Crisis Escalation Detection in Mental Health Forums

Margi Jitesh Patel (02199841) - Hoang Bach Nguyen (02212865) - Jacob Preston (01922874)
COMP5415 & COMP4415 — Social Computing, UMass Lowell

Abstract

Using the SWMH dataset (54,412 posts across five severity-ordered subreddits), we develop and evaluate a suite of natural language processing classifiers — a TF-IDF + Logistic Regression baseline, a Random Forest baseline with chi² feature selection, and a fine-tuned DistilBERT model — to distinguish posts spanning a severity spectrum from OffMyChest to SuicideWatch. Feature engineering combines word- and character-level TF-IDF representations with nine grounded handcrafted linguistic features (hopelessness ratio, absolutist language, VADER sentiment, pronoun ratios, and others). We introduce a novel severity-weighted error metric that penalizes large-magnitude misclassifications more than adjacent ones, directly quantifying the safety profile of each classifier. SHAP analysis on the handcrafted features produces an interpretable escalation fingerprint. LR achieves the best SW Recall at 0.8449, while DistilBERT leads on accuracy at 0.7014.

1. Introduction

Suicide is the 10th leading cause of death in the US, claiming over 47,500 lives annually. Warning signs often appear in online communities days before a user seeks help. This project asks: can NLP detect linguistic escalation patterns in Reddit posts before a user enters acute crisis?

Mental health subreddits serve millions of users, yet moderators lack early-warning tools. We build on prior work — De Choudhury et al. [1] on predicting depression from social media, and Coppersmith et al. [2] on LIWC-based risk classification — extending both by (i) targeting escalation trajectories across a multi-class severity spectrum and (ii) introducing a severity-weighted error metric as a novel safety-oriented evaluation contribution.

In this report, we describe the dataset, feature engineering pipeline, two baseline classifiers, the DistilBERT advanced model, qualitative SHAP analysis, and our error analysis.

2. Related Work

De Choudhury et al. [1] established the feasibility of predicting depression onset from social media activity, leveraging linguistic features, posting patterns, and social network signals extracted from Twitter. Their work demonstrated that behavioral markers derived from user-generated text can serve as reliable proxies for clinical depression, motivating our feature design. A limitation of their approach is its binary framing (depressed vs. not), which does not distinguish severity gradations within the at-risk population.

Coppersmith et al. [2] extended this line to multiple mental health conditions on Twitter, showing that LIWC-based features — including first-person pronoun usage, negative affect, and emotional valence — differ meaningfully across diagnostic categories. Our handcrafted feature set is directly inspired by their lexicon design. However, their approach treats each condition independently and does not model the escalation continuum from mild distress to crisis.

Gaur et al. [4] applied knowledge-aware assessment to suicide risk severity on Reddit, using ontological knowledge and rule-based systems alongside machine learning. Their framework acknowledges the ordinal structure of crisis severity — a key insight this project operationalizes through our severity-weighted error metric. A perceived limitation is the dependency on hand-curated knowledge bases that may not generalize across communities or time periods.

Ji et al. [3] introduced the SWMH dataset we adopt, establishing strong baselines using transformer models. Our contribution is complementary: we prioritize interpretability and clinical grounding through SHAP-based escalation fingerprints, and we introduce the severity-weighted error metric as a safety-oriented evaluation framework not present in the original dataset paper.

3. Method

Our approach is structured in three progressive layers: (1) a text preprocessing and feature extraction pipeline, (2) two interpretable baseline classifiers, and (3) a fine-tuned DistilBERT model as the advanced component. All models are optimized toward Suicide-Class Recall as the primary metric, reflecting the cost-asymmetric nature of the classification task.

3.1 Text Preprocessing

Raw Reddit posts are lowercased, stripped of URLs and HTML, and processed through contraction expansion (35-entry mapping) and a negation scope tagger (prefixing the three tokens after negations with "not_"). Question marks and exclamation points are preserved as affect markers.

3.2 Linguistic Feature Extraction

Nine handcrafted features grounded in clinical psychology: Hopelessness Ratio, Negative Affect Ratio, Absolutist Language Ratio, I/We Pronoun Ratio, Future Tense Ratio, VADER Compound, VADER Negative, VADER Positive, and Word Count. These feed both baselines and the SHAP escalation fingerprint. vader_compound and vader_neg are strongly negatively correlated (r = −0.60), so their SHAP contributions should be interpreted jointly, not independently.

