Replicating "Detection of Depression-Related Posts in Reddit Social Media Forum"

by Tadesse et al.

Authors:

Natasza Siwinska and Jamie Bergen
Bar-Ilan University

Date: January 27, 2025

Abstract

We have replicated the paper *Detection of Depression-Related Posts in Reddit Social Media Forum* by Tadesse et. al. The paper outlines how researchers used natural language processing (NLP) techniques and machine learning approaches to examine Reddit users' posts and identify factors that could reveal depression attitudes. They analyzed the correlation significance, hidden topics, and word frequency extracted from the text. They then used five text classifying techniques and compared their performance based on three single feature sets and their multiple feature combinations to detect depression. We have attempted to recreate this entire process, so this paper will be divided into each of the steps and how our results for each step were similar or different.

Data Scraping/Pre-processing

Their Scraping

The researchers utilized a dataset provided by Inna Pirina et al., which contained 1,293 depression-indicative posts. These posts were not publicly available. According to the Pirina paper, the dataset was curated using specific phrases indicative of depression, such as "I was just diagnosed with" in the r/depression subreddit. Additional posts from the same authors, within a one-month timeframe, were also included. To construct their control group, the researchers collected an equal number of posts from a subreddit focused on breast cancer. However, the description of the control group in Tadesse et al.'s paper indicates it contained 548 "standard posts," with no explicit clarification regarding their origin or selection criteria. The lack of explicit criteria for selecting control group posts introduces ambiguity, which may impact the reproducibility and generalizability of their results.

Prior to feature selection and classification, the original study tokenized posts and performed standard NLP preprocessing steps, including URL removal, punctuation stripping, stopword elimination, and stemming to reduce words to their root forms. This methodology laid the foundation for their subsequent analysis and feature engineering steps.

Our Scraping

Due to insufficient details about the control group used in the original study, we implemented our best understanding of their data collection strategy. Posts were scraped from Reddit using the Python Reddit API Wrapper (PRAW). For the control group, we selected posts from popular reddit forums, and for additional control, a subreddit dedicated to breast cancer discussions. This approach was chosen to reflect a similar balance of non-depression-related content. Each category comprised approximately 1,000 posts.

The scraped data underwent pre-processing using the Natural Language Toolkit (NLTK), which included tokenization, stemming, stopword removal, and general text cleaning. All subsequent pre-processing steps mirrored those outlined in the original study to ensure consistency and enable direct comparison of results.

Feature Extraction

Their Features

Summary of Feature Extraction Methods from Original Paper

Feature Type	Methods	Selected Features
N-grams	Unigram	3000
N-grams	Bigram	2736
Linguistic Dimensions	LIWC	68
Topic Modeling	LDA	70

The authors extracted features from pre-processed Reddit posts using three primary methods. They measured word frequency with unigrams and bigrams, applying TF-IDF to identify the most informative terms. Bigrams were further refined using Pointwise Mutual Information (PMI) to filter for co-occurring terms indicative of depressive themes.

For linguistic features, the LIWC2015 dictionary was used to analyze 95 psycholinguistic categories, focusing on attributes like affective processes, cognitive states, and time orientation. Depression-related posts exhibited significantly higher word frequencies for terms associated with self-preoccupation, negative emotions, and feelings of hopelessness.

Finally, the newpage modeling approach identified latent themes prevalent in depression-related posts. The analysis revealed topics such as "Depression," "Broke," and "Tired," which underscored recurring themes of financial hardship, emotional fatigue, and personal suffering. This demonstrated the ability of LDA to extract meaningful insights from textual data in the context of mental health.

Our Features

Summary of Feature Extraction Methods from Our Replication

Feature Type	Methods	Selected Features
N-grams	Unigram	5000
N-grams	Bigram	5000
Linguistic Dimensions	Empath	47
Topic Modeling	LDA	20

Using similar extraction methods for n-grams (TF-IDF), we extracted the top 5,000 unigrams and bigrams instead of the original study's 3,000, as our dataset contained more posts. For linguistic analysis, we substituted the LIWC2015 tool with EMPATH, a publicly available alternative that provides comparable psycholinguistic features. EMPATH allowed us to categorize text based on affective, biological, social, cognitive, and other dimensions as outlined in the original study. We identified significant features using Pearson correlation coefficients, as described in the original paper.

