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. We employ multiple vectorization techniques (TF-IDF, N-grams, Word2Vec, GloVe), PCA, and evaluate classifiers (logistic regression, SVM, neural networks with SGD). Methodology includes preprocessing, SMOTE for class balancing, and evaluation using F1-score, precision, recall, AUC-ROC. Dataset: 10,325 depressed and 22,245 normal samples from X (Twitter).

1 Introduction and problem statement

Depression affects millions worldwide; early detection is crucial World Health Organization [2023]. Social media provides rich linguistic data revealing psychological states De Choudhury et al. [2013]. Previous work demonstrates feasibility of depression detection from social media Coppersmith et al. [2014], Rezapour et al. [2019]. We propose an ML system analyzing social media posts to identify at-risk individuals, addressing distribution drift challenges.

2 Dataset

We constructed a dataset via systematic collection and manual annotation from X (Twitter) through web scraping. A crawler extracts user posts with metadata (gender, age, follower counts, engagement metrics, timestamps) stored in a database. Raw data was manually annotated through a custom labeling interface, with each user's posts reviewed and assigned binary labels (depressed/normal). Final dataset: **Depressed**: 10,325 samples; **Normal**: 22,245 samples.

3 Proposed methodology

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

Text vectorization: We explore multiple representations: (1) **TF-IDF** Salton and Buckley [1988] for term weighting; (2) **N-grams** (unigrams, bigrams, trigrams) to capture local word dependencies; (3) **Word embeddings** using pre-trained Word2Vec Mikolov et al. [2013] and GloVe Pennington et al. [2014] for semantic representations. Detailed formulations are provided in the Appendix.

Dimensionality reduction: We apply Principal Component Analysis (PCA) Jolliffe and Cadima [2016] to reduce high-dimensional text features, retaining components explaining 95% variance. The optimization formulation is detailed in the Appendix.

Classification: We evaluate multiple classifiers: (1) **Logistic Regression** for probabilistic classification; (2) **Support Vector Machines (SVM)** for maximum-margin classification; (3) **Neural Networks** trained via stochastic gradient descent (SGD) Bottou [2010], also exploring Adam Kingma and Ba [2014] and RMSprop optimizers. Mathematical formulations are provided in the Appendix.

Evaluation: We report comprehensive metrics including F1-score, Precision, Recall, Accuracy, and Area Under the ROC Curve (AUC-ROC). Detailed metric definitions are provided in the Appendix.

4 Expected contributions

This project demonstrates fundamental ML techniques (PCA, gradient descent, regularization) applied to mental health, 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.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

A Mathematical Formulations

A.1 Distribution Drift Quantification

We quantify distribution drift using Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P_{\mathrm{test}} \| P_{\mathrm{train}}) = \sum_{x,y} P_{\mathrm{test}}(x,y) \log \frac{P_{\mathrm{test}}(x,y)}{P_{\mathrm{train}}(x,y)} \tag{1}$$

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:

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \log \frac{N}{df(t)} \tag{2}$$

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:

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

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:

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\|_2 = 1 \tag{4}$$

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

A.4 Classification Algorithms

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

$$P(Y=1|X) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x} - b)} \tag{5}$$

Support Vector Machines: Find the optimal hyperplane by solving:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{6}$$

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

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

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t; x_i, y_i) \tag{7}$$

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

A.5 Evaluation Metrics

We report comprehensive metrics:

$$\text{F1-score:} \quad F_1 = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

$$\text{Precision:} \quad P = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall:} \quad R = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Accuracy:} \quad A = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

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.