1-dimensional—though again, none surpass the best purely univariate solution (see 5). Here, we do find three *multivariate* solutions (cells (2, 1), (2, 2), and (4, 2)). Given that we already have three *univariate* candidate solutions, we wanted to consider a multivariate solution. Thus, we extract the `verb, focuspresent, auxverb` solution with $k = 2$ and $w = 1$ weeks post-disclosure as a candidate solution for the classification task, which corresponds to BFS First Valid candidate in Table 3.

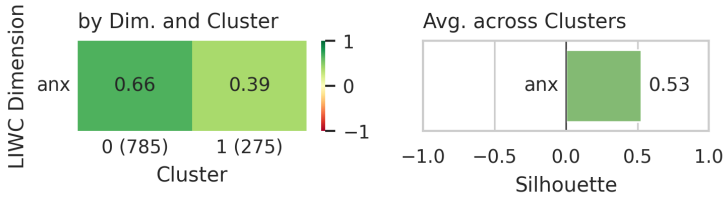
We expand on each of these candidate solutions in Figure 7.

Figure 7 reads as follows. For each candidate solution, the left panel depicts the silhouette scores by the LIWC dimension and cluster. The right panel depicts the silhouette score averaged across the clusters. *Perhaps, this should be a weighted average.* Consider the `filler` solution in Figure 7a. Here we see there is a dominant cluster (label 0) with roughly 73% of samples and a cluster-specific silhouette score of 0.63. Recall that this score is the average silhouette score of all samples in that cluster as measured on the `filler` projection. A score of 0.63 indicates that the average sample in cluster 0 is well-matched to its cluster and poorly matched to neighboring clusters. The smaller cluster (label 1) has 27% of samples with a cluster silhouette score of 0.49. The macro silhouette score is 0.56. The remaining candidate solutions are read similarly.

The cluster centers for the best performing `filler` solution are depicted in Figure 8a along with their 95% confidence intervals. We see that relative to the minority class (label 1), the majority of users (label 0) exhibit lower usage of `filler` words for 2 weeks post-disclosure. As reflected by the 95% confidence intervals (CI), there is some overlap between the two clusters, as is typical for clustering solutions: some samples are on the boundary between clusters. This is further illustrated in Figure 8b via silhouette plots for the `filler` solution. Here, we depict the *sample* silhouette score for each sample in the dataset, sorted by cluster and then silhouette score. The horizontal dashed line indicates the average silhouette score for that cluster. We see that the majority of samples are in cluster 0 and that the average silhouette score for cluster 0 is higher than that of cluster 1. This is consistent with the silhouette scores presented in Figure 7a. *Another thing we can try*

