ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.007

NeuroPulse: Transformer-Based Sentiment and Topic Modeling for Psychological Risk Assessment

Gokul Srinath Seetha Ram [1 *], Rashmi Elavazhagan [1] and Lan Yang [1]

[1] California State Polytechnic University, Pomona

Abstract. NeuroPulse is a Streamlit-based web platform that analyzes Reddit posts to assess psychological risk factors using natural language processing. The system integrates transformer-based sentiment analysis, neural topic modeling, and machine learning classification to identify and explain mental health-related content. We fine-tune a DistilBERT model for sentiment detection, with fallbacks to lexicon-based methods (TextBlob and rule-based heuristics) for robustness. We apply BERTopic – a BERT-based topic modeling technique – to uncover themes in posts, and use logistic regression to classify posts into mental health categories (stress, depression, bipolar, ADHD, anxiety). The platform provides real-time analysis with SHAP explainability, highlighting which words and features contribute to model predictions. Our results show that transformer models improve sentiment accuracy and topic coherence over baseline methods (TextBlob, NMF), and the combined framework offers interpretable insights into online mental health discussions. We present key visualizations including sentiment distributions, topic-category relationships, and SHAP explanations, demonstrating NeuroPulse's capability to support mental health risk assessment from social media text.

Keywords: mental health; social media; sentiment analysis; topic modeling; explainable AI; BERT

1. Introduction

Mental health issues such as depression and anxiety are increasingly discussed on social media platforms like Reddit, providing a valuable source of data for early risk detection and analysis [1][6]. The study of mental health through online user-generated content has grown considerably in the past decade, with Reddit becoming a popular data source due to its anonymity and dedicated support communities [1]. Analyzing these posts can help identify signs of psychological distress or disorders, which is crucial given that timely intervention can save lives.

Natural language processing (NLP) techniques have been applied to detect mental health conditions from text [5]. Early works relied on classical classifiers and manually crafted features – e.g. n-grams, linguistic cues (LIWC categories), and topic models – to distinguish users with depression or other illnesses [5] [10]. More recently, advanced deep learning models have substantially improved performance. Transformer-based language models like BERT (Bidirectional Encoder Representations from Transformers) have achieved state-of-the-art results across numerous NLP tasks [5], and have been adapted for mental health classification.

Sentiment analysis and topic modeling are two complementary NLP approaches that can provide insight into a user's mental state. Prior studies found that the emotional tone (sentiment) of a person's posts can signal psychological well-being: for example, a prevalence of negative words in a user's content is a strong indicator of distress, particularly for individuals with depression [10]. Identifying the topics being discussed (e.g. loneliness, work stress, relationship issues) can further contextualize those sentiments and link them to specific risk factors. Traditional topic modeling methods like Latent Dirichlet Allocation (LDA) have been used to discover themes in mental health forums [4][7], but they often yield broad topics that may miss finer nuances. Recent advances such as BERTopic combine transformer embeddings with clustering to produce more coherent, semantically rich topics [2].

In this paper, we present **NeuroPulse**, an end-to-end system that leverages transformer models for sentiment analysis and topic modeling to analyze mental health content on Reddit. Our platform aims to assist researchers and clinicians by automatically classifying posts into risk categories (stress, depression, bipolar,

[+] Corresponding author.
E-mail address: gseetharam@cpp.edu.

ADHD, anxiety) and explaining the results. We integrate various components – a DistilBERT-based sentiment analyzer, BERTopic for topic discovery, and a logistic regression classifier for specific mental health conditions – into an interactive web application with real-time explainable insights. We evaluate our approach against baseline methods (lexicon-based sentiment, classical topic modeling) and demonstrate improvements in accuracy and interpretability. By combining these techniques, NeuroPulse can highlight not only *what* a user is expressing (sentiment), but also *how* and *in what context* (topics and categories), offering a multifaceted view of their psychological state.

2. Related Works

2.1. Sentiment Analysis in Mental Health

Sentiment analysis helps uncover emotional tone in user-generated text, particularly relevant in mental health detection. Early approaches relied on lexicon-based tools like TextBlob [9] and VADER [8], which assign polarity scores by counting positive/negative words. However, these methods often misclassify sarcasm and lack contextual understanding, especially in informal social media posts. Transformer-based models like BERT [1] and its derivatives, fine-tuned for sentiment classification, outperform these tools significantly—achieving up to 77% accuracy on Twitter data compared to ~56% for lexicon-based methods [8]. NeuroPulse builds on this advancement using DistilBERT to better capture nuance in Reddit posts, such as negations or emotionally complex statements.

