Supplementary Information

Beyond General Awareness: A Global Empirical Framework for Behavior-Specific Policy Development Against Academic Bullying

1 Data Characteristics and Analytical Challenges

The primary objective of this study is to extract actionable, demographic-specific rules capable of predicting the concurrent manifestation of 15 Tepper Scale abusive behaviors and 10 contextual sabotage actions within academic environments.

1.1 Feature Space: Capturing Sociological Intersectionality

The independent feature space, $X \in \mathbb{R}^{n \times d}$, consists exclusively of categorical demographic variables: **Age, Ethnicity, Gender, Country, and Academic Position**. These are transformed into a sparse binary matrix via one-hot encoding.

However, predicting human behavior requires moving beyond isolated variables due to **sociological intersectionality**. In standard linear modeling, features are often assumed to have independent, additive effects. In sociological reality, variables interact non-linearly; the behavioral drivers of being a "Postdoc" may fundamentally change depending on whether the individual is in the "USA" versus "Japan," or "Age 25–30" versus "Age 40–50."

Intersectionality implies that the joint conditional probability

$$P(Y \mid \text{Gender, Position, Country}) \tag{1}$$

is fundamentally different from the sum of its isolated parts. Our predictive models must natively learn these multi-dimensional categorical interactions [1] (e.g., the joint probability of a specific Age $\times$ Gender $\times$ Academic Position) rather than just assigning isolated linear weights to individual columns.

1.2 Challenges: Severe Class Imbalance and the Accuracy Paradox

1.2.1 The Mathematical "Hostility" of Global Accuracy

The behavioral targets in this study exhibit a distribution where the "No" class (negative) vastly outnumbers the "Yes" class (positive). Based on our survey data ($N = 2,006$),

specific behaviors exhibit severe imbalance; for instance, critical incidents such as **"Cancelled or threatened to cancel visa"** and **"Cancelled or threatened to cancel current position"** were present in only **8.9%** and **11.0%** of cases, respectively. This means the majority ("No") class represents roughly **89–91%** of the total population for these critical metrics. In sociological survey data, it is common to observe baseline frequencies where a specific behavior is only present in **5% to 8%** of the total population.

Standard machine learning models—such as standard Logistic Regression or basic Decision Trees—are typically programmed to minimize a **Global Error Rate**. This objective function creates a significant mathematical bias toward the majority class, as the model seeks the path of least resistance to high accuracy. Consider a behavior with a **5% baseline frequency**:

- **The Naive Path:** If a model simply predicts "No" for every single participant, it will be correct **95%** of the time.
- **The Failure:** While a **95% accuracy** score appears statistically impressive, the model's **Recall** (the ability to actually identify the individuals performing the behavior) is **0%**. It has failed its primary scientific purpose: identifying the demographic drivers of the rare "Yes" response.
- **The Connection:** In this "hostile" environment, the model views the 5% "Yes" instances as "statistical noise" or "outliers" to be ignored in favor of a simpler, high-accuracy "No" rule. This phenomenon is known as the **Accuracy Paradox**.

To force the model to prioritize these rare signals, the chosen architecture moves away from global accuracy and utilizes **cost-sensitive learning** and **re-sampling** strategies [2]:

1. **Random Under-Sampling (RUS) in the 10-Behavior Set:** The algorithm utilizes **RUS** to balance the training distribution. If the data contains 950 "No" responses and 50 "Yes" responses, the model identifies the 50 "Yes" cases and then randomly selects only 50 "No" cases for a specific training iteration. By presenting the model with a **50/50 balanced view**, we strip away the majority's mathematical advantage and force the Gini Impurity calculation to focus on features that distinguish the two classes equally.

2. **Exponential Weight Updates:** We employ a boosting mechanism where misclassified "Yes" instances (False Negatives) are assigned a significantly higher **mathematical weight** (D) for subsequent iterations:

$$D_{t+1}(i) \propto D_t(i) \cdot \exp(\alpha_t) \tag{2}$$

   where α_t represents the penalty factor. This ensures that the rare demographic intersections—such as a specific minority ethnicity in a specific academic position—that trigger the "Yes" response are magnified until the model can no longer ignore them.

