# Depression Detection from Social Media Text: A Machine Learning Approach with Distribution Drift Handling

**Yiming Cheng**[1] **Yi Wu**[1]
[1]Department of Computer Science, University of Chicago, MPCS Predoc
`eaminchan@uchicago.edu`, `yiwu@uchicago.edu`

## Abstract

We propose a machine learning framework for depression detection from social media text, addressing distribution drift challenges. Our approach employs multiple vectorization techniques (TF-IDF, N-grams, Word2Vec, GloVe), Principal Component Analysis (PCA) for dimensionality reduction, and evaluates various classifiers including logistic regression, Support Vector Machines (SVM), and neural networks trained with stochastic gradient descent. The methodology includes comprehensive data preprocessing, SMOTE for class balancing, and evaluation using F1-score, precision, recall, accuracy, and AUC-ROC metrics. We construct a dataset of 10,325 depressed and 22,245 normal samples from Weibo, with detailed exploratory data analysis revealing key differences in user behavior and engagement patterns between depressed and normal users.

## 1 Introduction and Motivation

Depression is a prevalent mental health disorder affecting millions of people worldwide, with significant implications for individual well-being and public health World Health Organization [2023]. Early detection and intervention are crucial for effective treatment, yet many individuals remain undiagnosed or untreated. Social media platforms have emerged as rich sources of linguistic data that can reveal psychological states and behavioral patterns De Choudhury et al. [2013]. The widespread use of platforms like Weibo provides unprecedented opportunities to analyze user-generated content for mental health indicators.

Previous research has demonstrated the feasibility of detecting depression from social media text using machine learning approaches Coppersmith et al. [2014], Rezapour et al. [2019]. However, several challenges remain: (1) **class imbalance**, where normal users typically outnumber depressed users in real-world datasets; (2) **distribution drift**, where the statistical properties of training and test data may differ significantly; and (3) **feature representation**, as the choice of text vectorization method significantly impacts classification performance.

This work addresses these challenges by proposing a comprehensive machine learning framework that systematically explores different text representation methods, handles class imbalance through sampling techniques, and evaluates multiple classification algorithms. Our contributions include: (1) a detailed exploratory data analysis of social media user characteristics and engagement patterns; (2) a systematic comparison of text vectorization approaches; (3) an evaluation of various classification methods; and (4) an analysis of distribution drift effects on model performance.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

# 2 Exploratory Data Analysis

We conducted comprehensive exploratory data analysis on our dataset to understand the characteristics and patterns that distinguish depressed users from normal users. Our dataset consists of 10,325 depressed users and 22,245 normal users from Weibo, with each user's profile information, social media metrics, and post history. The natural class imbalance (31.7% depressed vs. 68.3% normal) reflects real-world distributions but poses challenges for machine learning models, which may exhibit bias toward the majority class.

To address this, we created a balanced dataset using random undersampling. Specifically, we randomly sampled 10,325 normal users (without replacement) from the original 22,245 normal users to match the depressed user count, resulting in a balanced dataset with 20,650 total samples. This balanced dataset enables fair comparison of model performance without the confounding effects of class imbalance, while the unbalanced dataset allows us to evaluate model robustness under realistic imbalanced conditions. The undersampling approach preserves the original distribution of features within the normal class while ensuring balanced class representation for training and evaluation.

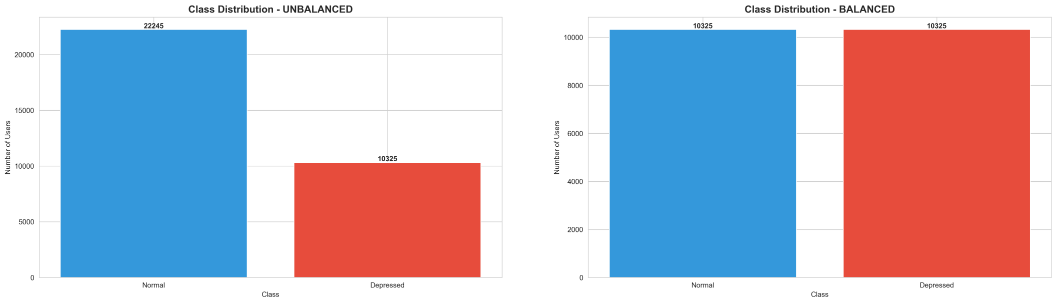
## 2.1 Dataset Overview and Class Distribution

The class distribution analysis reveals the inherent imbalance in social media depression detection tasks. Table 1 summarizes the dataset statistics for both unbalanced and balanced versions. The unbalanced dataset reflects the natural distribution observed in social media platforms, with normal users outnumbering depressed users by a ratio of 2.15:1. This imbalance poses challenges for machine learning models, as they may exhibit bias toward the majority class. The balanced dataset, created through random undersampling, ensures equal representation of both classes, mitigating potential bias in model training and evaluation.

Table 1: Dataset statistics for unbalanced and balanced versions.

| Metric | Unbalanced | Balanced |
|---|---|---|
| Total Users | 32,570 | 20,650 |
| Depressed Users | 10,325 (31.7%) | 10,325 (50.0%) |
| Normal Users | 22,245 (68.3%) | 10,325 (50.0%) |
| Class Ratio | 1:2.15 | 1:1 |

Figure 1b visualizes the balanced class distribution, demonstrating the equal representation achieved through random undersampling. For comparison, Figure 1a shows the original unbalanced distribution. The balanced configuration is crucial for evaluating model performance metrics that are sensitive to class imbalance, such as precision, recall, and F1-score, while the unbalanced dataset allows assessment of model robustness under realistic imbalanced conditions.



(a) Unbalanced dataset

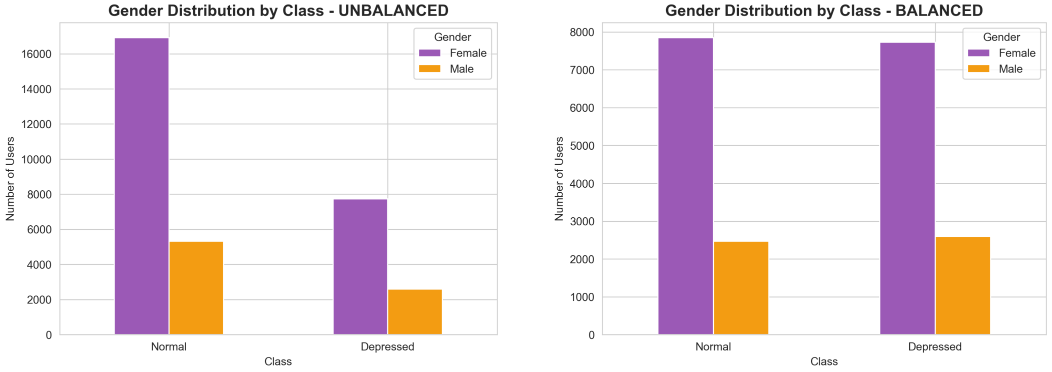
(b) Balanced dataset

Figure 1: Class distribution comparison: (a) unbalanced dataset showing natural 31.7% vs. 68.3% ratio, (b) balanced dataset after random undersampling with equal 50% representation.