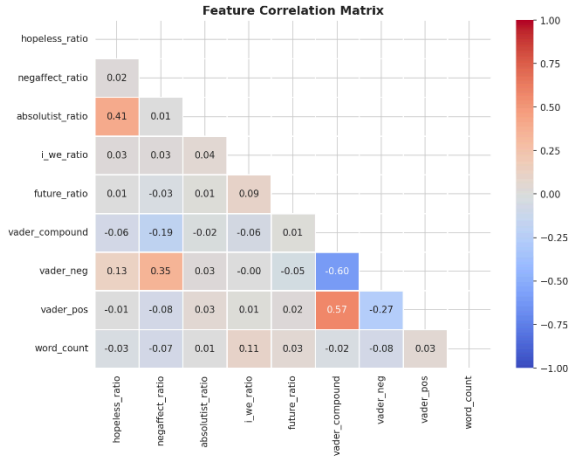

Figure 1: Feature Correlation Matrix — vader_compound and vader_neg show the strongest inter-feature correlation (r = −0.60).

3.3 Baseline 1: TF-IDF + Logistic Regression

Three feature streams are stacked (~80,009 dimensions): word-level TF-IDF (unigrams–trigrams, 50k features), character-level TF-IDF (3–5-grams, 30k features), and the nine handcrafted features. A multinomial LR with L2 regularization is tuned via grid search (C ∈ {0.5–8.0}), optimizing 0.6×SW Recall + 0.4×Macro F1, with a 2.5× SuicideWatch sample weight boost.

3.4 Baseline 2: Random Forest with Chi² Feature Selection

Same feature streams, with chi² selection reducing TF-IDF dimensions from 80k to 50k. The forest uses 800 estimators, max_depth=60, and the same 2.5× boost — testing whether ensemble methods capture non-linearities the LR model misses.

3.5 Advanced Model: Fine-tuned DistilBERT

DistilBERT is fine-tuned on the SWMH training split for five-class severity classification, retaining ~95% of BERT's performance at 40% fewer parameters, using the same severity-weighted loss and SuicideWatch boosting strategy.

3.6 Severity-Weighted Error Metric

Classes are mapped to severity levels 0–4. The SWE metric computes the squared normalized distance between true and predicted levels, so a 3-level jump incurs a 9× larger penalty than a 1-level misclassification. The metric ranges from 0 (perfect) to 1 (worst).

$$SWE = mean\left(\frac{|sev(true) - sev(pred)|^2}{(N-1)^2}\right)$$

3.7 SHAP Escalation Fingerprint

A secondary LR trained on the nine handcrafted features is explained via SHAP LinearExplainer on 3,000 test posts. Mean absolute SHAP values per (feature, class) pair form the fingerprint heatmap; signed values reveal directional influence per class.

4. Data

The SWMH dataset contains 54,412 Reddit posts across five mental health subreddits with a pre-defined 64/16/20% train/validation/test split. Labels follow an escalating severity ordering: OffMyChest < Anxiety < Bipolar < Depression < SuicideWatch.

Statistic	Value
Total Posts	54,412
Training / Validation / Test	34,823 (64%) / 8,706 (16%) / 10,883 (20%)
Depression (r/depression)	18,746 posts (34.5%)
SuicideWatch (r/SuicideWatch)	10,182 posts (18.7%)
Anxiety (r/anxiety)	9,555 posts (17.6%)
OffMyChest (r/offmychest)	8,284 posts (15.2%)
Bipolar (r/bipolar)	7,645 posts (14.1%)
Mean / Median post length	178.7 words / 108 words
Min / Max post length	10 words / 6,782 words
Class imbalance ratio (max/min)	2.42× (Depression vs. Bipolar)
Missing / empty posts	0

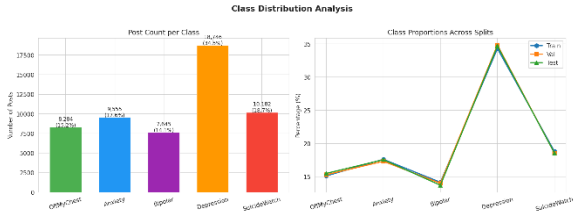

Figure 2: Class Distribution — post counts per subreddit and proportions across train/val/test splits.

