Identifying Toxic Language and Mental Health Signals in Twitter Data

Pujita Kodali
Texas Tech University
Lubbock, Texas, USA
pukodali@ttu.edu

Sarayu Parupalli
Texas Tech University
Lubbock, Texas, USA
sparupal@ttu.edu

Sai Charan Lanka
Texas Tech University
Lubbock, Texas, USA
sailanka@ttu.edu

Sathwik Tatiparthi
Texas Tech University
Lubbock, Texas, USA
statipar@ttu.edu

Abstract

Social networks such as Twitter create avenues for global communication and the spread of toxic behavior and emotional distress. Previous work highlights the harm caused by hate speech and online abuse [1, 2]. At the same time, social media contains signals of depression and suicidal ideation [4]. This project develops a dual-model system using RNNs and LSTMs [6] to detect toxic language and mental-health indicators. We merge nine datasets, apply extensive preprocessing, train deep learning models, and integrate explainability [8]. The results demonstrate strong performance on both classification tasks. In addition, we address the challenges of domain imbalance, informal language patterns, and overlapping symptom expressions in user-generated text. Our system is designed not only to classify but also to provide interpretable insights into prediction reasoning, making it suitable for practical decision support. Ultimately, this research contributes to safer digital platforms and advances the use of AI for mental well-being analysis.

CCS Concepts: **• Computing methodologies → Neural networks**; **Natural language processing**; **• Information systems → Data mining**; **• Applied computing →** *Health informatics*; *Computing in social science.*

Keywords: Toxic Language Detection, Mental Health Classification, RNN, LSTM, Explainability, Social Media NLP

ACM Reference Format:
Pujita Kodali, Sarayu Parupalli, Sai Charan Lanka, and Sathwik Tatiparthi. 2025. Identifying Toxic Language and Mental Health Signals in Twitter Data. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Online social platforms are widely used for communication, entertainment, and information sharing, yet they also act as channels for various forms of toxicity including hate speech, bullying, and abusive language. Research shows that exposure to such content results in measurable social and psychological harm [2, 3]. Moreover, social media often serves as a space where individuals express early symptoms of mental-health struggles such as depression, anxiety, and suicidal ideation [4].

Accurate automated detection of both toxic and mental-health–related language is thus crucial for enabling safer online environments. Our project implements a dual-model system for (1) toxicity detection and (2) mental-health signal classification, combining classical RNN ideas with the long-term memory capabilities of LSTMs [6]. Additionally, explainability is incorporated to allow transparency and trust.

2 Background / Related Work

2.1 Toxic Language Detection

Hate speech classification has been studied extensively using rule-based methods, statistical models, and deep learning techniques. Davidson et al. [1] distinguish between hate speech and offensive content, motivating the multi-label nature of toxic language classification. Fortuna and Nunes [2] provide a comprehensive survey and highlight challenges such as class imbalance and annotation subjectivity.

2.2 Mental-Health Signal Detection

Coppersmith et al. [4] quantify expressions of depression and suicidal ideation on Twitter and illustrate the importance of early identification. Social media provides natural, unstructured psychological signals, making it a valuable source for mental-health screening.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.3 Neural Architectures

RNNs process text one word at a time and can capture short connections in language [7]. However, they struggle to remember information that appears much earlier in a sentence or conversation. LSTMs solve this using special memory gates that help them keep important details for longer [6]. This makes them effective for understanding emotional tone and subtle feelings in longer messages, which often develop slowly over time.

2.4 Explainability

Explainability was integrated directly into the modeling pipeline using TF–IDF combined with a linear classification model [9]. This method provides token-level contribution estimates that indicate which words influenced the classifier toward toxic or mental-health categories. Such transparency is critical for sensitive tasks like identifying depression or suicidal ideation, where explainable outputs support trust, safety, and responsible decision-making [13].

Token Importance Calculation:

$$\text{token\_importance} = \text{tfidf\_weight} \times \text{linear\_coef}$$

3 Data Collection

We merged a total of **Nine Kaggle datasets**:

- **Toxicity:** Two comment-based datasets containing toxic, severe toxic, insult, threat, obscene, and identity hate labels.
- **Mental Health:** Seven datasets from Reddit, Twitter, and chatbot messages describing depression, anxiety, bipolar disorder, suicidal ideation, stress, personality disorder, and normal states.

