TriggerLens: Predicting Anxiety Triggers from Reddit Data

Team Members: Maria Mckay, Shen Shu, and Shuting He

Introduction

Mental health discussions on social media can be freeing and uniting, but they also risk amplifying anxiety-provoking content. With 19.1% of U.S. adults experiencing an anxiety disorder annually and 31.1% facing one in their lifetime (National Institute of Mental Health, n.d.), understanding how online content triggers anxiety has become increasingly crucial. Yet while these platforms have connected people across the globe, studies suggest that constant information exposure has left many feeling more isolated than ever (Primack et al., 2017).

The problem our team set out to solve is whether we can identify if a given Reddit post and the language surrounding it will "trigger" anxiety in readers. This subject presents a complex challenge that requires careful consideration of competing values. Attempts to remove or flag content raise essential questions about freedom of speech and censorship. However, we believe there is a middle ground: instead of censoring discussions, we can inform users that content may provoke uncomfortable sentiments. This approach empowers users to make informed choices about their consumption while validating their emotional responses and preserving open forum discourse.

Our motivation stems from the growing need to strike a balance between mental health support and user autonomy in online spaces. As social media becomes more central to discussions about mental health, understanding and addressing potential harms without resorting to censorship is essential. To address this challenge, we developed TriggerLens, an analytical framework that combines unsupervised and supervised machine learning to both identify discussion themes and predict "trigger scores" for Reddit posts. Our goal is to contribute research that informs future content moderation strategies, aiming to reduce harmful exposure while preserving the valuable aspects of online mental health communities. Beyond content moderation, this research has practical applications for mental health professionals seeking to understand online discourse patterns and for platform designers developing user-centered mental health tools.

Methods Summary

Our project combines both unsupervised and supervised machine learning approaches to address anxiety trigger detection in Reddit content. For unsupervised learning, we applied Non-negative Matrix Factorization (NMF) and BERTopic to discover latent thematic patterns in 6,283 unlabeled Reddit posts. NMF identified 15 coherent topics spanning clinical symptoms, interpersonal relationships, technical/productivity content, and discourse patterns, achieving NPMI coherence of 0.725. BERTopic achieved comparable coherence (0.730) with better topic diversity (0.97) but produced fewer topics (7), demonstrating complementary strengths between bag-of-words and semantic embedding approaches.

For supervised learning, we engineered features combining TF-IDF, NMF topic distributions, and metadata, then trained Logistic Regression, Random Forest, and DistilBERT models on a hybrid dataset of 1,006 posts. All three models were successfully implemented, with DistilBERT achieving the best performance (AUC 0.956), followed by Logistic Regression (AUC 0.901) and Random Forest (AUC 0.847).

Our approach improves the detection of mental health content through three main innovations. First, we develop a hybrid weak supervision framework that combines 599 human annotations with 407 GPT-based weak labels (OpenAI, 2023) to create a scalable 1,006-post dataset, addressing the scalability limitations of manual annotation approaches (KS, 2024) and existing labeled datasets (Shen & Rudzicz, 2017; Oryngozha et al., 2024). Second, we systematically compare NMF and BERTopic topic modeling approaches, achieving 0.725 NPMI coherence with 15 distinct NMF topics compared to previous LDA-based studies with fewer themes. We integrate these topics as supervised features, yielding interpretable representations (9,699 TF-IDF + 15 topics + 3 metadata) that provide measurable performance gains over TF-IDF alone (AUC 0.847 vs 0.842). This approach builds upon previous n-gram and embedding foundations, with explainable predictions essential for clinical deployment. Third, we extend beyond binary classification to predict continuous anxiety severity (0-5 scale) across eight diverse subreddits, revealing context-dependent anxiety expression patterns that could inform content moderation approaches across different online communities.

Main findings

For supervised learning, our main findings were that DistilBERT achieved the best performance with a test AUC of 0.956 and average precision of 0.958, demonstrating that transfer learning approaches outperformed traditional feature engineering methods. Logistic regression achieved a test AUC of 0.894 with 94.3% recall (33/35 true positives were identified) at the optimal threshold of 0.257, prioritizing high recall to minimize false negatives. Random Forest achieved an AUC of 0.847 with a comprehensive failure analysis revealing systematic under-prediction and feature representation limitations.

For unsupervised learning, NMF achieved NPMI coherence of 0.725, topic similarity of 0.204, and assignment purity of 71.5%. The 15 topics span clinical symptoms, interpersonal relationships, technical/productivity content, and discourse patterns, revealing thematic diversity across various online communities. BERTopic achieved slightly higher coherence (0.730) with better topic diversity (0.97) and low outlier rates (5%), but with less granularity (7 topics vs. 15), showing that both approaches capture genuine semantic patterns with complementary strengths.