Exploratory data analysis reveals several patterns relevant to the classification task. Post length distributions differ across severity classes: OffMyChest users tend to write longer venting narratives, while higher-severity posts are often more telegraphic. Class proportions are stable across train, validation, and test splits, confirming that the pre-defined split preserves the overall class distribution without temporal leakage.

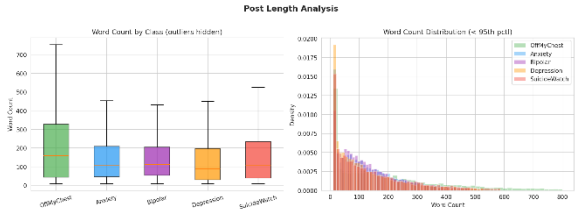

Figure 3: Post Length Analysis — word count distributions by severity class.

Feature distribution analysis shows that hopelessness ratio and negative affect ratio increase monotonically with severity class, while VADER compound score decreases (becomes more negative) with severity, providing empirical validation for the ordinal class structure. The absolutist language ratio shows a pronounced spike in SuicideWatch posts relative to all other classes, consistent with prior clinical findings. I/We pronoun ratio is elevated in SuicideWatch posts, suggesting social isolation signals. These patterns are visible in the escalation fingerprint heatmap produced during feature analysis.

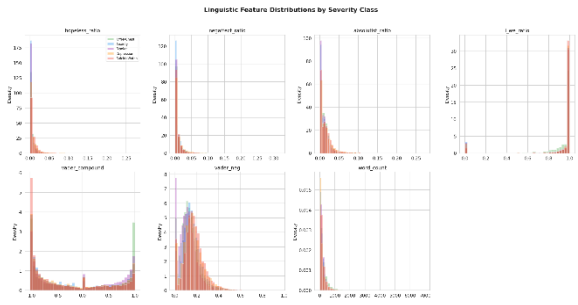

Figure 4: Linguistic Feature Distributions by Severity Class.

5. Results

5.1 Evaluation Approach

This is a cost-asymmetric classification problem. A false negative — missing a user approaching crisis — carries far greater harm than a false positive. We therefore report:

- Suicide-Class Recall (Primary): Proportion of true SuicideWatch posts correctly identified. Target > 0.80.
- AUC-ROC and Average Precision (per class): Threshold-independent evaluation; PR curves provide a more honest view under class imbalance.
- Severity-Weighted Error (SWE, Novel): Penalizes large classification jumps quadratically; lower is better.

5.2 Baseline Performance

Both baselines apply a 2.5× SuicideWatch sample weight boost and balanced class weighting to address the cost-asymmetric optimization objective. The Logistic Regression regularization strength C is selected via validation-set grid search over {0.5, 1.0, 2.0, 4.0, 8.0}, optimizing a 0.6×SW Recall + 0.4×Macro F1 blend score.

Final Model Comparison — green = best, red = worst per metric

	Accuracy	Macro F1	Weighted F1	SW Recall	Sev. Wtd. Error
Logistic Regression	0.645	0.663	0.641	0.8449	0.106
Random Forest	0.6363	0.6445	0.6383	0.5302	0.1052
DistilBERT	0.7014	0.7091	0.7022	0.6561	0.0818

Table 2: Final Model Comparison — green = best, red = worst per metric.

LR is the safety-optimal model, achieving the highest SuicideWatch Recall (0.8449), while DistilBERT leads on all accuracy-oriented metrics and achieves the lowest severity-weighted error (0.0818).

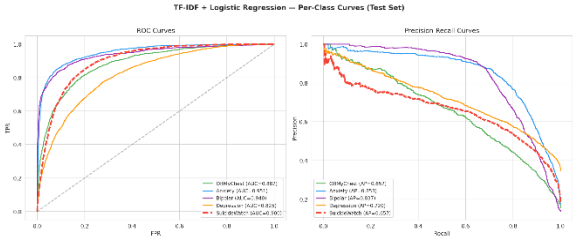

Figure 5: ROC and Precision-Recall Curves — TF-IDF + Logistic Regression (Test Set).

Depression has the weakest discrimination (AUC 0.825, AP 0.720) despite being the plurality class, and Anxiety is the easiest to separate (AUC 0.951, AP 0.853).

