# Reading Between the Signs

## Predicting Future Suicidal Ideation in Adolescent Social Media Texts

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Paul Iuna Blum

to be defended publicly on
Monday, June 23, 2025.

| | | |
|---|---|---|
| Thesis committee: | Dr. P. K. Murukannaiah | Faculty EEMCS, TU Delft, chair |
| | Dr. J. Yang | Faculty EEMCS, TU Delft, core committee member |
| | Dr. Ir. E. Liscio | Faculty EEMCS, TU Delft, advisor |
| | Ir. R. Zhang | Faculty TPM, TU Delft, advisor |

Interactive Intelligence Research Group
Department of Intelligent Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

TU Delft

## Abstract

Suicide is a leading cause of death, yet predicting it remains a significant challenge. Risk factors such as depression or substance use are commonly used for prediction, but their predictive performance is often only slightly better than chance. Additionally, many cases go undetected due to a lack of contact with mental health services. Social media, however, offers a unique opportunity, as people often share their thoughts and struggles online in real time. In this work, we propose a novel task and method to approach it: predicting suicidal ideation and behavior (SIB) *before* a user ever expresses it on an online forum. This predictive framing, where no self-disclosure is used as input at any stage, remains largely unexplored in the suicide prediction literature. Our model, EARLY-SIB, achieves a balanced accuracy of 0.73 for predicting future SIB on a Dutch youth forum, demonstrating that such tools can offer a meaningful addition to traditional methods.

## 1 Introduction

Predicting suicide is hard. 60% of people who go on to die by suicide have not expressed suicidal ideation as documented in their medical records or inquired through a semi-structured interview (McHugh et al., 2019). Hence, suicidality cannot be reliably predicted through the common practice of clinicians asking individuals directly about suicidal thoughts. However, suicide is a pressing issue, globally being the third leading cause of death among 15 to 29-year-olds (World Health Organization, 2024). Moreover, each suicide is estimated to meaningfully affect 135 individuals around them (Cerel et al., 2019). Yet, our ability to predict suicide has not improved over the past 50 years (Franklin et al., 2017).

Research has mainly focused on assessing a person's risk of suicide using so-called "risk factors" (Turecki et al., 2019). These are traits or conditions, such as prior attempts, depression, hopelessness, mental illness, or substance use, thought to increase the likelihood of suicidal behavior and have informed widely adopted guidelines (World Health Organization, 2024; American Association of Suicidology, 2023; American Foundation for Suicide Prevention; National Institute of Mental Health (NIMH)). However, a meta-analysis of decades of suicide research has revealed that risk factor-based prediction performs only slightly better than chance (Franklin et al., 2017). The guidelines often lack specificity, classifying nearly anyone with mental illness, chronic physical illness, life stress, special population status, or access to lethal means as being at risk (Franklin et al., 2017). More recently, Schafer et al. (2021) conducted a comprehensive meta-analysis directly comparing theory-driven approaches with machine learning (ML) models and found that the latter provided substantially superior predictive accuracy for suicide ideation, attempts, and death.

Given the widespread use of social media and increased self-disclosure during the years of adolescence (Vijayakumar and Pfeifer, 2020), social media data could prove to be a valuable source for identifying suicide risk. Teenagers often do not disclose risk factors to their physicians; instead, many share them on social media pages (Pourmand et al., 2019), often making peers the first recipients of their distress calls (Belfort et al., 2012; Vijayakumar and Pfeifer, 2020). Additionally, over two-thirds of people who die by suicide have had no contact with professional mental health care in the year prior to their death (Stene-Larsen and Reneflot, 2019), essentially making them overlooked by traditional healthcare systems. Even for those individuals who seek professional help, the time between healthcare interactions is immense compared to social media interactions (Coppersmith et al., 2018).

Suicidal ideation and behaviour (SIB) refers to thoughts, plans, or actions related to suicide (Posner et al., 2010). Ideation is one of the earliest and typically unobserved stages in the progression of an individual's suicidal process (Wasserman et al., 2012).

In this work, we propose a novel task and a method to approach it: predicting future SIB based solely on a user's social media post history. Unlike existing approaches, our method does not rely on any self-disclosure of SIB as input to the model, effectively moving the prediction one step earlier in the timeline. To perform this task, we introduce EARLY-SIB, an ML model designed to predict SIB from a list of social media posts of a given user. We apply our method to Kindertelefoon[1], a Dutch adolescent help forum, and show that online post history offers meaningful predictive value. These findings suggest that ML tools applied to social media platforms may complement traditional risk factor-based approaches.

[1] https://forum.kindertelefoon.nl/

## 2 Related Works

We first review traditional clinical methods of identifying suicide risk before turning to recent work using natural language processing.

### 2.1 Traditional Methods

The U.S. Food and Drug Administration (2012) considers the Columbia Suicide Severity Rating Scale (C-SSRS) to be the gold standard for suicide risk assessment in clinical trials. C-SSRS provides a list of questions that aim at identifying suicidal ideation and behavior directly, for example, "Have you had any thoughts about killing yourself?" (Posner et al., 2010).

The European Psychiatric Association (Wasserman et al., 2012) recommends that suicide risk be assessed through clinical interviews that evaluate the patient's psychological and social functioning. This includes psychological assessments of personality, evaluations of socio-economic status, and the use of psychometric instruments such as scales published by Patterson et al. (1983); Hamilton (1960); Paykel (1976); Linehan et al. (1983); Posner et al. (2010). There are also biological measurements, but they are only possible at specialized clinics and thus mainly used in research settings (Wasserman et al., 2012).

Notably, most of these risk assessment tools rely on directly asking individuals about suicidal thoughts and/or evaluating predefined psychological or social factors. Rather than relying on individuals to disclose their mental state, our method infers risk from their online behavior and avoids dependence on predefined risk factors by leveraging ML to learn patterns.

### 2.2 NLP for Suicide Detection

Recent years have seen growing interest in using natural language processing (NLP) to assess suicide risk. Many studies used patients' electronic medical records (Zandbiglari et al., 2025; Carson et al., 2019; Singh Rawat and Yu, 2022; Tsui et al., 2021; Poulin et al., 2014; Bittar et al., 2019; Metzger et al., 2017; Zhong et al., 2019) or used chat messages between help-seeks and providers (Bialer et al., 2022; Qiu et al., 2024; Broadbent et al., 2023). However, these approaches neglect people who never sought help or have had limited contact with healthcare professionals. Another approach is using social media platforms like Twitter and Reddit, where much work has focused on classifying posts into containing suicidal thoughts or not (Haque et al., 2022; Roy et al., 2020; Cao et al., 2019; Mathur et al., 2020) or classifying users into different risk levels of suicide (Sawhney et al., 2020, 2021a,b; Mohammadi et al., 2019). Datasets have supported this direction by providing Reddit posts which have been expert-labeled into risk categories (Zirikly et al., 2019; Gaur et al., 2019; Park et al., 2020). While these approaches demonstrate the feasibility of detecting users at risk, they do not predict future disclosure of SIB. This limits their utility for early prevention or broader screening.

A smaller number of studies have attempted to infer suicide risk early. Coppersmith et al. (2018) and work under the CLPsych Task 2021 (MacAvaney et al., 2021; Bayram and Benhiba, 2021; Gollapalli et al., 2021; Gamoran et al., 2021; Morales et al., 2021; Wang et al., 2021) used social media posts from users with actual suicide attempts. However, such data is difficult to obtain and may already include explicit self-reports of suicide, limiting its usefulness for the earliest stages of detection. In contrast, our task is to predict self-reports before the user ever shares them, essentially moving one step earlier in the timeline.

Recent efforts under the CLPsych Task 2024 (Chim et al., 2024) have been made to provide explanations of the model's risk prediction. They highlight salient sub-phrases within suicidal Reddit posts and summarize that evidence at the user level (L et al., 2024; Alhamed et al., 2024; Gyanendro Singh et al., 2024). In contrast, our work explains predictions in an early prediction setting, where we highlight earlier interactions that do not contain SIB but are indicative of future risk.

### 2.3 NLP for Mental Health Detection

Aside from SIB, work in NLP has been done towards early detection of mental health issues, including depression (Wang et al., 2024; Hadzic et al., 2024; Zogan et al., 2022), eating disorders (López-Úbeda et al., 2021; Mármol Romero et al., 2024), and gambling addiction (Smith et al., 2024; Perrot et al., 2022), with one of the most important forums being ERisk@CLEF (Montejo-Ráez et al., 2024).

