Interim Report – NLP Final Project

Analyzing Mental Health: Text Classification for Mental Health Conditions

Course Lecturer: Dr. Alexander(Sasha) Apartsin

Shahar Saadon | 206560526
Dudi Saadia | 318970944
Shanel Asulin | 205616014

Project Description

Data

Task Type: Multi-class Text Classification
Input: Free-form short English texts (social media statements)
Output: One of 7 mental health labels
Labels: Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, Personality Disorder
Dataset: Combined labeled dataset from Kaggle (53,000+ rows)
Format: CSV with two columns: statement (text), status (label)

Challenges

- Informal and varied mental health expressions
- Ambiguous/self-diagnosed language
- Imbalanced classes (e.g., fewer suicidal statements)
- Label noise and missing data

Evaluation Metrics

- **Accuracy** – Correct predictions rate
- **Precision** – Correctness of positive classifications
- **Recall** – Sensitivity to each class
- **F1-score** – Harmonic mean of precision and recall
- **Confusion Matrix** – Error analysis

Literature Review

Key Contribution	Best Result	Methods	Dataset & Task	Study
TF-IDF text features, simple yet effective	**77%** (LightGBM)	Naive Bayes, MLP, LightGBM	10K Reddit posts (6 conditions)	**Nova (2023)**
Multi-model benchmark with deep/transfer learning	**83%** (RoBERTa)	Traditional ML, DL, **RoBERTa**	17K Reddit posts (5 classes)	**Ameer et al. (2022)**
Detection from **general text**, not only support forums	**81%** (RoBERTa)	BERT, XLNet, RoBERTa	100K Reddit posts (9 DSM-5 conditions)	**Dinu & Moldovan (2021)**

Work Plan

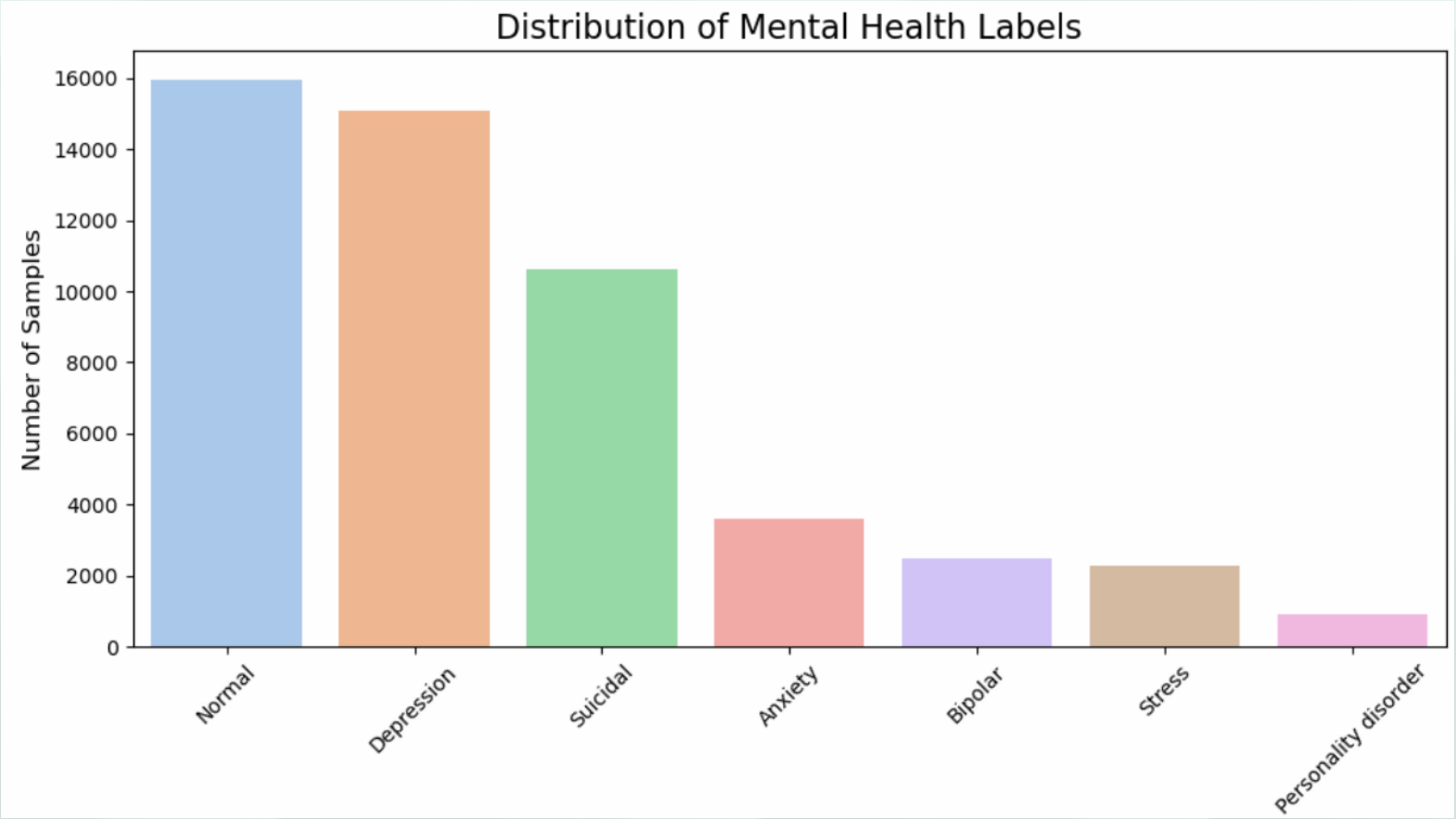
Completed so far

- Data preprocessing – loaded, cleaned and verified dataset (53K+ samples).
- Text vectorization – applied TF-IDF to convert text to numerical features.
- Baseline model – trained and evaluated Naïve Bayes classifier.
- Modeling – trained Logistic Regression and Linear SVM models.
- Evaluation – generated classification reports and confusion matrices.