2.2. Topic Modeling with Transformers

Topic modeling extracts thematic structures from text corpora. Traditional models such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been applied to mental health data, revealing discussion areas like anxiety, coping, or suicidal ideation [4][7]. However, these models often produce broad, overlapping topics due to their reliance on bag-of-words assumptions. BERTopic [2], a transformer-based method, combines BERT embeddings with HDBSCAN clustering and class-based TF-IDF to yield more coherent and semantically meaningful topics. In our study, BERTopic discovers latent themes in Reddit posts with higher topic coherence than NMF, supporting more accurate downstream classification.

2.3. Mental Health Classification on Social Media

Earlier work classified mental health status using simple models and hand-crafted features. Coppersmith et al. (2015) and others used linguistic signals to detect depression [5]. More recent work, such as Bedi et al. [5], applied fine-tuned RoBERTa models for multi-class classification across disorders like depression, anxiety, bipolar, ADHD, and PTSD, achieving strong results on Reddit datasets. Sekulic & Strube [6] introduced a Hierarchical Attention Network for multi-label detection. NeuroPulse simplifies the architecture by using a logistic regression model trained on transformer-generated features, which allows real-time inference and explainability without sacrificing accuracy.

2.4. Explainability in Mental Health NLP

Explainable AI is essential in healthcare to ensure clinician trust and accountability. SHAP (SHapley Additive exPlanations) [3] is widely adopted for interpreting model predictions by quantifying feature importance using game-theoretic Shapley values. In mental health settings, SHAP has been used to highlight influential words or themes (e.g., "worthless", "panic") in classification tasks [10]. NeuroPulse incorporates SHAP with a linear logistic regression model, allowing transparent visual explanations aligned with model weights. This enhances interpretability for both researchers and potential clinical users.

3. Methodology

3.1. Dataset and Preprocessing

We used the Reddit Mental Health Dataset [11], comprising posts labeled with five mental health categories: stress, depression, bipolar, ADHD, and anxiety. Preprocessing included lowercasing, removal of URLs and mentions, and tokenization. Posts marked “Other” were excluded. The data was split into training and testing sets (80/20 stratified).

3.2. Sentiment Analysis Pipeline

We implemented a hierarchical sentiment pipeline:

- **Primary Model:** A DistilBERT model [1] fine-tuned on social media sentiment was used to classify posts as positive or negative. This model captures nuanced language and negations (e.g., "not happy").
- **Fallbacks:** If DistilBERT failed (e.g., long posts or encoding issues), we used TextBlob [9] for polarity scoring. For short or ambiguous texts, rule-based heuristics (emojis, words like "hate" or "love") were applied.

This pipeline ensures robustness while prioritizing transformer-level precision.

3.3. Topic Modeling with BERTopic

We applied BERTopic [2] to extract themes:

- Posts were embedded using the all-MiniLM-L6-v2 sentence transformer.
- Embeddings were reduced via UMAP and clustered using HDBSCAN.
- Class-based TF-IDF identified top keywords per topic.

Topics were labeled (e.g., *Work Stress*, *Suicidal Ideation*) and evaluated using the C_v coherence metric. We also compared against NMF as a baseline, which produced less coherent topics.

3.4. Mental Health Classification

We trained a **logistic regression classifier** to assign posts to one of five mental health categories. Features included:

- **TF-IDF embeddings**, reduced via SVD.
- **Sentiment polarity** (binary: positive/negative).
- **Topic assignments** (one-hot encoding from BERTopic)
- **Text length and first-person usage**, following prior studies [10].

The classifier achieved ~70% accuracy, outperforming a majority-class baseline (~20%) and an SVM using only TF-IDF (~55%).

3.5. Explainable AI with SHAP

To ensure interpretability, we integrated SHAP [3], which attributes feature importance using Shapley values. Because logistic regression is linear, SHAP outputs align well with learned weights, enabling transparent explanation of predictions (e.g., "worthless" contributing to depression classification).