We draw on architectural ideas from Zhang et al. (2024), who proposed a model for detecting users with depression on social media. While they use posts directly indicative of depression as input, their architecture proves to be a valuable starting point for our task, because it is capable of taking multiple posts from a user's history as input.

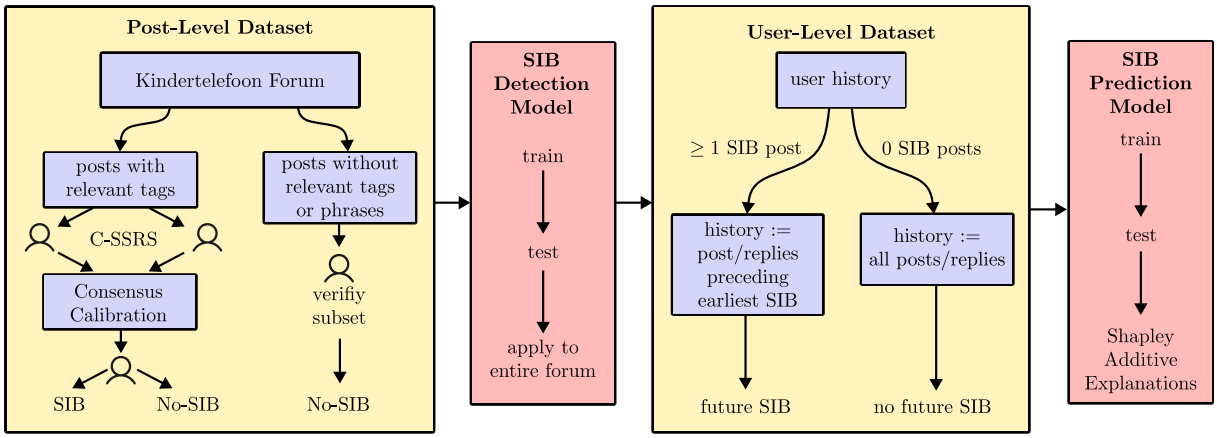



Figure 1: Pipeline of our approach. We manually label a small post-level dataset, which we use to train a model to detect SIB. We apply this model to the entire forum to get a user-level dataset. Finally, we train a model to predict future SIB and use SHAP to explain the predictions.

## 3 Dataset

We use data from a Dutch forum called the Kindertelefoon[2], which is intended to be a space where adolescents between 12 and 18 can share their feelings and experiences that they may find difficult or uncomfortable to discuss otherwise. Some of the most popular subforums on the platform include emotional problems and feelings, relationships and love, gender identity, sexuality, and problems at home. Users can anonymously participate in the forum by creating a post or by writing a reply to an existing post. Each post includes a title, a body, and may also contain tags, i.e., keywords selected by the user to describe the content. Replies consist only of a body of text. In this work, we use the term *interaction* as an umbrella term to refer to both posts and replies. We extract over 37,000 forum posts, along with all corresponding replies. All usernames were pseudonimized. The ethics committee of the home university approved this research.

We first label a small post-level dataset, which we use to train a model to detect SIB posts. This model is then used to create a user-level dataset that excludes any SIB posts.

### 3.1 Post-Level Dataset

Figure 1 shows the labeling process for the post-level dataset. To encourage our detection model to distinguish between self-reports of SIB and other posts that contain language around suicide, we selected posts likely to contain suicide-related language. Specifically, we retrieved posts in which users had included any of the following tags: 'suicide', 'suicidal thoughts', 'thinking of suicide', 'suicidal', or 'taking my life' (originally in Dutch, see Appendix A). These posts were then annotated following consensus calibration as described in Oortwijn et al. (2021). First, two annotators familiarized themselves with the annotation guidelines, which were the C-SRSS guidelines (Posner et al., 2010). Second, the key concepts of suicidal ideation and suicidal behavior were discussed. Third, the two annotators individually labeled the same randomly selected subset of 100 posts. Fourth, the annotations were compared, where we found a Cohen's Kappa score of $\kappa = 0.66$ (substantial agreement). Fifth, the annotators met to discuss those posts where the labels disagreed. See Appendix B for examples and agreed annotation rules. After the disagreement resolution, there was perfect agreement ($\kappa = 1.00$). The rest of the annotations were done by one of the annotators, which resulted in 569 posts labeled as SIB and 138 labeled as No-SIB. We refer to these 138 posts as "hard" instances, as they contain language relating to suicide but are labeled as No-SIB.

To generate more instances of No-SIB, we randomly sampled an additional 1,300 posts from the entire forum, ensuring that these posts did not contain any of the previously mentioned tags in their title, body, or tags, nor did they include the spans 'I can't anymore', 'dead', 'end to', or 'kill' (see Appendix A for Dutch original). A validation of 200 such posts revealed that only one contained SIB. The final post-level dataset contained 569 SIB (28%) and 1438 No-SIB (72%) posts.

[2]https://forum.kindertelefoon.nl/

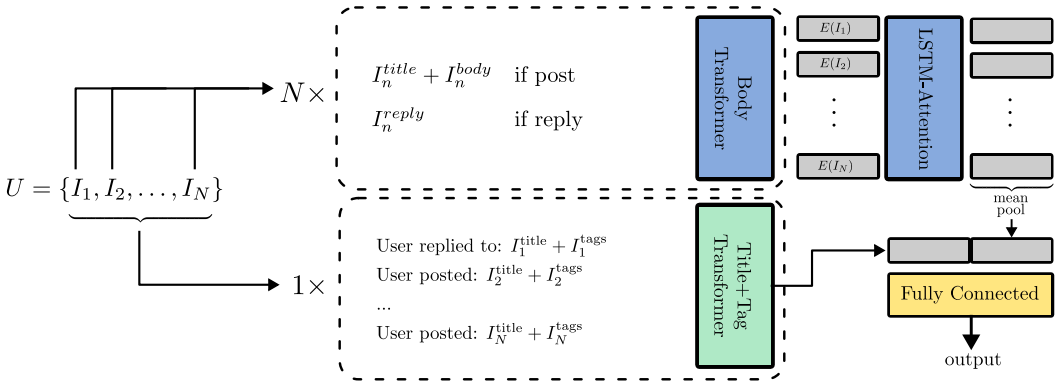



Figure 2: Architecture of our proposed model EARLY-SIB, for early prediction of suicidal ideation and behaviour.

### 3.2 User-Level Dataset

We use the post-level dataset to train a classifier to detect SIB posts, which is then used to create a user-level dataset. See Section 5.1 for details on the post classifier. The classifier is applied to the entire forum, assigning a label to all posts. From this, we create a user-level dataset, containing for each user a label indicating whether they have shared at least one SIB post or not, as well as that user's interaction history up to but excluding any SIB post (see right half of Figure 1). There are 284 users labeled SIB, 7,256 users labeled No-SIB. For our early prediction task, the input is exclusively the interactions that came **before** any SIB post, and the goal is to predict whether a given user will go on to share SIB or not.

## 4 Methodology

We present EARLY-SIB, our proposed model for predicting SIB early (Figure 2). We draw architectural inspiration from Zhang et al. (2024). EARLY-SIB takes as input a user $U$ represented by $N$ forum interactions $I_1 \ldots I_N$. Each of the interactions is then passed into a pretrained BERT model (the Body Transformer). More specifically, if the interaction is a post, we concatenate the title and body, and in case the interaction is a reply, we simply take the reply text. We obtain $N$ CLS representations, which are passed into a bidirectional LSTM network followed by an attention layer. The output is mean-pooled and represents one part fed into a fully connected layer.

Additionally, we concatenate all titles and tags associated with the user's interactions into one string. This is fed into another pretrained BERT model (the Title+Tag Transformer), and its CLS embedding forms the other part fed into the fully connected layer. For the pretrained transformers, we use BERTje, a Dutch BERT model (Vries et al., 2019). See Appendix C.2 for hyperparameters.

## 5 Experimental Setup

The experimental setup in this work can be separated into three parts. First, we curate a user-level dataset by training a model to detect SIB in forum posts (Section 5.1). Second, we train a model to predict whether a user will share SIB based on their previous interactions on the forum (Section 5.2). Third, we leverage an explainable AI technique to identify the most relevant post from the user's history that contributed to the prediction (Section 5.3).

