Project Report
QMBU-420

Group 4

Emir GÖCEN
Enes AKÇAKOCA
Mert TUNA
Yağmur KIZILÖZ

**all authors contributed equally*

1. Business Understanding

Mental health disorders represent an increasingly critical concern in global public health, with substantial societal, economic, and individual consequences. Given the growing prevalence of psychological distress and limited access to timely diagnostic resources, the early detection of mental health conditions through non-traditional, digital means is of paramount importance. Contemporary discourse in computational psychiatry has identified social media platforms as fertile terrain for extracting real-time, ecological data reflective of users' affective and cognitive states.

This project explores the application of machine learning (ML) and natural language processing (NLP) frameworks to the classification of user-generated social media content into clinically relevant categories: Depression, Anxiety, Stress, Bipolar Disorder, Suicidal Ideation, Personality Disorder, and a normative baseline (Normal). Our objective is to model language patterns associated with psychopathology to support scalable, ethically responsible tools for digital mental health triage, public health surveillance, and psycholinguistic research.

The justification for this initiative is twofold. First, there exists institutional demand for accessible and low-cost screening systems that do not rely on clinical gatekeeping. Second, methodological advances in representation learning and neural language modeling afford unprecedented capability for semantic interpretation of informal, heterogeneous textual data.

2. Data Understanding

The dataset employed in this study is an amalgamation of several public corpora hosted on Kaggle, aggregating approximately 51,000 labeled text samples. Data sources include, but are not limited to, Reddit posts related to depression and anxiety, Twitter content expressing suicidal ideation, and stress-related chat logs. Each instance includes a unique identifier, the raw text of the post, and an associated mental health label.

These texts reflect authentic, unfiltered expression, replete with vernacular variation, emojis, acronyms, and syntactic anomalies. Such characteristics present substantial challenges for NLP systems, particularly those reliant on rigid linguistic assumptions. A major concern in the dataset is class imbalance, with the Normal category significantly overrepresented. This imbalance complicates model training and evaluation, necessitating tailored metric selection and algorithmic techniques to mitigate bias toward the dominant class.

3. Data Preparation

Data preprocessing was implemented using a two-pronged approach combining the visual data flow architecture of KNIME with Python-based scripting. The preparation sequence included:

Lexical normalization: Lowercasing, removal of non-standard symbols, punctuation, and emojis.
Stemming and tokenization: Application of Snowball stemmer to reduce morphological variance; tokenized via whitespace and delimiters.
Stopword elimination: Discarding semantically null functional words.
Part-of-speech (POS) filtering: Retention of content-bearing lexical categories (nouns, verbs, adjectives, adverbs).
Vectorization: Employed TF-IDF and Word2Vec embeddings for classical models; utilized transformer token embeddings for BERT and integer encoding for LSTM.

Due to computational constraints, a balanced sample of 5,000 instances was selected, with terms appearing in fewer than five documents excluded from the vocabulary. This preprocessing pipeline was optimized for performance fidelity and model generalization under noisy, real-world conditions.

4. Modeling

In the initial modeling phase, we evaluated two baseline models — **Decision Tree** and **XGBoost** — using the KNIME Analytics Platform. These models were trained on a **randomly stratified subset of 5,000 samples** drawn from the full dataset to ensure faster iteration and workflow modularity. However, this sampling also limited their exposure to rare classes, making performance sensitive to **class imbalance**. To mitigate this, we incorporated the **SMOTE (Synthetic Minority Oversampling Technique)** node in KNIME, which synthetically augments underrepresented categories such as *Personality Disorder* and *Suicidal Thoughts*, thereby improving recall for these classes (see the performance change in Appendix B).

We also experimented with Support Vector Machines (SVM) using KNIME's SVM Learner node on a binary classification subset of 5,000 samples. Despite reducing the class complexity (only taking Depression vs Suicidal) and dataset size, the model performed poorly and was computationally expensive to train. This is largely due to the high dimensionality of text data after TF-IDF transformation and the inefficiency of the polynomial kernel on sparse features.

Additionally, KNIME's SVM implementation lacks optimization compared to libraries like scikit-learn. This trial highlighted the limitations of using non-linear SVMs in KNIME for text classification tasks and reinforced our decision to favor models like XGBoost and to explore Python-based approaches for more scalable, efficient modeling.

