Digital Footprints of Substance Use: Decoding Offline Behaviors from Writing Patterns Online Using Machine Learning

Raad Bin Tareaf, Alex Maximilian Korga, Sebastian Wefers, Keno Hanken
XU Exponential University of Applied Sciences
August-Bebel-Str. 26-53, 14482 Potsdam, Germany
Email: r.bintareaf@xu-university.de, {a.korga, s.wefers, k.hanken}@student.xu-university.de

***Abstract*—This study explores the use of Linguistic Inquiry and Word Count (LIWC) features from the myPersonality dataset to model alcohol, cigarette, and drug consumption behaviors. We employ machine learning classifiers—Logistic Regression, Support Vector Machines, Random Forest, XGBoost, and LightGBM—and benchmark their performance against a Baseline Classifier using accuracy, precision, recall, and f1-score. To address the class imbalance present in our target variables for smoking and drug consumption, we purposefully employ macro-averaged recall to ensure equal emphasis on both classes within these binary targets. Conversely, for the alcohol consumption target, which exhibits a more balanced distribution, we optimize our models using balanced accuracy. This dual approach enables precise model tuning, prioritizing the discovery and interpretation of patterns for all outcomes with equal significance. Furthermore, our experiments demonstrate that strategically combining oversampling of the minority class with undersampling of the majority class significantly enhances predictive model performance, effectively addressing the challenge posed by imbalanced target classes in our dataset. Our analysis reveals distinct LIWC categories correlating with substance use behaviors in writing patterns and expressions. Logistic Regression emerges as the top performer for smoking habits and drug use, achieving overall macro-averaged recall rates of approximately 66% and 65% respectively across varying train-test splits. Conversely, XGBoost exhibits the highest overall balanced accuracy of approximately 66% for predicting alcohol consumption. Our findings underscore the complexity of predicting consumption behavior from textual data and highlight the potential of machine learning in understanding and modeling intricate human behavioral patterns, contributing to computational linguistics, psycholinguistics, and personality psychology.** **Keywords: Big Five Traits, Personality Psychology, Computational Modeling, Psycholinguistics**

I. Introduction

The profiling discipline is constantly evolving. Since the introduction of electronic systems, that allow the utilization of more and more complex analytical methods for human characteristics, different profiling mechanisms have been used in a variety of use cases, from low-level ones like criminal investigation, where an individual is profiled using i.e., observations made at the crime scene, down to high-level profiling like simple customer analytics through revenue numbers per category. In our research, we will introduce the discipline of *natural language processing* (NLP) and the currently available and used tools throughout this paper. The development of NLP-supported methods for profiling or predicting certain behavioral patterns originates from 1969, when written content about a particular topic was used to predict mental abnormalities such as anxiety, hostility or social alienation with certain probabilities. Later developments with computer-aided capabilities allowed James W. Pennebaker and his team to shift from content-centric analytics towards cognitive and emotional analytics, making recent psychoanalytical processes and NLP-centric diagnostic processes less dependent on what an individual is writing. This framework is known as the LWIC framework and is discussed in section II. In the times of rising, and meanwhile sophisticated, digital social platforms like Facebook and X (former Twitter), a unique possibility has opened up for researchers and businesses to observe and research human social interaction among others and individual social behavior through, i.e., posts, likes, shares and many more actions. In 2007, David Stillwell created an application for Facebook users, which allowed a user through the answer of questions and access to their Facebook data and history to gain insights about their psychological characteristics. Participants in this study could donate their data for research purposes, resulting in the creation of the myPersonality dataset [1]. As a contribution to the dataset, there was a survey conducted in 2010 by David J. Stillwell and Richard J. Tunney, which analyzed the short- and long-term reward thinking of smokers and non-smokers [2]. With the already-present data and the new data added, this combination allowed us to analyze possible patterns and make predictions about drug, alcohol and cigarettes substance use just through textual information. Due to heavy class imbalances and a non-parametric form of the dataset, we utilized methods like alpha-trim and aggregation methods for data preprocessing, the Mann-Whitney-U-test for feature selection, feature normalization, and the Adaptive Synthetic Sampling Approach for dealing with imbalances, which are all described in Section III. Our research and experiments on the development of a statistical model for substance use are being supported by machine learning classifier algorithms, such as L1-Regularized Logistic Regression, Support Vector Machine, Random Forest, XGBoost and LightGBM, which were all trained individually for one of the three present labels. The setup of these algorithms is described in Section III-D, as are the training and testing results in Section IV. The rest of the paper contains the following content: The results Section IV will also show a detailed comparison of different AI models and metrics, as well as individual feature importance. A critical comparison between our findings and possible use-cases, as well as further research opportunities and ideas, are being presented in the discussion and future work section V. The conclusion in Section VI will provide a final summary of our conducted research.

II. Literature Review

As our hypothesis poses a novelty within the existing BIG 5 dataset and research landscape the idea of predicting a certain behavior, as the consumption of certain substances like alcohol, drugs and cigarettes, the concept of utilizing NLP in order to predict certain (social) behavior patterns or i.e. calculating likelihoods of an job applicant getting accepted into a new position, poses a common technique to draw conclusion about an individual's psychological profile. The foundation of the creation of the so called *LWIC-features* reaches back to 1992, where James W. Pennebaker introduced an approach to use written text by individuals in order to analyze their state of mental health and learn about the corresponding effects of traumatic experiences. Patients of a mental facility were asked to continuously write about 15 - 20 minutes of text on 3 - 5 consecutive days. The topics to write a text about changed day wise but uniquely ranged between writing about the most traumatic experiences a study participant has ever experienced, to writing about the experience to enter college. Due to the nature of a qualitative approach of data acquisition, the questionnaires way of answering the posed question varied from person to person, meaning answers could be expressed in the following ways, as Pennebaker wrote in his research paper: *"In one case, two students wrote about problems they were having with their roommate at college: one simply noted that her roommate was a bitch and listed all of her roommate's many faults; another person tried to analyze the conflict with her roommate in a deep, self-reflective manner."* Due to this issue in dealing with different expressed answers, there was an urge to advance from the commonly used *primary content dimension* in order to generate useful and analyzable data, leading to the development of the so-called *primary cognitive* and *primary emotional* dimensions. The primary content dimension was developed in 1969 and focused on the measurement for different pathological behaviors, such as anxiety, hostility or social alienation, via a scale, also known as the *compartmentalization of measurements*, for the individual psychological abnormalities. The scales worked via counting the with a corresponding behavior affiliated words of an written essay, which had to be written in a certain way in order to utilize this approach of psychoanalysis. The advancements made by Pennebaker and Martha Francis, were that they would not focus directly on mental abnormalities, but more on cognitive aspects, like causation or certainty, and emotional aspects, like positive or negative emotions and their portions of all written words of a given essay. This was achieved by mapping English words to word categories and sentiments. The mapping would map words like loving, hateful, joyful or guilt to their appropriate category, which would be in this case ether the positive or negative sentiment. Other word classes, such as 1st person pronouns, family associated words, and many more were also mapped accordingly to their definition. Today, the LIWC framework is being abstracted and used in other languages as well, such as German, French, Spanish and many more, as well as available software for analyzing natural language. In 2019, Pia M. Brandt and Philip Y. Herzberg utilized LWIC in order to analyze and calculate the probability of an job applicant getting hired based on extracted features from their cover letter or CV. Brandt and Herzberg were able to extract text patterns of (un-)successful applicants but pointed out, that just through LWIC analytics, reliable probability calculations couldn't be created [3].