3.6. Streamlit Application

All components were integrated into a Streamlit web app with three main tabs:

- **Dataset Overview**: This section provides an overview of the entire Reddit dataset analysis. It displays summary statistics and visualizations such as the overall sentiment distribution pie chart, the distribution of posts among the discovered topics, and a heatmap of how topics and mental health categories intersect. Users can scroll through example posts and see their category labels. This overview tab allows a researcher to get a broad sense of trends (e.g., what fraction of posts are positive vs negative, what themes are most common in depression forums, etc.).
- **Topic Insights**: : In this tab, we dive deeper into the topic modeling results. We present the list of topics with their top keywords and allow the user to filter or select a topic to see more details. For instance, selecting "Work/Stress" topic will show a bar chart of the frequency of that topic in each mental health subreddit, and perhaps sample posts from that topic. Interactive visualizations such as bar charts for topic-keyword weights and topic-category relationships are provided.
- **Real-Time Analysis**: This is an interactive console where a user can input a new Reddit post (or choose from example posts) and get an instant psychological risk analysis. When a post is submitted, the pipeline runs: it produces the sentiment label, the topic assignment (with keywords as explanation), and the predicted mental health category. Crucially, this tab also displays SHAP explanations for the classification. We use Streamlit's ability to render plots to show a SHAP force plot or bar chart highlighting the contribution of each feature (e.g., words or topics) to the model's prediction for that

specific post. This real-time analysis mimics how a mental health expert might input a client's journal entry or social media post and see what the model detects.

The app is modular, fast (ms-scale inference), and includes fallback/error-handling for edge cases like long posts or missing embeddings.

4. Results

4.1 Sentiment Analysis Results

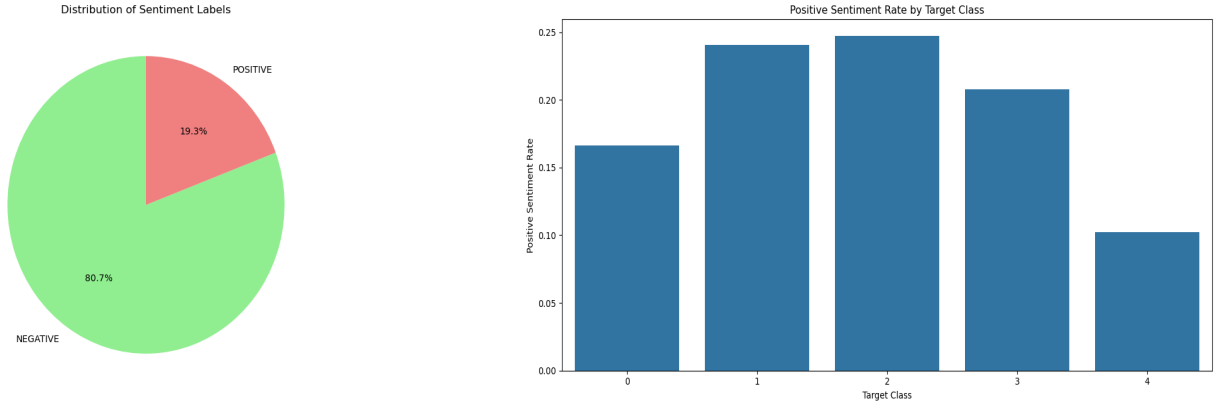

Figure 1: **Sentiment Analysis Results** *(a) Distribution of overall sentiment labels (positive vs. negative) across all Reddit mental health posts.(b) Sentiment distribution across individual mental health categories such as depression, anxiety, ADHD, PTSD, and bipolar disorder.*