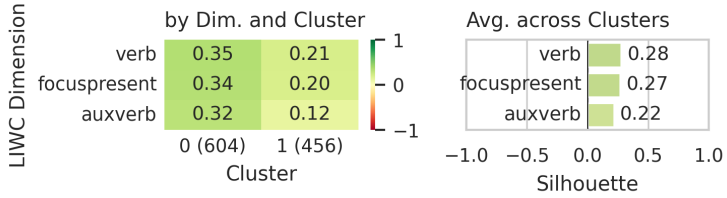


(a) Univariate Solution #1: `filler`



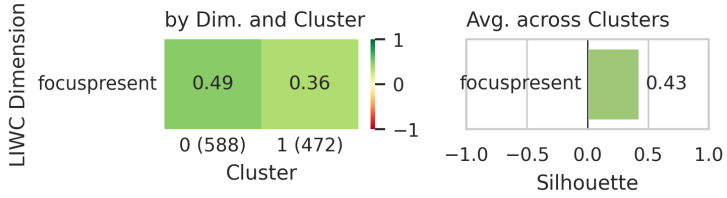
(b) Univariate Solution #2: `anxiety`

MTSC Solution after BFS (first valid model)
Macro-Silhouette Score: 0.25



(c) BFS Highest Scoring: `focuspresent`

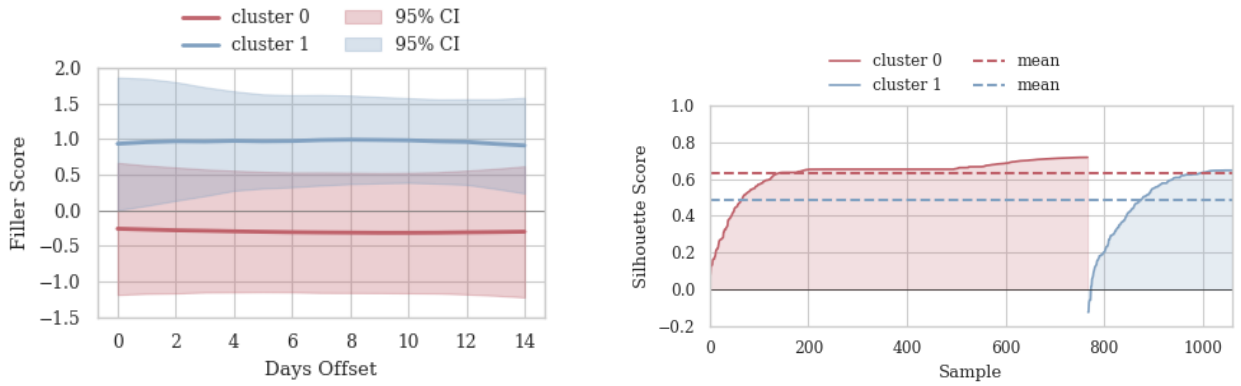
MTSC Solution after BFS (highest scoring model)
Macro-Silhouette Score: 0.43



(d) BFS First Valid: `verb, focuspresent, auxverb`

Figure 7: Timeseries clustering results aggregated by dimension and cluster. *I need to beautify these figures.*

for the classification phase is to drop the samples on the boundary between clusters and see if that improves classification performance. For example, drop all samples with sample silhouette score < some threshold.



(a) Cluster centers for the `filler` dimension.

(b) Silhouette plot for the `filler` dimension.

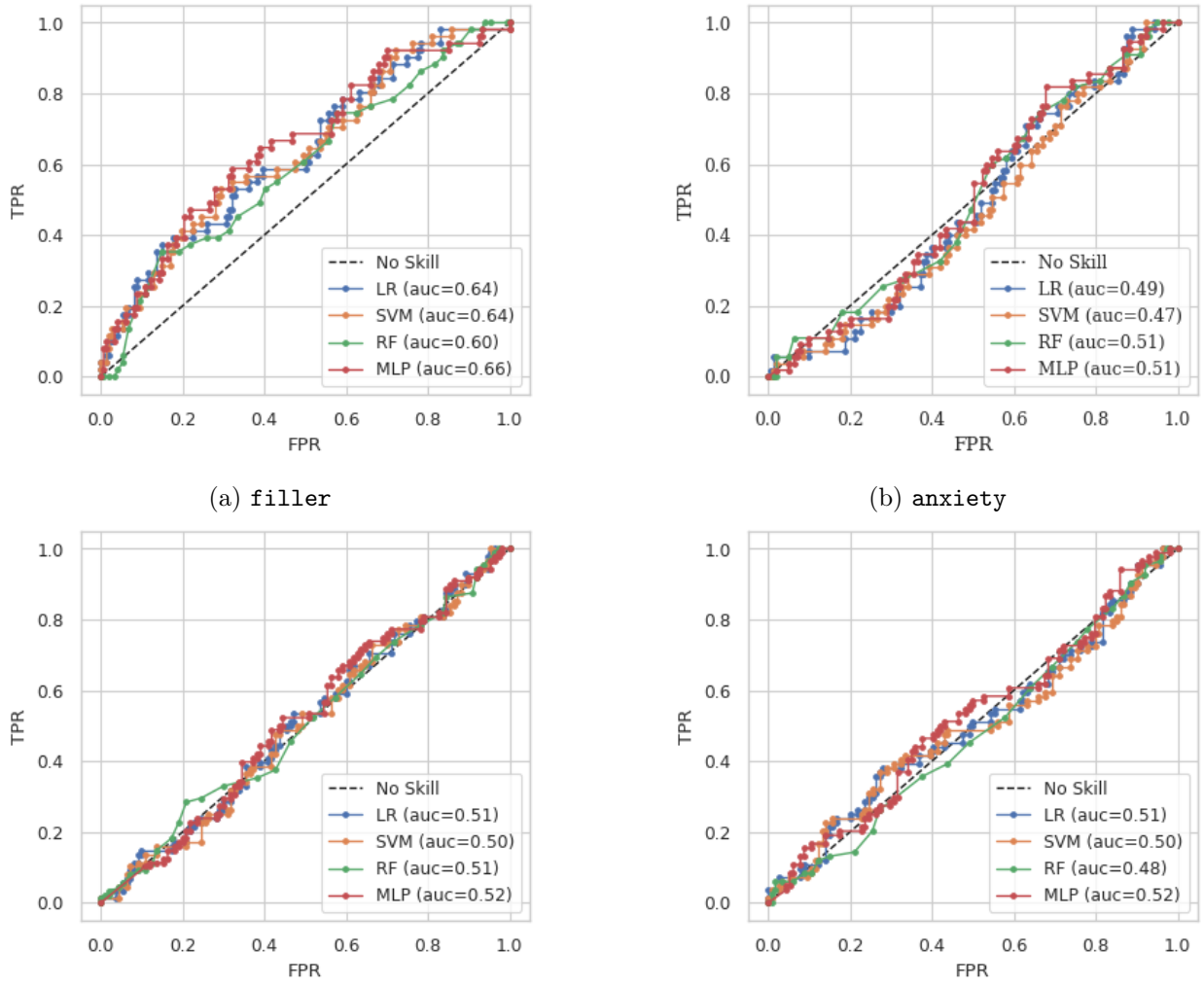
Figure 8: Cluster centers and silhouette plot for the `filler` dimension.