One dataset that we used in our research was created in 2007 during his PhD studies in psychology of David J. Stillwell. Stillwell has created the Facebook app *myPersonality* that would ask the user certain quantitative questions in order to create a psychological profile with the so-called *big 5 personality traits*. The 5 personality traits have their origin in 1949, where the psychologist Donald W. Fiske would develop an empirical model for a taxonomic system for personality traits until 1989. The 5 personality traits were known as *openness*, *conscientiousness*, *extraversion*, *agreeableness* and *neuroticism* (in short *OCEAN*) and well consensualized among other researchers [4]. The myPersonality Facebook app would take the answers from the conducted survey and return the result to the user in form of how much the user is associated with the OCEAN traits, as well as ask them to share their then psydonumized answers for research. Later, the answers included Facebook data as well, such as likes, posts, events, liked pages, etc. According to Stillwell and his later co-author Michal Kosinski disclosed, that 6 million individuals have taken the myPersonality test until 2012 [1]. The myPersonality dataset was continuously expanded through questions and surveys about things like political orientations, star sign believes, etc. One of these expansions was the *Delay reward discounting* (DRD) dataset, that was created by the myPersonality study founder David J. Stillwell himself and Richard J. Tunney, which we are including in our research work as well. They have run as a study about the measurement of the degree to which a person ether prefers a smaller reward right now, or a greater reward in ether 1 or 2 weeks, 1 or 6 months, 1 or 5 years later under a monetary aspect. The monetary rewards would be compartmentalized into 15 non-symetrical amounts ($1,000, $950, ..., $20, $10). Also during this study, Stillwell and Tunney included the question about the consumption of cigarettes in their survey, which offered a novelty for large scale research about the connection between the individual concept of reward-based thinking and (addictive) consumption of substances [2].

In the past, different research papers have worked on utilizing LWIC features for psychoanalytic profiling. The literature review paper *Recent trends in deep learning based personality detection* by Yash Mehta et al. compared different algorithms and data media types (text, audio, visual) to do personality prediction [5]. Concerning the textual data input, LWIC features were used throughout all text based classifiers for ether the most present OCEAN personality type, or one of the 16 *Myers-Briggs-typeindicator* (MBTI) types [6]. Among the compared machine and deep learning algorithms and their respective performance were, CNNs [7], SVCs [8] and RNNs [9]. In their a direct comparison, the mean accuracy for the big 5 classification models ranged between 57.99% and 63.6%, whilst the only MBTI classifier achieved 67.77% of accuracy.

However, since recent developments, especially in the discipline of large language models, the paper *BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions* by Jacopo Biggiogera et al. proposed a different approach for behavioral coding and utilized Googles' in 2018 developed language model BERT [10], as well as TF-IDF [11], which is a statistical method of measuring the relevance of terms throughout a given text. Biggiogera et al. compared the coding performance of LWIC and BERT on a dataset which consisted out of 368 German-speaking couples and an 8-minute text capture conflict interaction, divided into 10 seconds long sequences for each conversation. The end goal was the classification of ether positive or negative coded interaction within the 10 second sequence. The results there show, that for this particular problem, LWIC achived 65.4% of accuracy and was outperformed by the more complex BERT model with 69.4% and slightly by the more simpler TF-IDF method together with another statistical method called n-gram [12] for disassembling and counting words and syllables, with 65.6%.

Concerning the prediction of substance consumption of alcohol for instance: In the year 2021, the paper *A Deep Learning Algorithm to Predict Hazardous Drinkers and the Severity of Alcohol-Related Problems Using K-NHANES* by Suk-Young Kim et al. classified quantitative data, which was gathered through the South Korean national health and nutrition examination survey, into 4 different degrees of hazardous drinkers, where the deep neural network showed the best performance with 0.870, measured as *area under curvature* (AUC) [13]. A similar paper in terms of a direct quantifiable approach but for the prediction of drug consumption was conducted by Peng Han in 2021 with their research paper *The Application of Machine Learning Methods in Drug Consumption Prediction* where they utilized 12 features such as age, education, OCEAN traits, etc. and trained 3 different classifiers each for amphetamine, benzos and cocaine consumption. The trained algorithms were a Logistic Regression, a descision tree and a Random Forest classifier, where the Logistic Regression Classifier performed best on the unseen 30% test split with an AUC of 0.754 [14]. A different approach in prediction consumption behavior was done by Maryam Abo-Tabik et al. in their research paper *Towards a Smart Smoking Cessation App: A 1D-CNN Model Predicting Smoking Events* from 2019. Abo-Tabik et al. conceptualized an mobile app that is supposedly able to learn a user's smoking routines based on the devices acceleration sensor data and GPS location information. From the 3 compared machine and deep learning algorithms (CNN, SVM and Decision Tree), the CNN classifier performed best with an overall accuracy of 87% [15].

III. Methodology & Implementation

A. Data

This paper utilizes the *Delay Discounting* dataset, consisting of about 15000 samples, provided by [2] and is part of the psychological profiles in the MyPersonality Dataset. While the datasets focus is the topic of delay discounting, our work mainly focuses on the provided background questions on smoking, alcohol and drug-taking behaviour. As the part of the background questions featuring substance use was optional, depending on each specific category, a significant amount of questions were left unanswered, which results in the distribution shown in Figure 1 with 1185 remaining samples.

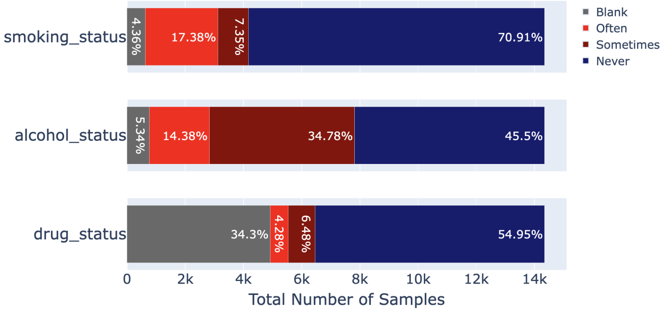


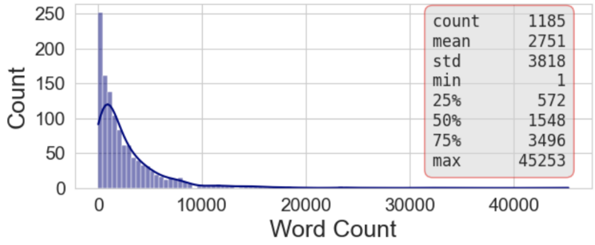
Fig. 1: Initial Class Distributions

1) *Data Preprocessing*

In addressing the challenges of class imbalance and non-responsive *blank* data in our dataset, we binarized the target variables for consumption behavior. This approach simplified our analysis by merging the *often* and *sometimes* responses into a *True* category, indicative of any consumption, and classifying *never* responses as *False* for non-consumption. This strategy not only eliminated non-informative left out responses but also mitigated class imbalances, especially improving the balance for the target variable*alcohol status*. The binarization facilitated a streamlined, binary classification framework, enhancing the interpretability and comparative analysis across the smoking, alcohol, and drug use variables, as depicted in Figure 3.

To ensure the integrity and robustness of our linguistic analysis, we evaluated the dataset's word count distribution, which ranged from 1 to 45,253 words, with an average of 2,751 and a standard deviation of 3,818 as displayed in Figure 2. This indicates a diverse array of text lengths and expressions among the participants. To manage this variability and maintain a high-quality dataset, we set a threshold of 50 words, excluding shorter entries while preserving texts rich enough for reliable linguistic profiling. This step ensured a sufficient volume of data for a robust analysis.

Observing that the 25th percentile of word counts was 572, we were confident that our threshold would retain a significant portion of meaningful entries. This cutoff was designed to omit texts lacking in substantial linguistic content, thus enhancing the dataset for LIWC analysis. Our approach balanced the representation of text lengths, laying a solid foundation for investigating substance use's digital footprints and ensuring accurate analysis.



```
count     1185
mean      2751
std       3818
min          1
25%        572
50%       1548
75%       3496
max      45253
```

Fig. 2: Word Count Distribution

2) *Handling Class Imbalance*

Furthermore, the fact that all target variables except for alcohol consumption exhibited a severe class imbalance, as illustrated in figure two, the combination of oversampling the minority class and downsampling the majority class had been applied.