For topic modeling, we employed Latent Dirichlet Allocation (LDA) with the same hyperparameters specified in the original study. Specifically, we generated 70 topics to represent latent themes. However, instead of using the Mallet toolkit as in the original paper, we opted for the gensim library, which offers a streamlined implementation of LDA. This reduced the number of hyperparameters available for tuning but yielded comparable results.

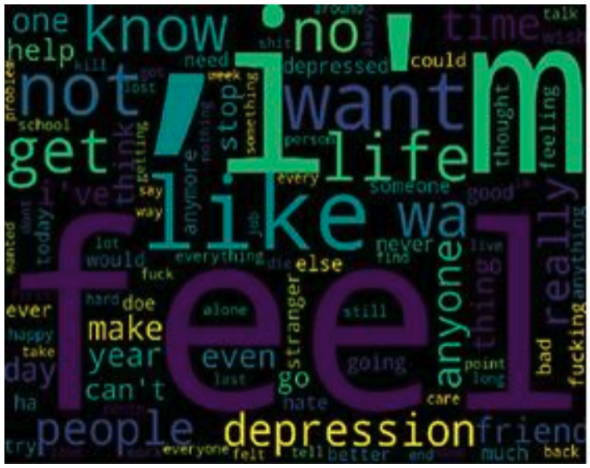
Word Clouds and N-Gram features

Their Word Clouds

Wordclouds generated from unigrams and bigrams from the original paper.

(a) Standard Post Unigrams

(b) Standard Post Bigrams



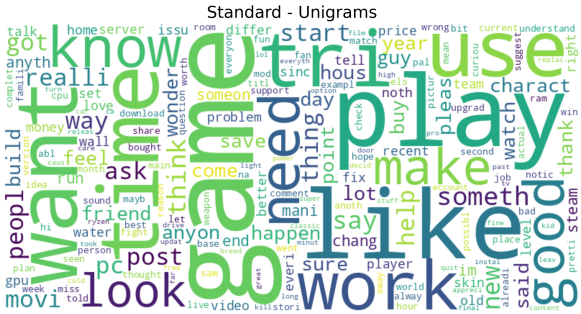
(c) Depression Post Unigrams

(d) Depression Post Bigrams

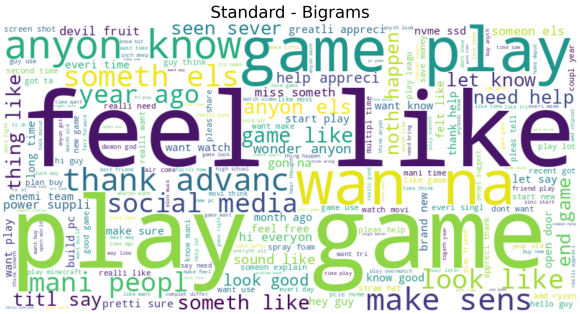
The original paper's analysis emphasized "signs of meaninglessness" (e.g., "pointless," "empty," "senseless") and identified frequent use of phrases reflecting social and emotional challenges. While many of our findings align with these observations, differences in pre-processing steps, such as the treatment of self-referential terms, contributed to slight variations in identified patterns. These distinctions underscore the influence of methodological choices on textual analysis outcomes and highlight opportunities for further refinement of feature extraction techniques.

Our Word Clouds

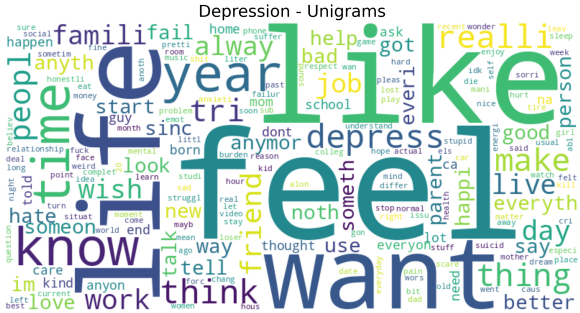
Wordclouds generated from unigrams and bigrams found in each Reddit post set, based on our replication.