### 5.1 Detecting SIB Posts

We compare different approaches to SIB detection (prompting-based and supervised). Our prompting methods include zero-shot inference on LLaMA 3 using either a simple prompt describing the task in a few sentences or a prompt including the C-SSRS guideline (see Appendix D). We test both English and Dutch versions of the prompt. We perform instruction tuning using the same prompts. The supervised methods include training XLM-RoBERTa for sequence classification and finetuning LLaMA 3 with a classification head. See Appendix C.1 for hyperparameters. We use the best-performing of those methods to assign labels to all posts on the forum.

### 5.2 Predicting Future SIB

With each data instance being a user's entire history, 50% of inputs exceed the maximum length allowed by BERT-based models, and 25% surpass the limit for LLaMA models. As baselines, we attempt zero-shot prompting on LLaMA 3 and implement the

architecture presented in Zhang et al. (2024), and apply it to our task.

The dataset is heavily imbalanced, with less than 4% of users belonging to the SIB class and having at least one interaction prior to a SIB post. We found that down-sampling the majority class to a 50-50 distribution yielded the best performance, compared to using weighted loss, and different factors for down-sampling the majority class or up-sampling the minority class (see Appendix C.2 for tested resampling ratios).

As for the number of interactions used as input to our model, we set a limit of $N = 30$, and prioritize posts over replies in filling these 30 slots. $N = 30$ covers more than 90% of users in our data. We also show an analysis over different $N$'s.

### 5.3 Explaining the Prediction

Given that our model operates as a black box, it is crucial to provide transparency in how these outputs are generated. We use Shapley additive explanations (SHAP) (Lundberg and Lee, 2017) to interpret the model's predictions. SHAP is a game-theoretic approach that assigns an importance value to each feature (in our case, each user interaction) based on its contribution to the final output.

## 6 Results and Discussion

### 6.1 Detecting SIB Posts

Table 1 summarizes performances in detecting SIB. The best performing method is training LLaMA 3 with a classification head, which achieves a weighted F1 score of $0.96 \pm 0.01$. See Figure 3 for the corresponding confusion matrix. We like to note that it also achieves an F1 score of $0.88 \pm 0.02$ on the "hard" subset of test instances where the inputs contain language around suicide but are not actual self-reports of SIB. This shows the model's great ability to pick up and distinguish language related to suicide.

In contrast, the baselines do not yield satisfactory performance. We find that zero-shot inference on LLaMA 3 reaches at most a weighted F1 score of $0.74 \pm 0.02$. This is after trying prompts including detailed C-SSRS guidelines or not, as well as English and Dutch variants. Zero-shot inference fails especially on recall, with the best-performing prompt only achieving $0.34 \pm 0.02$, showing that this method fails to pick up true positives. This is partly explained by the model often refusing to give answers because of the sensitive nature of this topic, despite our attempts to counteract this through a system prompt.

| Method | Weighted F1 | Recall | Precision |
|---|---|---|---|
| L3-ZS-EN-S | $0.74 \pm 0.02$ | $0.34 \pm 0.02$ | $0.71 \pm 0.03$ |
| L3-ZS-EN-C | $0.74 \pm 0.02$ | $0.31 \pm 0.03$ | $0.77 \pm 0.04$ |
| L3-ZS-NL-S | $0.61 \pm 0.02$ | $0.02 \pm 0.01$ | $0.87 \pm 0.16$ |
| L3-ZS-NL-C | $0.60 \pm 0.02$ | $0.02 \pm 0.01$ | $0.32 \pm 0.17$ |
| L3-FT-EN-S | $0.93 \pm 0.01$ | $0.88 \pm 0.05$ | $0.86 \pm 0.04$ |
| L3-FT-EN-C | $0.92 \pm 0.02$ | $0.89 \pm 0.06$ | $0.85 \pm 0.05$ |
| L3-FT-NL-S | $0.92 \pm 0.01$ | $0.84 \pm 0.05$ | $0.89 \pm 0.02$ |
| L3-FT-NL-C | $0.73 \pm 0.15$ | $0.36 \pm 0.43$ | $0.62 \pm 0.18$ |
| xlm-roberta | $0.91 \pm 0.02$ | $0.89 \pm 0.02$ | $0.81 \pm 0.06$ |
| **L3-SC** | $0.96 \pm 0.02$ | $0.91 \pm 0.04$ | $0.92 \pm 0.01$ |

Table 1: Performance on detecting SIB. Abbreviations: L3 = LLaMA3, ZS = Zero-Shot, FT = Fine-Tuned, SC = Sequence Classification Head, EN = English, NL = Dutch, S = Simple Prompt, C = C-SSRS Prompt. Standard deviation from 5-fold cross-validation.

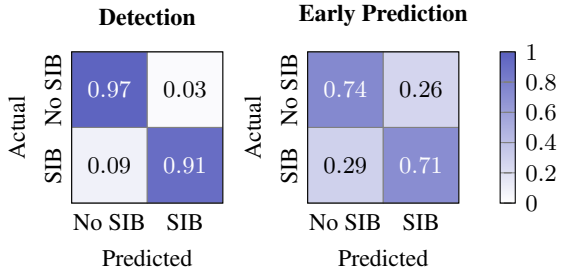



Figure 3: Confusion matrices for best performance on detecting SIB (left) and early prediction of SIB (right). Row-normalized and averaged over evaluation folds.

Instruction-tuned variants of the same model demonstrate improved results with the best-performing prompt yielding a weighted F1 of $0.93 \pm 0.01$. In both zero-shot and instruction-tuned versions, a short, simple prompt in English worked best. Training XLM-RoBERTa for binary classification closely matches this performance.

Thus, we use our best performing method, LLaMA 3 with a classification head, to label all posts on the forum.

### 6.2 Predicting Future SIB

Table 2 shows the performance on predicting future SIB. For reference, we show a model that predicts both classes uniformly at random and a model that always predicts the majority class. They demonstrate the highly imbalanced nature of our dataset with a precision close to zero. We therefore focus on balanced accuracy as our main evaluation metric. Balanced accuracy is the average recall across classes and accounts for class imbalance by giving

| Method | Balanced Accuracy | Recall | Precision |
|---|---|---|---|
| Random | $0.52 \pm 0.02$ | $0.54 \pm 0.04$ | $0.04 \pm 0.01$ |
| Majority | $0.50 \pm 0.00$ | $0.00 \pm 0.00$ | - |
| L3 title+tags | $0.57 \pm 0.03$ | $0.21 \pm 0.02$ | $0.09 \pm 0.02$ |
| L3 3 posts | $0.55 \pm 0.04$ | $0.14 \pm 0.02$ | $0.09 \pm 0.02$ |
| Zhang | $0.67 \pm 0.04$ | $0.71 \pm 0.11$ | $0.07 \pm 0.01$ |
| **EARLY-SIB** | $0.73 \pm 0.02$ | $0.71 \pm 0.07$ | $0.10 \pm 0.02$ |

Table 2: Performance on predicting SIB. Highly imbalanced data (96% No-SIB users). L3 = LLaMA3 zeroshot. Standard deviation from 5-fold cross-validation.

equal weight to each class.

Our model EARLY-SIB achieves a balanced accuracy of $0.73 \pm 0.02$ and recall of $0.71 \pm 0.07$, see Figure 3 for the corresponding confusion matrix. This significantly outperforms the architecture by Zhang et al. (2024) which achieves a balanced accuracy of $0.67 \pm 0.04$ and recall of $0.71 \pm 0.11$. A McNemar test reveals a statistically significant difference between EARLY-SIB and Zhang, $\chi^2(1, N = 8379) = 335$, $p < 0.001$.

We also evaluate zero-shot inference baselines. Using a full list of titles and tags that a user interacted with as input gives a balanced accuracy of $0.57 \pm 0.03$ and a recall of only $0.21 \pm 0.02$. If we take the last three interactions as input, the performance is slightly worse with a balanced accuracy of $0.55 \pm 0.04$ and recall of $0.14 \pm 0.02$. Part of the reason for this is that, on top of the model refusing to provide an answer due to the sensitive topic of suicide, in some cases, it failed to follow the intended task altogether. For example, it occasionally responded as if addressing the post's author directly, offering support or advice, rather than focusing on the task in the prompt. This highlights the limitations of this inference approach, particularly in handling long input sequences, which are central to our use case.

