Profiling Psychopathic Behavior Using Machine Learning

Avraham Treistman, Tehilla David, Sivan Levi, Dror Mughaz
Jerusalem COllege of Technology
Department of Computer Science, Jerusalem, Israel, 9372115 , Address2, Address3
treistma@jct.ac.il, sivanlevi94@gmail.com, tehila859@gmail.com, myghaz@gmail.com
{Avraham Treistman, Tehilla David, Sivan Levi, Dror Mughaz

Abstract

Psychopathy is a complex personality disorder characterized by persistent deficits in empathy and manipulative behavior. Traditional diagnostic methods often rely on subjective clinical assessments, which are susceptible to deception. This research proposes an objective, non-invasive computational framework for profiling psychopathic traits using Natural Language Processing (NLP) and Machine Learning. We developed a systematic pipeline utilizing transcribed interviews from confirmed criminal psychopaths and a balanced control group. To address data sparsity and noise, we employed the Dynamic Variance Thresholding (DyVaT) algorithm to construct a semantically dense vocabulary of over 1,300 features. The methodology integrates advanced preprocessing, TF-IDF vectorization, and synonym-based data augmentation to enhance model generalization. Among the evaluated classifiers, a Linear Support Vector Machine (SVM) achieved the highest performance, with an accuracy of 0.8081 and an F1-score of 0.7957. Our findings demonstrate the efficacy of linguistic biomarkers and feature importance analysis in distinguishing psychopathic speech patterns. This study provides a scalable methodology for early screening and diagnostics, with significant implications for forensic psychology, security, and ethical AI deployment in mental health.

Keywords: Computational Linguistics, Machine Learning Application, Psychopathy Profiling, Linguistic Biomarkers, Data Augmentation, Dynamic Variance Thresholding

1. Introduction

Psychopathy is a severe personality disorder with major social and clinical consequences. It involves low empathy, manipulation, and persistent antisocial behavior. Many individuals show a convincing “mask of sanity” in everyday interactions. The diagnosis often relies on the PCL-R checklist plus lengthy face-to-face interviews. These procedures are difficult to scale and can vary between clinicians. They also depend on cooperation, which deceptive subjects may deliberately undermine.

We therefore need complementary screening methods that are more objective and auditable. Computational pipelines can support clinicians without replacing clinical judgment. They fit remote data collection and monitoring workflows that are increasingly realistic. They also encourage shared resources, clear protocols, and repeatable evaluation. Still, the tools must be interpretable, safe, and cautious in high-stakes use. Ethics, consent, privacy, and bias management must be designed from the beginning.

NLP is promising because language leaks subtle patterns during spontaneous narration. Speakers rarely control these patterns consistently throughout an interview. Machine learning has identified “Dark Triad" traits from text (Yeasmin et al., 2024). Psychopathy detection remains difficult because labeled, verified transcripts are scarce. Legal barriers and sensitivity restrict access, creating a severe class imbalance. Domain shifts across sources can also distort models unless explicitly handled.

Our study presents a pipeline designed around these practical data constraints. We focus on distinguishing criminal psychopaths from a non-psychopathic control group. The offender interviews were collected from YouTube, while the controls came from NPR interviews. We matched conversational style and cleaned transcripts using consistent annotation rules. This reduced cues from formatting, editing, or interviewer structure. It also supports reproducibility and future sharing of datasets and tools.

Representation choices were critical, so we avoided using every observed token. Instead, we built a seed list of psychologically relevant words by manual selection. We expanded that list using Dynamic Variance Thresholding, or DyVaT (Treistman et al., 2022). DyVaT retains semantically related terms while filtering high-variance noise from embeddings. The final lexicon contains about 1,357 focussed features for modeling. This feature design improves interpretability and helps address sparsity and dimensionality issues.