Related Work

Previous research in mental health content analysis has focused on specific aspects of anxiety and stress detection in online communities. KS (2024) analyzed 3,765 Reddit posts to identify stress and anxiety themes using LDA and TextBlob sentiment analysis. While this project uses Reddit data, our study focuses on a different dataset and improves on this approach by using

NMF and BERTopic. We further extend the task by using the NRC Emotion Lexicon to score anxiety levels within each topic.

Shen and Rudzicz (2017) developed a Reddit dataset distinguishing anxiety from non-anxiety posts using n-grams, word embeddings, and lexicon-based features for classification. In contrast, our study employs NMF and BERTopic to enhance interpretability and leverages transfer learning models like DistilBERT along with advanced machine learning classifiers (Random Forest and Logistic Regression) for anxiety-level prediction.

Oryngozha et al. (2024) focused on detecting and analyzing stress-related content within academic Reddit communities, classifying posts as stressful or not using NLP and machine learning techniques. Our work instead targets a broader range of communities (e.g., economy, unpopular opinions) and integrates both human and AI-assisted annotations to improve the accuracy of anxiety-level labeling.

Data Source(s)

We pull unlabeled Reddit posts from mental-health subreddits via the Reddit API using the collectors in *src/pull_reddit.py* and *src/pull_reddit_compliant.py*. The raw exports (JSONL) live under data/raw/, while the cleaned and merged datasets (Parquet) used for modeling are saved in data/processed/. We totally extracted 6,283 posts and covered a time range from 2024-09-27 to 2025-09-18.
The Processed table *data/processed/reddit_anxiety_v1.parquet* contains variables used across our unsupervised and supervised models, notably subreddit (e.g., r/Anxiety) and *text_all*, the primary modeling text formed by concatenating cleaned *title*, *selftext*, and top-K comments. We applied English-only filtering and excluded NSFW/removed content; comments were aggregated from top-scored replies (top-K, up to 5), and posts with fewer than 7 tokens in *text_main* were removed. *text_all* served as the canonical text for topic modeling (TF-IDF+NMF, BERTopic) and downstream classifiers.

Feature Engineering

We collected Reddit posts using the Reddit API, storing titles, text, and comments as JSON data from eight subreddits: r/Anxiety, r/HealthAnxiety, r/mentalhealth, r/GetMotivated, r/TrueOffMyChest, r/economy, r/OpenAI, and r/unpopularopinion. The preprocessing pipeline transformed raw Reddit data into structured features with several key steps:

First, we applied domain-specific text cleaning, including expansion of contractions (e.g., don’t → do not), removal of Reddit-specific artifacts ([removed], [deleted], [text], [image]), and normalization of Unicode characters (curly quotes → straight quotes). We preserved negation terms (not, never, no) for sentiment analysis while removing URLs and standardizing punctuation.

Second, a custom tokenization function was implemented, creating domain-specific rules for mental health content: strip punctuation from word boundaries, lowercase all text, and remove words < 3 characters except for domain-specific short terms (ecg, sad, ptsd, mom, dad, anx) and filter out Reddit-specific noise tokens.

Third, for each post, we extracted the top-5 comments ranked by score (depth=0 only) and concatenated them with the post title and self-text to create the comprehensive `text_all` field. This enabled rich contextual information while balancing quality with score-based filtering.

To handle noisy data, we applied a minimum token threshold (7 tokens per post) to filter out mostly empty posts. This resulted in 6,283 high-quality posts from the initial 6,507 submissions. Posts with missing or corrupted text fields were excluded from the final dataset. For graceful handling of posts with missing comments, we flagged these posts with `comments_missing=1`. In this case, Reddit may have reported comments, but none were captured, which ensures transparency about data completeness. To allow for more robustness, we also used fillna("") operations to handle null text fields and applied nan_to_num() transformations to get numerical stability in sparse matrix operations.

Our feature engineering pipeline produces 9,716 total features per post across three categories.

- The lexical component consists of 9,699 TF-IDF scores generated using a fitted vectorizer with custom tokenization and domain-specific stopword filtering. These features represent word importance scores that capture specific anxiety-related vocabulary (panic, attack, worried) and symptom language.
- The semantic component comprises 15 NMF topic distributions extracted using a fitted Non-negative Matrix Factorization model. Each feature represents the post's affinity to one of 15 discovered themes spanning clinical symptoms, interpersonal relationships, technical/productivity content, and discourse patterns.
- The metadata component includes 2-3 contextual features: document length (character count of combined title, selftext, and comments), URL presence (binary indicator for posts originally containing URL links), and an optional NRC anxiety score (0-1 range) derived from NRC Emotion Lexicon word-emotion associations (Mohammad & Turney, 2013). The feature count varies by ±1 because the NRC anxiety score was not consistently available across all model implementations, representing a limitation in the feature engineering pipeline.