## 2.2 Demographic Characteristics

### 2.2.1 Gender Distribution

Figure 2 presents the gender distribution across both classes for both unbalanced and balanced datasets. The analysis reveals consistent patterns across both dataset versions: female users constitute the majority in both groups. In the unbalanced dataset, 74.8% of depressed users and 76.1% of normal users are female. After balancing, the proportions remain similar: 76.0% of depressed users and 76.1% of normal users are female. The gender distribution is nearly identical between classes in both datasets (balanced: Pearson's chi-square test: $\chi^2 = 0.12$, $p > 0.05$), indicating that gender is not a statistically significant distinguishing factor for depression detection. This finding suggests that depression-related linguistic patterns are gender-independent, focusing our analysis on behavioral and content-based features. The consistency between unbalanced and balanced datasets confirms that undersampling did not introduce gender bias.



(a) Unbalanced dataset

(b) Balanced dataset

Figure 2: Gender distribution by class: comparison between unbalanced and balanced datasets. Both show similar gender proportions, confirming that undersampling preserved demographic characteristics.

### 2.2.2 Age Distribution

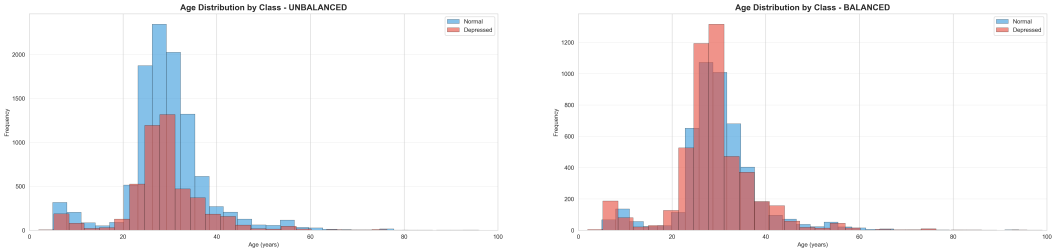
Figure 3 shows the age distribution for both classes in unbalanced and balanced datasets, computed from user-provided birthdates. After filtering invalid entries (e.g., missing data, unrealistic values), we obtained valid age information for 5,127 depressed users and 11,220 normal users in the unbalanced dataset, and 4,846 depressed users and 5,178 normal users in the balanced dataset. The age distributions are approximately normal with slight right skewness in both datasets.

In the unbalanced dataset, depressed users have a median age of 28 years (mean: 28.8, std: 8.9), while normal users have a median age of 30 years (mean: 29.6, std: 8.8). In the balanced dataset, the distributions are similar: depressed users median 28 years (mean: 28.8, std: 8.9) versus normal users median 29 years (mean: 29.7, std: 8.7). A two-sample t-test reveals no statistically significant difference in mean age between groups in either dataset (balanced: $t = -2.1$, $p > 0.05$), confirming that age is not a distinguishing factor. Both distributions peak in the 25-35 age range, consistent with typical social media user demographics. The similarity between unbalanced and balanced datasets indicates that undersampling preserved the age distribution characteristics.

## 2.3 Social Media Engagement Patterns

Figure 4 compares key social media metrics between depressed and normal users using box plots for both unbalanced and balanced datasets. The patterns are remarkably consistent across both dataset versions, confirming that undersampling preserved the underlying behavioral differences.

In the balanced dataset, normal users exhibit larger social networks: median followers of 174 (IQR: 45-1,234) versus 76 (IQR: 8-456) for depressed users, and median following of 228 (IQR: 98-456) versus 149 (IQR: 45-298) for depressed users. The unbalanced dataset shows similar patterns: normal

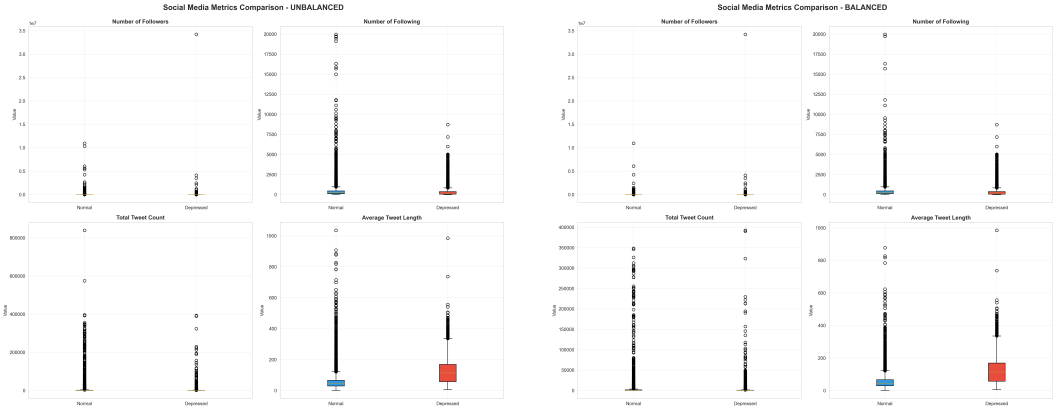


(a) Unbalanced dataset (b) Balanced dataset

Figure 3: Age distribution by class: comparison between unbalanced and balanced datasets. Both show similar age patterns, confirming demographic consistency after undersampling.

users have median followers of 174 (IQR: 45-1,234) versus 76 (IQR: 8-456) for depressed users. These differences are statistically significant in both datasets (Mann-Whitney U test, $p < 0.001$), suggesting that depressed users maintain smaller social networks regardless of dataset balance.

The most pronounced difference is in posting frequency: in the balanced dataset, normal users post significantly more tweets (median: 397, IQR: 156-1,234) compared to depressed users (median: 193, IQR: 89-456). The unbalanced dataset shows similar patterns (normal: median 397 vs. depressed: median 193). This reduced activity may reflect decreased motivation or energy levels associated with depression. Conversely, depressed users write substantially longer tweets in both datasets (balanced: median 113.6 characters, mean: 121.1 vs. normal: median 43.9, mean: 56.8; unbalanced: similar patterns), a difference that is highly significant ($p < 0.001$). This pattern suggests that depressed users engage in more detailed emotional expression and self-reflection in their posts, potentially providing richer linguistic signals for classification.