As a consequence, the subsequent models — **BERT**, **LSTM**, and **SVM** — were implemented in **Python** using the **entire dataset (51,000 samples)**. This allowed these models to better capture semantic patterns and long-range dependencies, especially in the context of imbalanced classes. The larger dataset and deep learning capabilities offered improved generalization, especially for underrepresented mental health conditions.

Decision Tree Classifier

Justification: High transparency and fast training times, ideal for initial benchmarking.
Limitations: Susceptibility to overfitting and poor performance on complex, high-dimensional data.
Results: Accuracy = 62.54%, F1 Score = 0.619, Normal recall = 0.90. Performance on minority classes was markedly inferior.

XGBoost (Extreme Gradient Boosting)

Justification: Robust ensemble method with inbuilt mechanisms for handling imbalance and noise.
Advantages: Incremental learning across weak classifiers enhances performance. Showed superior recall across all categories, notably raising recall for Personality Disorder to 0.548.
Results: Accuracy = 70.7%, F1 Score = 0.70, Cohen's Kappa = 0.6113.

BERT (Bidirectional Encoder Representations from Transformers)

Justification: Pre-trained language model with bidirectional context capturing, currently the gold standard in NLP.
Implementation: Fine-tuned `bert-base-uncased` model on stratified data split; employed Trainer API for reproducibility.
Results: Accuracy = 83.6%, Macro F1 = 0.8106, Cohen's Kappa = 0.7871. Particularly strong recall for Suicidal Thoughts (0.731) and Anxiety (0.859).
Interpretation: BERT demonstrated superior contextual sensitivity and robustness to informal linguistic artifacts.

LSTM (Long Short-Term Memory Network)

Justification: Captures temporal dependencies in text, effective for sequence learning.
Architecture: Vocabulary size = 10,000; embeddings = 128-dimensional; one-layer LSTM + fully connected classifier.
Training: 10 epochs, Adam optimizer, CrossEntropyLoss; input padded to uniform sequence length.

Results: Accuracy = 74.89%, Macro F1 = 0.6630, Cohen's Kappa = 0.6727. Notably, Normal recall = 0.917; Suicidal = 0.691.
Interpretation: LSTM performed well with moderate resource use, offering a computationally feasible alternative to transformer models.

Support Vector Machine (SVM)

Justification: Strong baseline model for text classification; robust to high-dimensional feature spaces.
Implementation: Utilized TF-IDF features and linear kernel.
Results: Accuracy = 77.07%, Macro F1 = 0.7228, Cohen's Kappa = 0.6992. High recall for Normal (0.955), Anxiety (0.783), and Depression (0.711). Lower performance for Stress (0.501) and Personality Disorder (0.658).
Interpretation: SVM offers a competitive alternative with high precision and balanced class-wise performance, particularly where model interpretability is desired.

Following the feedback for the progress report, to narrow our scope and dive deeper into specific class-level distinctions, we focused on the binary classification task of identifying posts labeled as either **Depression** or **Suicidal Thoughts**. These two classes are particularly significant from a mental health perspective, and differentiating them accurately can support better-targeted interventions. We filtered the dataset to include only instances of these two classes and retrained our models using KNIME. As shown in the confusion matrices in Appendix C.

The results reflect a meaningful gain in recall for Depression, though recall for Suicidal Thoughts remains comparatively lower. Despite both models still facing challenges due to overlapping features and potential annotation noise, XGBoost remains the stronger candidate.

This analysis highlights the difficulty of separating related but distinct mental health categories, and suggests further improvements might involve deeper linguistic modeling or psychological feature engineering tailored to expressions of suicidal ideation vs. depression.

5. Evaluation

Evaluation prioritized **recall** and **F1 score**, metrics aligned with the clinical imperative to minimize false negatives. In mental health screening, missing a genuine case is far more deleterious than flagging a false positive.

The results in Table 1 reinforce the superiority of BERT in multi-class mental health classification, especially for nuanced and infrequent conditions. SVM and XGBoost offer compelling performance in settings where computational constraints or transparency are key. LSTM remains a balanced alternative, while Decision Tree is best suited for interpretability over precision.