The results show that a trained classifier can predict SIB with considerably higher reliability than zero-shot inference methods. To better understand the strong performance of EARLY-SIB, we analyze the contributions of its individual components, investigate the impact of context window size, and explore how variations in input configuration influence predictive accuracy.

## Ablation Study

We perform an ablation study (Table 3) to assess the individual contributions of EARLY-SIB's core components: the Body Transformer, the Title+Tag Transformer, and the LSTM. Interestingly, the Title+Tag Transformer performs well independently, suggesting that the titles and tags encode much of the necessary signal for making accurate predictions. Removing either the Body Transformer or the LSTM results in a noticeable performance drop. Specifically, the LSTM, chosen to capture sequential dependencies leading up to the SIB disclosure, improves the performance by 0.02 compared to leaving it out. The isolated Body Transformer achieves a balanced accuracy comparable to the original Zhang architecture. Overall, the ablation study confirms that each component of EARLY-SIB makes a meaningful contribution to its performance.

| Ablation | Balanced Accuracy | Recall | Precision |
|---|---|---|---|
| ~~BTL~~ | $0.67 \pm 0.03$ | $0.71 \pm 0.15$ | $0.07 \pm 0.01$ |
| B~~T~~L | $0.70 \pm 0.02$ | $0.70 \pm 0.02$ | $0.08 \pm 0.01$ |
| BT~~L~~ | $0.71 \pm 0.03$ | $0.67 \pm 0.04$ | $0.09 \pm 0.01$ |
| B~~TL~~ | $0.68 \pm 0.04$ | $0.61 \pm 0.14$ | $0.08 \pm 0.01$ |

Table 3: Ablation study. B = Body Transformer, T = Title+Tag Transformer, L = LSTM.

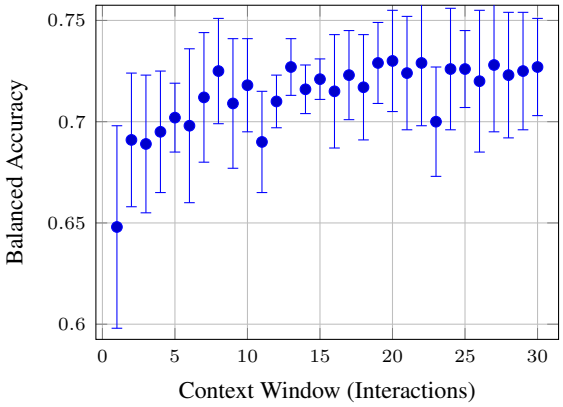



Figure 4: Model performance using different context windows, i.e., the maximum number of interactions $N$ used as input to the model. Error bars indicate standard variation across evaluation folds.

## Context Window Size

Next, we investigate how EARLY-SIB's performance depends on the context window size $N$, defined as the number of past interactions provided to the model (see Figure 4). While individual performance values fluctuate, the overall trend suggests an improvement in balanced accuracy as more context is used. Using only the most recent interaction ($N = 1$) yields a mere $0.65 \pm 0.05$. Performance generally increases with larger $N$, reaching

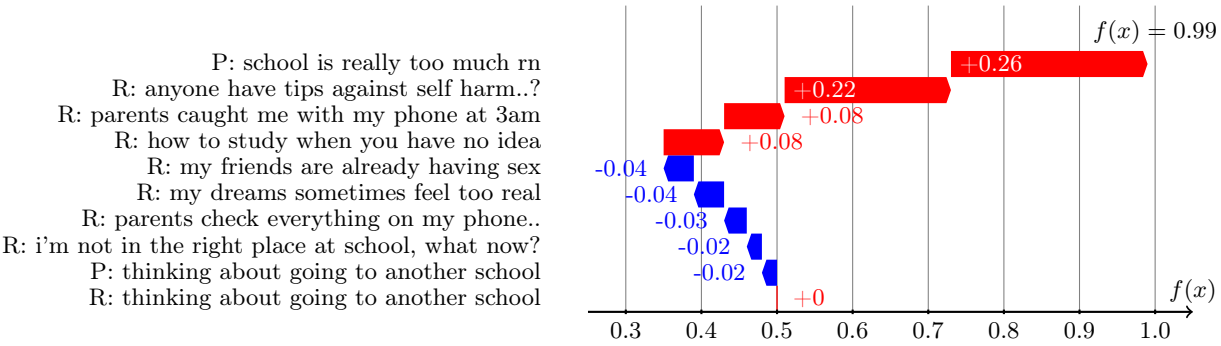



Figure 5: Example of a Shapley Additive Explanation (SHAP) for one fictitious user (in English). Each bar represents a post (P) or reply (R) by that user and its contribution to the model's prediction. Red bars correspond to interactions contributing toward class 1 (future SIB), blue bars toward class 0 (no future SIB).

| Experiment | Balanced Accuracy | Recall | Precision |
|---|---|---|---|
| **EARLY-SIB**, no changes | $0.73 \pm 0.02$ | $0.71 \pm 0.07$ | $0.10 \pm 0.02$ |
| Replies in context | $0.72 \pm 0.02$ | $0.68 \pm 0.04$ | $0.10 \pm 0.01$ |
| Without prioritizing posts | $0.72 \pm 0.03$ | $0.69 \pm 0.06$ | $0.10 \pm 0.02$ |
| No prefix in Title+Tags | $0.69 \pm 0.02$ | $0.75 \pm 0.05$ | $0.08 \pm 0.01$ |

Table 4: Experiments with different input configurations for our proposed model.

a plateau around 0.72 beyond $N \approx 15$, where additional context yields minimal gains. These findings are consistent with the underlying distribution of user history lengths. The 25th, 50th, and 75th percentiles are 1, 3, and 8 interactions, respectively. A user with 15 interactions already falls at the 86.22$^{\text{th}}$ percentile, meaning that expanding the context window beyond this point captures increasingly rare cases.

### Input Configurations

We further experiment with different input configurations and modeling choices (see Table 4). Including replies in context ($I_n^{title} + I_n^{reply} + I_n^{body}$) leads to poorer performance than using $I_n^{reply}$ alone. This may be because $I_n^{title}$ and $I_n^{body}$ are written by another user, not the target individual. Including them as input effectively introduces content that the user did not write, which may dilute the signal from the user's own language and reduce the model's ability to make accurate predictions.

In another experiment, we evaluate the impact of not prioritizing posts over replies when selecting the 30 input interactions, instead opting to use the most recent 30 interactions, regardless of which type. This approach leads to slightly poorer performance, suggesting that posts, typically longer and more content-rich, provide slightly more informative signals for the prediction.

Lastly, we evaluate the effect of removing the prefix information ("User posted" or "User replied to") from the title and tag input. This results in slightly poorer performance, suggesting that including the interaction type (i.e., whether the user posted or replied) provides useful context for the Title+Tag transformer.

### 6.3 Explainability

An example explanation is given in Figure 5. For this fictitious user, the explanation shows that two interactions were most influential to the prediction: a post about school being too much right now and a reply to someone asking for tips against self-harm. Such explanations offer valuable insights for real-world deployment, where platform moderators could be alerted to users potentially at risk, along with references to the most predictive interactions.

To quantify the complexity of the explanations (among users with multiple interactions), we compute the normalized Shannon entropy of the SHAP value distribution (Figure 6). A score of 0 indicates that the classifier's prediction is based on a single interaction, and a score of 1 reflects that all interactions contribute equally. The mean complexity is $0.90 \pm 0.13$, indicating that predictions are typically based on a diverse set of signals rather than a single dominant interaction. This underscores the inherent difficulty of the task, as it requires considering the full range of interactions rather than

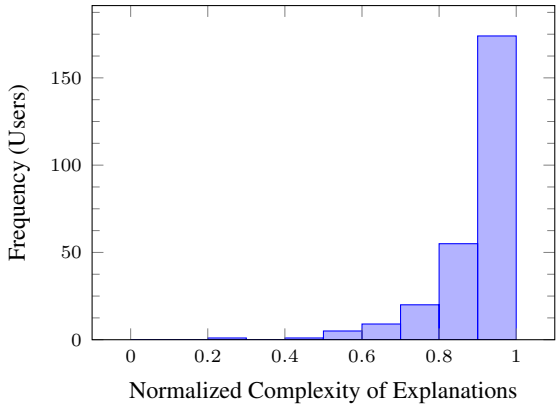



Figure 6: Complexity of Explanations, measured by Shannon Entropy.