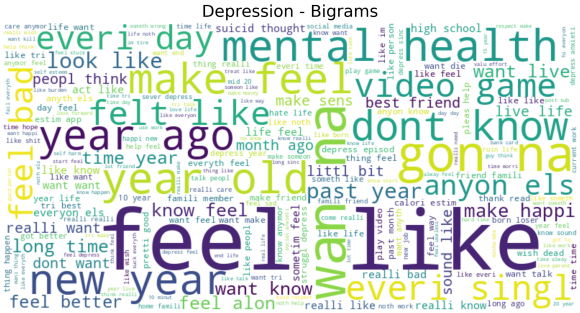
(a) Standard Post Unigrams



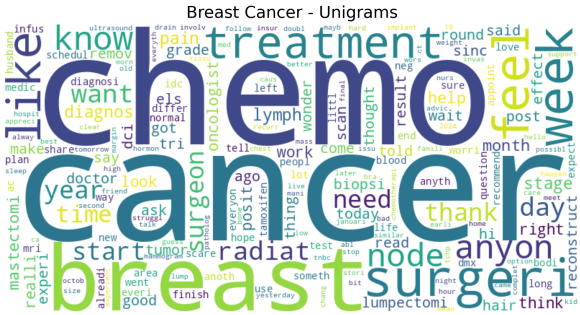
(b) Standard Post Bigrams



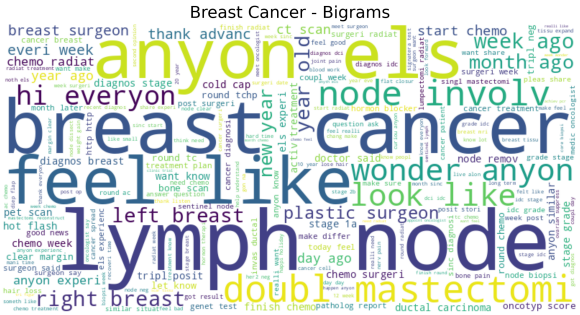
(c) Depression Post Unigrams



(d) Depression Post Bigrams



(e) Breast Cancer Post Unigrams



(f) Breast Cancer Post Bigrams

The top words for both unigrams and bigrams that we extracted were converted to word clouds, separated by depression-labeled posts, standard posts, and breast cancer posts.

Our analysis identified strong similarities with the original study, though some differences arose due to preprocessing. For instance, our tokenization removed self-referential terms (e.g., "I'm"), while the original study retained them. This distinction may account for variations in specific expressions captured. Nevertheless, predictors of depression in our analysis, such as "greater discrepancies" ("want," "wish," "wanna"), "suicidal thoughts" ("suicide," "want die," "end"), and "anger and hostility" ("fuck," "hate"), aligned with the original findings.

Our replication also noted additional terms not emphasized in the original study, such as words related to values ("respect," "value effort") and specific categories in non-depression posts. For example, standard posts included past-time-related terms ("game play," "game character"), and breast

cancer posts overlapped with depression posts in terms of interpersonal processes ("feel like," "wonder," "think"). To quantify these differences, we analyzed the frequency distributions of key terms. Words associated with emotional distress, such as 'worthless' and 'suicidal,' showed higher prominence in depression-labeled posts across both studies, affirming their predictive value. However, variations in tokenization approaches, such as the inclusion of self-referential terms, may explain differences in patterns observed.

Predictive power of LIWC features

Their Analysis

LIWC Category	Example Word	P-value
Linguistic Dimensions		
Personal pronoun	I, them, her	0.15*
1st person singular	I, me, mine	0.17***
Negations	no, not, never	0.16**
Psychological Processes		
Social Processes	buddy, mate, talk	0.17**
Affective Processes	happy, cry, hate	0.19***
Cognitive Processes	think, know, always	0.14*
Personal Concerns		
Work	job, majors, xerox	0.16**
Money	audit, cash, owe	0.14**
Death	win, success, better	0.11*

*P¡0.05, **P¡0.01, ***P¡0.001. All the correlation coefficients meet the P¡0.05 threshold.

The analysis of 68 psycholinguistic features revealed significant correlations between textual data and depressive or non-depressive posts. The strongest correlations were observed in psychological processes (0.19), linguistic dimensions (0.17), and personal concerns (0.16). Within psychological processes, affective processes (e.g., words like "kill," "worthless," "cry") had the highest correlation (0.19), particularly with negative emotions (0.19) and, to a lesser extent, positive emotions (0.09). Social processes (e.g., "brother," "friend," "mom") also showed a correlation of 0.17, aligning with linguistic dimensions. In terms of personal concerns, topics such as work (0.16), money (0.14), and death (0.11) were prominent. Depressed users displayed a higher use of self-referential words (e.g., "I," "me," "mine") (0.17), reflecting self-focus, and their language was characterized by negative emotions, sadness, anxiety, and a greater emphasis on the present and future.