Next steps

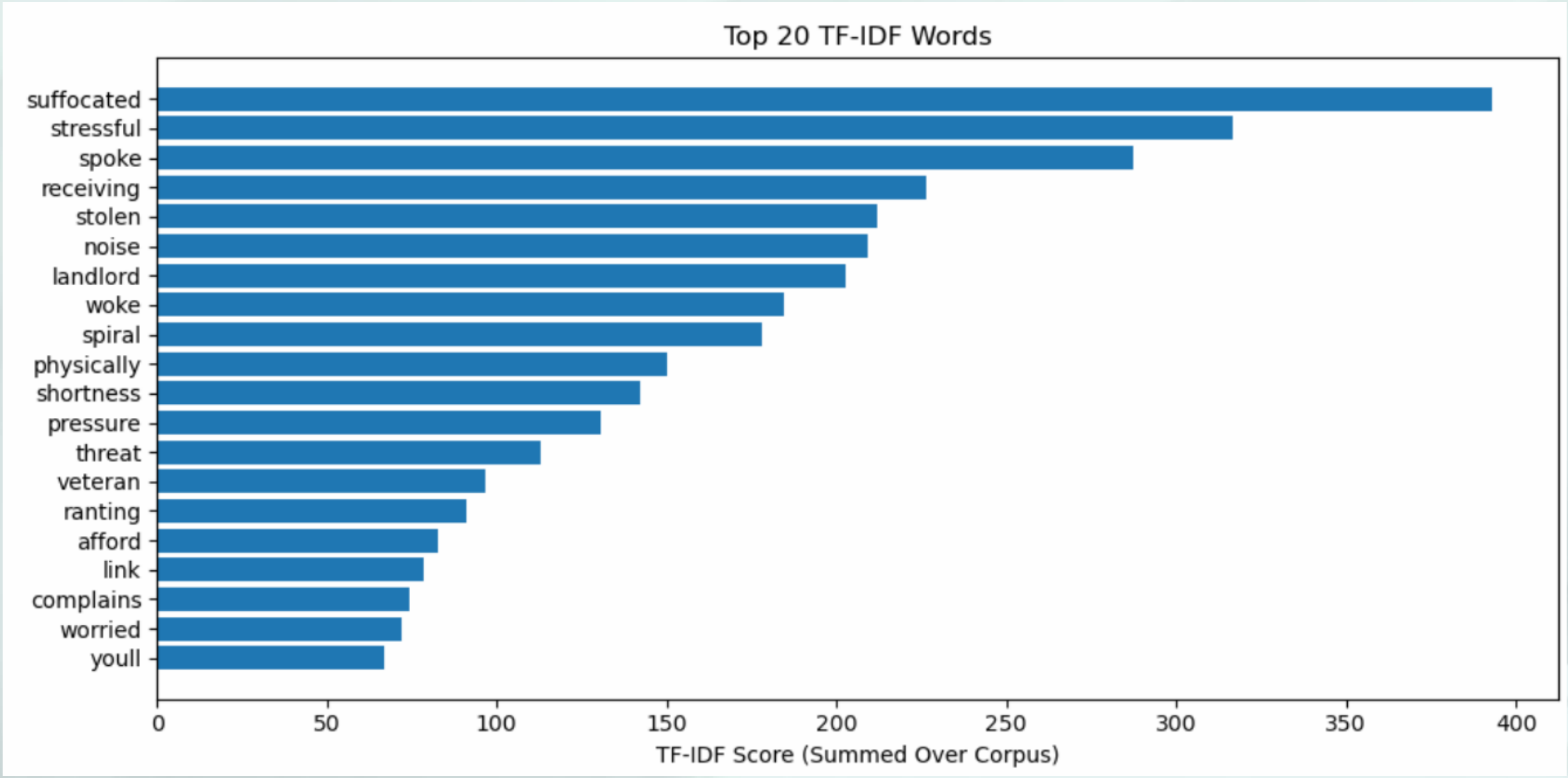
- Feature optimization – try n-grams, max_features, etc.
- Advanced modeling – experiment with BERT or RoBERTa.
- Error analysis – inspect misclassified samples, refine labels if needed.
- Visualization – add label distributions, top words, and confusion matrix plots.

.1



.2

Metric	Naïve bayes	Logistic regression	SVM (LinearSVC)
Accuracy	0.67	0.75	0.74
Marco F1 score	0.51	0.68	0.69
Weighted F1 score	0.65	0.74	0.74


Top 20 TF-IDF Words
suffocated
stressful
spoke
receiving
stolen
noise
landlord
woke
spiral
physically
shortness
pressure
threat
veteran
ranting
afford
link
complains
worried
youll
0
50
100
150
200
250
300
350
400
TF-IDF Score (Summed Over Corpus)

Key Insights & Next Steps

- **Baseline models** like Naive Bayes, Logistic Regression, and SVM performed well on clearer categories (e.g., Normal, Depression) but struggled with more ambiguous ones like Stress and Personality Disorder.

- **Class imbalance** had a noticeable impact on performance, especially in the **Macro F1-score**, emphasizing the importance of using multiple evaluation metrics.

- **TF-IDF helped identify high-impact words**, but it lacks the ability to understand deeper semantic context or word meaning in relation to surrounding words.

- Moving forward, we'll explore **context-aware models like BERT or RoBERTa**, which are better suited for capturing emotional subtleties and complex phrasing.

Thank You ☺