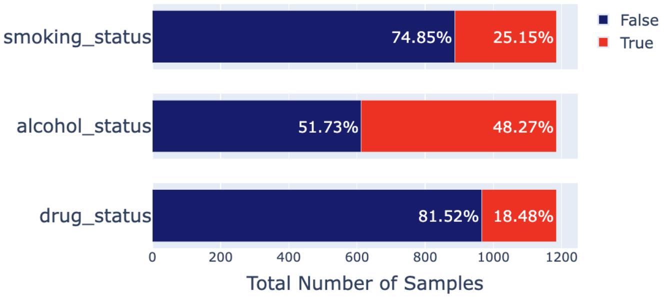


Fig. 3: Target Variable Ratios

By judiciously leveraging both undersampling and oversampling, we ensured the preservation of a significant number of realistic samples without overproducing synthetic data, thus achieving the ideal trade-off. This is critical to ensure that the machine learning models are able to discern patterns, that are not biased towards the majority class.

To effectively counteract this imbalance, we devised comprehensive pipelines that combined both traditional and innovative techniques to ensure an unbiased class representation. The most promising data pipeline for modelling the imbalanced targets, as shown in Figure 4, commenced with Tomek Links [16] for undersampling, effectively eliminating overlapping samples between classes to clarify the decision boundaries. The application of Tomeklinks was investigated in detail in a study of Swana et al. [17]. This critical step reduced majority class dominance, thereby increasing the model's sensitivity to minority classes. Random undersampling was then employed to further balance the distribution. Subsequently, the standard scaler [18] was applied to normalize all features, achieving zero mean and unit variance, thus effectively synthesizing samples for the minority class. The Adaptive Synthetic Sampling Approach (ADASYN [19]) was introduced to strategically generate artificial samples in regions where the classifier's learning was most challenged. This approach not only augmented the minority class representation but also focused on those areas that are crucial for improving classification performance. Masruriyah et al. [20] examined the effectiveness of ADASYN against SMOTE in their study.

It's imperative to note that these sampling techniques were applied exclusively to the training dataset to prevent information leakage and ensure the test set remained an unbiased evaluation ground.

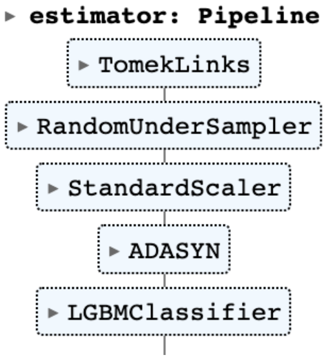


Fig. 4: Example Pipeline to Address Target Imbalance

B. Feature Selection

In our approach to reduce the feature space, we opted for the Mann-Whitney-U-test [21] due to several considerations:

- **Consistency across all models and targets:** This ensures uniformity in feature selection, regardless of the specific model or target variable.
- **Non-assumption of normality and homoscedasticity:** The Mann-Whitney-U-test does not require the data to follow a normal distribution (normality) or to have equal variances across groups (homoscedasticity), which are prerequisites for classical ANOVA. This makes it suitable for our data, since these assumptions are violated.
- **Highlighting discriminatory features:** By leveraging the Mann-Whitney-U-test, we focus on pinpointing those features that starkly separate the two groups. This approach not only highlights the most influential factors in writing styles but also enriches our understanding of what distinctly characterizes each class, aligning closely with our objective to decode the digital expressions related to substance use.
- **Informative features:** By selecting only those features that present discriminative power and are the most informative for the models, we can distinguish the binary targets and counteract the model's tendency to overfit to noisy patterns in the training data.

For our study, we set an alpha significance level at 0.05 to filter out the most informative features, aiming to concentrate on those that significantly distinguish between classes. Additionally, an alpha accumulation correction was considered to address the problem of multiple testing for the 93 linguistic features.

In the case of the balanced alcohol consumption target, we employed the Benjamini-Hochberg procedure [22] to mitigate the false discovery rate (FDR), enhancing the robustness of our findings related to alcohol-related linguistic patterns.

Conversely, for the imbalanced targets, smoking habits and drug use, we opted against the Benjamini-Hochberg correction. This decision was influenced by the challenges posed by their imbalanced nature, which complicated the identification of discriminative features and resulted in the loss of significant amount of insightful attributes. During our experimental phase, the application of the alpha correction resulted in the retention of almost all variables for the balanced alcohol target, while the smoking and drug use targets saw a reduction of 10 and 8 features, respectively.

Moreover, FDR control methods like Benjamini-Hochberg tend to be conservative, especially in imbalanced scenarios, risking the exclusion of important features that, despite not meeting strict FDR thresholds, could provide meaningful insights into smoking and drug use behaviors. We decided not to employ alternative methods for addressing the alpha accumulation issue in these imbalanced targets to avoid introducing additional complexity and potential biases into our analysis. Crucially, this choice helped retain important feature interactions, essential for uncovering the subtle linguistic patterns tied to substance use. By prioritizing a broad yet detailed examination of linguistic cues, our approach ensured a nuanced analysis without being hampered by overly strict statistical adjustments.

The selected features for each target can be seen in Figure 5.

More information about LIWC and detailed descriptions about each variable can be found on the myPersonality Project website under list of variables [23].

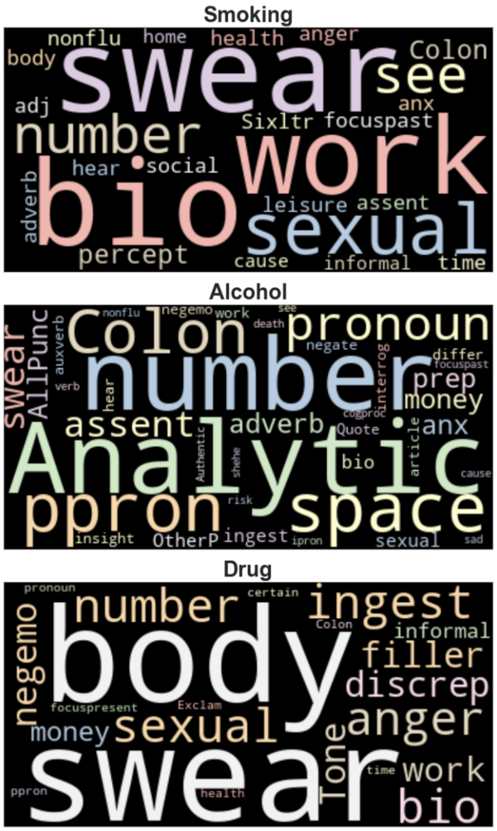


Fig. 5: Feature Spaces per Target Variable

C. Principle Component Analysis

To assess the separability of the three target variables in the dataset, we conducted the pairwise controlled manifold approximation projection (PaCMAP [24] [25]). This technique, applied to the selected feature space, aimed to provide an initial insight into the distinctness of each dependent variable within a reduced three-dimensional space.

The PaCMAP analysis, as shown in Figure6), indicated the lack of a clear decision boundaries among all targets, suggesting that defining a separable decision boundary in the high-dimensional feature space—determined by features selected through the Mann–Whitney-U-test presents a significant challenge for the chosen algorithms. This demonstrates the difficulty of finding meaningful patterns in the data to accurately predict the dependent variables, even with the most separable independent variables.

D. Machine Learning Models

In this study, we aim to model the interactions between variables within each feature space as they relate to consumption habits. We selected five candidate models, including Logistic Regression, Support Vector Machines, LightGBM, and XGBoost, based on their proven efficacy in handling complex datasets and their ability to model non-linear relationships. These models were compared against sklearn's dummy classifier, which serves as a simple baseline to establish the minimum performance threshold. The purpose of this comparison is to determine if our carefully tuned models can significantly outperform this basic benchmark, thereby demonstrating their value and effectiveness beyond what could be achieved with a naive approach to classification.