Predictive power of EMPATH features

Our Analysis

Category	Example Word	Corr.	P-value
Personal Concerns			
Death	"suicide"	0.16	2.06×10^{-18}
Sports	"run"	-0.157	1.14×10^{-17}
Psychological Processes: Affective			
Anger	"hate"	0.199	7.33×10^{-28}
Fear	"scared"	0.185	3.59×10^{-24}
Sadness	"cry"	0.164	2.38×10^{-19}
Shame	"failure"	0.141	1.18×10^{-14}
Psychological Processes: Biological			
Health	"health"	0.353	1.46×10^{-87}
Body	"calorie estimate"	0.23	1.26×10^{-36}
Injury	"hurt"	0.19	2.42×10^{-25}
Pain	"suffer"	0.152	9.46×10^{-17}

Top EMPATH Features Correlated with Depression, Based on Our Replication.

Note: Correlation values and P-values indicate the strength of association with depression based on our replication.

Since we had fewer features found than the original paper (48 compared to 95), we did not minimize this amount in the same way that the original paper shrank theirs from 95 to 68. From the correlations found in every category, we can compare table outputs to showcase the subcategory correlations for linguistic categories that are more indicative of depression. Since the pronoun 'I' was removed in our pre-processing, we have no words or correlations for linguistic dimensions about personal pronouns, which the original paper did have. Also, the categorizations came out slightly differently due to the different models and differences in our datasets.

One difference between our results and those in the original paper is that we suspect the "P-values" column in the original paper was intended to represent correlations. This is because their reported values don't align with the significance levels they assigned. We have created a list of the most influential categories (selecting a few from each of the top categories) along with their correlations and P-values.

We can see that our correlations are much higher in general than the ones found in the original paper. Perhaps due to the increased number of posts, different model usage (EMPATH vs LIWC), or alterations in pre-assigned categories.

Predictive power of LDA

Their LDA Results

Named Topics and Most Representative Words in the Original Paper.

Topics	Representative words
Job	boss, boring, broke, company, fired, handjob, jobless, pay, money, quit, stress, left, time, unemployed, unhappy, want, work, year
Depression	always, better, die, depressed, feeling, hate, isolate, long, mind, meaning, guilt, myself, negative, over month, pain, suffer, something, thoughts
Tired	abusive, bullied, burden, can't, hurt, ill, live, loneliness, mentally, myself, neg thought, sleep, started, time, think, wrong one, world, wanted
Friends	best, date, dude, chill, care, encourage, happy, help, insecure, relationships, spend, support, no friends, roommate, request, unfollow, worthless
Broke	find, help, heartbroken, emotional, inside lost, love, marriage, often, relationship, rejected, spouse, together, sex, problems, ugly
F***	break, death, drunk, damn, die, done, exercise, long, lonely, pain, kill me, removed, phone, stupid, shit, school, sleep, suck, text, treatment