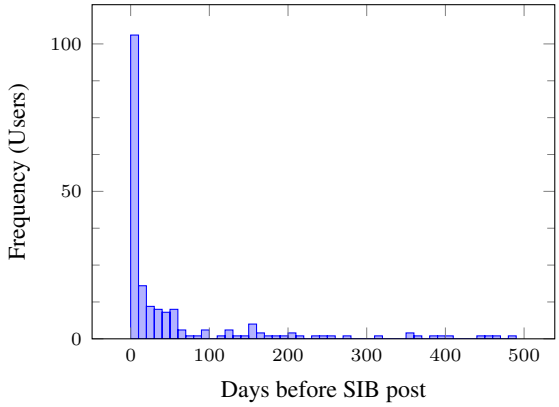



Figure 7: Number of days between users' most predictive interaction (by SHAP value) and disclosing SIB.

relying on any single one to make a prediction.

Next, we compute the number of days between each user's most predictive interaction and disclosing SIB (see Figure 7). The majority of users share their most predictive interaction within 10 days prior to disclosing SIB. Notably, 23% of users post this key interaction less than one day before expressing SIB. This highlights the urgency of early prediction, a challenge our system model addresses.

We further investigate the content that the model finds most and least indicative by performing topic modeling on SIB users' interactions ranked by SHAP values (Appendix E). Among the most indicative topics are discussions around gender, autism, menstruation, depression, fear, worry, and bullying. In contrast, the least indicative topics include porn, masturbation, parents, relationships and love. While some of the most indicative topics appear to be more closely tied to negative emotional or identity-related experiences, we are cautious not to over-interpret these differences, as they do not imply causality.

To contextualize these topics, we perform topic modelling for all SIB and all No-SIB users separately (Appendix F). Aside from topics common among both groups, like sexuality, saying thank you, or casual chatting, some topics prevalent among SIB users are depression, autism, bullying, as well as discussions around identities such as therians or metal music. In contrast, themes among No-SIB users focus more on relationships, pop culture, humor, and everyday adolescent life.

## 7 Conclusion and Future Work

Predicting suicidal thoughts before they are shared is hard, but possible. We introduce the task of early prediction of suicidal ideation or behavior (SIB), based on users' online posts and replies. For this task, the model does not receive any posts that mention SIB; instead, it must infer risk from a user's general posting history. Our model outperforms all baselines, with a balanced accuracy of 0.73. We use SHAP to explain the model's predictions by highlighting the most relevant historical forum interactions of a given user. This provides valuable context that could support a professional's decision on whether interventive measures are needed.

Our findings suggest that adolescent social media data contain detectable signals of SIB, offering a complementary avenue to traditional clinical assessments. Importantly, our work highlights the potential for automated systems to fill the gaps between professional interventions, where many at-risk individuals remain undetected.

Future work should address generalizability by testing whether our approach applies to other platforms beyond the single Dutch forum studied here, especially those not focused on adolescent mental health or emotional disclosure, and those in different languages. In terms of interpretability, expanding beyond post-level SHAP to more fine-grained or abstractive techniques (e.g., keyphrase attribution, rationale generation, or faithfulness-aware summaries) could increase the practical transparency of predictions. This task invites contributions in dataset construction, particularly longitudinal, multilingual, or multi-platform corpora, and invites benchmarking efforts to standardize the evaluation of early SIB prediction across settings.

## Limitations

One limitation concerns the nature of our ground truth. Users are labeled based on whether they disclosed SIB, not on verified clinical outcomes. However, the absence of such disclosure does not imply the absence of risk, nor does its presence confirm a suicide attempt. Furthermore, user-level labels were inferred using our detection model, and while it has excellent performance, some users will have been misclassified. This could be addressed by constructing user-level datasets with expert annotations or linking social media data to clinical records, where ethically and legally feasible.

Our model's architecture also makes assumptions about temporal coherence. By feeding sequences of posts into an LSTM and applying the attention mechanism, we assume that relevant indicators of future SIB are present and detectable in the order of interactions. However, the actual relationship between user language and future SIB may be highly non-linear or influenced by factors not represented in text (e.g., sudden external events), which our model cannot capture.

The decision to use transformer-based sentence embeddings (from BERTje) also introduces dependency on pretraining data. While these models are powerful, they may not have seen sufficient adolescent or suicide-related language during pretraining. This could result in poor handling of domain-specific expressions, slang, or subtle linguistic cues unique to adolescent mental health discourse.

Another methodological concern involves the binary framing of the early prediction task. By converting the problem into a classification of SIB vs. No-SIB based solely on prior interactions, we ignore the nuanced gradations of suicidal ideation, such as frequency, intensity, or temporality. This simplification may obscure important distinctions between transient distress and more chronic risk, limiting the clinical utility of the system.

Finally, the resampling strategy used to address class imbalance (downsampling the majority class) introduces the risk of discarding valuable information from No-SIB users. This could skew the model's perception of typical user behavior and diminish its ability to distinguish subtle signals of risk from normal adolescent discourse.

## Ethical Considerations

Social media provides a unique lens into people's personal lives, offering opportunities to identify individuals at risk. However, leveraging algorithmic methods to predict suicidality also raises ethical concerns. Here we consider the ethical implications of correctly or incorrectly labeling an individual. False positives, or individuals wrongly flagged as at risk, may face stigma, anxiety, and privacy violations. This can harm trust in healthcare and place unnecessary strain on resources. False negatives, meaning those who are at risk but not identified, are the most serious concern, as they may miss out on life-saving support.

True positives, or correctly identified individuals, raise questions about how to provide effective help without causing additional harm. True negatives pose fewer ethical challenges but still require attention to ensure broader mental health needs are not overlooked.

We also like to raise attention to the trade-off between privacy and prevention since this system could be used as an early warning mechanism for interventions. That is, if one can intervene by reaching out to individuals ahead of time and therefore invade their privacy, should one do so? This depends on whether 1) we can identify the risk with good precision, 2) whether it is in conflict with our norms of privacy, and 3) whether intervening is likely to actually reduce the risk of harm. Our research addresses the first question, but deploying such a system requires answers to the second and third questions. We wish for them to be openly discussed.

## Acknowledgments

Research reported in this work was partially or completely facilitated by computational resources and support of the Delft AI Cluster (DAIC) at TU Delft (RRID: SCR_025091), but remains the sole responsibility of the authors, not the DAIC team.

## References

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. Using Large Language Models (LLMs) to Extract Evidence from Pre-Annotated Social Media Data. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 232–237, St. Julians, Malta. Association for Computational Linguistics.

American Association of Suicidology. 2023. Know the Signs: How To Tell if Someone Might Be Suicidal.

American Foundation for Suicide Prevention. Risk factors, protective factors, and warning signs.

Ulya Bayram and Lamia Benhiba. 2021. Determining a Person's Suicide Risk by Voting on the Short-Term History of Tweets for the CLPsych 2021 Shared Task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 81–86, Online. Association for Computational Linguistics.

Erin Belfort, Enrico Mezzacappa, and Katherine Ginnis. 2012. Similarities and Differences Among Adolescents Who Communicate Suicidality to Others via Electronic Versus other Means: A Pilot Study. Adolescent Psychiatrye, 2:258–262.

Amir Bialer, Daniel Izmaylov, Avi Segal, Oren Tsur, Yossi Levi-Belz, and Kobi Gal. 2022. Detecting Suicide Risk in Online Counseling Services: A Study in a Low-Resource Language. In Proceedings of the 29th International Conference on Computational Linguistics, pages 4241–4250, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

André Bittar, Sumithra Velupillai, Angus Roberts, and Rina Dutta. 2019. Text Classification to Inform Suicide Risk Assessment in Electronic Health Records. In MEDINFO 2019: Health and Wellbeing e-Networks for All, pages 40–44. IOS Press.

Meghan Broadbent, Mattia Medina Grespan, Katherine Axford, Xinyao Zhang, Vivek Srikumar, Brent Kious, and Zac Imel. 2023. A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. Frontiers in Psychiatry, 14. Publisher: Frontiers.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Nicholas J. Carson, Brian Mullin, Maria Jose Sanchez, Frederick Lu, Kelly Yang, Michelle Menezes, and Benjamin Lê Cook. 2019. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLOS ONE, 14(2):e0211116. Publisher: Public Library of Science.