Class	BERT	XGBOOST	LSTM	DECISION TREE	SVM
Depression	0.762	0.669	0.694	0.566	0.711
Anxiety	0.859	0.630	0.751	0.510	0.784
Bipolar	0.794	0.709	0.686	0.519	0.723
Normal	0.962	0.916	0.917	0.874	0.955
Suicidal Thoughts	0.731	0.615	0.691	0.485	0.658
Stress	0.747	0.356	0.464	0.397	0.501
Personality Disorder	0.684	0.548	0.400	0.451	0.658

Table 1

Ethical Considerations

Our project raises significant ethical considerations regarding the use of personal social media data for mental health prediction. While our model aims to detect conditions like depression or bipolar disorder based on language patterns, labeling someone as "bipolar" or "depressed" based solely on a tweet may be inaccurate, stigmatizing, or even harmful. Should individuals be notified or warned before such classifications are made? Moreover, it raises the issue of responsibility. Should this information be shared with public health authorities, used by companies, or remain private? If the questions are shared, should it without the approval of the user? These questions highlight the need for careful boundary setting between technological capabilities and ethical obligations.

We believe that such models should be deployed, if at all, under strict oversight by public health professionals, not commercial entities, and with transparency, consent, and privacy safeguards in place.

Conclusion

This investigation affirms the capability of modern NLP techniques, particularly transformer-based architectures, to discern complex affective states embedded in informal social media text. BERT emerges as the most promising solution for deployment scenarios demanding high accuracy, recall, and robustness to linguistic noise.

Nevertheless, ethical considerations—such as the potential for algorithmic bias, data misuse, and stigmatization—necessitate careful governance. Models should be subjected to continual re-evaluation,
fairness auditing, and used only in tandem with human oversight.

Future Work

* Expand to binary classification frameworks for triaging (e.g., Normal vs. At-risk).
* Integrate advanced resampling and loss adjustment (e.g., SMOTE, focal loss).
* Conduct specificity validation to ensure precision in flagging.
* Develop transparent deployment protocols with embedded ethical and legal safeguards.

Appendices

Appendix A: Model Performance Summary

Model	Accuracy	F1 Macro	F1 Weighted	Cohen's Kappa
Decision Tree	62.54%	0.619	0.625	0.5096
XGBoost	70.7%	0.7	0.707	0.6113
BERT	83.6%	0.8106	0.8367	0.7871
LSTM	74.89%	0.663	0.7477	0.6727
SVM	77.07%	0.7228	0.7673	0.6992

Table 2

Appendix D: Tools & Libraries

KNIME: Modular preprocessing and sampling pipeline
Python Libraries: Hugging Face Transformers, Scikit-learn, PyTorch
Visualization Tools: Pandas, Matplotlib, Seaborn

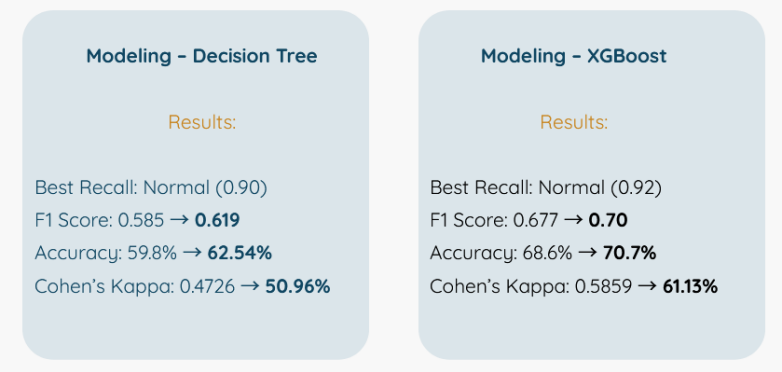
Appendix E: Data Provenance

Meta-collection: Kaggle repository by Suchintika Sarkar
Sources: Depression Reddit, Suicidal Tweets, Human Stress Chat, etc.

Appendix B: After reconfiguring the preprocessing nodes and addition of SMOTE node

Class	XGBoost Recall	Decision Tree Recall
Depression	0.669→0.669	0.533→0.566
Anxiety	0.673→0.63	0.57→0.51
Bipolar	0.611→0.709	0.429→0.519
Normal	*0.923→0.916*	*0.903→0.874*
Suicidal Thoughts	0.518→0.615	0.448→0.485
Stress	0.391→0.356	0.297→0.397
Personality Disorder	0.323→0.548	0.161→0.451

Table 3



Appendix C: Confusion Matrices

For Decision Tree

Scorer View

Confusion Matrix

	Anxiety (Pr...	Bipolar (Pre...	Depression...	Normal (Pre...	Personality ...	Stress (Pre...	Suicidal (Pr...	
Anxiety (Ac...	0	0	0	0	0	0	0	undefined
Bipolar (Ac...	0	0	0	0	0	0	0	undefined
Depression...	0	0	303	0	0	0	132	69.66%
Normal (Ac...	0	0	0	0	0	0	0	undefined
Personality...	0	0	0	0	0	0	0	undefined
Stress (Act...	0	0	0	0	0	0	0	undefined
Suicidal (A...	0	0	117	0	0	0	184	61.13%
	undefined	undefined	72.14%	undefined	undefined	undefined	58.23%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
66.17%	33.83%	0.306	487	249