We then applied a rigorous text-processing workflow to ensure consistent input. Transcripts were segmented into standardized 15-sentence chunks to stabilize sample length. Lemmatization reduced sparsity by mapping inflected forms to base roots. Data scarcity persisted, so we augmented the minority class with synonym substitution. Such augmentation is common for imbalance in personality classification (Pradana and Suhartono, 2024). We

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 115–125
12 May 2026.

treated augmentation conservatively, since it can introduce small semantic drift.

For classification, we compared Logistic Regression, Random Forest, and Support Vector Machines. Choosing the right classifier is central in text categorization (Allam et al., 2025). Random Forest often handles high-dimensional data well (Venkateshwarlu et al., 2024). However, our best results came from a Linear SVM in this setting. It achieved 80.81% accuracy and a 0.7957 F1 score (Alzoubi et al., 2023). We also tracked feature significance to support transparent screening and diagnostics.

Feature inspection revealed systematic differences that align with psycholinguistic expectations. Psychopathic speech overused basic-need terms like “money,” “food,” and “sex.” It also used more words related to violence, dominance, and authority. Controls used more language about social connection, norms, and morality (Adkins et al., 2025). High-stakes use demands transparency, consent, privacy safeguards, and bias checks (Zhou and Chen, 2023). Next, we will test LLMs, acoustic cues, and sensor-ready multimodal designs on lightweight assessment platforms.

2. Related Works

Classic machine learning remains competitive for high-dimensional text features. Random Forest often performs well in sparse spaces (Venkateshwarlu et al., 2024). Support Vector Machines can be robust across languages and settings (Alzoubi et al., 2023). A survey of these methods and common evaluation practices guides model selection under real deployment constraints (Gasparetto et al., 2022).

Work on personality and “Dark Triad” traits motivates psychopathy-oriented screening . Standard classifiers can separate higher-risk profiles from controls (Yeasmin et al., 2024). There are reported gains from ensembles on non-linear personality signals (Maxim et al., 2025). Regex rules can be mixed with NLP for phase-aware disorder detection (Patel and Johnson, 2025). Research connects deception and emotion signals to forensic text analysis (Adkins et al., 2025).

Data collection and labeling protocols strongly shape what models actually learn. Many studies depend on secondary data, which complicates consent and reuse. Shared schemas, annotation rules, and audit trails help build usable infrastructure. These steps also support domain adaptation when sources differ in style.

Preprocessing decisions are not neutral in psychological profiling tasks. Preprocessing can change accuracy, even for sentiment pipelines (Alam and Yao, 2019). Stop word removal is especially tricky for self-reference and social framing (Kaur and Buttar, 2018), and psychological settings need extra caution. Tokenization quality also matters for mental health signals (Dixit et al., 2023).

Feature weighting and representation are still common baselines in clinical text modeling. Various TF-IDF variants are compared for use on unstructured datasets, such as interview transcripts (Das et al., 2023). However, “noise” can contain the most diagnostic information linking psychopathic deviation to narrative style and lexical choices (Gawda, 2022).

A central innovation of our work is the method of vocabulary construction. We did not simply use all the words available in the transcripts. Instead, we used a "seed list" of manually selected words that are psychologically relevant. We expanded this list using the Dynamic Variance Thresholding (DyVaT) algorithm (Treistman et al., 2022). This algorithm helps identify semantically relevant words while filtering out high-variance noise. The result is a focused lexicon of approximately 1,357 features. This approach allows us to target the specific narrative structures associated with psychopathy. Deep learning is increasingly used to model subtle and contextual personality signals. A review of machine and deep learning for trait detection notes the growing use of ensembles and larger architectures (Naz et al., 2025). Graph approaches can model relations between words, users, and contexts, such as LL4G for self-supervised, dynamic graph-based personality detection (Shen et al., 2025).

Network science offers another angle on discourse and psychological structure, such as textual forma mentis networks for adolescents and psychopathology levels (Carrillo et al., 2025). Knowledge-guided filtering can focus models on clinically informative segments such as PsyTEx for refining text for psychological analysis (Bhandarkar et al., 2025). Emotion dynamics can serve as biosocial markers beyond static sentiment (Teodorescu et al., 2023).

