Universitat Pompeu Fabra Barcelona

Deep Learning Final Project

Sentiment Analysis with LSTMs for Mental Health Detection

05/06/2025

Pau Peirats, Mireia Pou, and Stuart Lance

Table of contents

01 Introduction

02 State of the Art

03 Methodology

04 Results

05 Conclusions

06 Future work

01

Introduction

Problem definition

Problem

- Identify potential mental health issues based on text inputs.
- LSTM-based model to classify into mental health categories

Motivation

- Critical issue nowadays
- Exploration of AI in a clinical environment
- Positive social impact

Dataset characteristics

- **Columns** = [Unique ID, Statement, Mental Health Status]
- 53.042 data entries with 51.704 unique values

1	trouble sleeping, confused mind, restless heart. All out of tune	Anxiety

Figure 1: Screenshot of a dataset entry

- Mental health status:

1 Normal

2 Suicidal

3 Anxiety

4 Personality Disorder

5 Depression

6 Bipolar

7 Stress

02

State of the Art

State of the Art

- RNNs (LSTMs, GRUs, BiLSTMs)
- Transformers (BERT-based)
- CNNs + RNNs

- Pre-trained embeddings
- Dropout
- Accuracy and F1-score

Model	Classification	Accuracy
LSTM	7 classes	75-88%
MentalBERT	4 classes	93%
CNN+GRU	3 classes	93%

Table 1: State of the Art comparison.

Benchmark

Transformers

- Transformer used= *nateraw/bert-base-uncased-emotion*
- Accuracy = 0.849

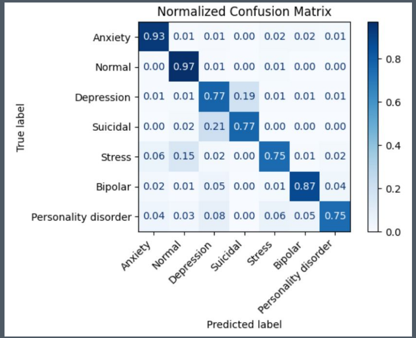


Figure 2: Transformer-based confusion matrix [1]

LSTM

- LSTM + ReLu + Softmax
- Accuracy = 0.723

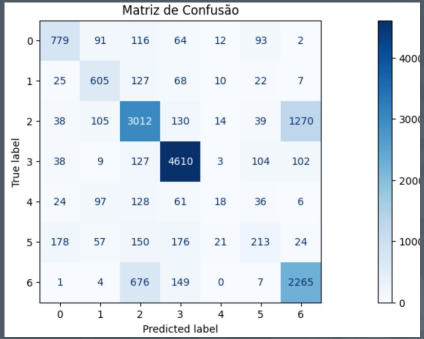


Figure 3: LSTM confusion matrix [2]

Machine Learning

- Model = XGBoost
- Accuracy = 0.808

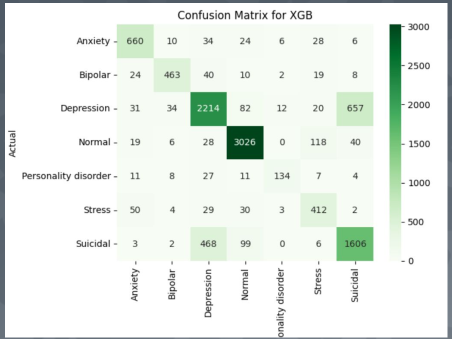


Figure 4: XGBoost confusion matrix [3]

03

Methodology

Exploratory Data Analysis (EDA)