The sparse TF-IDF matrix is horizontally stacked with dense topic and metadata features using scipy.sparse.hstack(). The final matrix is stored in CSR (Compressed Sparse Row) format for memory efficiency during model training. This yields 9,717 total features for Random Forest and Logistic Regression classifiers (9,699 TF-IDF + 15 topics + 3 metadata). DistilBERT, by contrast, processes raw text directly through its transformer architecture and generates contextual embeddings internally, bypassing the explicit feature engineering pipeline described above.

Part A: Supervised Learning

We developed supervised models using two labeled datasets: human-labeled (599 posts) and AI-labeled (1,000 posts), both stratified by subreddit. Three models were tested: DistilBERT, Random Forest, and Logistic Regression. DistilBERT directly compared the two datasets and achieved the best results using the AI-labeled data with raw text input. Random Forest and Logistic Regression used TF-IDF, NMF topic features, and meta features (document length, URL presence, optional NRC score). Both compared AI-, human-, and combined datasets, with the combined dataset performing best.

Learning Methods

Our supervised learning approach employs three complementary machine learning methods to capture different aspects of anxiety-related content detection. DistilBERT serves as our pretrained, distilled version of BERT, retaining most of BERT's language understanding capabilities while being smaller, faster, and more computationally efficient. In our study, this efficiency enables fine-tuning a transformer-based model for anxiety-level classification with strong semantic accuracy and reduced training time, making it particularly suitable for capturing nuanced contextual relationships in mental health discussions.

We selected Random Forest for its ensemble robustness, interpretable feature importance, and ability to capture non-linear interactions in our high-dimensional feature space. The model is configured with 600 estimators, balanced subsample class weighting to address the 17.3% positive class imbalance, and unlimited depth to handle complex feature relationships. This approach provides valuable insights into which features contribute most to anxiety detection while maintaining strong predictive performance through bootstrap aggregation and random feature selection.

Finally, we include a sparse linear baseline using L2-regularized Logistic Regression with class weighting to handle the 17.3% positive class imbalance. Logistic Regression is fast and robust on high-dimensional sparse text features, providing calibrated probabilities that we can tune to favor high recall, minimizing missed triggers. In our pipeline, Logistic Regression consumes TF-IDF vocabulary, 15 NMF topic features, and 2-3 metadata features, yielding a compact yet expressive representation for reliable ranking of anxiety-triggering posts.

Feature Representations

Our supervised learning models employ different feature representation strategies tailored to their specific architectures and strengths. DistilBERT uses its own tokenizer to convert text into tokens, embeds them into contextualized vectors that capture semantic meaning, and provides these representations as input features to downstream classifiers. This approach leverages pre-trained language understanding to capture nuanced contextual relationships that may not be apparent in traditional bag-of-words representations.

Random Forest operates on 9,716 total features, combining 9,699 TF-IDF scores capturing anxiety-specific vocabulary, 15 NMF topic distributions representing semantic themes, and two metadata features (document length, URL presence). The sparse TF-IDF matrix is horizontally stacked with dense features using scipy.sparse.hstack() in CSR format for memory efficiency, enabling the ensemble method to capture both lexical patterns and broader thematic structures.

Logistic Regression operates on a horizontally-stacked feature matrix combining TF-IDF (10,000 n-grams), 15 NMF topic distributions, and 2-3 metadata features (document length, URL flag, optional NRC anxiety score). In our implementation, this configuration produced a (1006 × 10,018) matrix; elsewhere in the report, we also use a 9,699-term TF-IDF vocabulary (9,717 total features) for tree models, though both configurations follow the same design principles.

Hyperparameter Tuning

Our hyperparameter optimization approach varied across models based on their characteristics and computational requirements. For DistilBERT, due to limited time and computational resources, we tested only three hyperparameter configurations: varying epoch size, learning rate, and weight decay, and proceeded with the dataset that demonstrated better performance (AI-labeled data). This constrained approach was necessary given the computational intensity of transformer model training.

Random Forest used fixed hyperparameters without extensive tuning due to its inherent robustness to overfitting. We configured the model with 600 estimators, unlimited max depth, minimum 2 samples per leaf, and balanced subsample class weighting to address the 17.3% positive class imbalance. The model leverages bootstrap aggregation and random feature selection to handle the high-dimensional, sparse feature space. It maintains interpretable feature importance rankings, making it suitable for our ensemble approach without requiring extensive hyperparameter search.

Logistic Regression underwent comprehensive hyperparameter optimization using class_weight="balanced" and max_iter=4000. Our search space included C ∈ {0.01, 0.03, 0.1, 0.3, 1, 3, 10}, penalty ∈ {l2, l1}, and solver ∈ {liblinear, saga}. Validation employed RepeatedStratifiedKFold (5×2) with CV AUC = 0.840 ± 0.039 and CV AP = 0.529. The best parameters were C=3.0, penalty=l2, and solver=liblinear. All features were fit within each CV fold to avoid leakage, and model selection used mean ROC-AUC across splits. After cross-validation, we refit the full training set, apply isotonic calibration, and set the threshold at 0.257 by maximizing F1 on the calibration curve.