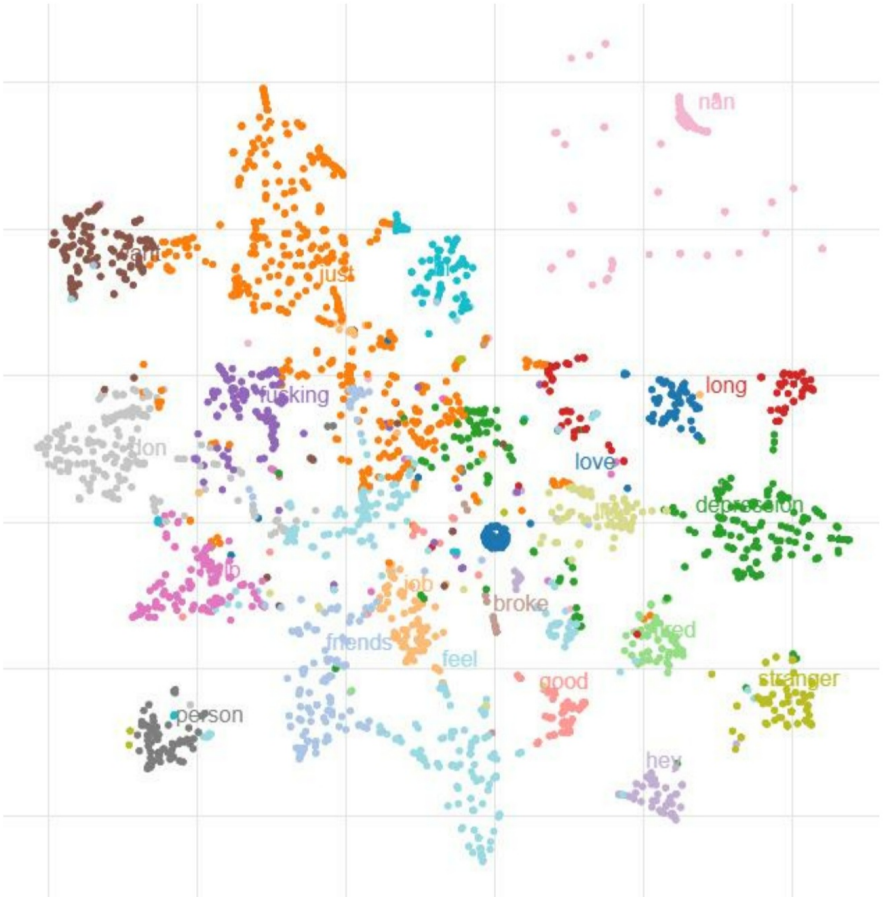
Top LDA Topics found with posts from r/Depression, Based on the original paper

The study originally identified 70 topics, which achieved a 70% validation accuracy during topic modeling—a measure indicating that the model successfully captured coherent and meaningful groupings of words within topics. However, rather than analyzing all 70 topics, the researchers prioritized the 20 most prominent topics for further exploration. These prominent topics were likely chosen because they had the highest interpretability and relevance, thus providing a manageable yet representative subset for deeper analysis.

To better understand the relationships between these topics, the researchers employed dimensionality reduction techniques using t-distributed Stochastic Neighbor Embedding (t-SNE). This method is particularly effective for visualizing high-dimensional data by projecting it onto a lower-dimensional space—in this case, clustering related topics into a two-dimensional plot. The hyperparameters used in this process included a perplexity of 50 and 500 iterations, which influence the balance between local and global structure in the visualization.

By clustering the topics, t-SNE provides insights into which topics are conceptually similar and how they are distributed across the data, facilitating an intuitive understanding of the model's output.

Here is the t-SNE cloud which they generated based on these 20 most prominent topics.



t-SNE with two-dimensions and 20 topics from original paper for depression-related posts.

As mentioned under our replication section for the t-SNE, we are unsure of the original paper's naming process behind the clusters found. The spatial proximity of clusters in the t-SNE plot provides insights into potential semantic relationships. For example, clusters such as "friends" and "feel" are positioned near one another, suggesting overlapping or related themes. In contrast, clusters like "person" and "broke," which are more distant, likely represent distinct topics with minimal overlap. However, the original paper underlines the aparent distinctions between word categories used in depression-related posts.

Our LDA Results

Named Topics and Most Representative Words Found in Our Replication.