1 Class Distribution

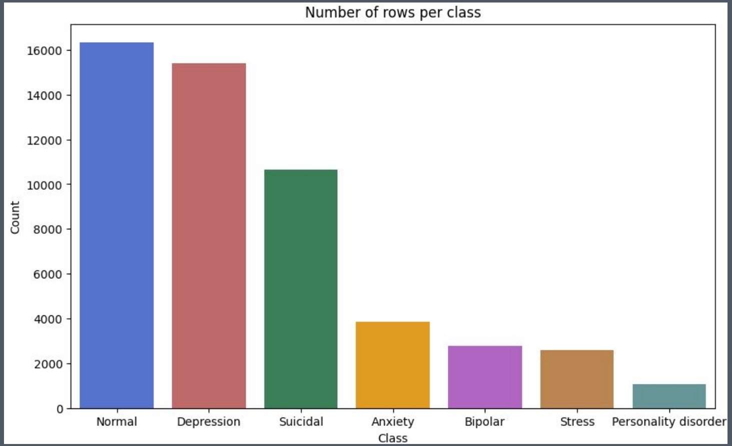


Figure 5: Barplot showing class distribution

2 Statement length

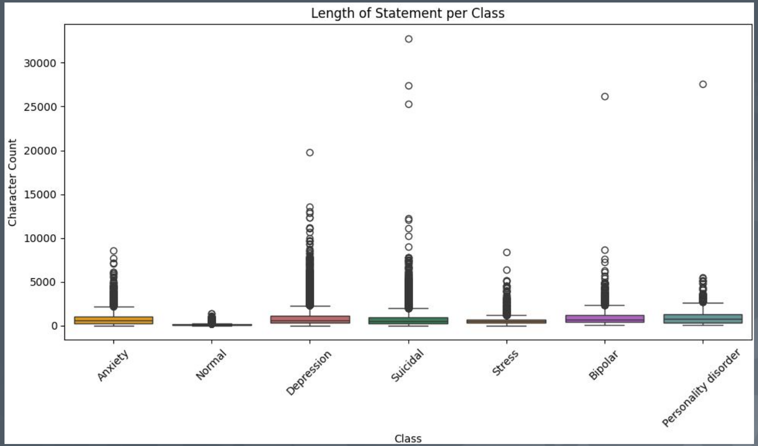


Figure 6: Boxplot showing statements' length

3 Text noise

	0
hashtag	1754
mention	1094
url	902
emoji	539952

Figure 7: Text noise count

Exploratory Data Analysis (EDA)