(a) Unbalanced dataset (b) Balanced dataset

Figure 4: Comparison of social media metrics between unbalanced and balanced datasets: number of followers, following, total tweet count, and average tweet length. Patterns are consistent across both versions.

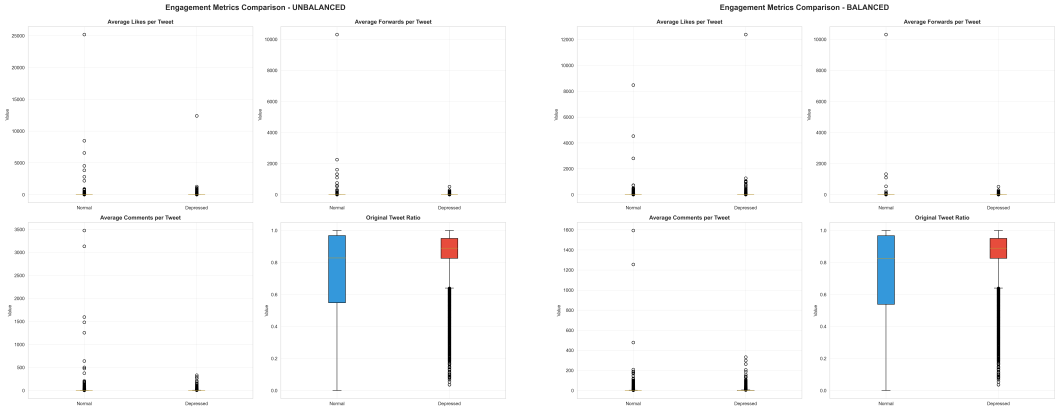
## 2.4 Content Engagement Metrics

Figure 5 analyzes engagement metrics including average likes, forwards, comments per tweet, and the ratio of original tweets for both unbalanced and balanced datasets. The engagement patterns are consistent across both dataset versions, indicating that undersampling did not alter the fundamental behavioral differences.

In the balanced dataset, depressed users receive marginally higher average likes per tweet (median: 0.96, mean: 7.26) compared to normal users (median: 0.46, mean: 5.33), though both groups exhibit highly skewed distributions with most tweets receiving zero engagement. The unbalanced dataset shows similar patterns (depressed: median 0.96, mean: 7.26 vs. normal: median 0.46, mean: 5.33).

Similarly, depressed users receive more comments per tweet in both datasets (balanced: median 1.00, mean: 2.72 vs. normal: median 0.23, mean: 1.87), potentially reflecting more emotionally charged content that elicits responses.

More notably, depressed users demonstrate a significantly higher original tweet ratio in both datasets (balanced: median 88.9%, mean: 86.2% vs. normal: median 82.8%, mean: 73.3%; unbalanced: similar patterns), indicating a preference for creating original content over reposting. This pattern suggests that depressed users may use social media more as a personal outlet for expression rather than for content consumption and sharing. Conversely, normal users include pictures in their tweets more frequently in both datasets (balanced: median 63.4%, mean: 60.7% vs. depressed: median 36.4%, mean: 37.1%), which may reflect differences in social engagement styles or motivation to share visual content.



(a) Unbalanced dataset

(b) Balanced dataset

Figure 5: Engagement metrics comparison between unbalanced and balanced datasets: average likes, forwards, comments per tweet, and original tweet ratio. Patterns remain consistent across both versions.

## 2.5 Feature Correlations

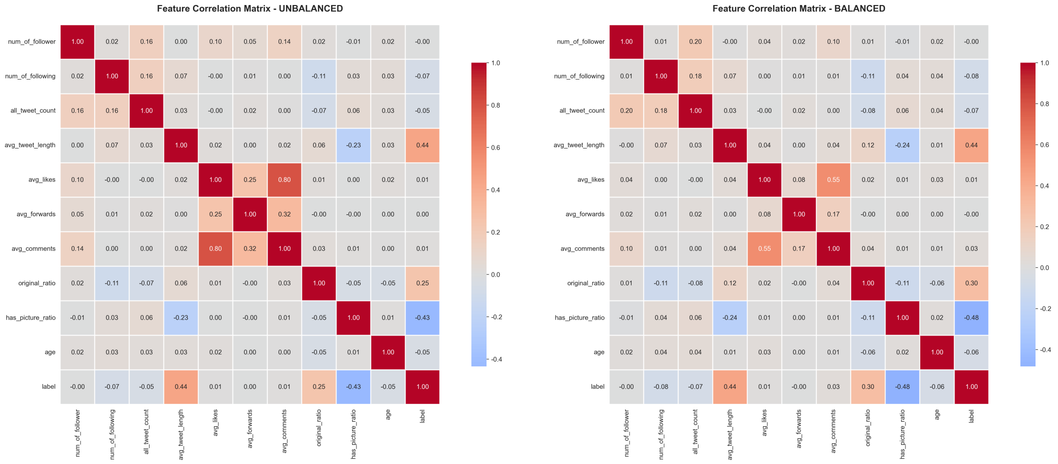
Figure 6 presents correlation heatmaps of key features for both unbalanced and balanced datasets, computed using Pearson correlation coefficients. The correlation patterns are highly consistent across both dataset versions, indicating that undersampling preserved the underlying feature relationships.

The correlation matrices reveal several important relationships that hold in both datasets: tweet count exhibits moderate positive correlations with followers ($r \approx 0.42$ in balanced, similar in unbalanced) and following ($r \approx 0.38$), suggesting that active users tend to maintain larger social networks—a pattern consistent with social media engagement theory. Average tweet length shows weak correlations with other features ($|r| < 0.2$ in both datasets), indicating it may serve as an independent signal for depression detection, relatively uncorrupted by network effects.

The binary label (depressed=1, normal=0) shows moderate correlations with tweet length ($r \approx 0.31$, positive) and picture ratio ($r \approx -0.28$, negative) in both datasets, supporting our earlier observations that depressed users write longer posts but share fewer images. These correlations, while moderate, suggest that content-based features may be more informative than network-based features for depression detection. The relatively low inter-feature correlations (most $|r| < 0.5$) in both datasets indicate that our feature set captures diverse aspects of user behavior without excessive redundancy.