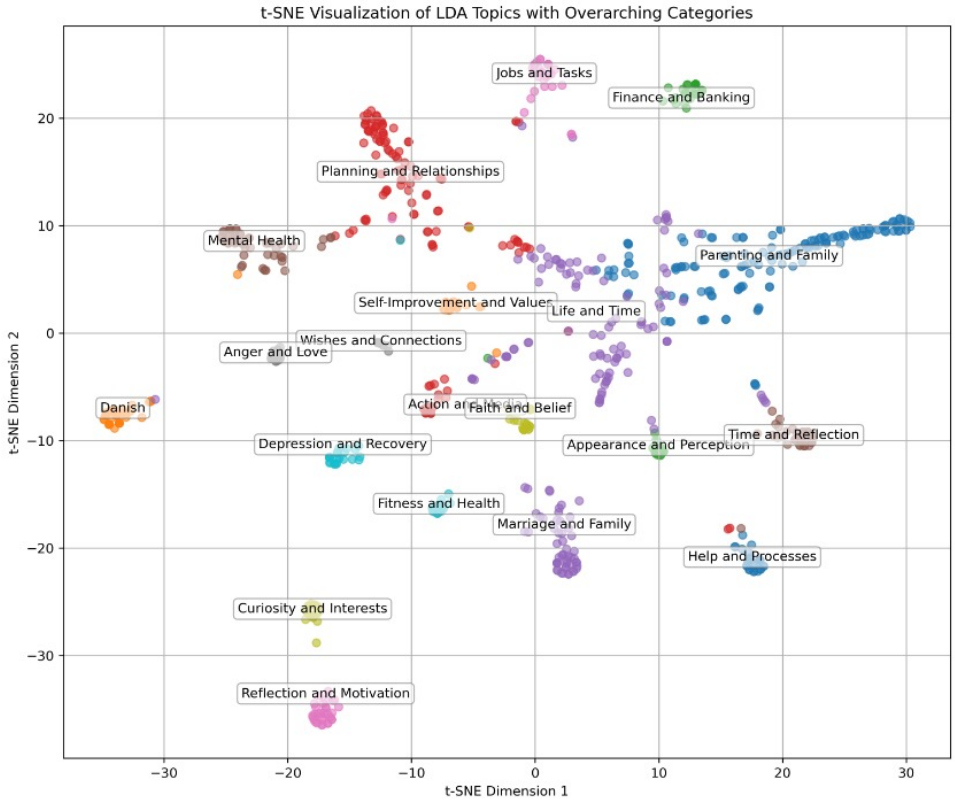
Topics	Representative words
Family and Parenting	time, parent, people, year, show, face, make, need, like, kid
Values and Respect	kind, make, value, respect, element, time, effort, real
Banking and Finance	bank, ask, told, back, card, nearby, said, look, question, go
Future Planning	want, family, plan, would, also, get, people, really, since, go
Marriage and Relationships	family, marry, even, like, would, time, ready, tell, said, told
Depression and Emotions	feel, want, get, go, year, try, like, I'm, time, depress

Top LDA Topics found in r/Depression, Based on Our Replication.

In our replication of the original study, we adopted the same parameters outlined as optimal for the Latent Dirichlet Allocation (LDA) analysis. Specifically, we also focused on 20 topics, selected based on interpretability and relevance, while setting the perplexity to 50 and the number of iterations to 500 for dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE). These parameters were reported as producing the most coherent and meaningful clustering of topics in the original study.

The original paper was unclear on how topics were named based on their representative words, rather than using numerical labels like Topic 1 or Topic 2. These names, such as "Family and Parenting" or "Depression and Emotions," reflect the main themes identified in the analysis. The table shows the most representative words from the LDA found in depression-related posts, highlighting these central themes. Since the original paper didn't specify how the topics were named, we used ChatGPT to assign relevant names to the representative words.

As a comparison, the clusters that we have outputted appear much more compact and separated than the original paper. We have topics which are closer together in the visualization such as 'Jobs and Tasks' and 'Finance and Banking', which seem to have semantic relation, but topics that are further apart but then have some convergence points towards the middle where there is more variance, such as 'Planning and Relationships' and 'Parenting and Family', can also be seen. This is not as evidently seen within the original paper.



t-SNE with two-dimensions and 20 topics from our replication for depression-related posts

Model Training

Their Training

Model Performance Metrics from Original Paper (Rounded to Integers)

Feature Set	LR (%)	SVM (%)	RF (%)	AdaBoost (%)	MLP (%)
LIWC	69/80/95/69	74/74/75/72	78/84/77/74	67/81/68/99	70/72/74/71
LDA	77/83/82/84	72/80/75/88	75/80/84/82	66/74/61/95	75/74/75/72
Unigram	68/80/93/79	70/81/70/89	72/81/73/92	72/81/73/92	70/81/71/95
Bigram	80/79/81/78	80/80/79/80	74/73/75/63	71/68/68/75	79/78/80/76
LIWC+LDA+Unigram	80/84/88/81	79/81/81/81	83/80/84/82	73/72/87/62	78/81/84/79
LIWC+LDA+Bigram	89/89/89/92	90/91/89/93	85/80/85/83	79/81/72/93	91/93/90/92

The columns for each model are in the format such that the percentages represent Accuracy/F1/Precision/Recall

The original study also used five classifiers to evaluate single and combined feature sets. Their results demonstrated that the MLP classifier with the LIWC + LDA + bigram feature set achieved the best performance, with 91% accuracy and an F1 score of 0.93. Other classifiers, such as SVM and RF, showed strong performance on single feature sets, particularly bigrams.