4 Top 20 words per class

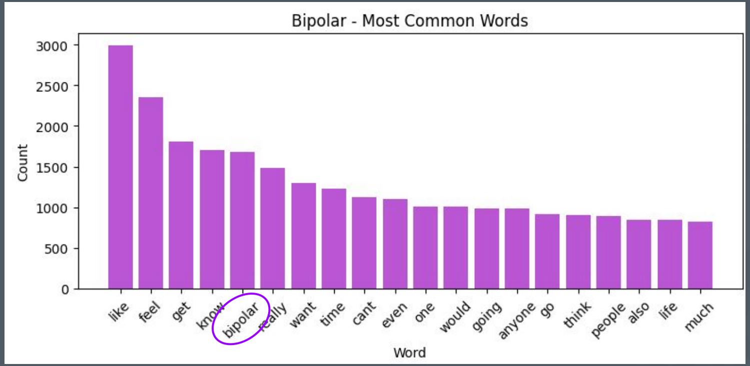


Figure 8(1): Barplot showing top 20 words in bipolar class

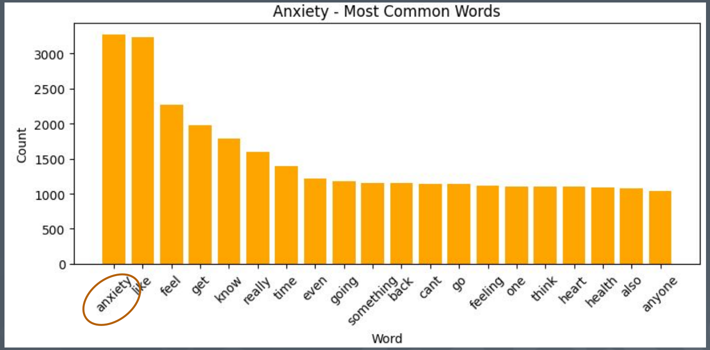


Figure 8(2): Barplot showing top 20 words in anxiety class

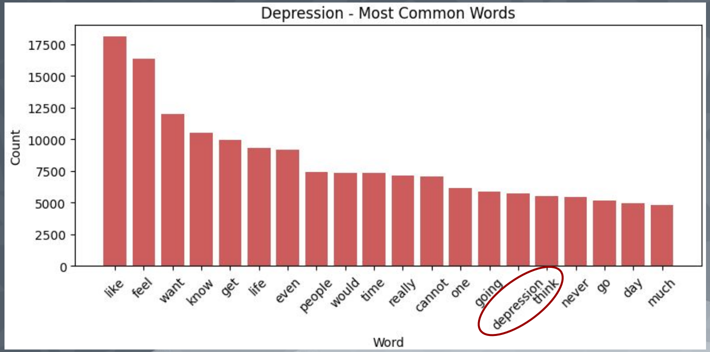
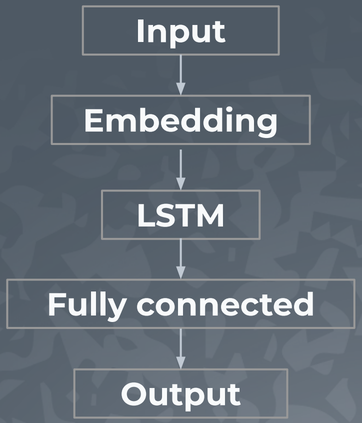


Figure 8(3): Barplot showing top 20 words in depression class

Baseline Model

- Single layer LSTM
- Trained with **raw text**
- Good performance but improvable
- Poor generalisation



Model optimization through input data preprocessing

1. Data cleaning
2. Stopword removal
3. Embedding Selection
4. Data augmentation
5. Grid search

Data processing models

Model 1: Data Cleaning

1. Length outlier removal
2. Text noise removal:

URL

HTML

Emojis

\n
Newlines

#
Hashtags

@
Mentions

a b
Extra spaces

! ?
Basic punctuation

Model 2: Stopword removal

- Articles (a, an, the)
- Prepositions (of, in, for, through)
- Pronouns (it, their, his)