For XGBoost Tree Ensemble

Scorer View

Confusion Matrix

	Anxiety (Pr...	Bipolar (Pre...	Depression...	Normal (Pre...	Personality ...	Stress (Pre...	Suicidal (Pr...	
Anxiety (Ac...	0	0	0	0	0	0	0	undefined
Bipolar (Ac...	0	0	0	0	0	0	0	undefined
Depression...	0	0	324	0	0	0	111	74.48%
Normal (Ac...	0	0	0	0	0	0	0	undefined
Personality...	0	0	0	0	0	0	0	undefined
Stress (Act...	0	0	0	0	0	0	0	undefined
Suicidal (A...	0	0	123	0	0	0	178	59.14%
	undefined	undefined	72.48%	undefined	undefined	undefined	61.59%	

Overall Statistics

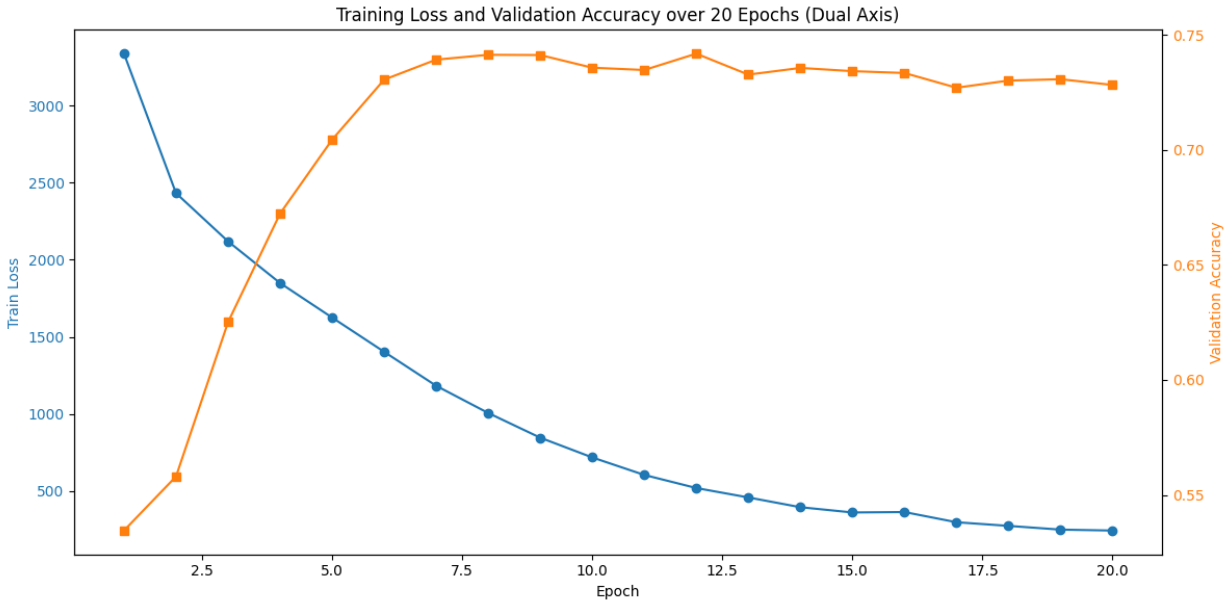
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
68.21%	31.79%	0.338	502	234