Supervised Evaluation

We report ROC–AUC for ranking quality and Average Precision (AP) to reflect performance under class imbalance. To avoid leakage, we select a single operating threshold on a held-out calibration/validation split, then apply it to the test set to report Precision/Recall/F1 and a confusion matrix. The test set prevalence is ~17.3% positives (35/202), so AP is especially informative.

Summary of Results

Model Family	ROC–AUC	Average Precision
DistilBERT	0.956	0.958
Random Forest	0.847	0.564
LogReg	0.9013	0.5714

DistilBERT leads on both AUC and AP, confirming the advantage of transfer learning. Logistic Regression is competitive, especially on AUC and with a recall-oriented threshold, minimizes missed triggers while remaining transparent and low-latency. Random Forest trails DistilBERT on AUC but contributes complementary non-linear signals over sparse text features.

In-Depth Evaluation

Feature Importance: DistilBERT's self-attention layers capture contextual relationships between anxiety related terms and surrounding linguistic context. This enables the model to distinguish genuine anxiety triggers from casual mentions. This semantic understanding explains the superior performance compared to traditional approaches that rely on lexical features alone.

Ablation Analysis: We evaluated whether stop-word removal affects DistilBERT's performance, we tested three sets: 1) ENGLISH_STOP_WORDS plus high-frequency words identified via TF-IDF; 2) ENGLISH_STOP_WORDS only; 3) Raw text only. We found that Set 3 achieved the highest performance (AUC: 0.93), outperforming both Set 1 and Set 2. This result aligns with existing findings that transformer models rely on contextual relationships among all tokens via self-attention. Using stop words to remove "function words" (e.g., the, is) can disrupt grammatical and semantic cues crucial to the model's understanding (Hugging Face Forums, 2024). This also explains why Set 1, which applies more aggressive text pruning, performs worse than Set 2, which uses a lighter pruning approach.

Sensitivity Analysis

We evaluated Random Forest performance across three datasets: human-labeled, AI-labeled, and combined (prioritizing human labels and removing duplicates). The model's AUC ranged from 0.81 to 0.85 and AP ranged from 0.44 to 0.57, where AI-labeled dataset showed poorer performance and combined dataset outperformed than others. This variation indicates that Random Forest is sensitive to different dataset quality and labeling consistency.

We conducted hyperparameter tuning on the logistic regression model to examine its sensitivity to different parameter settings. By varying the combination of C, penalty, and solver, we observed that AUC ranged from 0.836 to 0.840 and average precision (AP) from 0.522 to 0.531 and gained the best parameters set: C=3, penalty = l2 with solver = liblinear. Although these variations are small, they indicate that the model's performance is moderately sensitive to hyperparameter choices.

We analyzed DistilBERT's sensitivity to different hyperparameter settings and training fractions to understand how these factors influence generalization. The AUC across three hyperparameter sets ranged from 0.91 to 0.93, suggesting limited sensitivity to hyperparameter changes. However, performance varied more with the training fraction – as it increased from 0.3 to 0.9, the AUC showed greater variation (0.88 to 0.96), with the best performance at 0.9, indicating that the model is more sensitive to data availability than to moderate hyperparameter adjustments.

Identified Tradeoffs

During hyperparameter tuning for the DistilBERT model, we observed that configuration #3 had a slightly higher evaluation loss (~1.25) than configuration #1(~0.62). Evaluation loss represents the average prediction error on the validation set, where lower values typically indicate better probability calibration. However, it does not directly reflect the model's ranking ability. The higher loss in configuration #3 likely results from minor probability miscalibrations, even though its ranking metrics (ROC-AUC and Average Precision) are superior. Because configuration #3 more effectively distinguishes between classes, we accept this trade-off and align with our performance objective.