NLTK

Data processing models

Model 3: Embedding selection

Word2Vec

- Easiest and simplest
- Does not improve baseline

GloVe

- State-of-the-art embedder
- Improves baseline

Model 4: Data augmentation

1. **Back translation**

   English → French → English

2. **Synonym replacement**

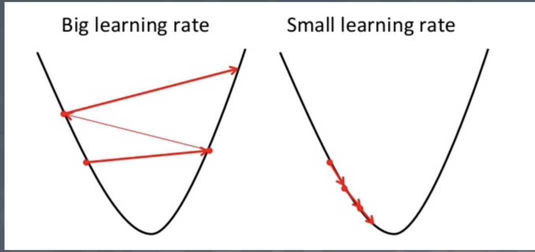
   WORD → SYNONYM
   stress → pressure

Data processing models

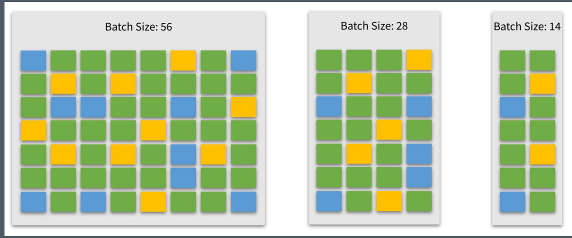
Model 5: Grid Search

- Maximize performance
- Brute-force

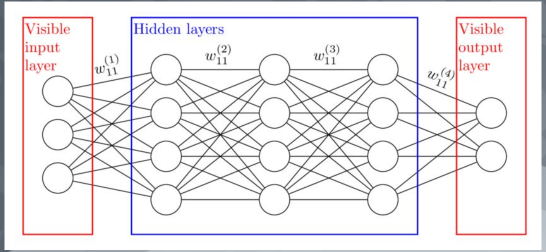
Learning rate



Batch size



Hidden layers



Model optimization through architecture refinement

6 Dropout

7 Stacked LSTM

8 Attention layer

9 GRU

Architecture refinement models

Model 6: Dropout

Input
↓
Embedding
↓
LSTM with dropout
↓
Fully connected
↓
Output

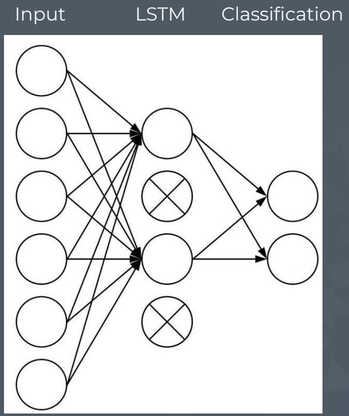


Figure 9: Dropout representation

Model 7: Stacked LSTM

Architecture refinement models

Model 8: Attention Layer

Input

↓

Embedding

↓

LSTM

↓

Attention

↓

Fully connected

↓

Output

Model 9: GRU

LSTM gates		GRU gates
Input gate Output gate Forget gate	→	Reset gate Update gate