Related mental health domains also shape methods we can reuse or adapt, such as enhanced TextGCN for depression detection using emotion representations (Mao and Han, 2025) or DepGLM to recognize degrees of depression with LLM support (Liu et al., 2025). These tasks differ from psychopathy, but the modeling patterns are transferred. They also raise similar needs for calibration and clinically meaningful metrics. The scarcity of labeled psychopathy data remains a practical bottleneck for supervised learning. The class imbalance is severe, and verified labels are rarely available to share. This imbalance can be addressed by augmentation, such as intent-aware (Saleem and Kim, 2024) or synonym replacement(Pradana and Suhartono, 2024).

Other pipelines generate synthetic samples for downstream interventions and support tools, such as ERNIE-based augmentation for CBT-related applications (Sambana et al., 2025). Topic modeling plus synthetic generation improved suicidal ideation detection (Ghanadian et al., 2025). These results suggest augmentation can help generalization when clinical data is constrained. Still, synthetic text can drift and must be validated carefully. Large language models are changing baselines, but introduce new risks, such as the applicability of LLMs for the classification of health text using public social media data (Guo et al., 2024), whether GPT-3 exhibits psychopathic traits under psychological perspectives (Li et al., 2022), questioning whether human personality tests can be applied to algorithmic agents (Sühr et al., 2025), or LLM blurring of linguistic markers used for trait inference (Sourati et al., 2024).

Safety and governance are central when models touch sensitive psychological labels (Li et al., 2024). Ethical principles must be applied to engineering practice (Zhou and Chen, 2023; Mittelstadt, 2019) while addressing regulatory gaps in AI-driven profiling on social media (Bose et al., 2025).

Bias is another recurring threat in clinical and assessment settings, such as racial bias in AI-mediated psychiatric diagnosis with large models (Bouguettaya et al.), video interviews (Mujtaba and Mahapatra, 2025), or emotional AI (Chavan et al., 2025).

Deployment discussions increasingly connect to law, clinical workflows, and patient safety. The Draft EU AI Act has implications for profiling (Veale and Zuiderveen, 2021). One must consider the legal risks of AI in mental healthcare (Rahman et al., 2025), as well as safe data management (Raygoza-L et al., 2025). These works collectively push for careful deployment, monitoring, and clinician involvement.

2.1. The DyVaT Algorithm

The DyVaT (Dynamic Variance Thresholding) algorithm is a novel technique designed to reduce the dimensionality of word embeddings in NLP tasks by adaptively removing high-variance noisy dimensions using the cosine distance metric. Using the kneedle algorithm to determine an optimal threshold, DyVaT retains low-variance features that contribute to forming tighter semantic clusters. This process enhances the quality of the vector representations without substantial loss of information, thereby improving downstream tasks such as classification and clustering. Experiments demonstrate DyVaT's ability to generate semantically coherent word collections with less noise compared to other methods, making it a powerful, scalable tool for feature selection in text analysis (Treistman et al., 2022).

3. Methodology

This section presents the project's methodology, including data collection and preprocessing, ML model selection, and system architecture. It outlines the approach used to accurately and reliably classify psychopathic traits from text data.

3.1. Data Collection

The efficacy of the proposed classification framework relies on a structured approach to data handling and feature engineering. To this end, we first define the linguistic corpora used for training and evaluation. Following the dataset description, this section outlines the methodology for generating a semantically dense vocabulary, the subsequent extraction of discriminative features, and the application of synonym-based augmentation to enhance the model's generalizability.

Psychopathic Dataset

- Text samples were gathered from approximately 77 YouTube interviews featuring individuals who were legally and psychologically confirmed as criminal psychopaths.
- The transcripts were generated using the Python library youtube-transcript-api.
- After extraction, a manual review was conducted on all transcripts to verify their accuracy.