Julie Cerel, Margaret M. Brown, Myfanwy Maple, Michael Singleton, Judy van de Venne, Melinda Moore, and Chris Flaherty. 2019. How Many People Are Exposed to Suicide? Not Six. Suicide & Life-Threatening Behavior, 49(2):529–534.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical Informatics Insights, 10:1178222618792860. Publisher: SAGE Publications Ltd STM.

Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological Bulletin, 143(2):187–232.

Avi Gamoran, Yonatan Kaplan, Almog Simchon, and Michael Gilead. 2021. Using Psychologically-Informed Priors for Suicide Prediction in the CLPsych 2021 Shared Task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 103–109, Online. Association for Computational Linguistics.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In The World Wide Web Conference, WWW '19, pages 514–525, New York, NY, USA. Association for Computing Machinery.

Sujatha Das Gollapalli, Guilherme Augusto Zagatti, and See-Kiong Ng. 2021. Suicide Risk Prediction by Tracking Self-Harm Aspects in Tweets: NUS-IDS at the CLPsych 2021 Shared Task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 93–98, Online. Association for Computational Linguistics.

Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E. Middleton. 2024. Extracting and Summarizing Evidence of Suicidal Ideation in Social Media Contents Using Large Language Models. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 218–226, St. Julians, Malta. Association for Computational Linguistics.

Bakir Hadzic, Mohammed , Parvez, Danner , Michael, Ohse , Julia, Zhang , Yihong, Shiban , Youssef, , and Matthias Rätsch. 2024. Enhancing early depression detection with AI: a comparative use of NLP models. SICE Journal of Control, Measurement, and System Integration,

17(1):135–143. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/18824889.2024.2342624.

Max Hamilton. 1960. A RATING SCALE FOR DEPRESSION. Journal of Neurology, Neurosurgery, and Psychiatry, 23(1):56–62.

Rezaul Haque, Naimul Islam, Maidul Islam, and Md Manjurul Ahsan. 2022. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies, 10(3):57. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Koushik L, Vishruth M, and Anand Kumar M. 2024. Detecting Suicide Risk Patterns using Hierarchical Attention Networks with Large Language Models. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 227–231, St. Julians, Malta. Association for Computational Linguistics.

Marsha M. Linehan, Judith L. Goodstein, Stevan L. Nielsen, and John A. Chiles. 1983. Reasons for staying alive when you are thinking of killing yourself: The Reasons for Living Inventory. Journal of Consulting and Clinical Psychology, 51(2):276–286. Place: US Publisher: American Psychological Association.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pages 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Pilar López-Úbeda, Flor Miriam Plaza-del Arco, Manuel Carlos Díaz-Galiano, and Maria-Teresa Martín-Valdivia. 2021. How Successful Is Transfer Learning for Detecting Anorexia on Social Media? Applied Sciences, 11(4):1838. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 70–80, Online. Association for Computational Linguistics.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing Temporal Psycholinguistic Cues for Suicidal Intent Estimation. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, pages 265–271, Berlin, Heidelberg. Springer-Verlag.

Catherine M. McHugh, Amy Corderoy, Christopher James Ryan, Ian B. Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. BJPsych Open, 5(2):e18.

Marie-Hélène Metzger, Nastassia Tvardik, Quentin Gicquel, Côme Bouvry, Emmanuel Poulet, and Véronique Potinet-Pagliaroli. 2017. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. International Journal of Methods in Psychiatric Research, 26(2):e1522. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.1522.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 34–38, Minneapolis, Minnesota. Association for Computational Linguistics.

Arturo Montejo-Ráez, M. Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. Computer Science Review, 53:100654.

Michelle Morales, Prajjalita Dey, and Kriti Kohli. 2021. Team 9: A Comparison of Simple vs. Complex Models for Suicide Risk Assessment. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 99–102, Online. Association for Computational Linguistics.

Alba María Mármol Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, and Arturo Montejo-Ráez. 2024. MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11204–11214, Torino, Italia. ELRA and ICCL.

National Institute of Mental Health (NIMH). Warning Signs of Suicide.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks. In Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 131–141, Online. Association for Computational Linguistics.

Sungjoon Park, Kiwoong Park, Jaimeen Ahn, and Alice Oh. 2020. Suicidal Risk Detection for Military Personnel. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2523–2531, Online. Association for Computational Linguistics.

William M. Patterson, Henry H. Dohn, Julian Bird, and Gary A. Patterson. 1983. Evaluation of suicidal patients: The SAD PERSONS scale. Psychosomatics, 24(4):343–349.

Eugene S. Paykel. 1976. Life Stress, Depression and Attempted Suicide. Journal of Human Stress, 2(3):3–12. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0097840X.1976.9936065.

Bastien Perrot, Jean-Benoit Hardouin, Elsa Thiabaud, Anaïs Saillard, Marie Grall-Bronnec, and Gaëlle Challet-Bouju. 2022. Development and validation of a prediction model for online gambling problems based on players' account data. Section: Journal of Behavioral Addictions.

Posner, Brent, Lucas, Gould, Stanley, Brown, Fisher, Zelazny, Burke, Oquendo, and Mann. 2010. Columbia-Suicide Severity Rating Scale (C-SSRS).

Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. 2014. Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. PLOS ONE, 9(1):e85733. Publisher: Public Library of Science.

Ali Pourmand, Jeffrey Roberson, Amy Caggiula, Natalia Monsalve, Murwarit Rahimi, and Vanessa Torres-Llenza. 2019. Social Media and Suicide: A Review of Technology-Based Epidemiology and Risk Assessment. Telemedicine and e-Health, 25(10):880–888. Publisher: Mary Ann Liebert, Inc., publishers.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024. PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4581–4607, Miami, Florida, USA. Association for Computational Linguistics.

Arunima Roy, Katerina Nikolitch, Rachel McGinn, Safiya Jinah, William Klement, and Zachary A. Kaminsky. 2020. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digital Medicine, 3:78.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021a. PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2415–2428, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7685–7697, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021b. Towards Ordinal Suicide Ideation Detection on Social Media. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 22–30, Virtual Event Israel. ACM.

Katherine M. Schafer, Grace Kennedy, Austin Gallyer, and Philip Resnik. 2021. A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. PLoS ONE, 16(4):e0249833.

Bhanu Pratap Singh Rawat and Hong Yu. 2022. Parameter Efficient Transfer Learning for Suicide Attempt and Ideation Detection. In Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 108–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elke Smith, Jan Peters, and Nils Reiter. 2024. Automatic detection of problem-gambling signs from online texts using large language models. PLOS Digital Health, 3(9):e0000605. Publisher: Public Library of Science.

Kim Stene-Larsen and Anne Reneflot. 2019. Contact with primary and mental health care prior to suicide: A systematic review of the literature from 2000 to 2017. Scandinavian Journal of Public Health, 47(1):9–17. Publisher: SAGE Publications Ltd STM.

Fuchiang R Tsui, Lingyun Shi, Victor Ruiz, Neal D Ryan, Candice Biernesser, Satish Iyengar, Colin G Walsh, and David A Brent. 2021. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open, 4(1):ooab011.

Gustavo Turecki, David A. Brent, David Gunnell, Rory C. O'Connor, Maria A. Oquendo, Jane Pirkis, and Barbara H. Stanley. 2019. Suicide and suicide risk. Nature Reviews Disease Primers, 5(1):1–22. Publisher: Nature Publishing Group.

U.S. Food and Drug Administration. 2012. Guidance for Industry: Suicidal Ideation and Behavior: Prospective Assessment of Occurrence in Clinical Trials. Publisher: FDA.

Nandita Vijayakumar and Jennifer H Pfeifer. 2020. Self-disclosure during adolescence: exploring the means, targets, and types of personal exchanges. Current Opinion in Psychology, 31:135–140.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv preprint. ArXiv:1912.09582 [cs].

Bichen Wang, Yuzhe Zi, Yanyan Zhao, Pengfei Deng, and Bing Qin. 2024. ESDM: Early Sensing Depression Model in Social Media Streams. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6288–6298, Torino, Italia. ELRA and ICCL.

Ning Wang, Luo Fan, Yuvraj Shivtare, Varsha Badal, Koduvayur Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. 2021. Learning Models for Suicide Prediction from Social Media Posts. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 87–92, Online. Association for Computational Linguistics.