We subjected each model to rigorous hyperparameter tuning, combined with cross-validation to ensure robustness and generalizability. Here the data was split in a startified fashion with respect to the binary target ensuring that each training fold included the same ratio as the test folds.

It is imperative to note that during this tuning phase the resampling techniques had only been applied to the training folds, hence the test set remained unaltered as a realistic evaluation ground.

Specifically, for the balanced target variable of alcohol consumption, we leveraged the Optuna framework [26], utilizing Bayesian optimization for a more intelligent hyperparameter space search. In this context, we chose balanced accuracy as our optimization metric for the balanced target variable, to ensure fair performance across classes by treating them with equal importance, thus eliminating bias in our balanced dataset. This metric is mathematically denoted as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$$

where TP and TN represent correctly predicted positive and negative instances, respectively, while FP and FN denote incorrect positive and negative predictions. This metric equally weighs the model's ability to identify all classes, ensuring a balanced evaluation.

For the imbalanced targets of smoking habits and drug abuse, we found that exhaustively searching through a narrower parameter space with k-fold cross-validation was not only more efficient but also crucial in mitigating overfitting, a common challenge in imbalanced datasets. To ensure our models accurately identify relevant patterns across all classes without bias, we optimized for macro average recall. This metric is essential for datasets with imbalanced classes, as it averages the recall scores across all classes, thus treating every class with equal importance:

$$\text{Macro Average Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$$

where N is the number of classes, TP_i the true positives, and FN_i the false negatives for each class i. This approach

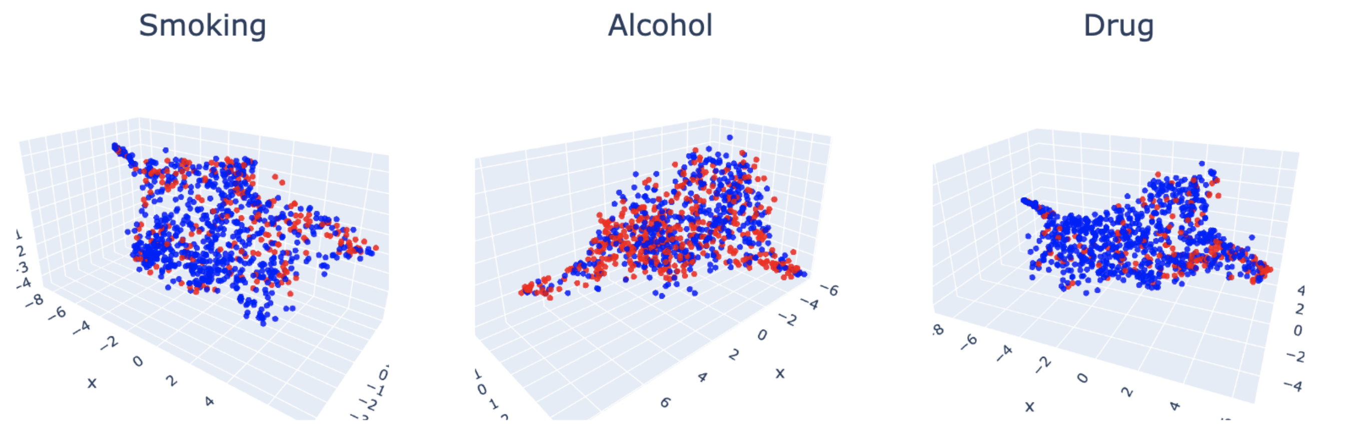


Fig. 6: Principle Component Analysis

ensures that our models are effectively tuned to recognize key patterns for both the less frequent (positive) and more frequent (negative) outcomes, which is vital for our subsequent analysis of variable importance in relation to smoking habits and drug abuse.

After discussing our model selection and optimization strategies, we provide a brief overview and key formulas for each model used in our study, highlighting their foundational aspects relevant to our research:

- **Dummy Classifier (Base)**: The Dummy Classifier is a simplistic model that serves as a benchmark for comparison with more sophisticated algorithms. It generates predictions based on simple rules that are independent of the input features. For the "stratify" strategy, the prediction function respects the training set's class distribution, making random predictions with the probability of each class equal to its relative frequency in the training data. The prediction function for class c_i can be expressed as:

  $$P(c_i) = \frac{N_{c_i}}{N}$$

  where $P(c_i)$ is the probability of predicting class c_i, N_{c_i} is the number of samples of class c_i in the training set, and N is the total number of samples in the training set. This strategy ensures that the class proportions in the classifier's predictions match those in the training data.

- **L1-Regularized Logistic Regression (LR)**: L1-Regularized Logistic Regression adds an L1 penalty to the Logistic Regressioncost function, encouraging sparsity in the model parameters and thus mitigating the risk of fitting to noisy patterns in the data. The optimization objective is as follows:

  $$\begin{aligned}\hat{y} = -\frac{1}{m}\sum_{i=1}^{m}&\left[y^{(i)}\log(h_\theta(x^{(i)}))\right.\\ &\left.+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]\\ &+\lambda\sum_{j=1}^{n}|\theta_j|\end{aligned}$$

  where $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ is the logistic function, θ represents the parameter vector, m is the number of training examples, n is the number of features, $y^{(i)}$ is the actual class label of the i-th example, $x^{(i)}$ is the feature vector of the i-th example, and λ is the regularization strength.

- **Support Vector Machine (SVC)**: Support vector classifiers work by finding the best fitting hyperplane that separate distinct instances from one another in the defined feature space. It is described as:

  $$\hat{y} = \text{sign}\left(\sum_{i=1}^{n} y_i \cdot \alpha_i \cdot K(x_i, x) + b\right)$$

  where α_i are the Lagrange multipliers, y_i are the class labels, $K(x_i, x)$ is the kernel function, b is the bias term, and sign function returns the class label based on the sign of the argument.

- **Random Forest (RF**: Random Forest is an ensemble learning technique that builds multiple decision trees and merges their predictions to improve classification accuracy. In classification tasks, the model's prediction is determined by taking the mode of the predictions from all the individual trees. The formula for classification is given as:

$$\hat{y} = \text{mode}\left(\{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}\right)$$

where $\hat{y}$ represents the predicted class label, and $y^{(i)}$ is the prediction of the i-th decision tree in the Random Forest ensemble. The model selects the class label that occurs most frequently among the individual tree predictions.

- **XGBoost (XGB)**: XGBoost (Extreme Gradient Boosting) is a high-performance gradient boosting framework that builds trees level-wise, where subsequent trees aim to to reduce the errors of its predecessors. It is known for its efficiency and scalability, incorporating features like sparsity-aware splitting and regularization to prevent overfitting. The model's prediction y is the sum of the predictions from K decision trees:

  $$\hat{y} = \sum_{k=1}^{K} f_k(x)$$

  where f_k represents the k-th decision tree, and K is the total number of trees.

- **LightGBM (LGBM)**: LightGBM optimizes gradient boosting by constructing trees leaf-wise, allowing for deeper, more effective tree structures. The ensemble reduces computational cost with a binning technique that discretizes continuous variables. Its prediction y is also formulated as the sum of the outcomes from K boosting iterations:

  $$\hat{y} = \sum_{k=1}^{K} g_k(x)$$

  where g_k denotes the k-th boosted tree, and K signifies the number of boosting rounds.

IV. Results and Evaluation

In this section, we present a comprehensive and comparative analysis of model performance for each target variable, focusing on key classification metrics such as accuracy, precision, recall, F1-score, precision macro average, and recall macro average. These metrics transparently show the performance of each model on unseen data with optimal hyperparameter settings.