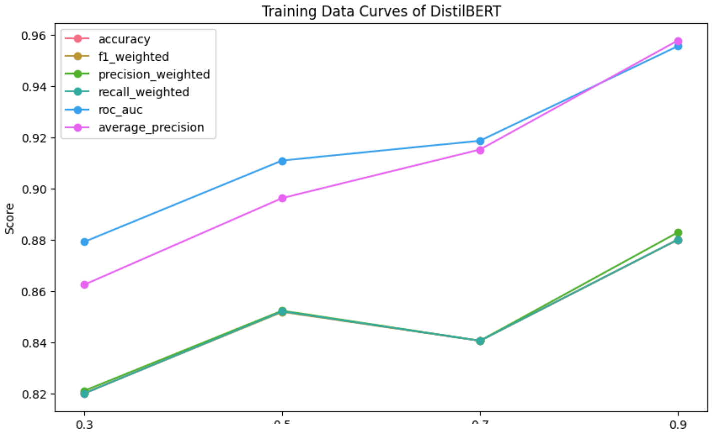


Figure 1: Training Data Curves of DistilBERT

Training Data Curves

We conducted a training data size variation experiment on the DistilBERT model. As the training size increased from 30% to 90%, the model's performance improved steadily. Evaluation loss dropped from 0.44 to 0.31, and ranking metrics AUC climbed from ~0.88 to ~0.96. These consistent gains indicate that adding more data helps the model learn richer patterns and generalize better. Overall, the model has not yet reached a performance plateau, and further increases in high-quality training data are likely to yield additional improvements.

Failure Analysis

Failure Example 1: We extracted the worst prediction from DistilBERT (training fraction = 0.7) using the negative log-likelihood of the true class. A higher value indicates the model was more confident in the wrong answer. The SHAP plot shows that almost all tokens in this data point pushed the model strongly toward the wrong class (high-anxious). Economically related tokens such as “disposable” (0.933) and “anymore” (0.934) had a stronger impact on the model’s decision, suggesting it was misled by these cues. We identified this as a systematic error because the model consistently overweights economically negative phrases (e.g., “disposable income,” “luxury items,” “dead”) and treats them as anxiety indicators. Therefore, we recommend adding rule-based or auxiliary features to detect genuine anxiety cues (e.g., physiological or emotional expressions) while down-weighting general complaints to improve model calibration.

Failure Example 2: Random Forest analysis revealed a concerning high confidence false positive, where the model predicted 62.6% anxiety probability for a low anxiety post. SHAP analysis showed unanimous positive contributions from topic features (topic_9: +0.0174, topic_13: +0.0171) and TF-IDF features (tfidf_4867: +0.0120, tfidf_f3865: +0.0095), indicating the model’s vulnerability when topic distributions align with anxiety themes despite lacking actual triggers.

This represents a fundamental representation limitation. Random Forest’s bag-of-words approach cannot capture emotional tone, sentence structure, or negation patterns, creating a critical gap between lexical matching and semantic understanding that explains the performance difference between DistilBERT. Improvements for this would include incorporating sentiment analysis, syntactic features, and transformer-based extraction for enhanced semantic understanding.

Failure Example 3: Logistic Regression (TF-IDF n-grams, calibrated; threshold ≈ 0.25). The model produced a high-confidence false positive (p = 1.000) on a post labeled not anxiety: “idk how much more of this life I can take… I have no one else to talk to…”. Despite the ground-truth label, the text expresses strong despair and isolation, which semantically overlap with anxiety triggers. Token-level contributions show heavy positive weights for hard, sorry, afford, and contact, revealing the model’s tendency to over-interpret generic hardship or emotional distress as anxiety. In this case, the prediction is psychologically plausible but violates the dataset’s stricter definition of “anxiety trigger,” illustrating annotation ambiguity rather than pure model failure. Future refinement should include clearer labeling criteria, hard-negative examples (emotional but non-anxious posts), and contextual features to help the classifier distinguish genuine anxiety cues from general life complaints.

Part B: Unsupervised Learning

Unsupervised Methods Description

We first used NMF as the baseline model and tuned its hyperparameters to achieve the best performance (coherence = 0.73, mean cosine similarity = 0.204, purity = 0.715) with 15 distinct topics. We then explored BERTopic, expecting improved results given its stronger semantic representation. However, achieving a balance between high metric scores and a larger number of distinct topics proved difficult. After tuning, the best BERTopic model reached coherence = 0.73, mean cosine similarity = 0.95, and purity = 0.009, producing only 7 distinct topics. The high similarity between topics 0 and 1 (0.8–1.0) likely reflects the dataset's heavy focus on anxiety-related content, which limited topic diversity.

Learning Methods

Our unsupervised learning approach employs two complementary topic modeling algorithms to capture different aspects of semantic structure in mental health discussions. Non-negative Matrix Factorization (NMF) serves as our classical baseline algorithm that identifies latent topics through matrix decomposition. A set of high-weight words characterizes each topic, and each document is represented as a mixture of these topics. We use NMF as a baseline because it relies purely on word co-occurrence patterns rather than contextual semantics, allowing us to observe the lexical structure of topics in a transparent, interpretable way and providing a helpful reference point before applying more complex, embedding-based models.

BERTopic represents our modern, transformer-based approach designed to capture semantic similarity among documents rather than just word frequency patterns. This algorithm represents each document using contextual embeddings generated by pre-trained language models and then clusters those embeddings into coherent groups that share similar meanings (Nguyen & Ramdas, 2025). We employ BERTopic to detect deeper conceptual relationships between texts, providing a complementary and more context-aware perspective for topic generation that complements our baseline NMF approach.