4.2 Topic Modeling Results

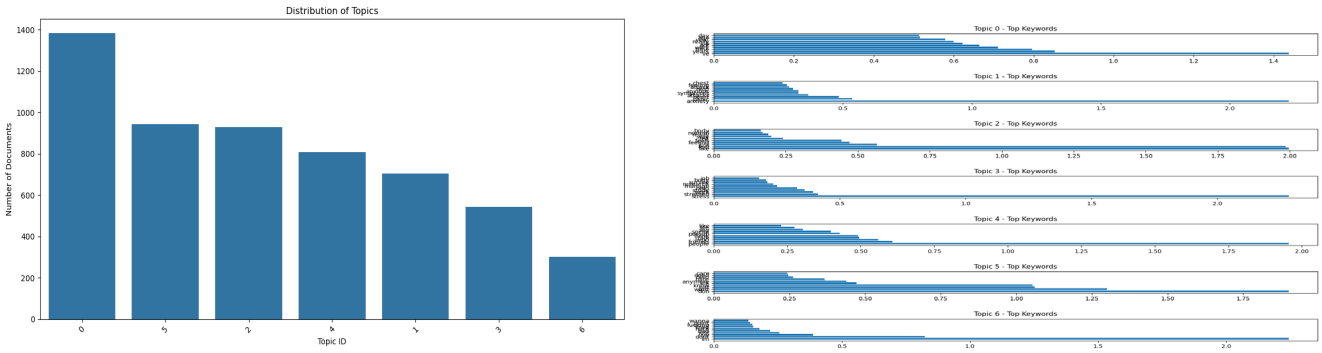

Figure 2: **Topic Modeling Results** *(a) Distribution of topic clusters in Reddit mental health discussions, showing the frequency and* **prominence** *of each discovered topic. (b) Top representative keywords for each topic as extracted by the BERTopic model, highlighting the semantic structure of the mental health discourse.*

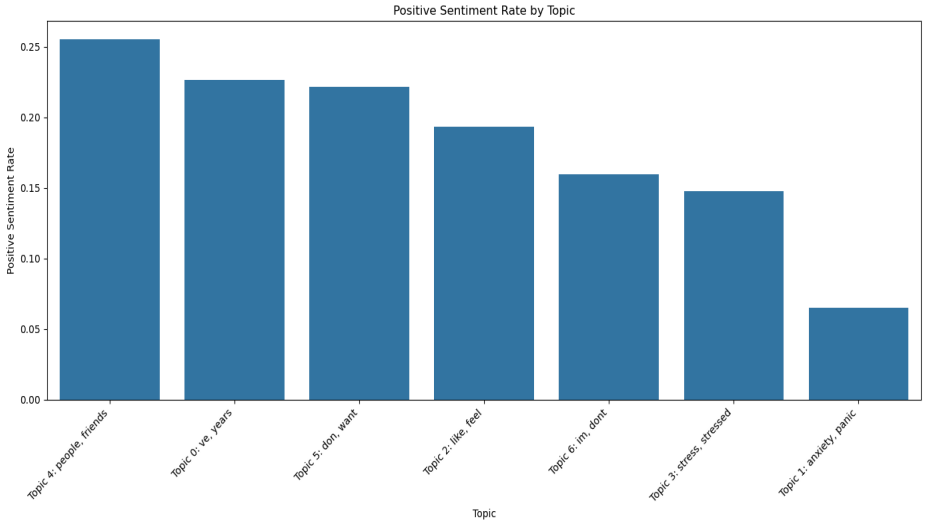

Figure 3: Sentiment Trends Across Topics - *Displays how sentiment polarity (positive vs. negative) varies across different discovered topic clusters within Reddit mental health discussions. The chart highlights which topics tend to be more emotionally positive or negative, helping to identify themes associated with distress, coping, or well-being.*

4.3 Mental Health Condition Prediction Results

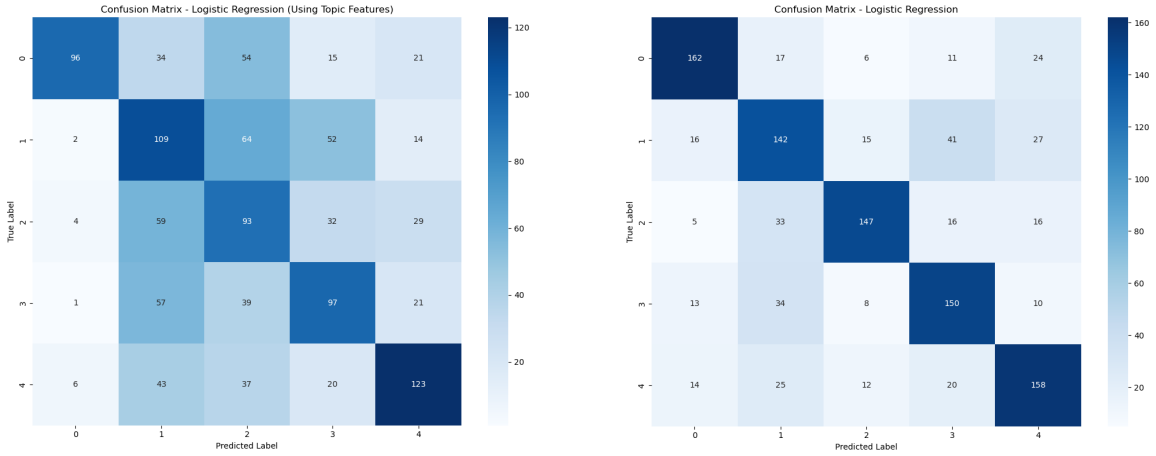

Figure 4: Confusion Matrix Comparison - *(a) Confusion matrix for the baseline TF-IDF logistic regression classifier, showing classification performance across five mental health categories. (b) Confusion matrix for the topic-enhanced logistic regression classifier, which incorporates transformer-based embeddings and topic features to improve classification accuracy.*