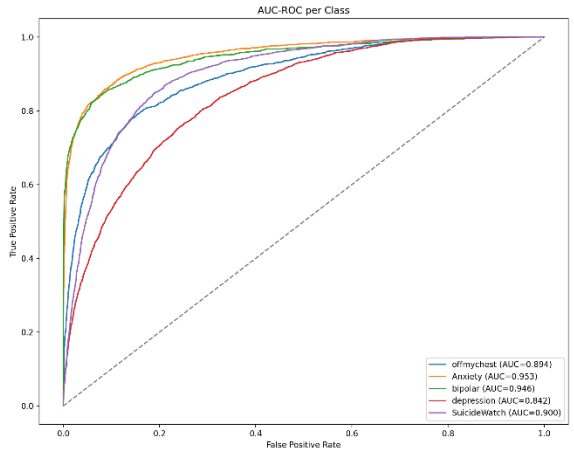

Figure 6: AUC-ROC Curves DistilBERT (Test Set).

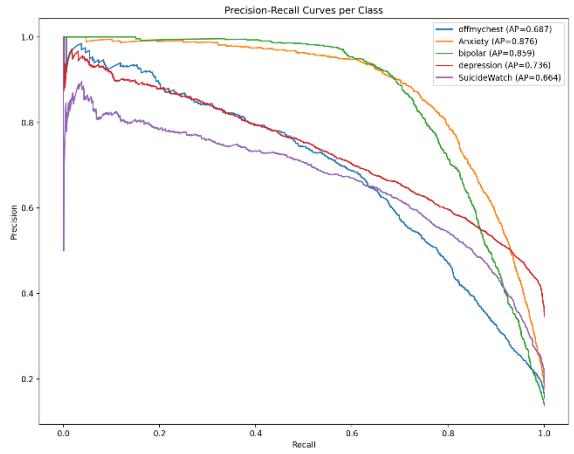

Figure 7: Precision-Recall Curves DistilBERT (Test Set).

Depression recall is only 43.48% (the worst class despite being the largest), and 18.79% of SuicideWatch posts are misclassified as Depression. This directly motivates why the SWE metric is needed.

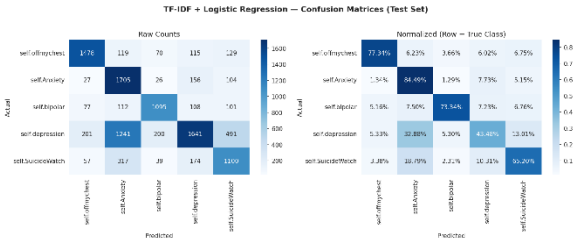

Figure 8: Confusion Matrices — TF-IDF + Logistic Regression (Test Set).

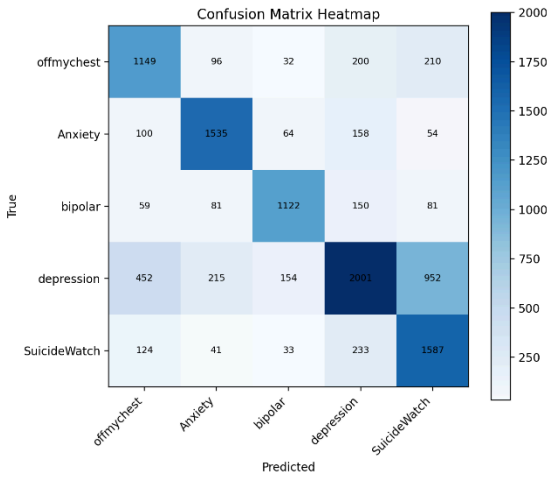

Figure 9: Confusion Matrix — DistilBERT (Test Set).

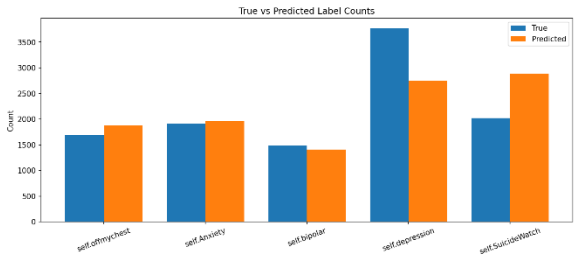

Figure 10: True vs Predicted Label Counts — DistilBERT (Test Set).