3.2 Phase II: Textual Classification Results

As a reminder, the classification task is to predict *post-disclosure* response—using the aforementioned clustering solutions as *labels*—using only *pre-disclosure* information.

Figure 9 shows the classification results in the form of Receiver Operator Characteristic (ROC) curves for each candidate solution and all models. Recall that the ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for various possible thresholds. (The default threshold is typically 0.5 where logit scores above this threshold are assigned 1, and 0 if below.) The area under the ROC curve (AUC) is a measure of the overall performance of the model. It ranges from 0 to 1, with higher values indicating better performance. A model with no skill, indicated by the diagonal dashed line, would receive an AUC score of 0.5. The closer the ROC curve is to the top left corner, the better the model.

From Figure 9 we see that of all candidate clustering solutions identified in Table 3, the `filler` is the only clustering solution in which all ROC curves consistently surpass the no skill baseline—indicating that there is some predictive information in the pre-disclosure data that can predict the post-disclosure `filler` clusters. In contrast, the other candidate solutions shown in Figures 9b, 9c, and 9d, do not significantly outperform the baseline. Note that this does not suggest that these clustering solutions are not valid or useful, but only that the pre-disclosure data does not contain predictive information for these clusters. In other words, of the four temporal clusters identified in Table 3, only the `filler` solution can be preempted. As a side note, we see that despite the wide range of model complexities, all models perform similarly, with no one model starkly outperforming



(a) `filler` (b) `anxiety`

(c) `focuspresent` (d) `verb, focuspresent, auxverb`

Figure 9: Text classification performance via Receiver Operator Characteristic (ROC) curves for all models and for each candidate solution. *Need to fix the fonts here. Also need to add BERT :P*

the others.

These results for the `filler` solution are expanded in Table 5 via macro precision, recall, and F1 scores. All models are evaluated against three baselines (BL). BL1 always predicts the majority class, BL2 is a purely random classifier, and BL3 predicts the majority class with probability at parity with the class imbalance. We see that all models surpass the baseline models, with the best model (MLP) achieving a 16-point increase in macro F1 score over the best baseline model (BL1). This MLP model also achieves the highest macro F1 score across all models.

We provided the class-specific performance of the best-performing model in Table 6. As a reminder, the majority class (label 0), represents those users who exhibit lower usage of `filler`

words post-disclosure. For the majority class, our best model achieves class-specific precision and recall of 0.83 and 0.72, respectively. This indicates that of all samples predicted to be in the majority class, 83% of them are actually in the majority class, and that of all samples in the majority class, the model can correctly identify 72% of them. Like the other models, the MLP model performs better on the majority class (0) than the minority class (1), which, due to the class imbalance, is expected. That said, we saw 10-15 point increases in minority class precision and recall with the addition of data augmentation and optimal threshold tuning with little to no impact on the majority class performance. *I should provide evidence for this in the supplementary material.*

Model	Precision	Recall	F1	AUC
BL1	0.52	0.53	0.48	0.51
BL2	0.47	0.47	0.47	0.55
BL3	0.37	0.50	0.43	0.50
LR	0.61	0.63	0.61	0.64
SVM	0.58	0.60	0.58	0.64
RF	0.60	0.62	0.60	0.61
MLP	**0.60**	**0.63**	**0.64**	**0.66**
DistilBERT	0.XX	0.XX	0.XX	0.XX

Table 5: **Macro classification performance for all models evaluated against three baselines.** BL1 always predicts the majority class, BL2 is a purely random classifier, and BL3 predicts the majority class with probability at parity with the class imbalance.