Table 1 gives an overall view of the combined dataset.

Table 1. Summary of datasets used in this project.

Category	# Datasets	Source	Task
Toxicity	2	Kaggle	Multi-label Classification
Mental Health	7	Kaggle, Reddit, Twitter	Multi-class Classification

Figure 1 shows the work flow of classifying all the datasets and classifying the categories for Toxicity and Mental Health.

4 Data Cleaning

Data cleaning followed machine learning best practices [9]:

- Standardizing columns and label formats.
- Removing duplicates and empty texts.
- Adding a "Non-Toxic" label for binary normalization.
- Applying text cleaning to remove URLs, emojis, punctuation, and stopwords.

Figure 2 shows "Toxic" and "Non-Toxic" labels for binary normalization.

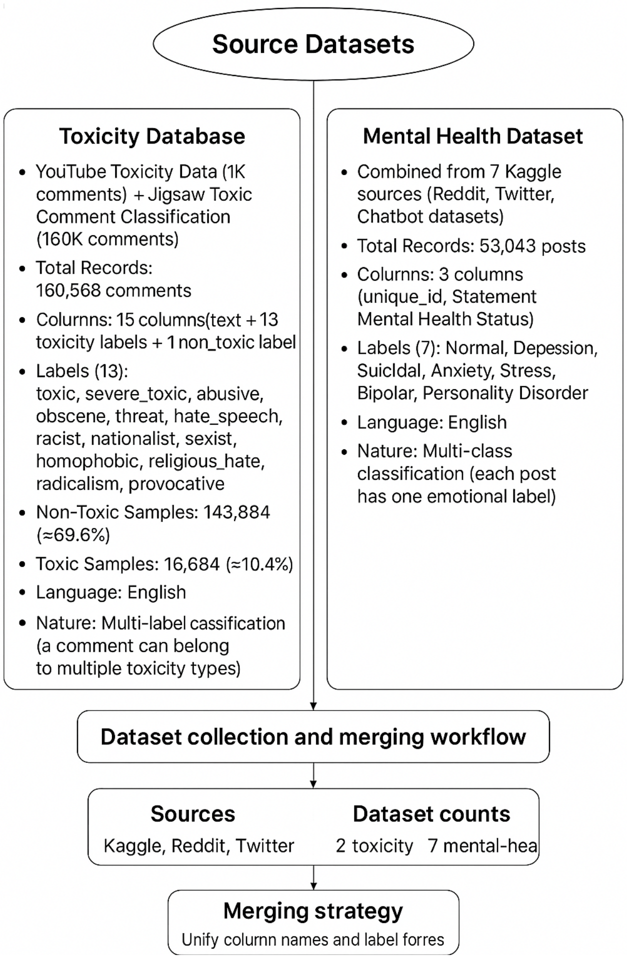

Source Datasets
Toxicity Database
• YouTube Toxicity Data (1K comments) + Jigsaw Toxic Comment Classification (160K comments)
• Total Records: 160,568 comments
• Columns: 15 columns(text + 13 toxicity labels + 1 non_toxic label
• Labels (13): toxic, severe_toxic, abusive, obscene, threat, hate_speech, racist, nationalist, sexist, homophobic, religious_hate, radicalism, provocative
• Non-Toxic Samples: 143,884 (≈69.6%)
• Toxic Samples: 16,684 (≈10.4%)
• Language: English
• Nature: Multi-label classification (a comment can belong to multiple toxicity types)
Mental Health Dataset
• Combined from 7 Kaggle sources (Reddit, Twitter, Chatbot datasets)
• Total Records: 53,043 posts
• Columns: 3 columns (unique_id, Statement Mental Health Status)
• Labels (7): Normal, Depession, Suicldal, Anxiety, Stress, Bipolar, Personality Disorder
• Language: English
• Nature: Multi-class classification (each post has one emotional label)
Dataset collection and merging workflow
Sources
Kaggle, Reddit, Twitter
Dataset counts
2 toxicity 7 mental-hea
Merging strategy
Unify column names and label forres