Non-Psychopathic Dataset

- The dataset was derived from a validated Kaggle resource, containing transcripts of National Public Radio (NPR) interviews with randomly selected (NPR).
- The non-psychopathic dataset was intentionally selected because it consisted of interview transcripts with a conversational style and structure similar to those in the psychopathic dataset. This deliberate matching ensured consistency in format and context across both datasets, minimizing bias and allowing for reliable comparison of linguistic features.
- The control sample selection was guided by accepted definitions of psychopathy, e.g., "a manipulative, cunning, and antisocial individual who, according to Hare (Hare, 2020), comprises about 1 percent of the general population" (Hancock et al., 2018).
- For this dataset, 10,000 records were initially sampled using two inclusion criteria: (1) each record had to contain a unique EPISODE identifier, and (2) the transcript text was required to include a minimum of 15 sentences. After

applying these criteria, 782 valid records were retained.

3.2. Text Preprocessing

To manage the variability in transcript length, the text was split into uniform segments of fifteen sentences each. Each segment was normalized, cleaned, tokenized, and lemmatized using standard methods (spaCy platform). Short, non-alphabetic or linguistically irrelevant tokens were filtered out (Vasiliev, 2020). Segments with fewer than nine relevant vocabulary terms were discarded to maintain dataset quality and ensure that texts contained sufficient linguistic material for meaningful analysis. Due to an inherent class imbalance in the dataset, data enhancement methods were applied to the minority psychopathic class to equalize representation and prevent biased model learning.

3.3. Vocabulary Construction

The vocabulary construction method (Figure 1) captures key linguistic markers distinguishing psychopathic and non-psychopathic traits. We began with a basic seed vocabulary for each class, manually selecting fifty words per class with the highest coefficients based on our dataset, validated using logistic regression. This vocabulary was then expanded using the DyVaT algorithm. The algorithm increased each class's vocabulary to approximately 1,000 words, and after removing duplicates and applying filtering, the final vocabulary totaled 1,357 words.

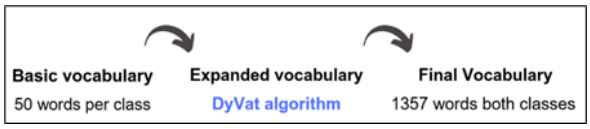

Figure 1: Vocabulary Construction Process

The initial seed vocabulary consisting of fifty terms per class was manually curated. The selection was tuned by both term frequency and the highest logistic regression coefficients observed in our dataset, ensuring that the chosen words captured the most discriminative linguistic markers for each class.

3.4. Expansion via DyVaT Algorithm

DyVaT was applied to expand the seed vocabulary by leveraging semantic similarity from GoogleNews Word2Vec embeddings, using cosine similarity distance thresholding to identify relevant terms. This approach increased the vocabulary coverage to approximately 1,357 words per class while maintaining interpretability. The base vocabulary expansion method relied on a fixed cosine similarity threshold from Word2Vec embeddings, while DyVaT's dynamic variance-based thresholding following initial cosine clustering introduced a broader set of semantically relevant terms, enhancing the model's ability to capture subtle linguistic distinctions. Unlike the base method's static cutoff, which limited coverage and generalization due to its rigid nature, DyVaT effectively balanced expansion with semantic relevance, providing a more robust feature set for psychopathy classification (Treistman et al., 2022).

3.5. DyVaT vs. Base Vocabulary

The visualization compares word expansions generated from the seed word 'control'. In the Base algorithm (Figure 2), related terms are spread with a fixed radius, producing a scattered distribution. In contrast, the DyVaT algorithm (Figure 3) yields tighter clusters of semantically related words, increasing density and better capturing nuanced relationships. This demonstrates that DyVaT provides a more coherent and semantically meaningful expansion compared to the Base method.

