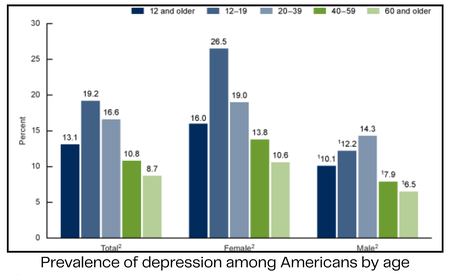
Depression Detection in Reddit Posts Using DistilBERT

Anh Huy Nguyen (anguyen6343@sdsu.edu), Thy Nguyen (tnguyen8119@sdsu.edu)
Course: CS 577 - Principles and Techniques of Data Science
Professor: Hajar Homayouni - Fall 2025
SAN DIEGO STATE UNIVERSITY

Overview

Problem

- About 1 in 7 youth aged 10–19 live with a mental health condition, with depression and anxiety among the most common, and suicide is now a leading cause of death in this group.
- In the US, roughly 20% of adolescents (12–17) experience at least one major depressive episode in a given year, yet a large fraction receive no formal treatment.



Prevalence of depression among Americans by age

Proposal

- Applies natural language processing (NLP) and machine learning to a Reddit dataset to detect passive signals of depression in user posts.

Research Goal

- Evaluate the accuracy and linguistic drivers of depression detection on Reddit by comparing traditional classifiers with transformer-based models.
- The findings assess the viability of these methods for digital mental-health screening tools.

Background

Dataset:

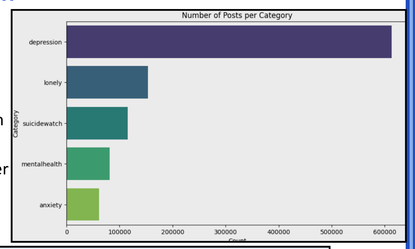
- Reddit Mental Health Dataset from Kaggle
- Number of records: About 1.3 million Reddit posts and comments (varies by version)
- Number of columns: about 6-8 columns.

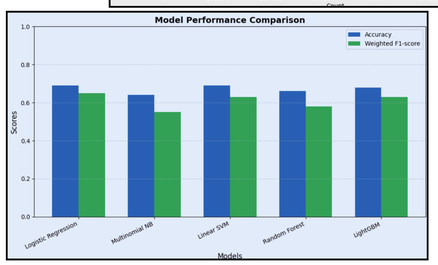
Columns	Description	Data Type
post_id	Unique ID of the Reddit post	String
subreddit	Name of the subreddit (e.g., r/depression, r/anxiety)	String
title	Title of the post	Text
body	Full post text	Text
author	Username	String
created_utc	The time the post was created	Datetime
label	Category (depressed, control, ...).	String

Column Data Types

Previous Research

The dataset has more than 3000 downloads after 12 months posted, but only one PTSD Detection work has been publicly released last month, while the number of depression-related posts is dominant.





Methods & Overall Pipeline

01 Data acquisition
Kaggle Reddit Mental Health dataset → raw monthly CSVs by subreddit

02 Data staging & cleaning
- Filter to relevant subreddits.
- Remove [deleted] / [removed], very short posts, duplicates.
- Build: Small 4-class "risk factor" labelled dataset. And large binary "depression vs not" dataset.

03 Author & time-aware splitting
Split into train / val / test with:
- No user leakage.
- Time-aware cut (train earlier, test later)

04 Text preprocessing
Lowercasing, remove URLs and usernames, normalize punctuation, no stopword removal for final pipeline (based on comparison).

05 Feature extraction (classical models)
- Word TF–IDF (1–2-grams)
- Optional character n-grams.

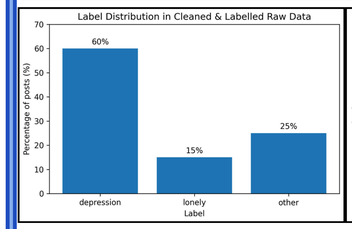
06 Classical models
- Logistic Regression (best baseline).
- Linear SVM (comparison).
- Hyperparameter tuning via cross-validation.

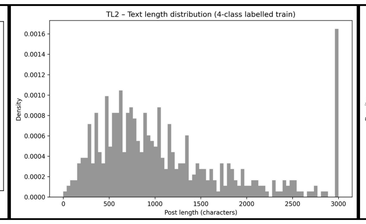
07 Transformer model
DistilBERT fine-tuned on GPU:
- Tokenization (max_len = 128 or 256),
- Batch size, learning rate,
- 1 epoch over ~800k posts.