Additionally, we examine the consistency of model performance across 100 distinct train-test splits to assess the robustness and reliability of our models. This analysis is designed to validate the stability of our models under varying data conditions, ensuring robustness of our undertaken study. Here, we report the overall macro average recall, which represents the aggregated performance of all models across varying data splits. This metric is calculated by first determining the macro average recall for each model within each data split, and then averaging these scores to provide a comprehensive measure of performance across our evaluations

Furthermore, we provide confusion matrices showing the classification probabilities, allowing us to visualize not just the successes (true positives and true negatives) but also the specific types of errors each model makes, such as false positives and false negatives. This detailed examination aids in understanding the nuances of model behavior in various scenarios, offering deeper insights into their predictive capabilities and limitations. Note, that the decision threshold of 0.5 remained unaltered for all candidates to ensure that both class labels were treated with equal importance.

Last but not least, we delve into the interpretability of the best-performing model for each target variable by presenting feature importance plots derived from global SHAP (SHapley Additive exPlanations) values [27]. These plots highlight the top 14 variables that contribute most significantly to the models' predictive capabilities, offering insights into the underlying patterns and factors driving the predictions. This analysis not only enhances our understanding of the models' decision-making processes but also identifies key variables that may inform further research and practical applications in understanding and influencing consumption habits.

A. Smoking Habits

For the target variable Smoking Habits, we observed that all models showed a considerable increase in prediction performance compared to our base dummy classifier (tab. **??**) for a consistent random state. Random Forest managed to outperform all other models in accuracy and macro-averaged precision and recall. Notably, the Logistic Regression and the Support Vector Classifier both had a better Recall for the target variable which was more difficult to predict (Class 1), while in macro-average being slightly weaker compared to Random Forest.

Analyzing the models across diverse train-test splits in Figure 7 reveals a markedly higher overall macro average recall for all tuned models compared to the baseline classifier. This significant improvement underscores the tuned models' effectiveness in uncovering meaningful patterns within the data, far surpassing the baseline, which performs no better than random chance, akin to the toss of a coin. Upon closer examination of the tuned models, the SVC stands out as the top performer in various data split scenarios, characterized

Metric	Base	LR	SVC	RF	XGB	LGBM
Accuracy	0.62	0.66	0.66	**0.70**	0.68	0.68
Precision (Class 0)	0.74	**0.86**	0.85	0.85	0.84	0.85
Recall (Class 0)	**0.76**	0.65	0.66	0.72	0.69	0.70
F1-Score (Class 0)	0.75	0.74	0.74	**0.78**	0.76	0.77
Precision (Class 1)	0.24	0.40	0.40	**0.44**	0.40	0.42
Recall (Class 1)	0.22	**0.69**	0.67	0.64	0.62	0.64
F1-Score (Class 1)	0.23	0.50	0.50	**0.52**	0.49	0.51
Precision Macro Avg	0.49	0.63	0.63	**0.65**	0.63	0.64
Recall Macro Avg	0.49	0.67	0.67	**0.68**	0.66	0.67

TABLE I: Classification Metrics Summary for Smoking Habits

not only by the highest overall macro average recall of 0.654 but also by the smallest performance variability, with a standard deviation of ±0.041. It is particularly striking that the more complex models exhibited higher variability in their performance, as evidenced by the standard deviations of their macro average recall scores: Random Forest (RF) at ±0.05, XGBoost at ±0.049, and notably, LightGBM at ±0.639. In contrast, the simpler models, specifically the Support Vector Classifier (SVC) with a standard deviation of ±0.041 and Logistic Regression (LogReg) at ±0.046, demonstrated greater robustness and efficiency across different data splits. These metrics underscore the stability and reliability of SVC and LogReg in handling data variability, contrasting with their more complex comrade-in-arms.

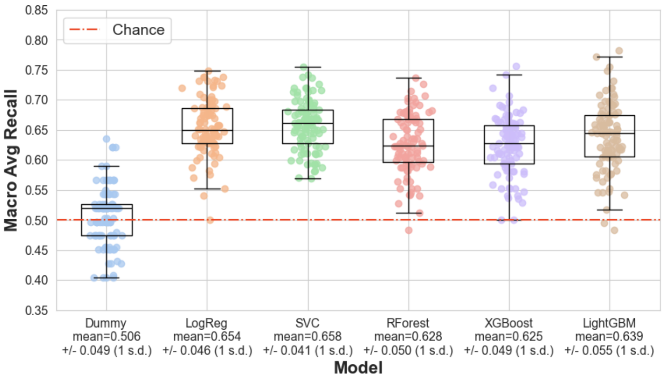


Fig. 7: Model Performances for Smoking Habits Across Varying K-Fold Splits

Having established SVC as the standout model for its consistent macro average recall and stability, as shown in Figure 7, we next delve into its detailed performance through a confusion matrix. This matrix not only reveal SVC's predictive accuracy and errors but also include probability scores, enriching our understanding of its confidence in various predictions. The analysis of the combined distributions of true negatives and false positives around the 0.5 decision threshold highlights a notable peak. This observation suggests a pivotal point in the model's classification process, where the differentiation between classes becomes less distinct. This convergence aligns with insights from the principal component analysis, which previously highlighted the lack of clear decision boundaries within the feature space. The very same interpretation aligns for the true positives and false negatives. Ultimately, the analysis reveals that true negatives (112) significantly exceed their corresponding false positives (58), indicating strong model specificity. Similarly, true positives (39) surpass the number of false negatives (19), demonstrating the model's sensitivity. However, it's noteworthy that in certain scenarios, the challenge of false positives becomes pronounced, at times even surpassing the true positives, pinpointing a critical area for enhancing the model's precision. Adjusting the decision boundary could mitigate classification errors, but such shifts require careful consideration to maintain a balance between reducing false positives and avoiding an increase in false negatives. The optimal adjustment of this boundary should be informed by the specific context of the application, where the relative costs of false positives versus false negatives are thoroughly evaluated to ensure the most beneficial trade-off for the model's intended use.

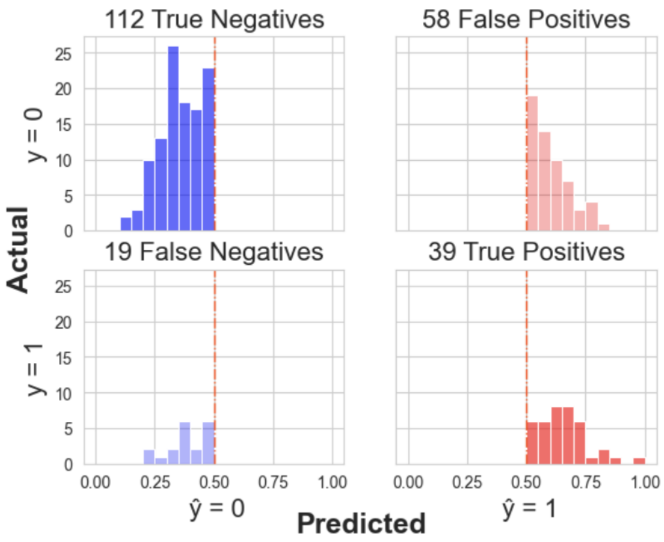


Fig. 8: Confusion Matrix Probabilities for Smoking Habits

The SHAP analysis in Figure 9 unveils linguistic predictors of smoking habits from Facebook writing styles. Swear words, numerical references, and work-related language are prominent, suggesting a correlation between profanity, number-centric discussions, and mentions of professional life with smoking behaviors. Emotional expressions linked to anger, structural communication elements like colons, and terms associated with visual perception and time also play significant roles, indicating that emotional tone, communication structure, and discussions of time and sight may reflect smoking tendencies.