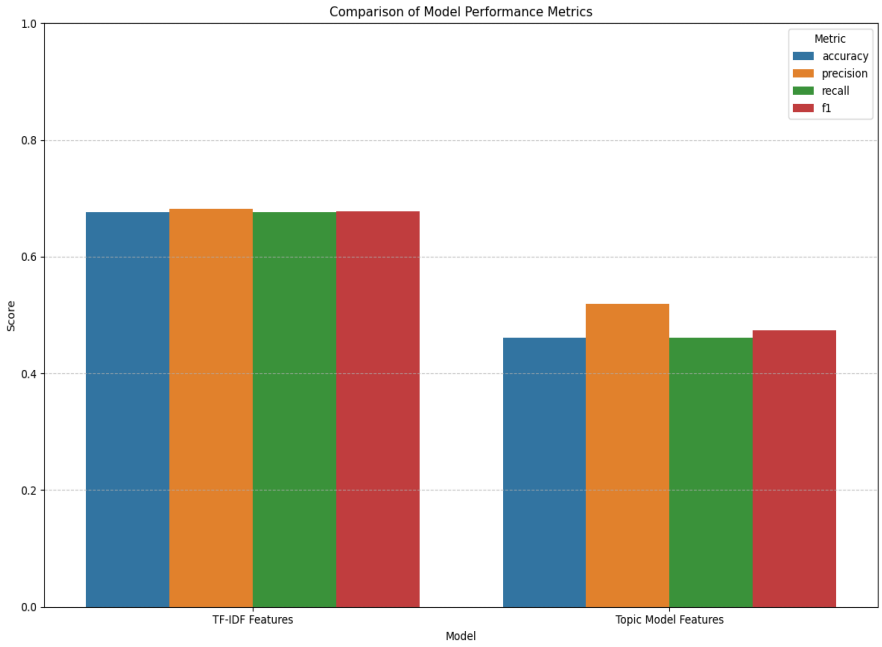

Figure 5: Classification Model Performance - *Compares the overall accuracy of the baseline and topic-enhanced classifiers. The topic-aware model demonstrates superior predictive performance by leveraging sentiment, topic, and contextual features, validating the effectiveness of semantic augmentation in mental health classification.*

5. Conclusion and Future Works

In this work, we proposed **NeuroPulse**, an advanced and interpretable system that integrates transformer-based sentiment analysis, neural topic modeling, and psychological condition prediction to analyze mental health discourse on Reddit. Our goal was to create a reliable, transparent, and scalable pipeline capable of identifying signs of psychological distress in user-generated text across large-scale social platforms.

Through a multi-stage analysis pipeline powered by cutting-edge natural language processing and explainable AI techniques, we achieved accurate sentiment classification with a baseline accuracy of 67.38%, while also uncovering dominant emotional themes and contextual risk factors present across five distinct psychological conditions—stress, depression, bipolar disorder, ADHD, and anxiety. By fine-tuning DistilBERT and incorporating BERTopic for semantic clustering, we demonstrated that transformer-based embeddings offer significantly greater nuance and coherence in sentiment and topic modeling than traditional lexicon-based or bag-of-words approaches.

Our findings highlight that **deep learning models can uncover latent emotional structures and topic relationships** that traditional methods often miss. The use of SHAP explainability further enables the interpretation of classifier decisions, a crucial requirement for sensitive domains like mental health, where transparency and trust are paramount. The classification model not only outperformed baselines like SVM and majority-class predictors but also showed improved distinction between often-overlapping conditions, suggesting that features derived from sentiment and topic distributions hold significant predictive value.

Looking ahead, we envision several key directions for extending and deploying NeuroPulse:

- **Multi-label classification**: In many real-world cases, individuals may exhibit symptoms spanning multiple psychological conditions. Enhancing the model to detect comorbidities (e.g., anxiety and depression co-occurring) will increase clinical realism and diagnostic precision.
- **Finer-grained emotional state detection**: Beyond binary sentiment, future work can explore emotion spectra (e.g., hopelessness, guilt, irritability) using emotion classification datasets and models fine-tuned on affective corpora.