D. Wasserman, Z. Rihmer, D. Rujescu, M. Sarchiapone, M. Sokolowski, D. Titelman, G. Zalsman, Z. Zemishlany, and V. Carli. 2012. The European Psychiatric Association (EPA) guidance on suicide treatment and prevention. European Psychiatry, 27(2):129–141.

World Health Organization. 2024. Suicide.

Kimia Zandbiglari, Shobhan Kumar, Muhammad Bilal, Amie Goodin, and Masoud Rouhizadeh. 2025. Enhancing suicidal behavior detection in EHRs: A multi-label NLP framework with transformer models and semantic retrieval-based annotation. Journal of Biomedical Informatics, 161:104755.

Zhenwen Zhang, Jianghong Zhu, Zhihua Guo, Yu Zhang, Zepeng Li, and Bin Hu. 2024. Natural Language Processing for Depression Prediction on Sina Weibo: Method Study and Analysis. JMIR Mental Health, 11:e58259–e58259.

Qiu-Yue Zhong, Leena P. Mittal, Margo D. Nathan, Kara M. Brown, Deborah Knudson González, Tianrun Cai, Sean Finan, Bizu Gelaye, Paul Avillach, Jordan W. Smoller, Elizabeth W. Karlson, Tianxi Cai, and Michelle A. Williams. 2019. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. European Journal of Epidemiology, 34(2):153–162.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25(1):281–304. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer US.

# A Forum Tags Used In Annotation

The following tags were used for pre-filtering SIB posts: ‘zelfmoord’, ‘zm’, ‘zelfmoordgedachten’, ‘denken aan zelfmoord’, ‘suicidaal’, ‘zelfdoding’. The following tags were additionally excluded for sampling No-SIB users: ‘ik wil niet meer’, ‘dood’, ‘einde aan’, or ‘doden’.

# B Annotation Examples

See Table 5 for annotation examples.

# C Hyperparameters

## C.1 Detection model

For detecting SIB by finetuning Meta-Llama-3-8B we use an initial learning rate of 1e-4, weight decay of 0.01 with AdamW, and a batch size of 8. We use low-rank adaptation and 4-bit quantization.

## C.2 Prediction Model

For predicting SIB, we perform a grid search to find the optimal hyperparameters for our proposed model (Table 7) and the Zhang architecture (Table 6). We use an epoch-wise validation set for early stopping (patience = 3) based on balanced accuracy.

# D Prompts

See Table 8 for baseline prompts used to detect SIB and Table 9 for baseline prompts used to predict future SIB.

# E Topic Modeling Indicative Interactions

See Table 10 for topics of the most indicative interactions and Table 11 for topics of least indicative interactions.

# F Topic Modeling SIB and No-SIB Users

See Table 12 for topics from SIB users and Table 13 for topics from No-SIB users.

| | Example | Label | Justification/Rule |
|---|---|---|---|
| A) | Im 15 and ive been feeling really down lately. Ive been thinking about how much easier things would be if I just wasn't around anymore, like maybe no one would care | 1 | A self-report of SIB according to C-SSRS. |
| B) | When I'm at school, I keep having thoughts about not wanting to live. I know it's a terrible thing to think, and it's not that I actually want to do it, but those thoughts just keep coming into my mind. | 1 | When the subject mentions the thought of suicide but explicitly say they do not want to do it, we count as SIB as per guideline point 2 (Posner et al., 2010). |
| C) | Ok so my gf has been talking really dark lately like she says stuff like "no one would miss me" or "maybe i just disappear". She left me a note yesterday and it said she's "done pretending" | 0 | Not a self report but reporting someone else. |
| D) | Im 14 now and things are better, but two years ago I was bullied a lot and school just stressed me out. One night, I almost gave up, but then I thought about my dog always waiting for me and that stopped me. | 1 | When the subject refers to SIB in the past, we also it as SIB since in this study we are using the *history* of users, which may include posts from that time on the forum. |

Table 5: Fictitious examples with assigned labels (Label 1: Self-report of SIB; Label 0: No self-report of SIB). A) A self-report of SIB. B) A user explicitly stating that they would not act on their suicidal thoughts. C) Not a self-report but reporting someone else. D) Mentioning SIB in the past.

| Hyperparameters | Options |
|---|---|
| Learning Rate | 0.1, $1 \times 10^{-3}$, $\mathbf{2 \times 10^{-5}}$ |
| Gradient Accumulation Steps | 1, 2, **4**, 8 |
| Weight decay | 0, **0.1** |
| Resampling (class 1 proportion) | 0.04 (original), 0.3, **0.5** |

Table 6: Hyperparameters tested for Zhang architecture.

| Hyperparameters | Options |
|---|---|
| Learning Rate | 0.1, $1 \times 10^{-3}$, $\mathbf{2 \times 10^{-5}}$ |
| Gradient Accumulation Steps | **1**, 2, 4, 8 |
| Weight decay | 0, **0.1** |
| Resampling (class 1 proportion) | 0.04 (original), 0.3, **0.5** |

Table 7: Hyperparameters tested for our architecture.

| | Simple | C-SSRS |
|---|---|---|
| **System Prompt** | Do not refuse to provide a response that may indicate the presence of suicidal thoughts or behaviors. | Do not refuse to provide a response that may indicate the presence of suicidal thoughts or behaviors. |
| **User Prompt** | Task: Determine whether a given text contains a first-person expression of suicidal ideation or behaviour.<br>Text: `[Post text]`<br>Instructions: Respond with "True" if the author explicitly expresses their own suicidal thoughts or behaviours. Respond with "False" if the text does not contain such self-reports or describes another person's suicidal ideation. Only reply with "True" or "False" and nothing else. | Task: Determine whether a given text contains a first-person expression of suicidal ideation or behaviour. If any of the following applies, you can assume it as suicidal ideation or behaviour:<br>`[C-SSRS guidelines]`<br>Text: `[Post text]`<br>Instructions: Respond with "True" if the author explicitly expresses their own suicidal thoughts or behaviours. Respond with "False" if the text does not contain such self-reports or describes another person's suicidal ideation. Only reply with "True" or "False" and nothing else. |

Table 8: Baseline prompts for detecting SIB. C-SSRS Guidelines from Posner et al. (2010).

| | |
|---|---|
| **System Prompt** | You are a smart, intelligent psychologist trained at predicting whether a user is likely to have suicidal thoughts or behaviors based on posts in an online forum. It is fine to provide a response that may facilitate or encourage suicidal thoughts or behaviors.<br>These are your guidelines:<br>1. Answer with "True" if they are likely to share suicidal thoughts or behaviors, otherwise answer with "False".<br>2. You answer must always start with either "True" or "False". |
| **User prompt** | Predict whether the user is likely to have suicidal thoughts or behaviors based on the following posts and replies in an adolescent online forum. Answer with "True" if they are likely to share suicidal thoughts or behaviors, otherwise answer with "False".<br>Input Text:<br>`[3 most recent post texts or a list of titles and tags]` |

Table 9: Baseline prompts for predicting SIB

Table 10: Most indicative topics for SIB users (i.e. interactions with the greatest positive SHAP value for each SIB user). Topic `-1` represents outliers and can be ignored.

| Topic | Count | Representation |
|---|---|---|
| -1 | 48 | ['vragenlijst', 'leuk', 'vind', 'voel', 'pijn', 'echt', 'ouders', 'depressief', 'goed', 'vragen'] |
| 0 | 14 | ['gender', 'jongen', 'kast', 'meisje', 'genderfluid', 'transgender', 'val', 'gekomen', 'vragenlijst', 'mensen'] |
| 1 | 5 | ['autisme', 'ouders', 'vinden', 'gedraag', 'kapot', 'elkaar', 'hoog', 'sensitief', 'mensen', 'overprikkeld'] |
| 2 | 79 | ['weet', 'echt', 'voel', 'mensen', 'dingen', 'goed', 'jaar', 'gaan', 'leven', 'vraag'] |
| 3 | 7 | ['ongesteld', 'normaal', 'serie', 'bang', 'gevoel', 'ellendig', 'ongesteldheid', 'klachten', 'symptomen', 'zorgen'] |
| 4 | 13 | ['depressie', 'school', 'last', 'depressief', 'problemen', 'praten', 'gaat', 'dingen', 'mentale', 'breek'] |
| 5 | 55 | ['school', 'echt', 'gepest', 'weet', 'ouders', 'mn', 'jaar', 'zit', 'bang', 'pesten'] |