Words related to social interactions, biological concepts, sensory perceptions, and anxiety reveal a connection to smoking habits, emphasizing how language reflects social behaviors, physical sensations, sensory experiences, and mental states. The unique focus on six-letter words indicates that specific patterns in word use may signal smoking tendencies.

In conclusion, this interpretation of the SHAP analysis provides nuanced insights into how various linguistic elements used in Facebook posts can serve as indicators of tobacco use, offering valuable perspectives for subsequent research and potential interventions focused on smoking behavior.

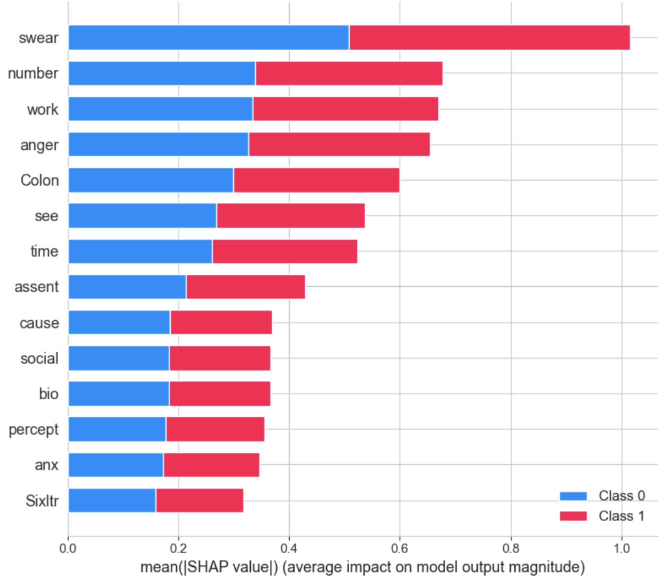


Fig. 9: Feature Importance for Smoking Habits

B. Alcohol Consumption

For the target variable Alcohol Consumption, we also observed that all models showed a considerable increase in prediction performance compared to our base dummy classifier (tab. **??**). The XGB Classifier managed to outperform all other models in all categories, reaching a accuracy and macro-averaged precision and recall of 0.69. All models have similar results between both classes, showcasing the importance of having a good class balance.

In Figure 10, when interpreting the model performances across varying k-fold splits, again the fine tuned models do outperform the base line classifier. Similar to smoking habits, the simpler models are dominating the performance rankings regarding alcohol consumption for both in terms of macro average recall and standard deviation. Here Logistic Regressionachieved a overall macro average recall of 0.654 and a standard deviation of ±0.037, followed by the Support Vector Machine with an overall macro average recall of 0.66 and standard deviation of ±0.039. Besides, XGBoost shows the highest maximum macro average recall for one specific data split and the highest median of the macro average recall. Both Random Forest and LightGBM did also perform better than random chance, however with slightly lower mean and median values for macro average recall as well as larger ranges or variability across varying data splits. The dummy classifier does no better than the flip of a coin and also highlights the largest variability across all splits.

Metric	**Base**	**LR**	**SVC**	**RF**	**XGB**	**LGBM**
Accuracy	0.51	0.65	0.66	0.65	**0.69**	0.65
Precision (Class 0)	0.53	0.66	0.68	0.65	**0.70**	0.66
Recall (Class 0)	0.47	0.68	0.64	0.69	**0.70**	0.66
F1-Score (Class 0)	0.50	0.67	0.66	0.67	**0.70**	0.66
Precision (Class 1)	0.50	0.65	0.64	0.65	**0.68**	0.64
Recall (Class 1)	0.56	0.63	**0.68**	0.60	**0.68**	0.65
F1-Score (Class 1)	0.53	0.64	0.66	0.63	**0.68**	0.65
Precision Macro Avg	0.51	0.65	0.66	0.65	**0.69**	0.65
Recall Macro Avg	0.51	0.65	0.66	0.65	**0.69**	0.65

TABLE II: Classification Metrics Summary for Alcohol Consumption

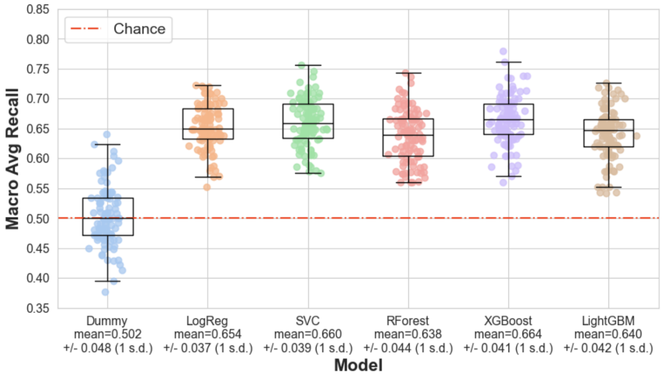


Fig. 10: Model Performances for Alcohol Consumption Across Varying K-Fold Splits

As Logistic Regressionranks first on the leader board across varying k-fold splits, its performance for the target variable alcohol consumption was further dissected using the confusion matrix. Similar to the observations with smoking habits, the probabilities for most cases hover around the 0.5 decision threshold. This pattern aligns with the findings from the principal component analysis (Figure 6), where a distinct decision boundary is absent, highlighting the complexity of distinguishing between individuals who consume alcohol and those who do not. Notably, in contrast to the results with imbalanced target variables, the balanced nature of the alcohol variable yields a higher proportion of correct predictions. Specifically, the confusion matrix shows 75 true negatives compared to 42 false positives, and similarly, 75 true positives against 36 false negatives, underscoring a more effective classification for alcohol consumption behaviors.

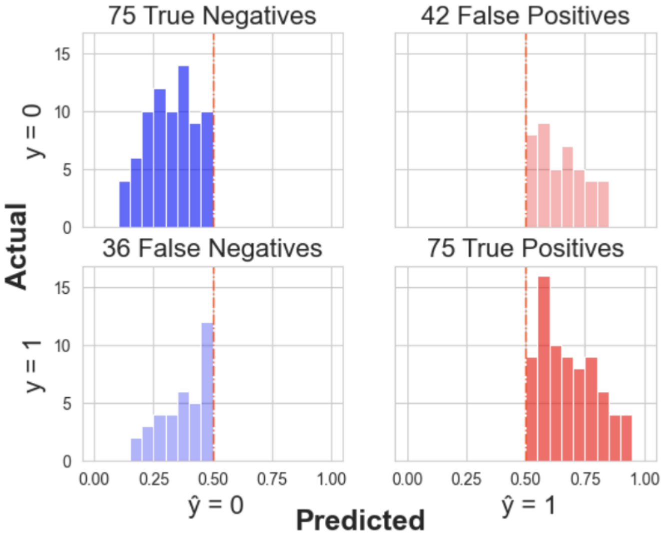


Fig. 11: Confusion Matrix Probabilities for Alcohol Consumption

The global beeswarm plot analysis sheds light on how various linguistic features in Facebook writing styles correlate with alcohol consumption. Swear words stand out as a significant predictor, suggesting a strong association between the use of profanity in posts and alcohol consumption. In contrast, fewer numerical references in posts tend to indicate alcohol consumption, implying that discussions involving alcohol might be less focused on numbers or quantitative details.

The analysis also shows that increased mentions of biological processes, financial terms, spatial references, and prepositions are linked with higher alcohol consumption. This suggests that Facebook posts discussing physical sensations, monetary matters, or locations might more frequently relate to alcohol consumption.

On the other hand, fewer mentions of work-related terms, first-person singular pronouns, and sexual content are associated with higher alcohol consumption predictions. This indicates a tendency for users who consume alcohol to write less about their professional life, personal experiences, or sexuality.

Furthermore, the presence of analytical thinking, expressions of anxiety, and agreement in posts tends to predict lower alcohol consumption. This finding suggests that more rational, anxious, or agreeable writing styles might be less common among users who discuss or engage with alcohol.