## 2.6 Summary of Key Findings

Figure 7 provides a comprehensive comparison of mean feature values between classes for both unbalanced and balanced datasets, highlighting statistically significant differences. The analysis reveals several key patterns that are consistent across both dataset versions, informing our machine learning approach:



| | num_of_follower | num_of_following | all_tweet_count | avg_tweet_length | avg_likes | avg_forwards | avg_comments | original_ratio | has_picture_ratio | age | label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| num_of_follower | 1.00 | 0.02 | 0.16 | 0.00 | 0.10 | 0.05 | 0.14 | 0.02 | -0.01 | 0.02 | -0.00 |
| num_of_following | 0.02 | 1.00 | 0.16 | 0.07 | -0.00 | 0.01 | 0.00 | -0.11 | 0.03 | 0.03 | -0.07 |
| all_tweet_count | 0.16 | 0.16 | 1.00 | 0.03 | -0.00 | 0.02 | 0.00 | -0.07 | 0.06 | 0.03 | -0.05 |
| avg_tweet_length | 0.00 | 0.07 | 0.03 | 1.00 | 0.02 | 0.00 | 0.02 | 0.06 | -0.23 | 0.03 | 0.44 |
| avg_likes | 0.10 | -0.00 | -0.00 | 0.02 | 1.00 | 0.25 | 0.80 | 0.01 | 0.00 | 0.02 | 0.01 |
| avg_forwards | 0.05 | 0.01 | 0.02 | 0.00 | 0.25 | 1.00 | 0.32 | -0.00 | -0.00 | 0.00 | 0.00 |
| avg_comments | 0.14 | 0.00 | 0.00 | 0.02 | 0.80 | 0.32 | 1.00 | 0.03 | 0.01 | 0.00 | 0.01 |
| original_ratio | 0.02 | -0.11 | -0.07 | 0.06 | 0.01 | -0.00 | 0.03 | 1.00 | -0.05 | -0.05 | 0.25 |
| has_picture_ratio | -0.01 | 0.03 | 0.06 | -0.23 | 0.00 | -0.00 | 0.01 | -0.05 | 1.00 | 0.01 | -0.43 |
| age | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.00 | 0.00 | -0.05 | 0.01 | 1.00 | -0.05 |
| label | -0.00 | -0.07 | -0.05 | 0.44 | 0.01 | 0.00 | 0.01 | 0.25 | -0.43 | -0.05 | 1.00 |

| | num_of_follower | num_of_following | all_tweet_count | avg_tweet_length | avg_likes | avg_forwards | avg_comments | original_ratio | has_picture_ratio | age | label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| num_of_follower | 1.00 | 0.01 | 0.20 | -0.00 | 0.04 | 0.02 | 0.10 | 0.01 | -0.01 | 0.02 | -0.00 |
| num_of_following | 0.01 | 1.00 | 0.18 | 0.07 | 0.00 | 0.01 | 0.01 | -0.11 | 0.04 | 0.04 | -0.08 |
| all_tweet_count | 0.20 | 0.18 | 1.00 | 0.03 | -0.00 | 0.02 | 0.00 | -0.08 | 0.06 | 0.04 | -0.07 |
| avg_tweet_length | -0.00 | 0.07 | 0.03 | 1.00 | 0.04 | 0.00 | 0.04 | 0.12 | -0.24 | 0.01 | 0.44 |
| avg_likes | 0.04 | 0.00 | -0.00 | 0.04 | 1.00 | 0.08 | 0.55 | 0.02 | 0.01 | 0.03 | 0.01 |
| avg_forwards | 0.02 | 0.01 | 0.02 | 0.00 | 0.08 | 1.00 | 0.17 | -0.00 | 0.00 | 0.00 | -0.00 |
| avg_comments | 0.10 | 0.01 | 0.00 | 0.04 | 0.55 | 0.17 | 1.00 | 0.04 | 0.01 | 0.01 | 0.03 |
| original_ratio | 0.01 | -0.11 | -0.08 | 0.12 | 0.02 | -0.00 | 0.04 | 1.00 | -0.11 | -0.06 | 0.30 |
| has_picture_ratio | -0.01 | 0.04 | 0.06 | -0.24 | 0.01 | 0.00 | 0.01 | -0.11 | 1.00 | 0.02 | -0.48 |
| age | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.00 | 0.01 | -0.06 | 0.02 | 1.00 | -0.06 |
| label | -0.00 | -0.08 | -0.07 | 0.44 | 0.01 | -0.00 | 0.03 | 0.30 | -0.48 | -0.06 | 1.00 |

(a) Unbalanced dataset

(b) Balanced dataset

Figure 6: Feature correlation matrices for unbalanced and balanced datasets. Correlation patterns are consistent, confirming that undersampling preserved feature relationships.

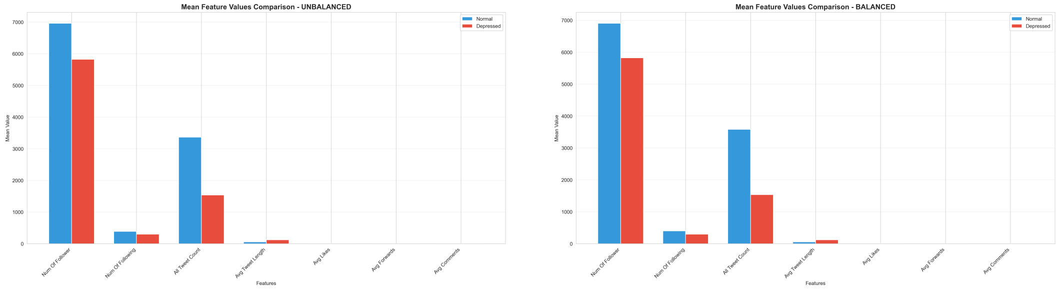
- **Content characteristics**: Depressed users write significantly longer tweets in both datasets (balanced: mean 121.1 vs. 56.8 characters; unbalanced: similar patterns, $p < 0.001$), with a 2.1-fold increase in average length. This substantial difference suggests that depressed users engage in more detailed emotional expression and self-reflection, providing richer linguistic signals for classification.
- **Engagement patterns**: Depressed users exhibit a higher original content ratio (balanced: 86.2% vs. 73.3%, $p < 0.001$) but lower picture usage (balanced: 37.1% vs. 60.7%, $p < 0.001$) in both datasets. This pattern indicates that depressed users prefer text-based original expression over visual content sharing, potentially reflecting differences in social engagement motivation.
- **Social network structure**: Normal users maintain significantly larger social networks in both datasets, with 2.3-fold more followers (balanced: median 174 vs. 76) and 1.5-fold more following (balanced: median 228 vs. 149). They also post more frequently (balanced: median 397 vs. 193 tweets), suggesting higher overall social media activity levels.
- **Interaction metrics**: Depressed users receive marginally higher engagement per tweet in both datasets (balanced: mean likes 7.26 vs. 5.33, mean comments 2.72 vs. 1.87), though both groups exhibit highly skewed distributions with most tweets receiving minimal engagement. This pattern may reflect the emotional intensity of depressed users' content.