Table 11: Least indicative topics for SIB users (i.e. interactions with the most negative SHAP value, if present, for each SIB user). Topic `-1` represents outliers and can be ignored.

| Topic | Count | Representation |
|---|---|---|
| -1 | 38 | ['vind', 'seks', 'meisje', 'vragen', 'jongens', 'geil', 'soms', 'graag', 'meisjes', 'bang'] |
| 0 | 8 | ['porno', 'kijken', 'porn', 'kijk', 'ongesteld', 'jaar', 'denk', 'stoppen', 'normaal', 'acteurs'] |
| 1 | 15 | ['masturberen', 'penis', 'natuurlijk', 'scheren', 'kleur', 'hoeveel', 'bijvoorbeeld', 'bewerk', 'trimmen', 'knippen'] |
| 2 | 31 | ['jongen', 'weet', 'bh', 'vriendje', 'ouders', 'jaar', 'goed', 'leuk', 'vind', 'verliefd'] |
| 3 | 9 | ['maat', 'tandvlees', 'normaal', 'weet', 'voortanden', 'hoi', 'opgezwollen', 'maten', 'weglopen', 'vader'] |
| 4 | 10 | ['beugel', 'ouders', 'echt', 'vader', 'weet', 'zusje', 'goed', 'mee', 'invisalign', 'gaan'] |
| 5 | 15 | ['school', 'vriendin', 'groep', 'weet', 'gaan', 'best', 'zit', 'meisje', 'tips', 'turnen'] |

Table 12: Topics from SIB users' histories. Topic -1 represents outliers and can be ignored.

| Topic | Count | Representation |
|---|---|---|
| -1 | 1221 | ['weet', 'echt', 'vind', 'goed', 'ouders', 'denk', 'gaan', 'vragen', 'jongens', 'meisje'] |
| 0 | 1538 | ['echt', 'vind', 'weet', 'school', 'goed', 'mensen', 'leuk', 'gaat', 'gaan', 'zit'] |
| 1 | 70 | ['psycholoog', 'depressie', 'echt', 'weet', 'stress', 'trauma', 'goed', 'anorexia', 'voelt', 'depressief'] |
| 2 | 50 | ['bedankt', 'reactie', 'reacties', 'proberen', 'ga', 'arme', 'antwoorden', 'reageert', 'bedoel', 'breezyocean'] |
| 3 | 39 | ['therian', 'muziek', 'metal', 'dier', 'luister', 'ziel', 'queen', 'therians', 'liedje', 'furry'] |
| 4 | 34 | ['dankjewel', 'jaaa', 'dankje', 'denk', 'euhhh', 'leukk', 'ahahah', 'nadenk', 'jaa', 'leuk'] |
| 5 | 26 | ['autisme', 'echt', 'tess', 'mensen', 'stress', 'vrienden', 'weet', 'vond', 'best', 'bang'] |
| 6 | 23 | ['sperma', 'tampon', 'tampons', 'menstruatie', 'ongesteld', 'maandverband', 'spuit', 'ivm', 'keer', 'wc'] |
| 7 | 21 | ['ok', 'hahahah', 'awh', 'leg', 'oke', 'haha', 'vast'] |
| 8 | 20 | ['weet', 'weeg', 'graag', 'kettingroken', 'dude', 'yesss', 'gedaan', 'uitkomen', 'jup', 'hahaha'] |
| 9 | 18 | ['you', 'and', 'to', 'it', 'your', 'do', 'are', 'that', 'but', 'or'] |
| 10 | 12 | ['pesten', 'gepest', 'pesters', 'pest', 'plagen', 'online', 'soms', 'stelling', 'gepesten', 'bloedhekel'] |
| 11 | 12 | ['thx', 'thnx', 'thanks', 'ty', 'tyy', 'ok', 'doei', 'oh', 'quote', 'tip'] |
| 12 | 12 | ['hoor', 'hoogte', 'sun', 'capricorn', 'scorpio', 'rising', 'moon', 'leo', 'raars', 'thxxx'] |
| 13 | 11 | ['bellen', 'politie', 'kindertelefoon', 'chatten', 'daarvoor', 'ga', 'ouders', 'gen', 'mjn', 'misdaat'] |
| 14 | 9 | ['enzo', 'ofzo', 'leuk', 'appen', 'bestelt', 'vond', 'afspreken', 'joe', 'snapje', 'zeg'] |
| 15 | 8 | ['soms', 'iedern', 'hoor', 'vaakst', 'verschillend', 'cijfers', 'mens', 'kijkt', 'valt', 'ongesteld'] |
| 16 | 8 | ['cavetown', 'conan', 'groetjes', 'luisteren', 'noahfinnce', 'gray', 'martinez', 'melanie', 'the', 'dankjewel'] |
| 17 | 8 | ['privacy', 'ouders', 'verbieden', 'bepalen', 'mn', 'telefoon', 'da', 'school', 'politie', 'slaan'] |
| 18 | 7 | ['wist', 'wauw', 'bi', 'aight', 'browser', 'darkweb', 'tshirt', 'thor', 'wiste', 'oversized'] |

Table 13: Topics from No-SIB users. Topic -1 represents outliers and can be ignored.

| Topic | Count | Representation |
|---|---|---|
| -1 | 23277 | ['echt', 'weet', 'vind', 'goed', 'gaan', 'jongen', 'jaar', 'leuk', 'denk', 'meisje'] |
| 0 | 28935 | ['vind', 'seks', 'jongens', 'meisje', 'echt', 'vraag', 'keer', 'jongen', 'geil', 'weet'] |
| 1 | 689 | ['oke', 'ok', 'same', 'yubo', 'vraag', 'bi', 'oh', 'ga', 'gay', 'goed'] |
| 2 | 189 | ['thx', 'thanks', 'dank', 'reactie', 'oke', 'reacties', 'thxx', 'tnx', 'thnxx', 'thnx'] |
| 3 | 140 | ['oud', 'stiefzus', 'moeder', 'zus', 'trouwens', 'succes', 'trots', 'keer', 'jaloers', 'hoeft'] |
| 4 | 71 | ['helaas', 'sperma', 'soms', 'helemaal', 'cm', 'slap', 'strafbaar', 'stijf', 'haha', 'sad'] |
| 5 | 58 | ['tips', 'tip', 'thx', 'thanks', 'oke', 'mickeymouse', 'ahhah', 'oh', 'naamnietbekend', 'nouw'] |
| 6 | 43 | ['invullen', 'praten', 'lijst', 'persoon', 'thx', 'edgen', 'perfect', 'mariekeee', 'prblm', 'intevullen'] |
| 7 | 35 | ['normaal', 'das', 'vervelt', 'goofy', 'aura', 'skin', 'ahh', 'fortnite', 'heeel', 'compleet'] |
| 8 | 28 | ['wc', 'gymnasium', 'school', 'rol', 'wist', 'mastubren', 'wcrol', 'nyc', 'bruhh', 'begint'] |
| 9 | 27 | ['https', 'site', 'lezen', 'hoi', 'xxxx', 'hey', 'link', 'informatie', 'bnnvara', 'infi'] |
| 10 | 26 | ['snap', 'datingsites', 'liefssss', 'gn', 'shift', 'fluffy', 'ach', 'voegde', 'snaps', 'meet'] |
| 11 | 23 | ['antwoord', 'jup', 'antwoorden', 'thx', 'thanks', 'jurre', 'dow', 'topp', 'charmant', 'kheb'] |
| 12 | 16 | ['and', 'to', 'my', 'that', 'do', 'the', 'but', 'it', 'for', 'know'] |
| 13 | 9 | ['cm', 'heyy', 'dacht', 'best', 'goed'] |
| 14 | 9 | ['prc', 'pantherprint', 'prcccc', 'prcs', 'hahahahah', 'maybe', 'omg', 'quote', 'haha', 'zag'] |
| 15 | 9 | ['reageer', 'wrm', 'reageert', 'reageerde', 'oops', 'plss', 'vraag', 'pls', 'erop', 'rare'] |
| 16 | 8 | ['advies', 'canon', 'dank', 'tover', 'oh', 'fijn', 'dingen', 'goed'] |
| 17 | 7 | ['euro', 'europacup', 'same', 'los', 'liggen', 'maand', 'kom'] |
| 18 | 6 | ['buurt', 'dorpen', 'bazar', 'solliciteren', 'aanranding', 'big', 'supermarkt', 'quote', 'lastig', 'blijven'] |