This concise interpretation of the beeswarm plot highlights the intricate relationship between Facebook writing styles and alcohol consumption, offering insights into how language use on social media can reflect lifestyle and behavioral patterns.

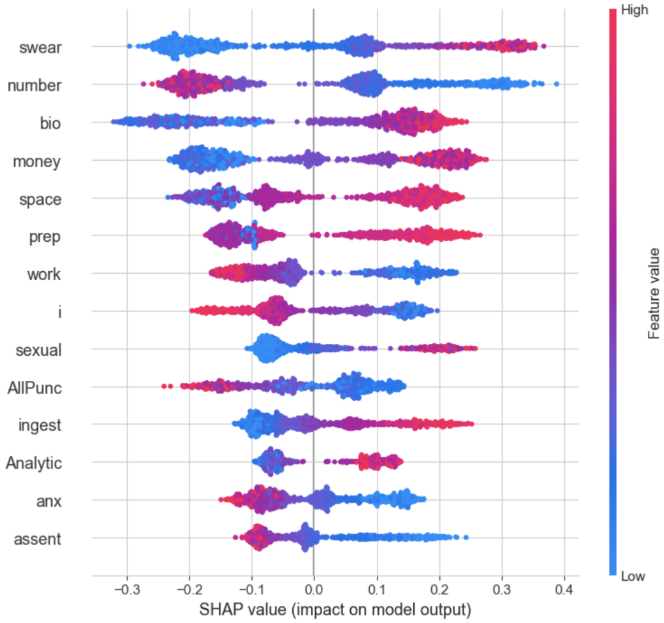


Fig. 12: Feature Importance for Alcohol Consumption

C. Drug Abuse

For the target variable Alcohol Consumption, we also observed that all models showed a considerable increase in prediction performance compared to our base dummy classifier (tab. **??**). Both the Logistic Regression and the Support Vector Classifier showed the best result in macro-averaged precision and recall, where the Logistic Regression performed better on Class 0 and the Support Vector Classifier performed better on Class 1. Additionally, the high accuracy of the base dummy classifier shows the importance of metrics other than accuracy in situations with a high class imbalance.

As illustrated in Figure 13, modeling drug use proved particularly challenging, evidenced by the minimum macro average recall values of the fine-tuned models hovering around or even falling below the 50% mark. Among these, Lo-

Metric	Base	LR	SVC	RF	XGB	LGBM
Accuracy	**0.71**	0.67	0.61	0.64	0.61	0.63
Precision (Class 0)	0.81	0.88	**0.90**	0.88	0.87	0.87
Recall (Class 0)	**0.83**	0.70	0.59	0.65	0.61	0.64
F1-Score (Class 0)	**0.82**	0.77	0.71	0.75	0.72	0.74
Precision (Class 1)	0.20	**0.30**	0.28	0.29	0.27	0.28
Recall (Class 1)	0.19	0.58	**0.70**	0.60	0.60	0.60
F1-Score (Class 1)	0.19	**0.40**	**0.40**	0.39	0.37	0.38
Precision Macro Avg	0.51	**0.59**	**0.59**	0.58	0.57	0.58
Recall Macro Avg	0.51	**0.64**	**0.64**	0.63	0.61	0.62

TABLE III: Classification Metrics Summary for Drug Abuse

gistic Regression, despite its simplicity, achieved the highest overall macro average recall. However, it exhibited a relatively high standard deviation of ±0.055, indicating more variability in performance compared to models for other consumption habits.

Tree-based models, particularly XGBoost and LightGBM, displayed notably lower performance, with minimum recall values dipping below the 50% threshold. Random Forest's minimum performance barely reached this critical point, underscoring the difficulty in predicting drug use with these models. Conversely, SVC demonstrated the least variability in performance across all splits, with a standard deviation of ±0.053, although achieving the lowest overall macro average recall among the models evaluated.

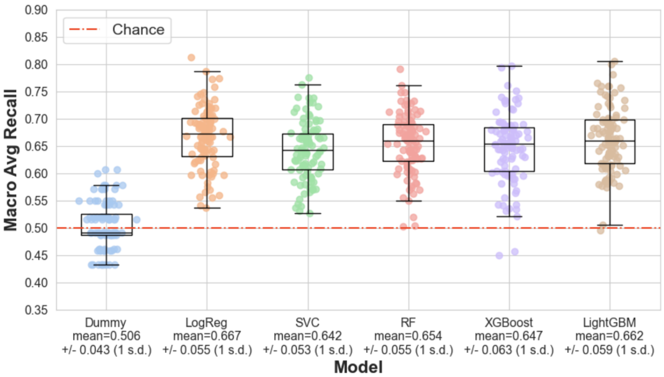


Fig. 13: Model Performances for Drug Use Across Varying K-Fold Splits

Logistic regression, with the highest overall macro average recall, was analyzed further through its confusion matrix. The distribution of true negatives and false positives around the 0.5 decision threshold indicates a balanced classification, albeit with instances extending to the tails, suggesting variability in prediction confidence. False negatives cluster near the 0.5 mark, while true positives, despite many centering around the midpoint, show a broader spread, indicating mixed confidence in positive predictions.

The model identified 128 true negatives and 25 true positives, showing a better capability to identify negative cases. However, the 57 false positives and 18 false negatives highlight the need for improvement, particularly in reducing incorrect positive predictions and better detecting positive cases.

This analysis reveals the Logistic Regression model's nuanced performance in drug use prediction, emphasizing the challenge of balancing precision and recall in an imbalanced dataset. The difficulty of accurately modelling drug use was again pronounced in the dimensionality reduction analysis III-C.

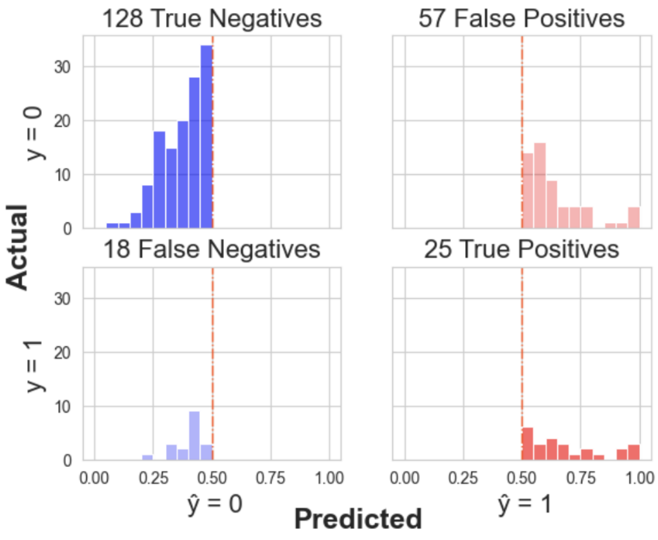


Fig. 14: Confusion Matrix Probabilities for Drug Abuse

The global SHAP barplot, displayed as Figure 15, offers valuable insights into the linguistic predictors for drug use in Facebook writing styles. It's noteworthy that numerical references and the use of swear words emerge as leading indicators, suggesting a potential link between the frequency of numerical discussions and profanity in posts with drug-related behaviors.

Moreover, sexual content and filler words also hold significance in predicting drug use. These less formal language elements imply that casual or expressive language might be associated with discussions or attitudes related to drug use.

The emotional tone of posts ranks high in importance, indicating that the overall emotional or attitudinal disposition in posts plays a role in predicting drug use. Work-related language and terms related to biological processes further contribute to the predictive model, suggesting that discussions about professional life and physical experiences or sensations are relevant when considering drug use predictions.

Punctuation, particularly the use of exclamations, references to time , discussions involving money, and expressions of anger provide additional layers of context. This implies that excitement, temporal references, financial discussions, and emotional expressions, especially those related to frustration or irritation, are pertinent to understanding drug use behaviors as reflected in social media language.