Figure 1. Dataset Classification workflow

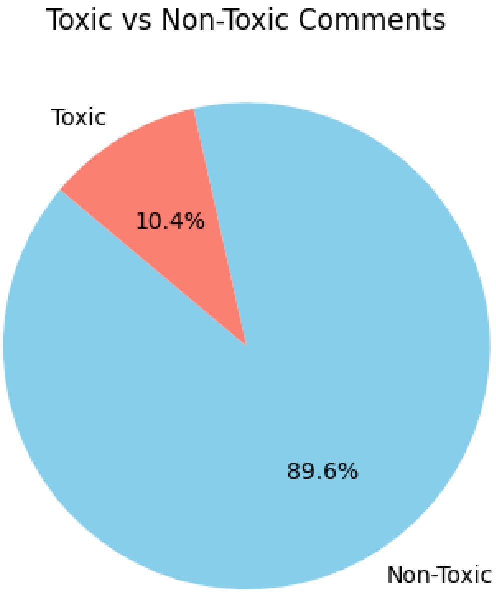


Figure 2. Labeling Toxic and Non Toxic

Figure 3 provides an overview of the project's conceptual structure. This figure summarizes the high-level flow from data collection through deployment.

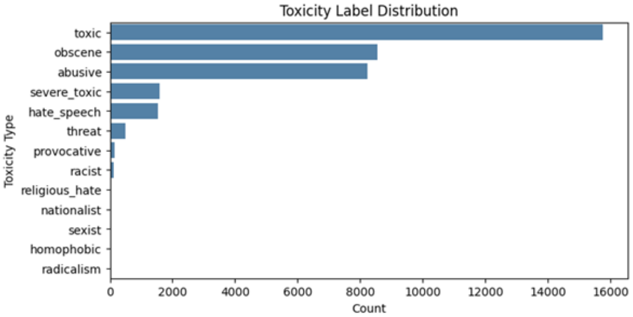


Figure 3. Parallel processing streams for toxicity detection (multi-label)

Figure 4 shows part of the preprocessing pipeline for Mental health datasets. The raw text normalization steps including URL removal, emoji stripping, punctuation normalization, tokenization, lowercasing, and stopword removal. This image demonstrates the preprocessing flow applied consistently across all merged datasets.

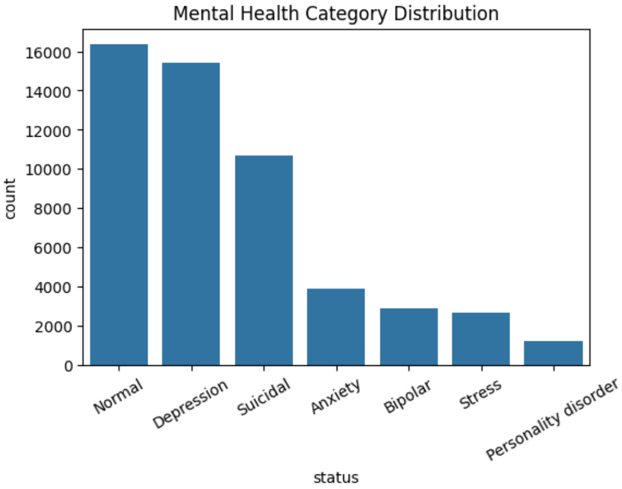


Figure 4. Parallel processing streams for mental-health classification (multi-class)

5 Exploratory Data Analysis (EDA)

5.1 Toxicity Dataset

Consistent with past research [1], the toxicity dataset exhibited heavy class imbalance, with nearly 90% of samples labeled non-toxic. Minor categories such as "threat" and "identity hate" had very low representation, posing challenges for model learning.

5.2 Mental-Health Dataset

The mental-health dataset was more balanced, with comparable representation across depression, anxiety, bipolar, and suicidal categories.

Table 2 shows the simplified class distribution.

Table 2. Simplified class distribution across datasets.