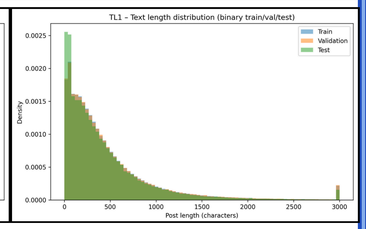
08 Evaluation
- Metrics: accuracy, macro-F1, ROC-AUC, PR-AUC.
- Confusion matrices.

Data Curation & Label Statistics

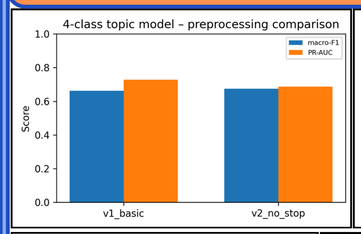
Dataset	Labels / Classes	Splits (rows) – Train / Val / Test	Label balance	Avg text length (chars)	Purpose in pipeline
4-class "Risk Factor" labelled set	4 classes: **Drug & Alcohol**, **Early Life**, **Personality**, **Trauma & Stress**	**479 / 160 / 160** (total 799)	~200 posts per class (almost perfectly balanced)	≈ **1,300** chars (train mean)	Small, expert-labelled subset used to prototype cleaning, text preprocessing variants (v1 vs v2_no_stop), and multi-class baselines.
Binary "Depression vs Not" Reddit posts	2 classes: **depression**, **not-depression**	**810,656 / 201,813 / 125,537** (≈ **1.14M** posts total)	Train ≈ **49%** dep, Val ≈ **50%** dep, Test ≈ **40%** dep (overall ~**49%** dep)	Train ≈ **477**, Val ≈ **479**, Test ≈ **438** chars	Large-scale dataset built from cleaned raw Reddit posts; used for final binary classifiers (TF-IDF + Logistic Regression/SVM, word+char models, and DistilBERT fine-tuning).

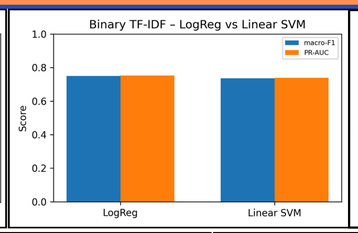


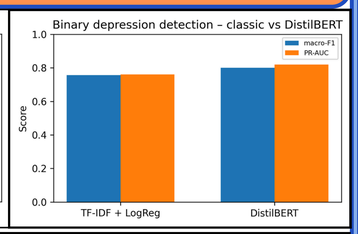


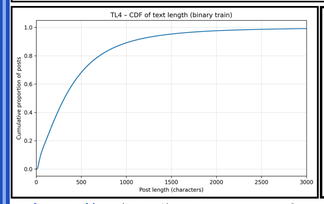


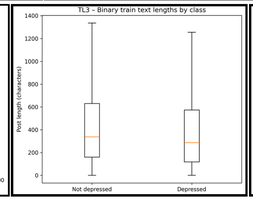
Classification Pipeline: TF-IDF Baselines and DistilBERT

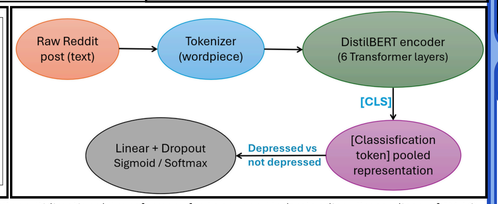












Reference: ***(I)*** *Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.* ***(II)*** *Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.* ***(III)*** *Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of EMNLP 2020: System Demonstrations. (Hugging Face Transformers library).* ***(IV)*** *Shao, H., Zhu, M., & Zhai, S. (2023). Mental Health Diagnosis in the Digital Age: Harnessing Sentiment Analysis on Social Media Platforms upon Ultra-Sparse Feature Content. arXiv:2311.05075. (Uses a Reddit mental-health dataset for classification.)* ***(V)*** *Kaggle. (2022). Reddit Mental Health Dataset (Depression, Anxiety, etc.). Kaggle Datasets. (Source of large-scale Reddit mental-health posts used, in related work; adapt to the exact dataset)*

Evaluation

Evaluation Goals:

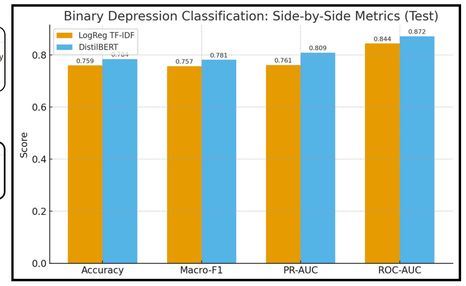
- Accurately detect posts from users likely experiencing depression.
- Prefer high recall / F1 for the depression class (missing someone in need is worse than a false alarm).
- Compare classical models vs DistilBERT and analyze where they still fail (error patterns).