Class	Precision	Recall	F1	Support
0	0.83	0.72	0.77	147
1	0.42	0.59	0.49	51
Macro Avg.	0.63	0.65	0.64	198
Weighted Avg.	0.73	0.69	0.70	198

Table 6: **Class-specific performance of the Multi-Layer Perceptron (MLP) for the classification task.**

4 Discussion

In this work, we have shown that within the set of Twitter users who have disclosed some form of suicidality, there are several distinct temporal clustering solutions that capture the different ways people respond to self-disclosures of suicidality. One such clustering solution is characterized by `filler` words (I mean, you know, like, uh, um). Our findings can be summarized as follows.

Among those who have disclosed suicidality, a majority (73%) of them exhibit and maintain lower usage of filler words following disclosure. Additionally, we have shown that for this filler clustering solution, there is some predictive information in the pre-disclosure data that can preempt to which post-disclosure cluster a user would belong. Together, these results suggest that (1) there is some psychological connection between `filler` word usage and suicidality, specifically in the temporal vicinity following self-disclosure, and (2) that it is possible to preempt how a user will respond post-disclosure using only pre-disclosure information.

4.1 Interpretation of results

The connection between `filler` word usage and suicidality is not immediately obvious, but there have been some studies that have explored this relationship. For example, a study by Coppersmith et al. [11] on Twitter data from users who have attempted suicide found that `filler` words are among the top LIWC constructs *used more often by people who have attempted suicide*. Furthermore, it has been argued that people who use more tentative words (e.g., filler words) may not yet have psychologically processed a prior event ([27] citing [32]) enough such that they cannot form a coherent narrative. In light of these studies, we offer two possible interpretations of our findings.

First, the latent variable separating the `filler` clusters may be related to self-harm or past suicide attempts. Like in [11], in which increased filler word usage was associated with past suicide attempts, the `filler` clusters identified in our work, may be reflecting a latent self-harm / past suicide attempt variable. The class distribution of the `filler` clusters agrees with this supposition. We find that only 27% of users exhibit significantly higher usage of `filler` words. This reflects the fact that within the set of those who desire suicide, comparatively few can act on it [34].

Reproduce Figure 1 but break down by filler cluster? This is just my hunch that there is something fundamentally different between 'sleep forever' and 'hang myself'.

Another possible interpretation is that the `filler` usage is in direct response to the self-disclosure. As per [32], the increased use of filler words may reflect a lack of processing of some prior event which manifests as suppressed coherency. In our work, we see that the majority of users do not exhibit increased filler word usage. The majority of users exhibit *lower than average* usage of filler words following disclosure. While no causal claims can be made, our results suggest there may be some therapeutic benefit to self-disclosing suicidality, particularly in the form of broad-

casted self-disclosures on social media. This finding supports prior work that has investigated the therapeutic benefit of such self-disclosures [17].

As a side note, it's pretty interesting to me the filler words manifest on written text. Perhaps this is just the casual nature of Twitter these days. I suspect this result would be amplified if we were to look at the spoken word.

4.2 Theoretical and practical implications

From the timeseries clustering analysis, our results suggest that there is some psychological connection between filler word usage and suicidality, specifically in the temporal vicinity following self-disclosure, and we hope this work encourages further study in this area. If the connection between filler words and suicidality were a manifestation of some other latent variable (e.g., self-harm / past suicide attempts), then perhaps filler word usage could be used as a modest proxy for that latent variable. Alternatively, if the change in filler word usage is in direct response to self-disclosures of suicidality, then we may have one metric by which to quantify the therapeutic benefit of such disclosures.

From classification analysis, our results suggest that not only are there distinct post-disclosure groups (characterized by different amounts of filler words), but that these groups can also be preempted using pre-disclosure information. Importantly, this means that if decreased filler word usage is reflective of therapeutic benefit to self-disclosures of suicidality, then it is possible to predict *whether someone would benefit from such disclosures* with an estimated precision of 83$ (see Table 6). This would enable targeted intervention strategies for those struggling with suicidality. For example, were our model to preempt that an individual would benefit from disclosing their suicidality, could we provide them with more resources to do so? Or, were our model to preempt that an individual would not benefit from disclosing their suicidality, could we provide them with alternative resources?