04

Results

Baseline model

- Final accuracy = 0.7332
- Weighted F1-score = 0.7355

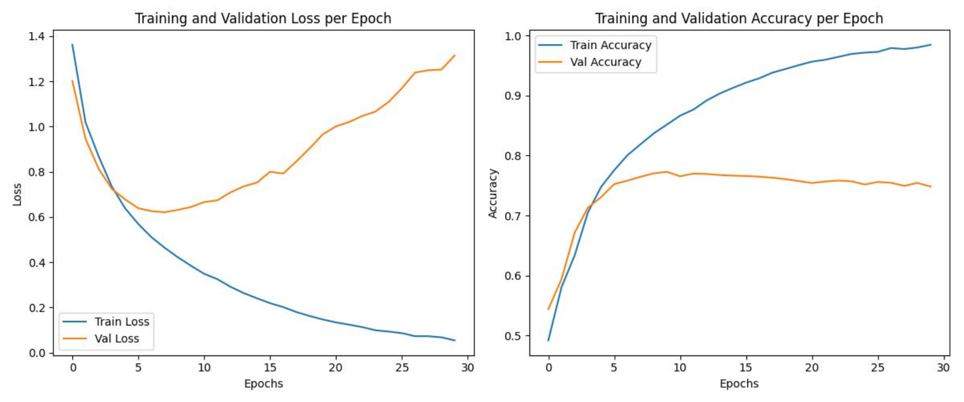


Figure 10: Baseline LSTM Training and Validation Loss and Accuracy plots

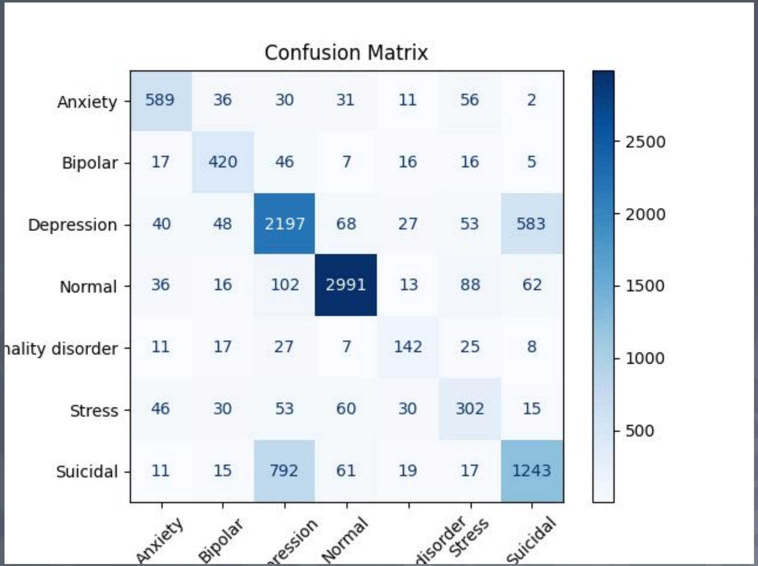


Figure 11: Baseline LSTM Confusion Matrix

Data preprocessing experiments

Model	Name	Best validation accuracy	Final validation accuracy	Final validation loss	Weighted F1-score
0	Baseline	0.7524	0.7332	1.5110	0.7355
1	Data cleaning	0.7566	0.7408	1.3936	0.7412
2	Stopword removal	0.7539	0.7394	1.4413	0.7384
3	Embedding Selection	0.7779	0.7744	0.6665	0.7753
4	Data Augmentation	0.7590	0.7442	0.8882	0.7476
5	Grid Search	0.7818	0.7735	0.6010	0.7769

Table 2: Results of the data preprocessing models

Data preprocessing experiments

	Baseline	Data cleaning
Anxiety	**0.78**	0.75
Bipolar	0.73	**0.75**
Depression	0.67	**0.69**
Normal	0.91	**0.92**
Personality Disorder	0.48	**0.52**
Stress	0.53	**0.54**
Suicidal	**0.62**	0.61

Data preprocessing experiments

Data cleaning

Embedding Selection

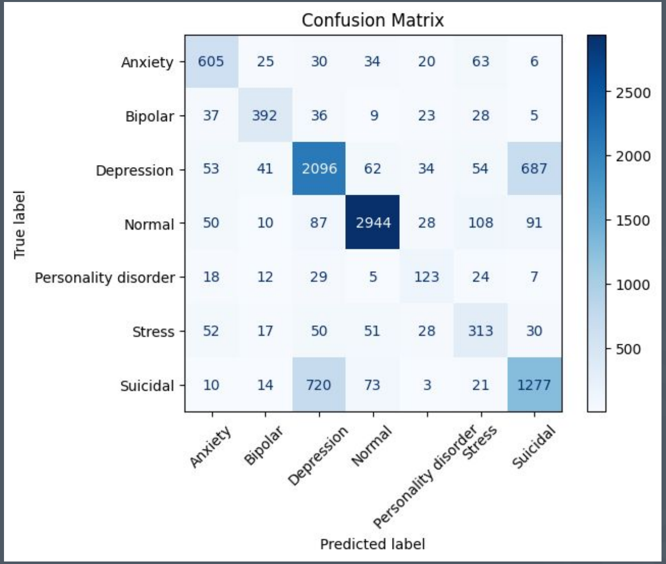


Figure 13: Data cleaning model. Confusion matrix.

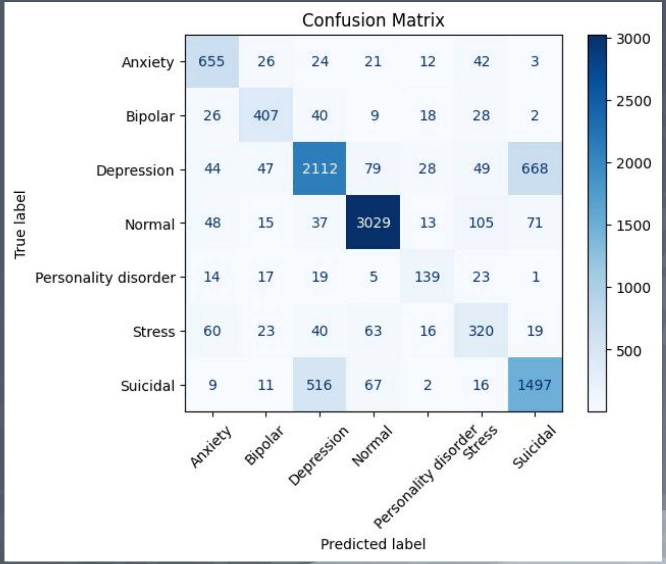


Figure 15: Embedding Selection model. Confusion matrix.