Core classification metrics:

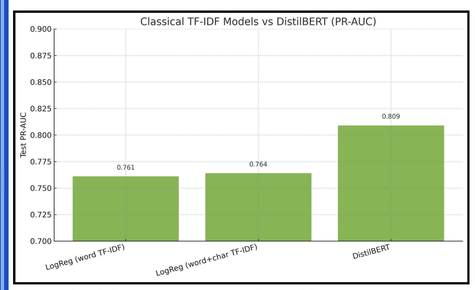
Accuracy	Precision (Depression)	Recall (Depression)
Overall % of posts the model classifies correctly.	When the model predicts "depressed", how often that prediction is actually true.	Of all truly depressed posts, how many the model correctly flags as "depressed".

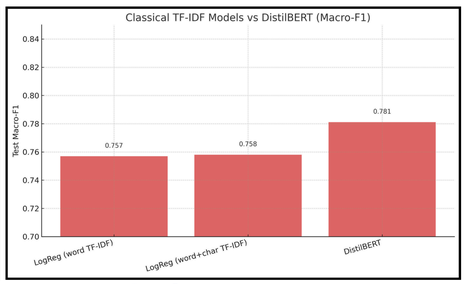
Advanced evaluation metrics:

Macro-F1	PR-AUC (Depression)	ROC-AUC
Single score balancing precision and recall across classes, treating each class equally.	How well the model separates depressed vs. not-depressed across all thresholds, focusing on the positive (depressed) class.	How well the model ranks depressed posts above non-depressed posts overall.

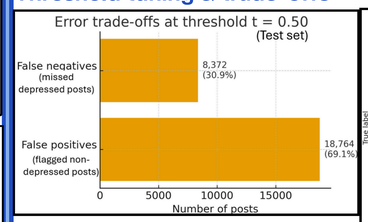


Overall model comparison:

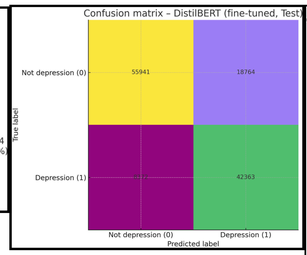


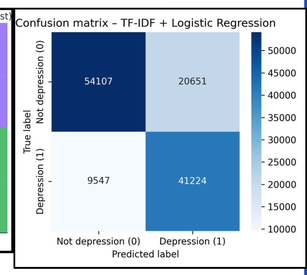


Threshold tuning & trade-offs

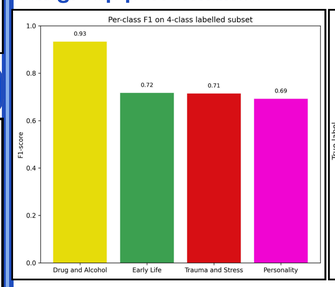


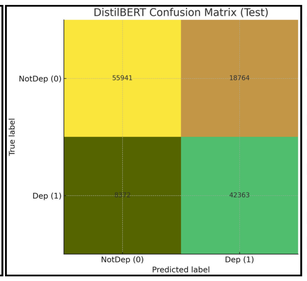
Error patterns: Confusion matrices





Per-class & subgroup performance





Evaluation Key Point

Best model: DistilBERT (fine-tuned on Reddit posts) is our best-performing model, slightly outperforming the TF-IDF + Logistic Regression baseline on macro-F1 and PR-AUC.

Biggest gain: Compared to classical TF-IDF baselines, DistilBERT gives better recall on depression posts, especially for harder borderline examples, while keeping "not-depressed" precision reasonably high.

Post snippet (shortened)	True label	Model prediction	Comment
depression bullsh t removed	Depressed	Depressed	Correct, clear depressive signal
collected ideas helped cope anxiety depression suffer anxiety depression ideas simple techniques really helped heal b...	Depressed	Not depressed	Missed subtle expression of depression
depression professionalism removed	Not depressed	Depressed	Over-sensitive to depressive keywords/context
anxiety nfndj	Not depressed	Not depressed	Correct, everyday stress but not depression

Conclusion

- Clean data + careful splits matter. Normalizing labels, removing duplicates/very short posts, and using author/time-aware splits gave more realistic, "deployment-like" performance estimates.
- Strong baselines, modest transformer gains. TF–IDF + Logistic Regression was already strong; fine-tuning DistilBERT gave consistent but modest improvements in macro-F1 / PR-AUC, especially for harder borderline depression posts.
- Errors reflect label overlap & subtle language. Most mistakes happened between conceptually similar labels (e.g., Early Life vs. Trauma and Stress) and on short, sarcastic, or context-poor posts where even humans may disagree.
- Useful as a screening tool, not a diagnosis. The model can help prioritize potentially at-risk Reddit posts when tuned for higher recall, but it should complement—never replace—professional clinical assessment.