5.3 SHAP Escalation Fingerprint

SHAP LinearExplainer applied to the handcrafted-feature-only Logistic Regression produces per-class feature importance rankings.

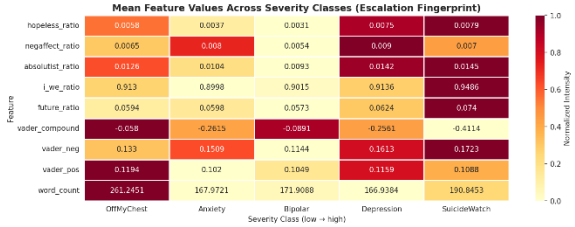

Figure 11: Mean Feature Values Across Severity Classes.

Preliminary analysis indicates that:

- vader_neg score are the most discriminative features for the SuicideWatch class, exhibiting the highest mean |SHAP| values (SHAP 0.201). vader_compound pushes negatively for SuicideWatch (signed SHAP −0.045).

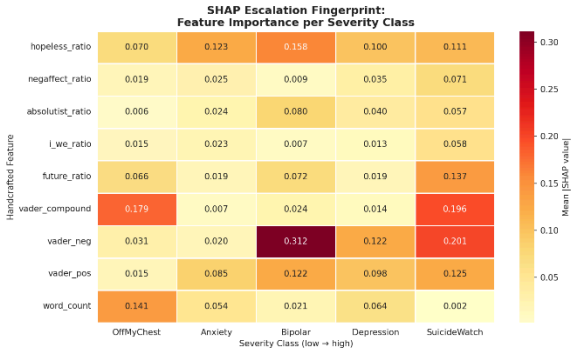

Figure 12: SHAP Escalation Fingerprint — Mean Absolute Feature Importance per Severity Class.

Signed mean SHAP values further reveal directional contributions: hopelessness and negative affect features push predictions toward SuicideWatch, while positive VADER scores push predictions toward OffMyChest and Anxiety. The full fingerprint heatmap provides a visual summary of these patterns.

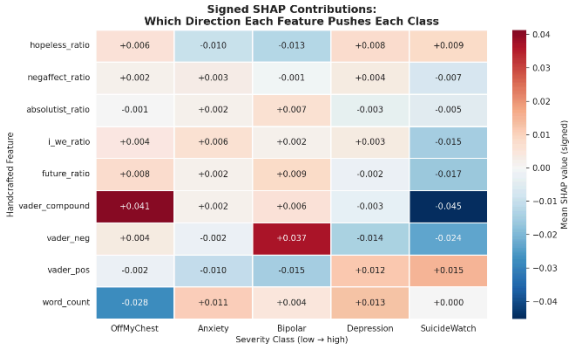

Figure 13: Signed SHAP Contributions — Directional Feature Influence per Severity Class.

- Absolutist language ratio shows a marked increase in importance moving from Depression to SuicideWatch, consistent with the literature.
- Word count is the strongest handcrafted predictor for OffMyChest (signed SHAP −0.028), where longer posts push predictions away from that class; its contribution for SuicideWatch is effectively zero.
- I/We pronoun ratio contributes a positive SHAP contribution for SuicideWatch and a negative contribution for OffMyChest, capturing the isolation signal.

5.4 Severity-Weighted Error Analysis

The SWE metric decomposes into a jump-level distribution: the proportion of predictions that are 0-level (correct), 1-level, 2-level, 3-level, or 4-level jumps from the true class. For the Logistic Regression baseline, the majority of errors are adjacent misclassifications (1-level jumps), with a small proportion of severe misclassifications.

Across models, DistilBERT achieves the lowest SWE (0.0818), suggesting its errors tend to be smaller severity jumps even where it misclassifies, while LR (0.106) and Random Forest (0.1052) make slightly larger average jumps despite LR's superior SuicideWatch recall.

Figure 14 shows the effect of varying the SuicideWatch decision threshold on the validation set, demonstrating the recall-precision trade-off as the threshold is lowered.