Preprocessing steps in the original study included tokenization, stopword removal, and feature normalization. Hyperparameter tuning via cross-validation was used to optimize classifier performance.

Our Training

Model Performance Metrics from Our Replication (Part 1: LR, SVM)

Feature Set	LR (%)	SVM (%)
Empath	66.84/65.75/67.44/66.84	68.19/66.98/70.33/68.19
LDA	70.73/70.74/70.76/70.73	70.73/70.80/71.12/70.73
Unigram	86.63/86.63/86.95/86.63	92.55/92.58/92.83/92.55
Bigram	70.39/70.66/71.30/70.39	68.19/68.38/72.08/68.19
Empath+LDA+Uni	86.80/86.84/87.12/86.80	91.88/91.89/92.18/91.88
Empath+LDA+Bi	90.69/90.77/90.90/90.69	99.66/99.66/99.67/99.66

Model Performance Metrics from Our Replication (Part 2: RF, AdaBoost, MLP)

Feature Set	RF (%)	AdaBoost (%)	MLP (%)
Empath	68.70/68.25/68.29/68.70	66.33/65.64/66.53/66.33	69.20/68.90/70.01/69.20
LDA	75.63/75.59/75.67/75.63	70.56/69.82/70.13/70.56	71.24/71.31/71.42/71.24
Unigram	97.80/97.81/97.81/97.80	100.00/100.00/100.00/100.00	76.82/76.87/77.12/76.82
Bigram	70.39/70.30/74.26/70.39	51.27/47.12/65.09/51.27	67.51/67.78/68.32/67.51
Empath+LDA+Uni	87.14/86.86/87.44/87.14	100.00/100.00/100.00/100.00	79.70/79.86/80.18/79.70
Empath+LDA+Bi	81.39/81.03/81.39/81.39	100.00/100.00/100.00/100.00	74.62/75.07/76.82/74.62

The columns for each model are in the format such that the percentages represent Accuracy/F1/Precision/Recall

We replicated the methodology outlined in the original paper using five classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost), and Multilayer Perceptron (MLP). Each classifier was tested on single and combined feature sets, with metrics for accuracy, precision, recall, and F1 score reported (rounded to two decimal places, shown as Accuracy/Precision/Recall/F1).

The best-performing model in our replication was the SVM classifier trained on the combined Empath, LDA, and bigram feature set, achieving 99.66% across all metrics. AdaBoost also performed perfectly (100.00%) on both combined feature sets. Differences in performance across classifiers and features may be attributed to dataset size, preprocessing variations, and hyperparameter tuning.

The superior performance of certain models, such as SVM and AdaBoost, with combined feature sets, underscores the value of integrating multiple linguistic and thematic features for improved classification. The perfect accuracy of AdaBoost warrants further investigation to rule out potential overfitting or data leakage

Concluding Thoughts

Our replication of Tadesse et al.'s study on detecting depression-related posts in Reddit demonstrates both the robustness of their original methodology and opportunities for methodological refinement. While we achieved comparable or superior classification performance across most metrics, particularly with our SVM classifier reaching 99.66% accuracy on combined feature sets, several important differences emerged in our implementation and results.

The substitution of EMPATH for LIWC2015 and our expanded dataset yielded different but complementary insights into the linguistic patterns associated with depression-related content. Our topic modeling analysis revealed new themes around family dynamics and future planning that were not prominent in the original study, suggesting temporal shifts in how depression manifests in social media discussions (the original study was conducted in 2019). These differences may highlight the dynamic nature of online mental health discourse and underscore the importance of periodic methodology updates to capture evolving patterns of expression.

This replication effort also reveals certain limitations in the original methodology, particularly regarding the control group selection criteria and feature extraction processes. Future research would benefit from more standardized approaches to control group selection and more transparent documentation of preprocessing steps. Nevertheless, the strong performance of our models validates the fundamental approach of using natural language processing and machine learning techniques for detecting depression-related content in social media, while suggesting potential avenues for methodological improvement and standardization.

Future studies should prioritize transparency in pre-processing and tool selection, as well as explore the integration of emerging NLP techniques, such as transformer-based language models, to capture the evolving linguistic patterns in mental health discourse.