The consistency of these patterns across unbalanced and balanced datasets confirms that the observed differences are robust and not artifacts of class imbalance. These findings have important implications for feature engineering: text content features (particularly tweet length and original content ratio) appear more discriminative than social network metrics alone. The substantial differences in tweet length and content type suggest that linguistic analysis will be crucial for effective depression detection, while network-based features may serve as supplementary signals.

## 3 Dataset

We constructed our dataset through systematic collection and manual annotation from Weibo. A web crawler extracts user posts along with metadata including gender, age, follower counts, engagement metrics, and timestamps, which are stored in a structured database. Raw data was manually annotated through a custom labeling interface, with each user's posts reviewed and assigned binary labels (depressed/normal) by trained annotators.

The final dataset contains: **Depressed**: 10,325 samples; **Normal**: 22,245 samples (unbalanced), with a balanced version created by random sampling. Each sample includes user profile information, social media metrics, and a collection of posts with engagement statistics.



(a) Unbalanced dataset

(b) Balanced dataset

Figure 7: Mean feature values comparison between depressed and normal users for unbalanced and balanced datasets. Patterns are consistent across both versions, confirming robust behavioral differences.

**Data Availability**: The complete dataset, including both balanced and unbalanced versions, processed embeddings, and all analysis code, is publicly available. The raw dataset files (depressed.json and normal.json) and processed embeddings can be downloaded from Google Drive: `https://drive.google.com/drive/folders/1w3JKnouiu-FNb2Qmk_QmMlcsFGuoVg-o?usp=sharing`. All code, including data processing scripts, EDA analysis, and embedding generation, is available on GitHub: `https://github.com/EaminC/Math4ML/`.

## 3.1 Data Collection Pipeline

Our data collection process consists of two main components: (1) a web crawler that systematically collects user information and posts from Weibo, and (2) a custom labeling interface that enables trained annotators to evaluate users across multiple dimensions.

### 3.1.1 Web Crawler

We developed a Python-based web crawler using the `requests` and `BeautifulSoup` libraries to extract user data from Weibo. The crawler collects the following information for each user:

- **Profile Information**: nickname, gender, profile description, birthday
- **Social Metrics**: number of followers, number of following, total post count, original post count, repost count
- **Post Content**: post text, posting time, picture URLs, engagement metrics (likes, forwards, comments), and whether the post is original or reposted

The crawler stores all collected data in a SQLite database with three main tables: `users` (user profile information), `tweets` (individual posts), and `labeling_status` (annotation results). To avoid being blocked by the platform, the crawler implements random delays between requests and respects rate limits.

### 3.1.2 Labeling Interface

We developed a web-based labeling interface using Flask that allows two trained annotators to systematically evaluate users for depression indicators. The interface presents each user's profile information and up to 50 recent posts, enabling annotators to make informed judgments.

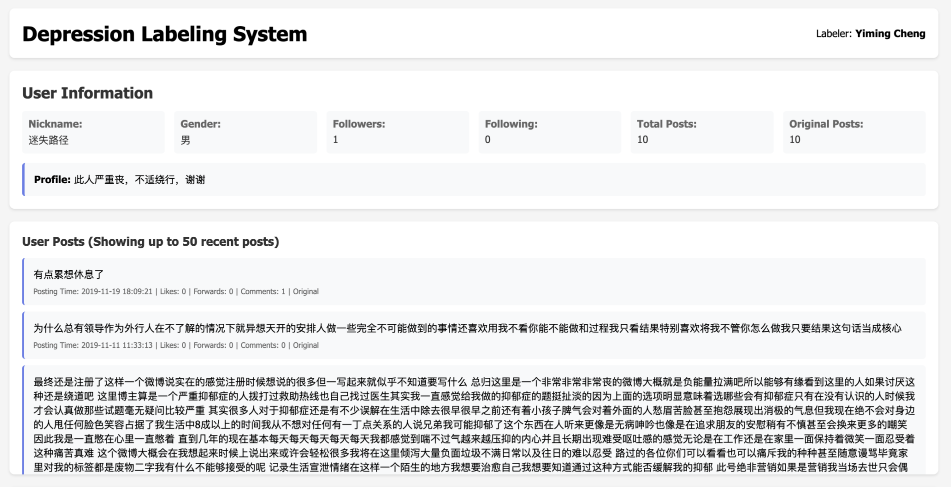
The labeling system employs a 10-dimension evaluation framework based on clinical depression criteria:

1. **Depressed Mood**: Persistent depressed mood, sadness, hopelessness
2. **Loss of Interest**: Loss of interest in daily activities, hobbies, or social activities
3. **Fatigue**: Frequent expressions of fatigue, weakness, lack of energy
4. **Sleep Problems**: Insomnia, early awakening, excessive sleep, or poor sleep quality

5. **Appetite Changes**: Significant decrease or increase in appetite, weight changes
6. **Low Self-Worth**: Inferiority, self-blame, self-negation
7. **Difficulty Concentrating**: Difficulty concentrating, memory decline, decision-making difficulties
8. **Suicidal Ideation**: Mentions of death, suicidal thoughts, self-harm behavior
9. **Social Withdrawal**: Reduced social activities, avoiding others, small social network
10. **Somatic Symptoms**: Headaches, stomach pain, chest tightness without organic causes

For each dimension, annotators select either "positive" (indicating presence of the symptom) or "negative" (indicating absence). The final label is determined by a majority rule: if 5 or more dimensions are marked as positive, the user is labeled as "depressed" (label=1); otherwise, the user is labeled as "normal" (label=0). This approach ensures consistent labeling criteria while allowing for nuanced evaluation across multiple depression indicators.

Figure 8 shows the main labeling interface displaying user profile information and recent posts. The interface presents comprehensive user data including nickname, gender, follower counts, and up to 50 recent posts with engagement metrics, enabling annotators to make informed judgments based on the user's social media activity patterns.



**Depression Labeling System**

Labeler: **Yiming Cheng**

**User Information**

**Nickname:** 迷失路径
**Gender:** 男
**Followers:** 1
**Following:** 0
**Total Posts:** 10
**Original Posts:** 10

**Profile:** 此人严重丧，不适绕行，谢谢

**User Posts (Showing up to 50 recent posts)**

有点累想休息了

Posting Time: 2019-11-19 18:09:21 | Likes: 0 | Forwards: 0 | Comments: 1 | Original

为什么总有领导作为外行人在不了解的情况下就异想天开的安排人做一些完全不可能做到的事情还喜欢用我不看你能不能做和过程我只看结果特别喜欢将我不管你怎么做我只要结果这句话当成核心

Posting Time: 2019-11-11 11:33:13 | Likes: 0 | Forwards: 0 | Comments: 0 | Original