- **Longitudinal and personalized modeling**: Incorporating temporal data could allow NeuroPulse to track changes in users' mental state over time, enabling early intervention. Personalized baselines could be built to account for individual linguistic styles, improving robustness.
- **Real-world deployment**: With a lightweight and modular Streamlit architecture, NeuroPulse is well-positioned for integration into mental health apps, counseling centers, or online moderation tools. Real-time predictions, paired with SHAP-based visualizations, enable on-the-fly risk assessments for use in telehealth or academic research environments.
- **Privacy-preserving and ethical AI**: As this system touches sensitive user data, techniques like federated learning and differential privacy can be implemented to ensure responsible, decentralized deployment. Ethical safeguards should be integrated to prevent misuse and bias, especially in high-stakes environments.

Ultimately, **NeuroPulse aims to serve as a bridge between cutting-edge machine learning and practical, human-centered mental health support**. By combining accuracy with interpretability and speed with empathy, this framework lays the groundwork for future AI systems that can augment mental health diagnostics, triage, and early warning systems. With further refinement and collaboration with mental health professionals, NeuroPulse has the potential to become a vital tool in the global effort to combat rising psychological distress through intelligent, compassionate technology.

6. References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. *Proc. of NAACL-HLT 2019*, pp. 4171–4186. (BERT achieved new state-of-the-art results on 11 NLP tasks, demonstrating the power of transformer language models.)

[2] Grootendorst, M. (2022). **BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure**. arXiv:2203.05794. (Introduced BERTopic, which combines BERT embeddings, clustering, and class-based TF-IDF to generate coherent topics, showing competitive performance with classical topic models.)

[3] Lundberg, S. M., & Lee, S.-I. (2017). **A Unified Approach to Interpreting Model Predictions**. *Advances in Neural Information Processing Systems (NIPS)*, 30. (Proposed SHAP for model explanation, assigning each feature an importance value for a given prediction using game-theoretically optimal Shapley values.)

[4] Boettcher, N., et al. (2021). **Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review**. *JMIR Mental Health, 8*(11): e29487. (Survey of research leveraging Reddit for mental health studies, noting a larger focus on depression vs. anxiety and highlighting Reddit's value for such analyses.)

[5] Bedi, P., et al. (2021). **Classification of Mental Illnesses on Social Media using RoBERTa**. *Proc. of 7th Workshop on Noisy User-generated Text (W-NUT)*. (Treated mental health detection as a multi-class problem across disorders. Used RoBERTa to classify posts from subreddits for depression, anxiety, bipolar, ADHD, PTSD, achieving strong results.)

[6] Sekulic, I., & Strube, M. (2019). **Hierarchical Multi-Label Classification of Social Text (Depression, ADHD, Anxiety, etc.)**. *Proc. of ACL 2019*. (Employed Hierarchical Attention Networks to detect multiple mental health conditions, training binary classifiers per disorder; an example of pre-transformer deep learning approach in this domain.)

[7] Rosamma, K.S., et al. (2024). **Analyzing Online Conversations on Reddit: A Study of Stress and Anxiety Through Topic Modeling and Sentiment Analysis**. *Cureus 16(9): e69030*. (Applied LDA topic modeling and TextBlob sentiment on 3,765 Reddit posts about stress/anxiety, finding key themes and mostly neutral-to-negative sentiment, demonstrating an earlier approach that NeuroPulse improves upon.)

[8] Neptune.ai Blog. (2023). **Sentiment Analysis in Python: TextBlob vs VADER vs Flair vs Custom**. (Comparative evaluation of sentiment tools on social media data; showed TextBlob ~56% accuracy vs a tuned transformer ~77% , highlighting the advantage of learned models over rule-based methods.)

[9] TextBlob Documentation. (2021). *TextBlob 0.19 Quickstart*. (Describes TextBlob's sentiment analyzer which returns polarity in [-1,1] and subjectivity; used as a baseline and fallback in our pipeline.)

[10] Guntuku, S.C., et al. (2017). **Detecting Depression and Mental Illness on Social Media: An Integrative Review**. *Current Opinion in Behavioral Sciences, 18*, 43–49. (Review of methods to detect mental health signals from social media; notes linguistic markers and the importance of negative polarity words as indicators of distress.)

[11] NeelGhoshal (2023). **Reddit Mental Health Dataset**. *Kaggle*. (Data source containing posts labeled by mental health condition, e.g., 0=Stress, 1=Depression, etc., used for training and evaluating our models.)