3. **Bipolar Margin Penalty in the 15-Behavior Set:** For the SVM classifiers, the C **parameter** acts as a misclassification penalty. In this case, C is tuned to be high enough that the cost of missing a "Yes" (+1) is far more "expensive" to the model's objective function than the cost of a wider margin. This forces the geometric boundary to pivot toward the rare +1 points, even at the expense of slight decreases in aggregate accuracy.

1.3 High Inter-Label Correlation

Behaviors are non-mutually exclusive and highly correlated. A sub-population exhibiting Behavior A may have a statistically significant predisposition to also exhibit Behavior B. If we model each behavior independently without capturing the conditional dependencies $P(Y_B \mid Y_A, X)$, we discard vital predictive power.

Depending on target dimensionality, the data is split into two distinct analytical frameworks:

- **15-Behavior Set:** 15 independent behaviors, each containing 5 ordinal/nominal classes.
- **10-Behavior Set:** A refined subset of 10 concurrent binary (No/Yes) behaviors exhibiting severe imbalance and strong pairwise correlations.

To account for behavioral correlation, the methodology diverges based on target dimensionality: for the **15-behavior set**, behaviors are modeled independently using a **One-vs-All (OVA) SVM** framework, which captures correlations implicitly through shared demographic support vectors in a high-dimensional feature space. Conversely, for the **10-behavior set**, correlations are explicitly internalized using an **Ensemble of Classifier Chains (ECC)** [3]. The details of these methods are provided next.

2 Algorithmic Methodology and Specific Case Justifications

Standard architectures, such as Deep Neural Networks, are systematically ill-suited for this specific dataset. Deep learning struggles with highly sparse, one-hot encoded categorical data and lacks the transparent explainability required to extract distinct demographic rules. Therefore, we deploy geometrically and tree-based architectures tailored to the specific dimensionality of the target spaces.

2.1 Methodology for 15 Behaviors: OVA decomposition with Support Vector Machines

The 15-behavior set represents a high-dimensional target space where each behavior consists of 5 distinct, mutually exclusive classes. Due to the high number of possible class combinations across all 15 behaviors, sequential dependency modeling (chaining) is computationally prohibitive. Therefore, we utilize an independent OVA decomposition strategy using **Support Vector Machines (SVM)** [4, 5].

2.1.1 Bipolar Target Mapping and Mathematical Symmetry

For each individual behavior j, we decompose the 5-class problem into 5 separate binary classification tasks. For each class $c \in \{1, \ldots, 5\}$, the original target $y_{i,j}$ is transformed into a bipolar space $y_{i,j}^{(c)} \in \{+1, -1\}$:

$$y_{i,j}^{(c)} = \begin{cases} +1 & \text{if } y_{i,j} = c \\ -1 & \text{if } y_{i,j} \neq c \end{cases} \tag{3}$$

Unlike 0/1 encoding, the $+1/-1$ bipolar mapping used in our MATLAB implementation ensures mathematical symmetry around the decision boundary. This is critical for the Hinge Loss function, as it treats the detection of a specific behavior class ($+1$) and the aggregate rejection of all other classes (-1) with equal geometric weight.

The model learns a decision hyperplane defined by $f_j^{(c)}(x) = w^T x + b$. We strictly employ a **Linear Kernel**. In this specific case study, our feature space X is already high-dimensional and sparse due to the one-hot encoding of demographic categories.

According to Cover's Theorem, a complex pattern-classification problem is more likely to be linearly separable when cast into a high-dimensional space. Since our demographic intersections already create a sparse, high-dimensional representation, using non-linear kernels (like RBF) would introduce unnecessary complexity and lead to overfitting on rare demographic anomalies.

2.1.2 Multi-Class Inference (Max-Margin Decision)

During the testing phase, for a new demographic profile x_{test}, all 5 binary classifiers are evaluated. The final class assignment follows the *argmax* rule, selecting the class whose classifier yields the highest confidence score (the greatest distance from the margin):

$$\hat{y}_j = \arg \max_{c \in \{1..5\}} (w_c^T x_{test} + b_c) \tag{4}$$

2.2 Methodology for 10 Behaviors: ECC with RUSBoost

For the 10-behavior set, targets are binary $y_j \in \{1, 2\}$ (No/Yes). We explicitly tackle the "No" imbalance and the inter-label correlations sequentially.