Label	Count	Category
Non-Toxic	High	Toxicity
Toxic/Abusive	Low	Toxicity
Depression	Medium	Mental Health
Anxiety	Medium	Mental Health
Bipolar	Medium	Mental Health
Suicidal	Low-Medium	Mental Health
Stress	Low	Mental Health

6 Model Training

6.1 Traditional Machine Learning Baselines

Before exploring deep learning architectures, we first implemented traditional machine learning classifiers as foundational baselines. The project slides highlighted two primary models:

- **Decision Tree Classifier**
- **Random Forest Classifier**

Both models were trained using TF–IDF vectorized features for the two tasks:

1. **Toxicity Detection** (multi-label classification)
2. **Mental-Health Detection** (multi-class classification)

This baseline stage allowed us to benchmark classical approaches before moving to sequential deep-learning models capable of understanding context.

6.1.1 Baseline Performance.

Tables 3 and 4 summarize the baseline evaluation results, as reported in the project slides.

Table 3. Baseline results for toxicity detection (multi-label).

Model	Micro F1	Weighted F1
Decision Tree	0.68	0.65
Random Forest	0.74	0.72

Both traditional models achieved noticeably lower scores, especially in categories involving subtle emotional cues or rare toxic labels. This confirmed that TF–IDF–based classical models were not sufficient for nuanced linguistic understanding.

Table 4. Baseline results for mental-health detection (multi-class).

Model	Accuracy	Macro F1
Decision Tree	0.61	0.58
Random Forest	0.67	0.63

6.1.2 Why Decision Trees and Random Forests Performed Poorly.

Although Decision Trees and Random Forests are strong general-purpose classifiers, they exhibit several limitations when applied to complex linguistic phenomena such as toxic language and mental-health signals:

Inability to Capture Short-Term Dependencies. TF–IDF treats text as an unordered bag of words, meaning that traditional ML models ignore local semantic patterns such as:

> "feeling hopeless today"
> "stop being an idiot"

Sequential meaning is essential for toxicity and emotional state detection, but tree-based methods have no notion of token order.

Failure to Model Long-Term Emotional Context. Mental-health language frequently relies on multi-clause expressions:

> "I look fine, but inside I feel empty."

Such expressions contain emotional progression that spans the entire sentence. Decision Trees and Random Forests cannot track evolving emotional signals across time, whereas LSTMs store and update context through memory cells.

Overfitting to Sparse TF–IDF Vectors. Decision Trees tend to overfit rare words, leading to brittle decision boundaries. Random Forests reduce variance slightly, but the underlying sparsity of TF–IDF still limits generalization.

Difficulty Handling Rare Toxic Categories. Multi-label toxicity datasets contain highly imbalanced categories such as *identity-hate* or *threat*. Traditional ML models struggle with these cases, producing low recall. Deep models, however, learn shared representations that improve generalization across rare labels.

6.1.3 Motivation for LSTM–RNN Models.

Given the limitations of traditional ML models, sequence-based deep learning architectures became a natural next step. LSTM networks offer several advantages:

- **Sequential processing:** tokens are read in order, allowing capture of short-term semantic dependencies.
- **Memory gates:** LSTMs store long-range emotional context, enabling them to interpret subtle expressions.
- **Better handling of imbalance:** shared latent representations provide improved performance on rare categories.
- **Contextual understanding:** phrases such as "I am tired of everything" or "you people are disgusting" require tone and sequence awareness rather than keyword matching.

Thus, the baseline results clearly indicated that deeper contextual modeling was required. These insights motivated the transition toward LSTM–RNN architectures in the next stage of the project.

We trained two independent models:

1. **Multi-label Toxicity Classifier** Six labels: toxic, severe toxic, obscene, threat, insult, identity hate.
2. **Multi-class Mental-Health Classifier** Seven labels: depression, anxiety, bipolar, suicidal, stress, personality disorder, normal.

6.2 Architecture

The final system architecture includes:

- Tokenization and Embedding Layer (100–300 dimensions)
- **LSTM Layer (128 units)** with dropout
- Dense output layers using Sigmoid (toxicity) and Softmax (mental-health)
- Adam optimizer (lr=0.001)
- Early stopping (6–12 epochs)

Table 5 summarizes the architecture.

Table 5. Model architecture summary.

Layer	Units/Type	Purpose
Embedding	100–300	Word vector representation
LSTM	128	Long-range sequence modeling
Dense	Variable	Final classification
Dropout	0.3–0.5	Regularization

Figure 5 illustrates the architecture.