Feature Representations

Our unsupervised learning models employ different feature representation strategies reflecting their algorithmic approaches. NMF decomposes a document-term frequency matrix into two non-negative matrices representing topics and topic weights per document, while BERTopic uses transformer-based language models (like BERT, Sentence-BERT, etc.) to create semantic embeddings, which are numerical vectors that capture meaning.

Hyperparameter Tuning

Our hyperparameter tuning approach differed between NMF and BERTopic due to their distinct algorithmic foundations. For NMF, we conducted iterative hyperparameter exploration starting with initial parameters (min_df=20, max_df=0.5, max_features=20,000) and refining to final values (min_df=12, max_df=0.40, max_features=10,000) based on coherence scores and topic quality. The final configuration used 15 topics (K=15) with KL divergence loss, multiplicative update solver, L1/L2 regularization (alpha_H=0.1, l1_ratio=0.7), and 1,500 maximum iterations. This tuning process improved topic coherence to NPMI=0.725 while maintaining interpretable topic separation (mean similarity=0.204).

For BERTopic, we tested five parameter configurations to improve performance, including slight improvements over the baseline and others emphasizing balanced performance, coherence, outlier handling, and diversity. The slightly enhanced baseline achieved the best overall performance, though with less distinct topic separation compared to our NMF results.

Unsupervised Evaluation

We evaluated our unsupervised models using four metrics: coherence score (word consistency within a topic), mean cosine similarity (overlap between topics), purity score (alignment with known labels), and number of topics (breadth of themes). Together, these metrics help identify a model that balances topic interpretability, diversity, clarity, and the overall breadth versus depth of insights.

Method	Coherence Score	Mean Cosine Similarity	Purity Score	Number of Topics
NMF	0.73	0.204	0.715	15
BERTopic	0.73	0.95	0.009	7

Figure 2: Topic Similarity Analysis - Visual representation of topic coherence and similarity patterns across the 15 NMF topics, demonstrating topic interpretability and the 0.204 mean cosine similarity score.

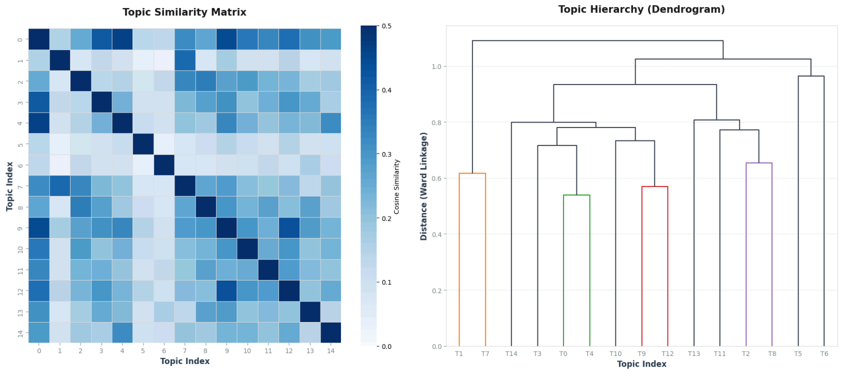
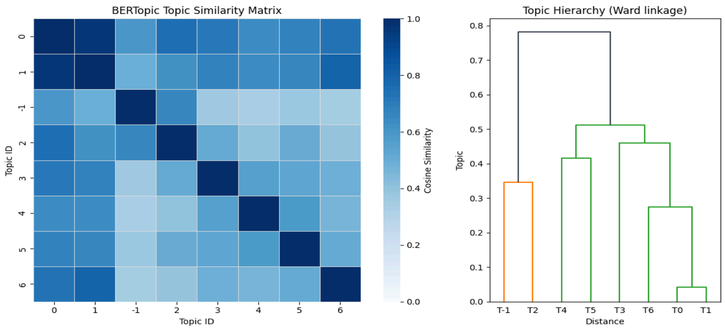


Figure 2

Figure 3: BERTopic Topic Similarity Matrix - The Topic Matrix and Topic Hierarchy visualize coherence and similarity patterns across the seven BERTopic topics, indicating



slightly higher similarity among them.

Sensitivity Analysis

We stress-tested the BERTopic pipeline along two axes (input features and model hyperparameters) and observed that performance is quite sensitive to both. Our feature sensitivity analysis examined stop-word strategies by comparing four preprocessing variants: (a) no extra filtering, (b) a light custom list, (c) the NMF-derived stop words, and (d) the NMF list augmented with TF-IDF top terms. Coherence ranged from ≈0.49 to ≈0.73, outlier fraction from ≈5% to ≈56%, and topic diversity from ≈0.88 to ≈0.97. The NMF-derived stop words yielded the best balance (coherence ≈0.73, outliers ≈5%, diversity ≈0.97), indicating that the model is highly sensitive to feature selection-small changes in token filtering dramatically alter cluster quality.