2.2.1 Encoding Correlation via Classifier Chains

For a random sequence of behaviors $L = \{l_1, l_2, \ldots, l_{10}\}$, the model predicting the j-th behavior dynamically concatenates the demographic features X with the predicted outcomes of all preceding behaviors in the chain [6]:

$$Z_j = [X, \hat{y}_{l_1}, \hat{y}_{l_2}, \ldots, \hat{y}_{l_{j-1}}] \tag{5}$$

This mathematically forces the model to evaluate the conditional probabilities of concurrent behaviors. To prevent sequential error propagation, we utilize an **Ensemble** of $M = 5$ separate chains with randomized internal orders.

2.2.2 Combating Imbalance via RUSBoost Mechanics

The base learners at each node are Classification and Regression Trees (CART). The trees calculate splits by minimizing the Gini Impurity G for a given node m with class probabilities p_k:

$$G(m) = 1 - \sum_{k=1}^{K} (p_k)^2 \tag{6}$$

To prevent the massive "No" majority from dominating these splits, the trees are wrapped in a **RUSBoost** (Random Under-Sampling Boosting) architecture [7].

In each boosting iteration t:

1. **Targeted Undersampling:** The algorithm isolates a balanced subset S_t' by randomly dropping majority "No" instances until the classes are equal.

2. **Weak Learner Training:** A shallow decision tree $h_t(z)$ is trained on this balanced subset.

3. **Global Exponential Weight Update:** Observation weights D are updated via AdaBoost logic:
$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_{i,j} h_t(z_i))}{Z_t} \tag{7}$$
where α_t is the tree's influence weight.

3 Model Evaluation and Sociological Justification

During evaluation, the ECC model for the 10-behavior set achieved an approximate global accuracy of 65% on the training data and 53% on the unseen testing data. While an accuracy of 53% may appear modest in deterministic disciplines, it is highly significant in the social sciences because human decisions are driven by massive unobserved latent variables. Furthermore, a model that identifies a demographic sub-population with a 55% probability of exhibiting a behavior (compared to a 5% baseline) represents an *11-fold increase* over the baseline odds.

4 Analysis Strategy and Rule Extraction

The ultimate goal of this methodology is to extract reliable, human-readable demographic rules.

4.1 Evaluating Correlation Capture

To prove the ECC architecture successfully learned the behavioral dependencies, we compute a 10×10 Cramer's V heatmap on the *true* training labels and compare it against the heatmap of the *predicted* labels. Cramer's V defines the association between two categorical variables, yielding a value between 0 and 1.

4.2 Combinatorial Demographic Extraction

Finally, we conduct an exhaustive extraction of predicted probabilities across 1-way to 5-way demographic combinations. We enforce two strict empirical guardrails:

- **Minimum Support ($N \geq 100$):** A demographic combination is only analyzed if ≥ 100 individuals match that exact profile.
- **Probability Threshold ($P \geq 0.50$):** The ensemble must predict a $\geq 50\%$ absolute probability that this specific group will exhibit a "Yes" response.

References

[1] Foulds, J. R., & Pan, S. (2020). An intersectional definition of fairness. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*.

[2] Zhao, Y., Lin, W., Jiang, Y., Luo, X., & Liu, X. (2024). Re-sampling strategies for imbalanced multi-label learning: A comprehensive review. *Information Sciences*, 652, 119742.

[3] Moyano, J. M., Gibaja, E. L., Ventura, S., & Cano, A. (2020). Speeding up classifier chains in multi-label classification. *In Proceedings of the 4th International Conference on Internet of Things, Big Data and Security*, 2019.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

[5] Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5, 101-141.

[6] Read, J., Pfahringer, B., Holmes, G., Frank, E. (2009). Classifier Chains for Multi-label Classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science(), vol 5782. Springer, Berlin, Heidelberg.

[7] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1), 185-197.