Ingestion-related terms, expressions of discrepancy, and the use of the second person pronoun ’you’ contribute to the list of predictors. These features shed light on the social and cognitive dimensions of drug use discussions, reflecting conversational dynamics, personal appeals or accusations, and cognitive dissonance present in narratives related to drug use.

This in-depth exploration of the SHAP analysis not only outlines the linguistic elements that influence the predictive model for drug use but also sheds light on the intricate interplay of language, emotions, and social dynamics that are an inherent part of such conversations on social media platforms.

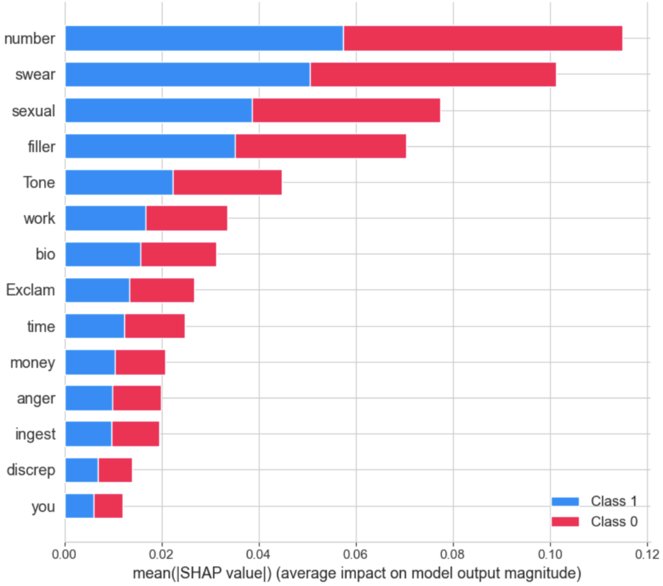


Fig. 15: Feature Importance for Drug Abuse

V. Discussion and Future Work

Our study, despite facing limitations due to a small sample size and severe class imbalance, demonstrated that our models could discern meaningful patterns to a reasonable, achieving a baseline performance of at least 60%.

To overcome these limitations, future research should prioritize obtaining a more extensive and diverse dataset. This expansion is crucial for uncovering more nuanced patterns that small samples may overlook, ensuring robust conclusions.

To further advance the research in the domain of substance use behaviors, one possible next step is to revisit the original data source to extract additional textual patterns. Incorporating transformer-based models for both feature extraction and classification will play a pivotal role in this advancement. These advanced models, due to their comprehensive understanding of language context, have the capacity to generate intricate feature sets through the creation of high-dimensional text embeddings. Moreover, they hold the potential to directly classify substance use behaviors from social media posts without the need for additional feature sets. These approaches are anticipated to significantly enhance the identification of linguistic patterns associated with consumption behaviors, leading to improved prediction accuracy and deeper insights into the textual indicators linked to substance use, ultimately generating more effective classifications.

In summary, future work aims to expand the dataset, extract additional patterns, and leverage transformer-based models to address the limitations encountered, advancing the research’s understanding and practical implications.

VI. Conclusion

To conclude, our research utilized the Mann-Whitney U test to reveal a variety of distinctive features between substance users and non-users, highlighting the most influential features to distinguishing consumption behaviors. PCA analysis further underscored the inherent challenges in predicting these behaviors, pointing to the intricate nature of the problem at hand. Despite these challenges, all models demonstrated a notable capability to identify patterns within the linguistic features of Facebook users’ writing styles, achieving an average performance exceeding 60% across various data splits. This was particularly evident for the balanced target of alcohol consumption, where the models made fewer errors compared to correct predictions. Even in the face of imbalanced targets, which introduced additional complexity, the models were still able to unearth meaningful patterns, underscoring their effectiveness. Additionally, SHAP (SHapley Additive exPlanations) plots provided insightful interpretations of the writing styles of survey participants on Facebook, offering a deeper understanding of how linguistic nuances correlate with substance consumption behaviors. This synthesis of analytical approaches not only affirms the potential of linguistic analysis in behavioral prediction but also sets a foundation for future exploration in the domain of psycholinguistics.

REFERENCES

[1] D. Stillwell and M. Kosinski, "mypersonality project website," 2015.

[2] D. J. Stillwell and R. J. Tunney, "Effects of measurement methods on the relationship between smoking and delay reward discounting," *Addiction*, vol. 107, no. 5, pp. 1003–1012, 2012.

[3] P. M. Brandt and P. Y. Herzberg, "Is a cover letter still needed? using liwc to predict application success," *International Journal of Selection and Assessment*, vol. 28, no. 4, pp. 417–429, 2020.

[4] P. E. Shrout and S. T. Fiske, *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Psychology Press, 2014.

[5] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2020.

[6] S. P. King and B. A. Mason, "Myers-briggs type indicator," *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*, pp. 315–319, 2020.

[7] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[8] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, and N. Howard, "Common sense knowledge based personality recognition from text," in *Advances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II 12*. Springer, 2013, pp. 484–496.

[9] R. Hernandez and I. Scott, "Predicting myers-briggs type indicator with text," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[11] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[12] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[13] S.-Y. Kim, T. Park, K. Kim, J. Oh, Y. Park, and D.-J. Kim, "A deep learning algorithm to predict hazardous drinkers and the severity of alcohol-related problems using k-nhanes," *Frontiers in psychiatry*, vol. 12, p. 684406, 2021.

[14] P. Han, "The application of machine learning methods in drug consumption prediction," in *Advances in Computer, Communication and Computational Sciences: Proceedings of IC4S 2019*. Springer, 2021, pp. 497–507.

[15] M. Abo-Tabik, N. Costen, J. Darby, and Y. Benn, "Towards a smart smoking cessation app: A 1d-cnn model predicting smoking events," *Sensors*, vol. 20, no. 4, p. 1099, 2020.

[16] imbalanced-learn developers, "imbalanced-learn: Tomek Links - imblearn.under_sampling.TomekLinks," https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.TomekLinks.html, 2023, accessed: 2024-01-29.

[17] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and smote approaches for machine fault classification with an imbalanced dataset," *Sensors*, vol. 22, no. 9, p. 3246, 2022. [Online]. Available: https://doi.org/10.3390/s22093246

[18] scikit-learn developers, "scikit-learn: StandardScaler - sklearn.preprocessing.StandardScaler," https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, 2023, accessed: 2024-01-29.

[19] imbalanced-learn developers, "imbalanced-learn: ADASYN - imblearn.over_sampling.ADASYN," https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html, 2023, accessed: 2024-01-29.

[20] A. F. N. Masruriyah, H. Y. Novita, C. E. Sukmawati, A. Fauzi, D. Wahiddin, and H. H. Handayani, "Thorough evaluation of the effectiveness of smote and adasyn oversampling methods in enhancing supervised learning performance for imbalanced heart disease datasets," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, Manado, Indonesia, 2023, pp. 1–7.

[21] R. W. Emerson, "Mann-whitney u test and t-test," *SageJournals*, vol. 117, 2023.

[22] A. Savchenko, "Facial expression recognition with adaptive frame rate based on multiple testing correction," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 30 119–30 129. [Online]. Available: https://proceedings.mlr.press/v202/savchenko23a.html

[23] D. Stillwell and M. Kosinski, "List of variables available," https://web.archive.org/web/20180428085709/http://mypersonality.org/wiki/doku.php?id=list_of_variables_available#facebook_activity, 2015, retrieved February 28, 2024, from mypersonality.org.

[24] Y. Wang, "Pacmap." [Online]. Available: https://github.com/YingfanWang/PaCMAP

[25] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1061.html

[26] optuna, "Optuna." [Online]. Available: https://github.com/optuna/optuna

[27] shap, "Shap." [Online]. Available: https://shap.readthedocs.io/en/latest/