Explainability was incorporated using a TF-IDF–Linear Model hybrid inspired by [8]. Token importance was computed using:

$$\text{token\_importance} = \text{tfidf\_weight} \times \text{linear\_coef}$$

7 Results / Findings

7.1 Mental-Health Classifier

- Accuracy: 0.81
- Macro-F1: 0.78
- Strongest classes: Normal (0.93), Bipolar (0.84), Anxiety (0.83)
- Suicidal detection: 0.73 F1
- Hardest class: Stress (0.66 F1)

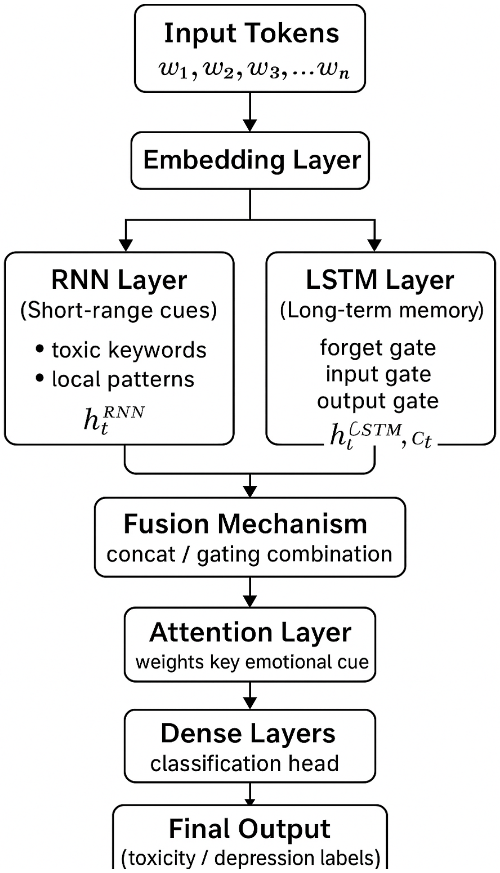


Figure 5. Final LSTM-based architecture.

7.2 Toxicity Classifier

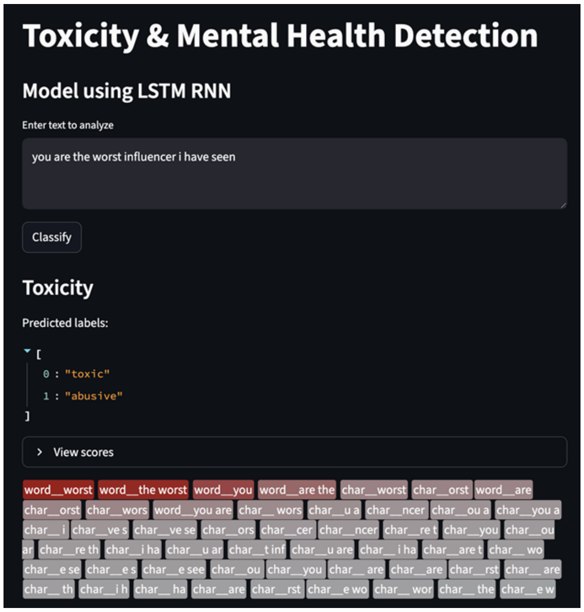
- Micro-F1: 0.92
- Weighted-F1: 0.93
- Strong detection: Obscene (0.82), Toxic (0.79), Insult (0.70)
- Non-Toxic: 0.98 F1
- Weakest: Rare classes (identity hate, racist)

Table 6 summarizes key metrics for Toxicity and Mental Health.

Table 6. Summary of model performance metrics.

Model	Accuracy/F1
Toxicity (Multi-label)	0.92–0.93
Mental Health (Multi-class)	0.78–0.81

Figure 6 and 7 shows part of the output visualization.