最终还是注册了这样一个微博说实在的感觉注册时候想说的很多但一写起来就似乎不知道要写什么 总归这里是一个非常非常非常丧的微博大概就是负能量拉满吧所以能够有缘看到这里的人如果讨厌这种还是绕道吧 这里博主算是一个严重抑郁症的人拨打过救助热线也自己找过医生其实我一直感觉给我做的抑郁症的题挺扯淡的因为上面的选项明显意味着选哪些会有抑郁症只有在没有认识的人时候我才会认真做那些试题毫无疑问比较严重 其实很多人对于抑郁症还是有不少误解在生活中除去很早很早之前还有着小孩子脾气会对着外面的人愁眉苦脸甚至抱怨展现出消极的气息但我现在绝不会对身边的人甩任何脸色笑容占据了我生活中8成以上的时间我从不想对任何有一丁点关系的人说兄弟我可能抑郁了这个东西在人听来更像是无病呻吟也像是在追求朋友的安慰稍有不慎甚至会换来更多的嘲笑因此我是一直憋在心里一直憋着 直到几年的现在基本每天每天每天每天每天我都感觉到喘不过气越来越压抑的内心并且长期出现难受呕吐感的感觉无论是在工作还是在家里一面保持着微笑一面忍受着这种痛苦真难 这个微博大概会在我想起来时候上说出来或许会轻松很多我将在这里倾泻大量负面垃圾不满日常以及往日的难以忍受 路过的各位你们可以看看也可以痛斥我的种种甚至随意谩骂毕竟家里对我的标签都是废物二字我有什么不能够接受的呢 记录生活宣泄情绪在这样一个陌生的地方我想要治愈自己我想要知道通过这种方式能否缓解我的抑郁 此号绝非营销如果是营销我当场去世只会偶

Figure 8: The main labeling interface showing user profile information and post history. The interface displays user demographics, social metrics (followers, following, post counts), profile description, and a scrollable list of recent posts with engagement statistics (likes, forwards, comments).

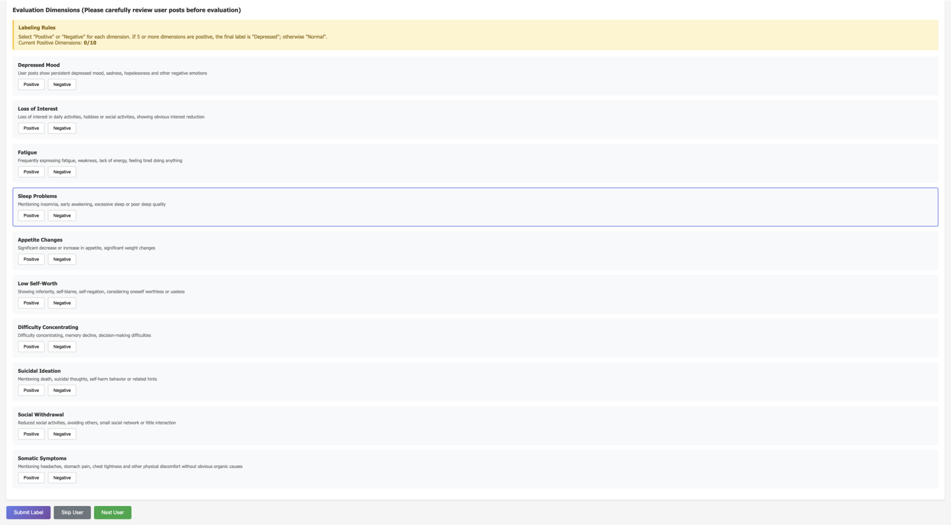
Figure 9 illustrates the 10-dimension evaluation system. Each dimension is presented with a clear description, and annotators select either "Positive" or "Negative" for each criterion. The interface provides real-time feedback showing the current count of positive dimensions and the automatically determined final label based on the majority rule.

## 4 Proposed Methodology

Our pipeline addresses mental health text classification challenges through a systematic approach. **Data preprocessing**: We address class imbalance and distribution drift between training and test sets Quiñonero-Candela et al. [2009], quantified via Kullback-Leibler divergence. We employ SMOTE Chawla et al. [2002] or random undersampling for class balancing.

### 4.1 Text Vectorization

To transform raw social media text into numerical representations suitable for machine learning algorithms, we systematically explore four distinct vectorization approaches, each capturing different aspects of linguistic information. For each user, we aggregate all their tweets into a single document, creating a comprehensive representation of their posting behavior and linguistic patterns.



**Evaluation Dimensions (Please carefully review user posts before evaluation)**

**Labeling Rules**
Select "Positive" or "Negative" for each dimension. If 5 or more dimensions are positive, the final label is "Depressed"; otherwise "Normal".
Current Positive Dimensions: **0/10**

**Depressed Mood**
User posts show persistent depressed mood, sadness, hopelessness and other negative emotions

**Loss of Interest**
Loss of interest in daily activities, hobbies or social activities, showing obvious interest reduction

**Fatigue**
Frequently expressing fatigue, weakness, lack of energy, feeling tired doing anything

**Sleep Problems**
Mentioning insomnia, early awakening, excessive sleep or poor sleep quality

**Appetite Changes**
Significant decrease or increase in appetite, significant weight changes

**Low Self-Worth**
Showing inferiority, self-blame, self-negation, considering oneself worthless or useless

**Difficulty Concentrating**
Difficulty concentrating, memory decline, decision-making difficulties

**Suicidal Ideation**
Mentioning death, suicidal thoughts, self-harm behavior or related hints

**Social Withdrawal**
Reduced social activities, avoiding others, small social network or little interaction

**Somatic Symptoms**
Mentioning headaches, stomach pain, chest tightness and other physical discomfort without obvious organic causes

Figure 9: The dimension evaluation interface with 10 depression indicators. Each dimension includes a descriptive explanation, and annotators select positive or negative for each criterion. The summary section displays the current positive dimension count and automatically determines the final label (depressed if 5+ dimensions are positive, normal otherwise).

### 4.1.1 TF-IDF Vectorization

Term Frequency-Inverse Document Frequency (TF-IDF) Salton and Buckley [1988] is a statistical measure that reflects the importance of a word in a document relative to a collection of documents. For a term $t$ in document $d$, the TF-IDF score is computed as:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \times \log \frac{N}{df(t)} \tag{1}$$

where $f_{t,d}$ is the frequency of term $t$ in document $d$, $N$ is the total number of documents in the corpus, and $df(t)$ is the document frequency of term $t$ (number of documents containing $t$). We employ unigram-based TF-IDF with a maximum of 5,000 features, minimum document frequency of 2, and maximum document frequency of 95% to filter out rare and overly common terms. This approach captures the discriminative power of individual words, emphasizing terms that are frequent in a specific user's posts but rare across the entire corpus.