Our hyperparameter sensitivity analysis examined UMAP + HDBSCAN + vectorizer parameters using the best preprocessing as a baseline. We varied UMAP n_neighbors, HDBSCAN min_cluster_size/min_samples, and vectorizer min_df/max_df to test five parameter sets aimed at improving coherence, diversity, or outlier control. Coherence spanned ≈0.45–0.73, outlier fraction ranged from near zero to ≈66%, and topic counts shrank from 52 to as few as 2. The "slightly improved baseline" achieved the strongest overall metrics (coherence ≈0.73, outliers ≈5%, diversity ≈0.97) but collapsed to seven topics, underscoring that the results are sensitive to these settings and constrained by the dataset's heavy concentration of anxiety content.

Discussion

Part A: Supervised Learning

For DistilBERT, we learned that the AI-labeled dataset outperformed the human-labeled one, likely due to its more consistent labeling. The model also performed best using raw, unprocessed text, with the optimal hyperparameters found at 2 epochs, learning rate = 5e-5, and weight decay = 0.05. Training data curve analysis showed that adding more data continues to improve model performance. For the Random Forest model, the combined dataset (prioritizing human labels) achieved the best results (AUC = 0.849, n = 1006). For Logistic Regression, the best configuration was C = 3.0, penalty = L2, and solver = liblinear, achieving AUC = 0.9013 and AP = 0.5714. The Ablation Study confirmed that combining TF-IDF and NMF features outperformed using either feature set alone, highlighting their complementary value in model training.

We were surprised that DistilBERT significantly outperformed both Random Forest and Logistic Regression, even when trained only on raw text. We initially expected that engineered features, such as TF-IDF (capturing anxiety-related keywords), NMF topic features, and emotion scores from the NRC Lexicon, would enhance classical models to perform comparably. However, DistilBERT achieved much higher scores (AUC = 0.956) than Random Forest (AUC = 0.847) and Logistic Regression (AUC = 0.9013). This result reflects DistilBERT's advantage as a context-aware transformer model that captures deep semantic relationships and word meanings within context, rather than relying solely on surface-level word frequencies (Sanh et al., 2019).

One key challenge we faced was that our training-size comparison for DistilBERT kept reusing the already fine-tuned model from earlier runs, leading to inconsistent metrics. Each rerun continued training from leftover weights instead of starting from a clean checkpoint, causing performance drift across experiments. We resolved this by reinitializing the model before each training-size test, either by reloading the original base checkpoint or deep-copying an unfine-tuned model. This fix ensured reproducible and fair comparisons across different dataset sizes.

With more time and resources, we could extend this work by improving data quality, model diversity, and computational depth. For DistilBERT, we would address subreddit imbalance and annotation noise by using stratified sampling, weighted losses, and hybrid human–AI label validation, while also expanding the hyperparameter search with larger compute resources (e.g., Great Lakes). For Random Forest and Logistic Regression, we could perform multi-annotator validation, refine AI label filtering, and engineer new temporal or engagement-based features to improve generalizability beyond Reddit. Together, these enhancements would strengthen model robustness, reduce bias, and enable more confident real-world deployment.

Part B: Unsupervised Learning

Our unsupervised learning experiments revealed several key insights about topic modeling approaches for mental health content analysis NMF achieved fine-grained topic discovery with 15 distinct topics achieving NPMI coherence of 0.725, which successfully captured nuanced thematic patterns from clinical anxiety symptoms to interpersonal relationships and technical productivity discussions. BERTopic achieved comparable coherence (0.730) with better topic diversity (0.97) but less granularity (7 topics vs. 15), suggesting complementary strengths between semantic embeddings and bag-of-words approaches. The topic diversity across subreddits was particularly revealing: mental health communities (r/Anxiety, r/HealthAnxiety) produced clinically-focused topics while neutral communities (r/economy, r/OpenAI) generated context-specific themes, validating our multi-community approach for comprehensive anxiety detection.

The most surprising finding was the comparable performance between NMF and BERTopic despite their different approaches. We expected BERTopic's modern transformer-based architecture to significantly outperform NMF's bag-of-words approach, but both achieved similar coherence scores (0.725 vs. 0.730) with complementary strengths: NMF providing more detailed topic discovery and BERTopic offering better topic diversity. The assignment purity of 71.5% was also higher than expected, indicating that most posts could be confidently assigned to single topics rather than requiring multi-topic distributions. The topic diversity across subreddits was particularly revealing: mental health communities (r/Anxiety, r/HealthAnxiety) produced clinically-focused topics while neutral communities (r/economy, r/OpenAI) generated context-specific themes, validating our multi-community approach for comprehensive anxiety detection.