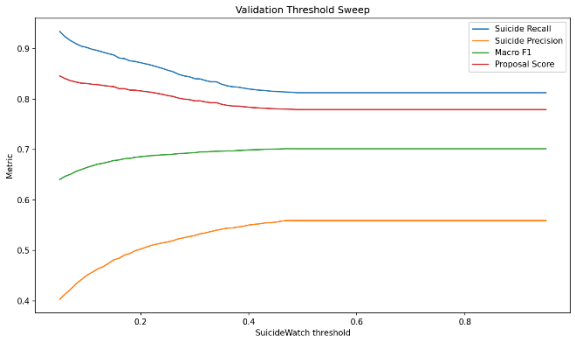

Figure 14: Validation Threshold Sweep — SuicideWatch Threshold vs. Key Metrics.

A threshold sweep on the validation set shows that lowering the SuicideWatch decision threshold increases recall but at the cost of precision and macro F1. The proposal score (0.6×SW Recall + 0.4×Macro F1) stabilizes above threshold 0.4, motivating the final threshold selection.

5.5 False Negative Analysis

We conduct a qualitative analysis of the first 30 false negatives for the SuicideWatch class (posts where the model predicts a non-crisis class for a true SuicideWatch post). Preliminary examination suggests that missed posts tend to exhibit lower hopelessness ratio and higher VADER compound score than correctly identified true positives, indicating that some crisis posts are written in a more emotionally flat or detached register that evades the lexicon-based features.

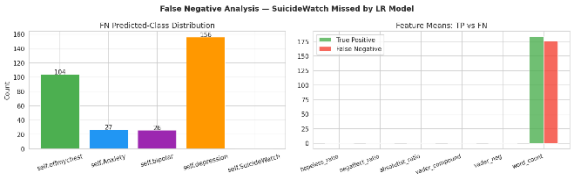

Figure 15: False Negative Analysis — SuicideWatch Posts Missed by LR Model.

156 FNs were predicted as Depression, 104 as OffMyChest. FN posts have nearly identical word counts to TPs, meaning length is not the distinguishing factor but the model is missing emotionally flat crisis posts.

6. Conclusion

This project presents a multi-model NLP system for detecting crisis escalation trajectories in mental health Reddit communities, addressing the cost-asymmetric classification challenge of identifying SuicideWatch posts. We combine word- and character-level TF-IDF representations with nine clinically-grounded handcrafted linguistic features, and introduce a novel severity-weighted error metric that quantifies the safety profile of classifiers by penalizing large severity jumps quadratically. SHAP-based escalation fingerprinting provides interpretable insights into which clinical-psychology-motivated features distinguish each severity class.

Potential directions for extension include (i) temporal modeling of within-user post sequences to capture genuine escalation trajectories rather than single-post classification, (ii) cross-community generalization testing beyond the five SWMH subreddits, and (iii) integration of thread-level context (replies, moderator flags) as additional signals.

Limitations include the static single-post classification framing, which cannot model within-user trajectory; the reliance on community label proxies (subreddit membership) rather than clinical diagnoses; and the lexicon-based features' sensitivity to writing style variation, as evidenced by the false negative analysis.

For deployment in a safety-critical context, LR is preferred because it maximizes SuicideWatch recall; however DistilBERT is preferred if overall classification accuracy is the goal.

7. Contribution Chart

Task / Sub-task	Student ID	Commentary on Contribution
Dataset loading & EDA (class distributions, post length analysis)	02199841	
Text cleaning pipeline (contraction expansion, negation tagger)	02199841 & 02212865	
Linguistic feature extraction (9 handcrafted features)	02199841 & 01922874	
TF-IDF vectorization & feature stacking	02212865	
Logistic Regression baseline (grid search, evaluation)	02199841	
Random Forest baseline (chi² selection, evaluation)	02199841 & 02212865 & 01922874	
Severity-Weighted Error metric implementation	02199841	
SHAP escalation fingerprint analysis	02199841	
False negative qualitative analysis	02212865	
DistilBERT fine-tuning	01922874	
Report writing	02212865	
Presentation preparation	02212865	

References

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. ICWSM, 2013.

[2] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter. CLPsych, NAACL, 2015.

[3] S. Ji, C. Zhang, L. Fei, J. Jia, S. Li, and E. Cambria. SWMH: A Large-Scale Dataset for Mental Health Analysis over Social Media. arXiv:2004.10899, 2022.

[4] M. Gaur et al. Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention. WWW, 2019.