Data preprocessing experiments

Embedding Selection → **Grid Search**

- Less overfitting
- Better accuracy

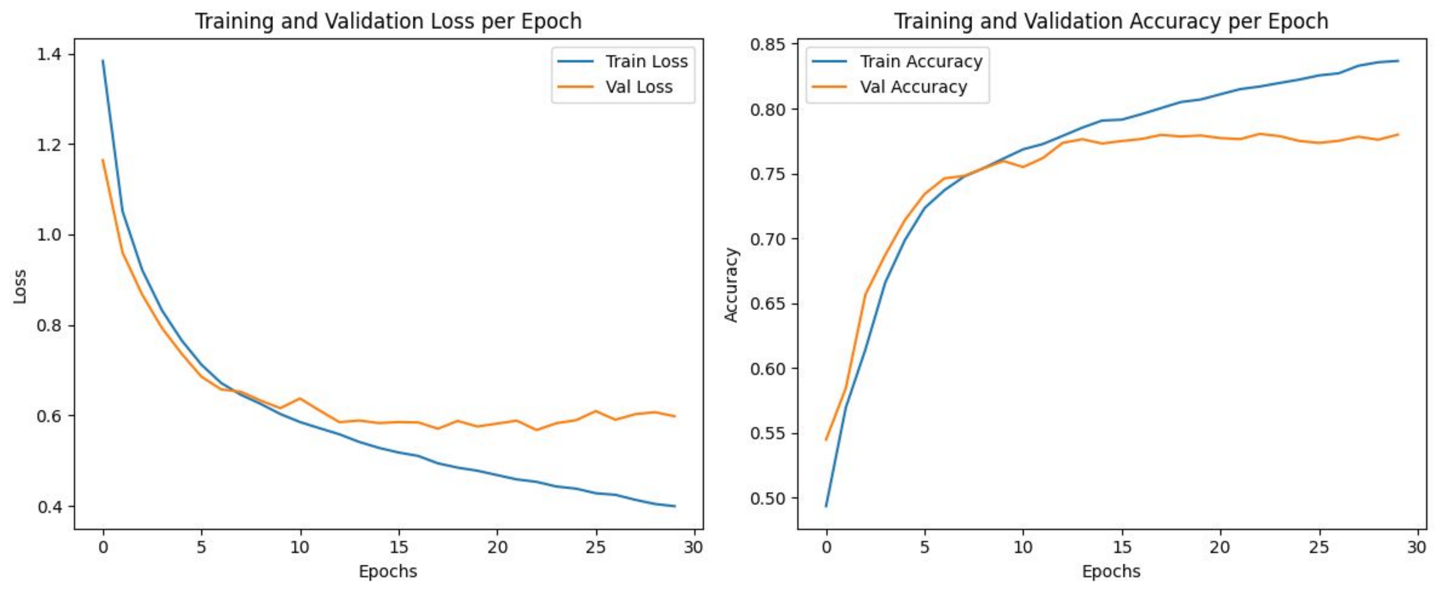


Figure 17: Grid Search model. Loss and accuracy curves.