Appendix F: Representative Sample Entries

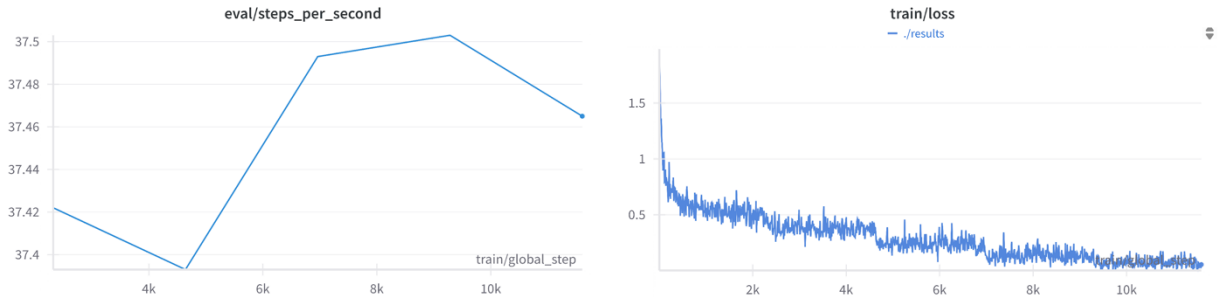
ID	Statement	Status
001	"I feel like nothing matters anymore..."	Depression
002	"Just got promoted today! So happy!"	Normal
003	"I want it to end. No one understands."	Suicidal Thoughts
004	"Exams are killing me. Can't sleep or eat."	Stress

Appendix G: Training Results

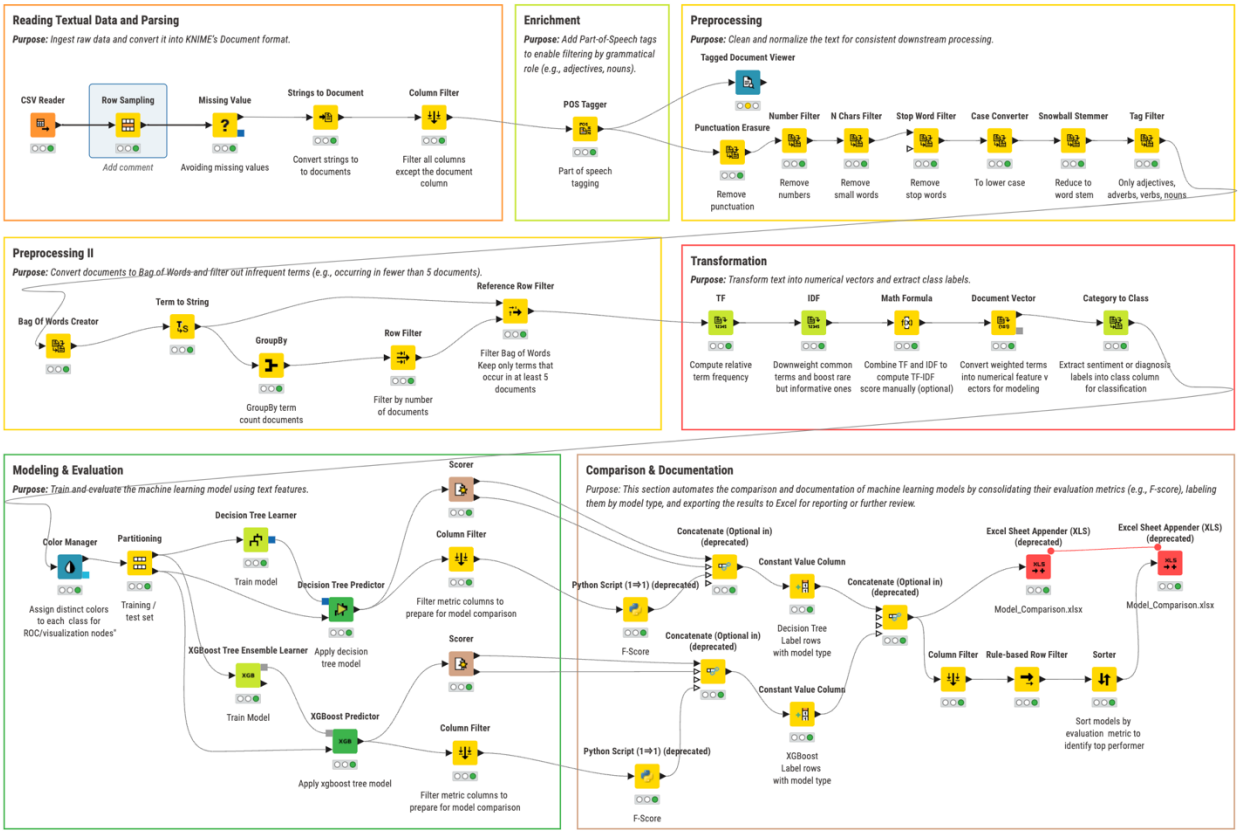
For LSTM



For BERT



Appendix G: KNIME Workflow



References

https://www.knime.com/blog/sentiment-analysis
https://www.knime.com/blog/visual-scoring-techniques-for-classification-models