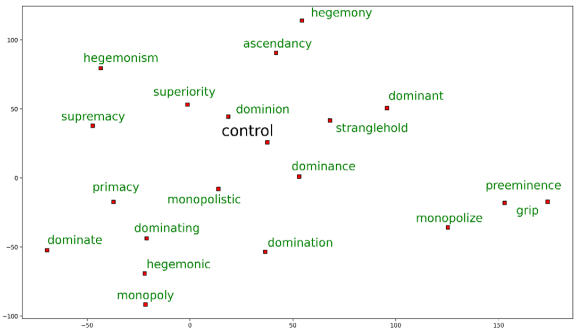

Figure 2: Base Algorithm

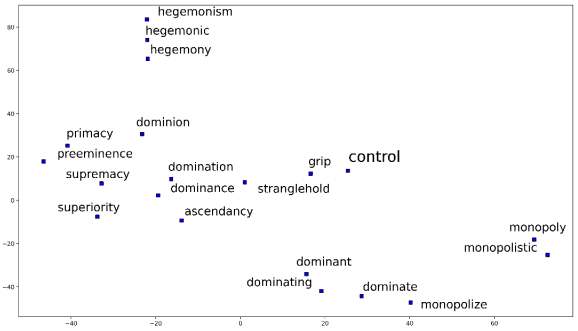

Figure 3: DyVaT Algorithm

3.6. Data Augmentation

To overcome the limitations inherent in small datasets and to improve classification performance, text-based data augmentation techniques were applied to preprocessed text records. Specifically, each original record was transformed, in which selected words, limited only to those present in the

vocab, were replaced by semantically similar synonyms. This was achieved through the use of the SpaCy library, integrating vocabulary vectors and part-of-speech filtering, thus ensuring both contextual fidelity and grammatical correctness. To prevent redundancy, augmented records with high semantic overlap to the originals were systematically removed based on configurable similarity thresholds. This process expanded the dataset from 224 to 436 unique records, resulting in measurable improvements in model performance.

Recent empirical research supports this approach, demonstrating through comprehensive analysis that token-level augmentations, particularly word replacement and random swapping most consistently enhance supervised NLP performance on limited data. These methods generate new text by substituting words or phrases with synonyms from dictionaries or embedding similarities, preserving semantic meaning and original labels while expanding linguistic diversity. Such techniques prove especially effective when training samples are scarce, as they intuitively maintain sentence intent through semantically equivalent token replacements (Chen et al., 2023).

3.7. Feature Extraction

The TF-IDF vectors were restricted to a custom vocabulary, manually curated and further refined using the Dynamic Variance Thresholding (DyVaT) algorithm. Following the DyVaT process, additional manual customization was performed to further optimize the feature set. This ensured that only lexically and semantically meaningful words were included. These vectors served as the input features for machine learning, allowing the algorithm to learn which words and patterns are most indicative of psychopathic versus non-psychopathic text. Feature extraction in this manner provides interpretability and insights into the key linguistic cues distinguishing the two classes.

3.8. Hyperparameter Tuning

Hyperparameter tuning was performed using a grid search approach combined with stratified 5-fold cross-validation to systematically explore combinations of model parameters. The stratification ensured balanced class distributions across folds, addressing potential imbalances in psychopathic versus non-psychopathic samples. The primary optimization objective was the F1-score, chosen to balance precision (reducing false positives) and recall (reducing false negatives), reflecting the critical need to identify all true cases without excessive false alarms.

4. Implementation and Results

Table 1 sumarizes the key differences between the standard baseline and our DyVat augmented approach.

4.1. Baseline vs, Augmented DyVaT Method

Aspect	Baseline	Augmented DyVaT
Pre-processing	Basic noise removal Simple tokenization	Advanced preprocessing including lemmatization Removal of short/irrelevant records
Vocabulary	No extensive feature engineering Relied on basic lexicon if at all	Curated vocabulary built with DyVaT algorithm Expanded manual lexicon to 1,233 semantically relevant features
Feature Extraction	Basic TF-IDF or bag-of-words, no focused filtering	Only records containing =9 VOCAB words retained TF-IDF vectors based on enhanced vocabulary
Data Augmentation	None	Systematic generation of new records by synonym substitution within curated vocabulary
Model Training	Baseline models trained directly on raw data	Models trained and validated on high-quality, filtered, and augmented data
Goals	Quick baselines to establish initial feasibility	Maximized performance, interpretability, and generalization (based on literature insights and pilot results)

Table 1: Comparison Between Original and Final Methods

4.2. Original vs. Augmented Dataset

Original Dataset Contained raw transcribed interviews labeled according to psychopathic and non-psychopathic traits, without advanced filtering or augmentation.