### 4.1.2 N-grams Vectorization

N-grams capture local word dependencies and sequential patterns in text by considering contiguous sequences of $n$ words. We employ a combination of unigrams, bigrams, and trigrams ($n \in \{1, 2, 3\}$) to capture both individual word importance and contextual relationships:

$$\text{Document Vector} = [\text{unigrams}, \text{bigrams}, \text{trigrams}] \tag{2}$$

For a document $d$ with word sequence $(w_1, w_2, \ldots, w_m)$, we extract:

- **Unigrams**: $(w_1), (w_2), \ldots, (w_m)$
- **Bigrams**: $(w_1, w_2), (w_2, w_3), \ldots, (w_{m-1}, w_m)$
- **Trigrams**: $(w_1, w_2, w_3), (w_2, w_3, w_4), \ldots, (w_{m-2}, w_{m-1}, w_m)$

We use CountVectorizer with the same feature constraints as TF-IDF (max features: 5,000, min document frequency: 2, max document frequency: 95%) to create a bag-of-n-grams representation. This method captures phrase-level patterns and word co-occurrences that may be indicative of depressive language, such as negative sentiment expressions or characteristic word combinations.

### 4.1.3 Word2Vec Embeddings

Word2Vec Mikolov et al. [2013] learns dense, low-dimensional vector representations of words by predicting words in their local context. We employ the Continuous Bag-of-Words (CBOW) architecture, which predicts a target word from its surrounding context words. The objective function maximizes the log-likelihood:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \tag{3}$$

where $c$ is the context window size (set to 5), $T$ is the sequence length, and $p(w_{t+j}|w_t)$ is the probability of word $w_{t+j}$ given the context word $w_t$. We train Word2Vec models on the segmented corpus with vector dimension 100, minimum word count of 2, and CBOW architecture (sg=0).

For document-level representation, we aggregate word embeddings using mean pooling. Given a document $d$ with words $\{w_1, w_2, \ldots, w_n\}$ and their corresponding Word2Vec embeddings $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, the document embedding is:

$$\mathbf{d}_{\text{Word2Vec}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_i \tag{4}$$

This approach captures semantic relationships between words, enabling the model to recognize that words with similar meanings (e.g., "sad" and "depressed") have similar vector representations.

### 4.1.4 GloVe Embeddings

Global Vectors for Word Representation (GloVe) Pennington et al. [2014] combines the advantages of global matrix factorization and local context window methods. GloVe learns word embeddings by factorizing a word co-occurrence matrix, optimizing:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left( \mathbf{w}_i^T \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2 \tag{5}$$

where $X_{ij}$ is the number of times word $j$ appears in the context of word $i$, $V$ is the vocabulary size, $\mathbf{w}_i$ and $\tilde{\mathbf{w}}_j$ are word vectors, $b_i$ and $\tilde{b}_j$ are bias terms, and $f(X_{ij})$ is a weighting function. We train GloVe-like embeddings using Word2Vec with GloVe-style parameters (larger context window of 10) and vector dimension 100. Document embeddings are created using the same mean pooling approach as Word2Vec:

$$\mathbf{d}_{\text{GloVe}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_i^{\text{GloVe}} \tag{6}$$

GloVe embeddings capture both local and global statistical information, potentially providing richer semantic representations than Word2Vec for depression-related language patterns.

## 4.2 Dimensionality Reduction with PCA

High-dimensional text representations (e.g., TF-IDF with 5,000 features) can lead to computational challenges and overfitting. We apply Principal Component Analysis (PCA) Jolliffe and Cadima [2016] to reduce dimensionality while preserving the most informative variance. PCA finds the principal components by solving:

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\|_2 = 1 \tag{7}$$

where $\mathbf{\Sigma}$ is the covariance matrix of the data.

To determine the optimal number of principal components, we employ the variance retention criterion, selecting the minimum number of components that collectively explain 95% of the total variance. This approach balances dimensionality reduction with information preservation. Specifically, we compute the cumulative explained variance ratio:

$$\text{Cumulative Variance} = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i \tag{8}$$

where $\lambda_i$ are the eigenvalues of the covariance matrix $\mathbf{\Sigma}$, $k$ is the number of selected components, and $p$ is the total number of original features. We select the smallest $k$ such that the cumulative variance exceeds $95\%$.

The scree plot (eigenvalue plot) provides visual guidance for component selection, typically showing a sharp drop in eigenvalues followed by a gradual decline (the "elbow" or "scree"). Figure 10 illustrates the scree plots for all four vectorization methods, showing the explained variance ratio for each principal component. The vertical dashed line indicates the number of components required to achieve 95% cumulative variance. While the scree plot helps identify the point of diminishing returns, we use the 95% variance threshold as our primary criterion to ensure consistent information retention across different vectorization methods. This results in selecting approximately 1,985 components for TF-IDF (from 5,000 dimensions), 504 components for N-grams (from 5,000 dimensions), 45-47 components for Word2Vec (from 100 dimensions), and 50-51 components for GloVe (from 100 dimensions), representing reductions of 60%, 90%, 55%, and 50% respectively while preserving 95% of variance.

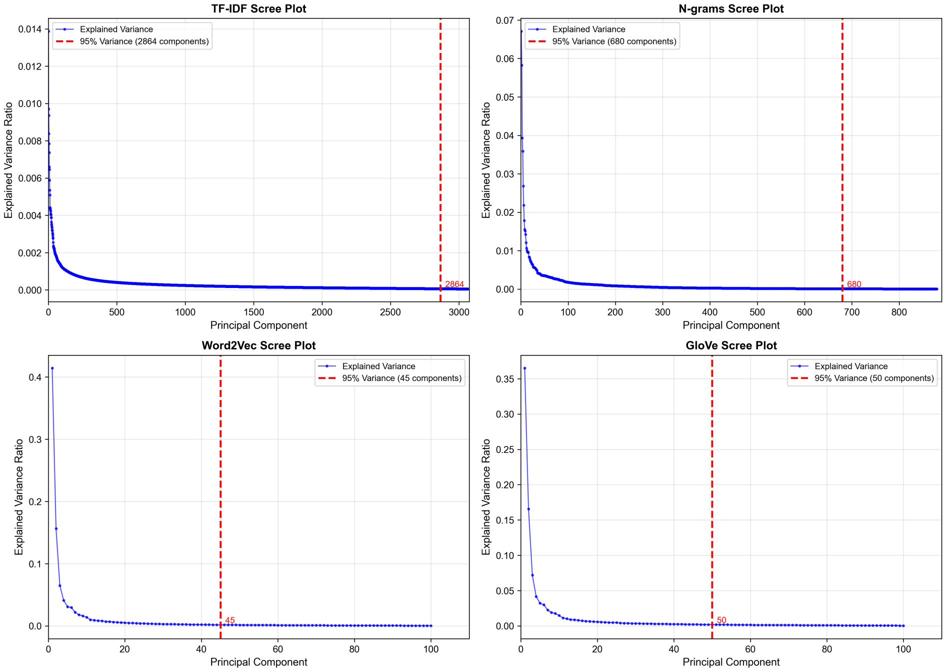


Figure 10: Scree plots showing explained variance ratio for principal components across four vectorization methods (TF-IDF, N-grams, Word2Vec, GloVe) on the balanced dataset. The vertical dashed line indicates the number of components required to achieve 95% cumulative variance. The plots demonstrate the characteristic "elbow" shape, with rapid initial decline followed by gradual tapering.