Architecture optimization experiments

Model	Name	Best validation accuracy	Final validation accuracy	Final validation loss	Weighted F1-score
0	Baseline	0.7524	0.7332	1.5110	0.7355
5	Grid Search	0.7818	0.7735	0.6010	0.7769
6	Dropout	0.7790	0.7773	0.5892	0.7774
7	Stacked LSTM	0.7820	0.7789	0.5835	0.7795
8	Attention	0.7804	0.7715	0.6420	0.7732
9	Baseline GRU	0.7627	0.7388	1.7622	0.7376
10	Improved GRU	0.7932	0.7684	0.9345	0.7682

Architecture optimization experiments

Grid Search → Dropout

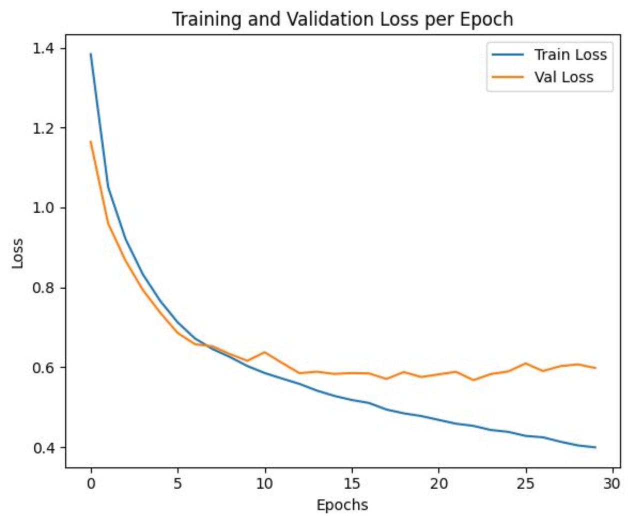


Figure 18: Grid Search model. Loss curve.

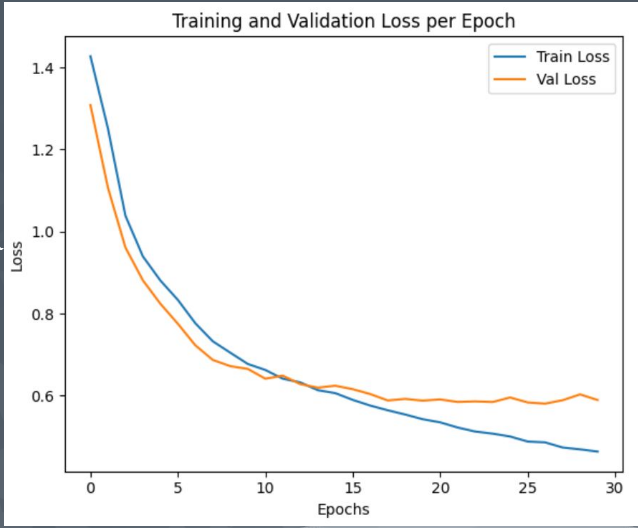


Figure 19: Dropout model. Loss curve.

Architecture optimization experiments

Dropout → Stacked LSTM

	Dropout	Stacked LSTM
Anxiety	0.81	**0.82**
Bipolar	0.77	**0.79**
Depression	**0.73**	0.72
Normal	0.92	**0.93**
Personality Disorder	**0.62**	0.60
Stress	0.57	**0.61**
Suicidal	0.68	0.68

05

Conclusions

Conclusions

1. Raw baseline model limited by noisy data and simple features.

2. Data cleaning and using embeddings like GloVe raise accuracy.

3. **Final best model:** Stacked LSTM with dropout. Accuracy = 0.7789. Weighted F1-score = 0.7795

4. Outperformed benchmark models, with exception of transformers.

06

Future work

Future work

1. Bidirectional LSTMs (BiLSTM)

2. **Transformer Models:** Experiment with SotA transformers (BERT, RoBERTa, etc)

3. **Ensembling models**: Combination of multiple models like LSTM and transformer classifiers. **CNNs + RNNs, like seen in SotA**

Resources

[1]
https://www.kaggle.com/code/grantgonnerman/mental-health-sentiment-analysis-eda-modeling

[2] https://www.kaggle.com/code/rafaeldrago/lstm-sentiment-prediction/notebook

[3] https://www.kaggle.com/code/hnfrmdhni/klasifikasi-jenis-depresi

Thank you!

Any question?

05/06/2025

Pau Peirats, Mireia Pou, and Stuart Lance