Augmented Dataset Expanded version created by applying the full augmentation pipeline, resulting in increased linguistic diversity, a more

balanced class distribution, and improved robustness for model training and evaluation.

4.3. Data Processing Pipeline

The data processing pipeline was systematically developed to ensure that raw interview data was transformed into high-quality, interpretable features optimized for ML classification. The pipeline includes these key stages:

Data Ingestion and Organization

Raw interview transcripts were extracted directly from MongoDB and imported into pandas DataFrames, establishing a structured, accessible format for all downstream analyses.

Text Preprocessing

- The transcript texts underwent thorough normalization and noise reduction, including lowercasing, removal of punctuation, stop words, and irrelevant symbols.
- Standardized input size by dividing transcripts into fixed-length segments.
- Further linguistic cleaning was completed by tokenization and lemmatization (via spaCy), transforming words to their canonical forms and ensuring linguistic consistency.
- Semantic filtering excluded segments with insufficient relevant vocabulary, minimizing noise and ensuring only linguistically meaningful samples advanced.

Vocabulary Construction And Restriction

Feature extraction was restricted to a carefully engineered vocabulary, expanded using the DyVaT algorithm. This method augmented a manual seed list with semantically similar terms, enhancing lexical coverage for both psychopathic and non-psychopathic classes while ensuring feature interpretability.

Data Augmentation And Balancing

Data augmentation strategies were employed for the minority class, generating new records by substituting words with synonyms within segments, thus increasing dataset size and diversity while preserving semantic integrity.

Feature Extraction

Texts were vectorized using the TF-IDF algorithm, resulting in numerical features that represent each record in terms of word importance and discriminative value.

Model Training, Selection and Evaluation

- Multiple ML models were trained using stratified k-fold cross-validation for robust and balanced evaluation.
- Model persistence was achieved by serializing the optimal models with joblib, supporting reproducibility and deployment.

Feature Importance And Visualization

- The key linguistic features distinguishing psychopathic vs. non-psychopathic speech were identified through LR coefficient analysis.

- The fifty most influential words per class were visualized as word clouds, highlighting dominant lexical markers and supporting transparent model interpretation.

4.4. Feature Importance

The analysis of the most influential features was conducted using the coefficient values derived from a logistic regression model, which allowed for clear interpretability of the linguistic markers associated with psychopathy. The logistic regression coefficients quantify the contribution of each feature to the classification decision by indicating the change in the log-odds of the target class.

Figures 4 and 5 present the top 50 linguistic features most strongly associated with each class (Psychopath vs. Non-Psychopath).

Figure 4: Top 50 Features of Psychopaths

Figure 5: Top 50 Features of Non-Psychopaths

5. Analysis

5.1. Baseline Method

The baseline model (Table 2) was trained without dataset balancing, without the DyVaT vocabulary, and without any filtering or augmentation. The TF-IDF vectorizer relied entirely on the raw, unfiltered lexical space, resulting in high noise levels and severe class imbalance.

This configuration struggled to capture meaningful psychopathy-related linguistic patterns, highlighting the need for improved data quality and class balance.

Model	F1-score	Vocabulary Size
SVM	0.5000	2881
Logistic Regression	0.3871	80182
Random Forest	0.3200	43028

Table 2: Performance Metrics for Baseline Method

- Training set size: 781
- Test set size: 194
- Total records: 975
- Psychopath records: 193
- Non-psychopath records: 782

5.2. Augmented DyVaT Method

Our method uses data augmentation was combined with DyVaT expansion (Table 3 and Figure 6), expanding the psychopathic class and enhancing generalization. Linear SVM achieved the highest performance. This result represents the culmination of progressive refinement, demonstrating that the combination of controlled vocabulary, balanced classes, semantic filtering, and targeted augmentation yields the most reliable psychopathy detection model.