Toxicity & Mental Health Detection

Model using LSTM RNN

Enter text to analyze

you are the worst influencer i have seen

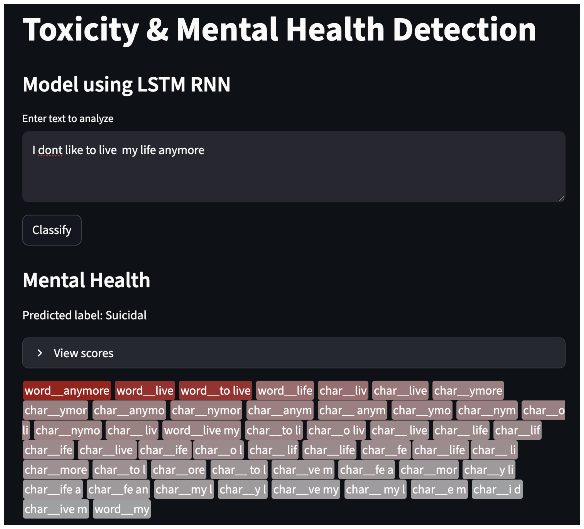
Toxicity

Predicted labels:

```
[
  0 : "toxic"
  1 : "abusive"
]
```

```
word__worst word__the worst word__you word__are the char__worst char__orst word__are
char__orst char__wors word__you are char__ wors char__u a char__ncer char__ou a char__you a
char__ i char__ve s char__ve se char__ors char__cer char__ncer char__re t char__you char__ou
ar char__re th char__i ha char__u ar char__t inf char__u are char__ i ha char__are t char__ wo
char__e se char__e s char__e see char__ou char__you char__ are char__are char__rst char__ are
char__ th char__i h char__ ha char__are char__rst char__e wo char__ wor char__ the char__e w
```

Figure 6. Model performance output for Toxicity.


Toxicity & Mental Health Detection

Model using LSTM RNN

Enter text to analyze

I dont like to live my life anymore

Mental Health

Predicted label: Suicidal

```
word__anymore word__live word__to live word__life char__liv char__live char__ymore
char__ymor char__anymo char__nymor char__anym char__ anym char__ymo char__nym char__o
li char__nymo char__ liv word__live my char__to li char__o liv char__ live char__ life char__lif
char__ife char__live char__ife char__o l char__ lif char__life char__fe char__life char__ li
char__more char__to l char__ore char__ to l char__ve m char__fe a char__mor char__y li
char__ife a char__fe an char__my l char__y l char__ve my char__ my l char__e m char__i d
char__ive m word__my
```

Figure 7. Model performance output for Mental Health.

8 Conclusion

This project successfully developed a dual deep-learning system capable of accurately identifying toxic and mental-health signals in social media content. Our LSTM-based models [6] achieved high performance across multiple categories, demonstrating robustness despite dataset imbalance. Incorporating explainability [8] improved interpretability for sensitive classifications.

The merged datasets, strong preprocessing pipeline, and exploratory analysis created a reliable foundation. Future work includes:

- Using transformer models such as BERT [11] and attention mechanisms [12]
- Incorporating multilingual capability
- Connecting to live social APIs for real-time monitoring
- Deploying as a production-grade microservice (FastAPI)

References

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. *Automated Hate Speech Detection and the Problem of Offensive Language.* ICWSM, 2017.

[2] Fortuna, P., & Nunes, S. *A Survey on Automatic Detection of Hate Speech in Text.* ACM Computing Surveys, 2018.

[3] Burnap, P., & Williams, M. L. *Cyber Hate Speech on Twitter.* Social Science Computer Review, 2015.

[4] Coppersmith, G., Dredze, M., & Harman, C. *Quantifying Mental Health Signals in Twitter.* CLPsych, 2014.

[5] Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. *From ADHD to SAD: Validity of Online Mental-Health Data.* CLPsych, 2015.

[6] Hochreiter, S., & Schmidhuber, J. *Long Short-Term Memory.* Neural Computation, 1997.

[7] Goodfellow, I., Bengio, Y., & Courville, A. *Deep Learning.* MIT Press, 2016.

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* KDD, 2016.

[9] Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python.* JMLR, 2011.

[10] Kim, Y. *Convolutional Neural Networks for Sentence Classification.* EMNLP, 2014.

[11] Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers.* NAACL, 2019.

[12] Vaswani, A., et al. *Attention Is All You Need.* NeurIPS, 2017.

[13] Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. *Towards an Empathetic AI: Using Machine Learning to Support Mental Health.* Journal of Medical Internet Research, 2020.