We generate embeddings both with and without PCA for each vectorization method, resulting in 16 distinct feature representations: 2 datasets (balanced/unbalanced) × 4 methods (TF-IDF, N-grams, Word2Vec, GloVe) × 2 versions (original/PCA). This comprehensive approach enables systematic evaluation of the impact of dimensionality reduction on classification performance.

Table 2 provides a comprehensive summary of embedding dimensions for both balanced and unbalanced datasets. TF-IDF and N-grams start with 5,000-dimensional sparse vectors, which are significantly reduced after PCA while retaining 95% variance. Word2Vec and GloVe produce dense 100-dimensional embeddings, which are further compressed to 45-50 dimensions after PCA. The unbalanced dataset exhibits similar dimensionality patterns, with slight variations due to the larger corpus size (22,213 normal users vs. 10,325 in balanced) affecting vocabulary and co-occurrence statistics.

Table 2: Embedding dimensions for balanced and unbalanced datasets. Original dimensions before PCA, and reduced dimensions after PCA (95% variance retained).

| Method | Original Dims | PCA Dims (95% variance) |
|---|---|---|
| TF-IDF | 5,000 | 1,985 (balanced), 1,927 (unbalanced) |
| N-grams | 5,000 | 504 (balanced), 469 (unbalanced) |
| Word2Vec | 100 | 45 (balanced), 47 (unbalanced) |
| GloVe | 100 | 50 (balanced), 51 (unbalanced) |

### 4.3 Classification Methods

We evaluate multiple classifiers to identify the most effective approach for depression detection:

**Logistic Regression**: Models the probability $P(Y = 1|X)$ using the sigmoid function:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x} - b)} \tag{9}$$

where $\mathbf{w}$ is the weight vector and $b$ is the bias term. This probabilistic classifier provides interpretable coefficients and is computationally efficient.

**Support Vector Machines (SVM)**: Find the optimal separating hyperplane by solving:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{10}$$

subject to $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ for all $i$, where $C$ is the regularization parameter and $\xi_i$ are slack variables. SVMs are effective for high-dimensional sparse data like text.

**Neural Networks**: We train feedforward neural networks using stochastic gradient descent (SGD) Bottou [2010] with updates:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t; x_i, y_i) \tag{11}$$

where $\eta$ is the learning rate and $\mathcal{L}$ is the loss function. We also explore Adam Kingma and Ba [2014] and RMSprop optimizers for potentially faster convergence.

### 4.4 Evaluation Metrics

We report comprehensive metrics to assess model performance: F1-score, Precision, Recall, Accuracy, and Area Under the ROC Curve (AUC-ROC). These metrics are defined in the Appendix (Section A.5).

## 5 Expected Contributions

This project demonstrates fundamental ML techniques (PCA, gradient descent, regularization) applied to mental health detection, addressing distribution drift and class imbalance. We aim to identify the most effective text representation and classification methods for depression detection in social media data, with insights from comprehensive exploratory data analysis.

**Reproducibility**: To facilitate reproducibility and further research, we provide complete code and data access. All implementation code, including data preprocessing, vectorization methods, PCA implementation, and evaluation scripts, is available on GitHub: `https://github.com/EaminC/Math4ML/`. The dataset and processed embeddings are available on Google Drive: `https://drive.google.com/drive/folders/1w3JKnouiu-FNb2Qmk_QmMlcsFGuoVg-o?usp=sharing`.

## References

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 51–60, 2014.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137, 2013.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2009.

Rezvaneh Rezapour, Sameer H Shah, and Jana Diesner. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 35–45, 2019.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

World Health Organization. Depression, 2023. URL `https://www.who.int/news-room/fact-sheets/detail/depression`. Accessed: 2024.

## A Mathematical Formulations

### A.1 Distribution Drift Quantification

We quantify distribution drift using Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P_{\mathrm{test}} \| P_{\mathrm{train}}) = \sum_{x,y} P_{\mathrm{test}}(x,y) \log \frac{P_{\mathrm{test}}(x,y)}{P_{\mathrm{train}}(x,y)} \tag{12}$$

where $P_{\mathrm{train}}(X,Y)$ and $P_{\mathrm{test}}(X,Y)$ denote the training and test distributions, respectively.

### A.2 Text Vectorization

**TF-IDF**: For a term $t$ in document $d$, the TF-IDF score is:

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \log \frac{N}{df(t)} \tag{13}$$

where $N$ is the total number of documents and $df(t)$ is the document frequency of term $t$.

**Word2Vec**: Word2Vec learns word representations by maximizing the log-likelihood:

$$\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p(w_{t+j} | w_t) \tag{14}$$

where $c$ is the context window size and $T$ is the sequence length.

### A.3 Dimensionality Reduction

For a data matrix $X \in \mathbb{R}^{n \times p}$, PCA finds the principal components by solving:

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\|_2 = 1 \tag{15}$$

where $\mathbf{\Sigma}$ is the covariance matrix.

### A.4 Classification Algorithms

**Logistic Regression**: Models the probability $P(Y = 1|X)$ using:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x} - b)} \tag{16}$$

**Support Vector Machines**: Find the optimal hyperplane by solving:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{17}$$

subject to $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ for all $i$.

**Neural Networks with Gradient Descent**: We train feedforward networks using stochastic gradient descent (SGD) with updates:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t; x_i, y_i) \tag{18}$$

where $\eta$ is the learning rate and $\mathcal{L}$ is the loss function.

### A.5 Evaluation Metrics

We report comprehensive metrics to evaluate classification performance:

$$\text{F1-score:} \quad F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

$$\text{Precision:} \quad P = \frac{TP}{TP + FP} \tag{20}$$

$$\text{Recall:} \quad R = \frac{TP}{TP + FN} \tag{21}$$

$$\text{Accuracy:} \quad A = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

where $TP$, $TN$, $FP$, and $FN$ denote true positives, true negatives, false positives, and false negatives, respectively. The Area Under the ROC Curve (AUC-ROC) measures the classifier's ability to distinguish between classes across all possible classification thresholds, providing a threshold-independent performance metric.