4.3 Limitations and future work

Our results should be considered in light of several limitations they share with other similar studies. First, this study was non-intrusive, meaning that we did not reach out to any users to verify suicidality or the authenticity of the disclosures. Second, and related to the first, we assumed that

all suicide disclosures obtained from different keywords were created equal—for example, the phrase "sleep and never wake up" was considered to be of the same nature as "shoot myself". Thus, it is possible that our sample contained a mixture of, using one theoretical suicide framework, *active* and *passive* suicidal individuals. Further analysis is warranted to investigate the connection between word choice in suicide disclosures and modes of suicidality (e.g. active vs. passive). Third, our sample was limited to 1060 predominantly English-speaking Twitter users. Further study is needed to generalize our results to larger populations and other platforms. Finally, in the absence of a carefully curated control group, no causal claims can be made. However, the fact that our findings are in line with supporting research (e.g., [11] and [27] citing [32]), lends credence to the connection between suicidality and filler words, though further research is warranted to better understand this relationship. For example, is filler word usage a manifestation of some other latent variable? Or are changes in filler word usage in direct response to self-disclosures of suicidality?

5 Conclusion

In this work, we identified several psycholinguistic patterns post self-disclosure of suicidality on social media. In particular, we found that a majority of users exhibit lower usage of filler words following disclosure. Furthermore, we found that there is some predictive information in the pre-disclosure data that can preempt to which post-disclosure cluster a user would belong. These results suggest that (1) there is some psychological connection between `filler` word usage and suicidality, specifically in the temporal vicinity following self-disclosure, and (2) that it is possible to preempt how a user will respond post-disclosure using only pre-disclosure information. We hope this work encourages further study in this area, and that it may be used to inform intervention strategies for those struggling with suicidality.

References

[1] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(5):1–35, 2018.

[2] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3906–3918, 2016.

[3] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. Sensitive self-disclosures, responses, and social support on instagram: the case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1485–1500, 2017.

[4] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022.

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[6] Daejin Choi, Steven A Sumner, Kristin M Holland, John Draper, Sean Murphy, Daniel A Bowen, Marissa Zwald, Jing Wang, Royal Law, Jordan Taylor, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly us suicide fatalities. *JAMA network open*, 3(12):e2030932–e2030932, 2020.

[7] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[8] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition. *J. Off. Stat*, 6(1):3–73, 1990.

[9] Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300, 2016.

[10] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39, 2015.

[11] Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. Quantifying suicidal ideation via language usage on social media. In *Joint statistics meetings proceedings, statistical computing section, JSM*, volume 110, 2015.

[12] Paul C Cozby. Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73, 1973.

[13] H Wayland Cummings and Steven L Renshaw. Slca iii: A metatheoretic approach to the study of language. *Human Communication Research*, 5(4):291–300, 1979.

[14] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137, 2013.

[15] Munmun De Choudhury and Emre Kiciman. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 32–41, 2017.

[16] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110, 2016.

[17] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–27, 2017.

[18] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[19] Adam N Joinson, Carina B Paine, et al. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, 2374252:237–252, 2007.

[20] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[21] John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5:245–257, 2020.

[22] Charles E Osgood and Evelyn G Walker. Motivation and language behavior: a content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, 59(1):58, 1959.

[23] Michelle O'reilly, Nisha Dogra, Natasha Whiteman, Jason Hughes, Seyda Eruyar, and Paul Reilly. Is social media bad for mental health and wellbeing? exploring the perspectives of adolescents. *Clinical child psychology and psychiatry*, 23(4):601–613, 2018.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

[26] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[27] Annika Marie Schoene, Alexander Turner, Geeth Ranmal De Mel, and Nina Dethlefs. Hierarchical multiscale recurrent neural networks for detecting suicide notes. *IEEE Transactions on Affective Computing*, 2021.

[28] Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.

[29] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.

[30] Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31, 2020.

[31] Amelia Strickland. Exploring the effects of social media use on the mental health of young adults. 2014.

[32] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[33] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *The Journal of Machine Learning Research*, 21(1):4686–4691, 2020.

[34] Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. The interpersonal theory of suicide. *Psychological review*, 117(2):575, 2010.

[35] Junwei Xiao, Jianfeng Lu, and Xiangyu Li. Davies bouldin index based hierarchical initialization k-means. *Intelligent Data Analysis*, 21(6):1327–1338, 2017.

A Appendix

A.1 Appendix A: Timeseries Clustering Methodology

A.2 Appendix B: Textual Clustering Methodology