The primary challenge was balancing competing objectives across both approaches. For NMF, we struggled with topic granularity (K=10 too broad, K=20 too fragmented) and vocabulary noise from Reddit artifacts, ultimately resolving these through systematic hyperparameter tuning and custom preprocessing. For BERTopic, no single configuration achieved all goals: the best-performing setup (baseline_improvement) achieved strong coherence (0.730) and diversity (0.97) but collapsed to only 7 topics. In contrast, other configurations suffered from high outlier rates (56-66% unassigned documents). We addressed these challenges by selecting optimal trade-offs: NMF's K=15 for fine-grained topics, and BERTopic's baseline_improvement for better coherence at the cost of topic count. Both approaches required extensive preprocessing to handle Reddit-specific noise and computational complexity from large vocabulary sizes.

With more time and resources, several extensions could enhance both NMF and BERTopic approaches. For NMF hierarchical topic modeling could capture multi-level relationships (e.g., “Mental Health” → “Clinical Symptoms" → “Panic Attacks”), while dynamic topic modeling could track anxiety theme evolution over time. For BERTopic, addressing the high outlier rates through improved HDBCAN parameter tuning or custom embedding models could increase topic coverage, while ensemble methods combining multiple BERTopic configurations could achieve both high coherence and sufficient topic count.

Both approaches could also benefit from semi-supervised learning methods that incorporate expert annotations to identify clinically-relevant themes. Interactive visualization tools with real-time filtering capabilities would also enhance the practical utility of discovered topics for content moderation and mental health research applications.

Ethical Considerations

Our research raises several ethical considerations related to bias, representation, and potential misuse of mental health content analysis. In supervised learning, human-annotated anxiety ratings may introduce labeling bias influenced by annotators' personal experiences, mental fatigue, and inconsistent judgment. While this limitation was not fully addressed due to time constraints, potential mitigation strategies include employing multiple annotators with aligned label standards (Artstein & Poesio, 2008) and implementing short labeling sessions to prevent cognitive fatigue (Snow et al., 2008).

For unsupervised learning, our dataset exhibits systematic bias through varying anxiety-related content density across subreddits. Mental health communities (r/Anxiety, r/HealthAnxiety, r/mentalhealth) may dominate topic discovery compared to neutral communities (r/economy, r/OpenAI), potentially leading to over-representation of anxiety themes and under-representation of positive content patterns. Additionally, our manual stopword filtering approach may inadvertently remove contextually important terms that could reveal nuanced mental health discussions, potentially limiting topic discovery sensitivity. Our current approach applies topic modeling collectively without addressing subreddit imbalance. Proposed mitigation strategies include stratified topic modeling within each subreddit before ensembling results, implementing community-aware weighting where each subreddit contributes proportionally to its user base, validating discovered topics across subreddit boundaries to ensure generalizability, and developing domain-specific stopword lists that preserve mental health terminology while filtering noise.

Statement of Work

- **Maria Mckay:** Initial data quality check, NMF topic modeling, Random Forest classifier implementation.
- **Shen Shu:** Repo structure, Reddit API pipeline, AI labeling, Logistic Regression model.
- **Shuting He:** Project Planning, BERTopic, Human-labeled dataset, DistilBERT.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555-596.

Hugging Face Forums. (2024, December 31). Do I need to perform 'Stop word removal' before feeding into Hugging Face pipeline or AutoModels? Retrieved from https://discuss.huggingface.co/t/do-i-need-to-perform-stop-word-removal-before-feeding-into-hugging-face-pipeline-or-automodels/133661

KS, R. (2024). Analyzing online conversations on Reddit: A study of stress and anxiety through topic modeling and sentiment analysis. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11464268/

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436–465.

National Institute of Mental Health. (n.d.). Any anxiety disorder. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder

Nguyen, M., & Ramdas, T. (2025). A comparative overview of BERTopic, LDA, and beyond. Chamomile AI. Retrieved from https://chamomile.ai/topic-modeling-overview/

OpenAI. (2023). GPT-3.5-turbo [Large language model]. https://platform.openai.com/docs/models/gpt-3-5-turbo

Oryngozha, T., Shmatova, I., & Kamalov, F. (2023). Detection and analysis of stress-related posts in Reddit's academic communities. arXiv preprint arXiv:2312.01050. https://arxiv.org/abs/2312.01050

Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., Colditz, J. B., Radovic, A., & Miller, E. (2017). Social media use and perceived social isolation among young adults in the U.S. American Journal of Preventive Medicine, 53(1), 1-8. https://doi.org/10.1016/j.amepre.2017.01.010

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* [Preprint]. arXiv. https://arxiv.org/abs/1910.01108

Shen, J., & Rudzicz, F. (2017). Detecting anxiety through Reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality (pp. 58–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3107

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008

Conference on Empirical Methods in Natural Language Processing (pp. 254-263). Association for Computational Linguistics.