Model	F1-score	Vocabulary Size
SVM	0.7957	1233
Logistic Regression	0.6882	1233
Random Forest	0.7391	1233

Table 3: Performance Metrics for Augmented DyVaT Model

- Training set size: 407
- Test set size: 99
- Total records: 506
- Psychopath records: 253
- Non-psychopath records: 253

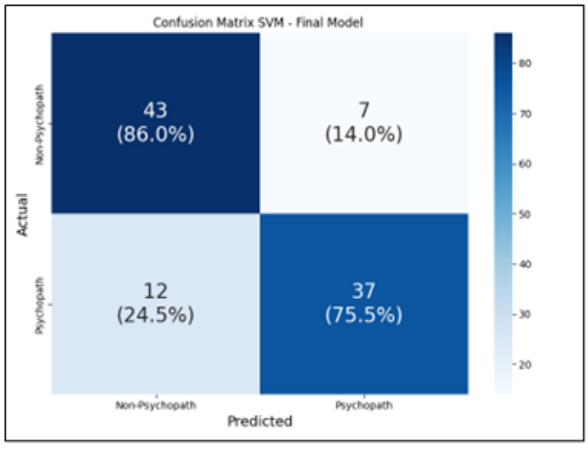

Figure 6: Confusion Matrix of the Augmented DyVaT Method

6. Discussion

Our research successfully achieved its objectives by developing a ML and NLP-based system capable of identifying psychopathic traits through linguistic analysis. After systematically evaluating multiple methods, the Support Vector Machine (SVM) trained with the DyVaT vocabulary and augmented dataset was identified as our optimal algorithm for psychopathy detection.

This model achieved the highest F1-score (0.7957), while maintaining strong generalization and interpretability. The integration of longer text records, controlled vocabulary filtering, and targeted data augmentation proved essential for capturing psychopathy-related linguistic structures without overfitting. The confusion matrix further demonstrated that the SVM achieved a high true positive rate while maintaining a low false positive rate, reflecting strong discriminative ability and balanced performance.

Thus, the findings confirm that computational methods, particularly SVM with carefully constructed linguistic features, are effective for early detection of psychopathic traits. These results strengthen the positive consideration of ML approaches in psychological assessment.

Future work may focus on expanding the dataset, incorporating deep learning architectures, integrating sentiment and emotion analysis, and exploring multimodal inputs (e.g., audio, video) to further improve detection accuracy and robustness.

6.1. Project Challenges and Limitations

During the execution of this research, several major challenges were identified, which influenced the planning, implementation, and evaluation stages:

Ambiguity in Defining Psychopathy

There is no universally accepted definition of psychopathy. Diagnosis relies on a complex combination of psychological, behavioral, and social criteria. Although academic literature offers several assessment tools, such as the Psychopathy Checklist-Revised (PCL-R), there is no full consensus on the construct's boundaries. This lack of uniformity complicates the process of labeling training data and may negatively impact model performance.

Dataset Collection and Preparation

Censored sources: Some available materials, such as YouTube recordings or interview transcripts, included censorship, omissions, or edits that reduced data completeness and reliability.

Scarcity of psychopathy-labeled texts: Texts authored or spoken by clinically confirmed psychopaths are rare, due to ethical, privacy, and data access constraints.

Non-psychopath text collection: This task presented additional complexity, as sufficiently long texts authored or spoken by non-psychopaths were difficult to obtain. Furthermore, care had to be taken to ensure that the selected individuals represented a general population rather than a narrowly defined or homogeneous group, in order to reduce potential sampling biases.

Selecting And Evaluating ML Model

Developing an accurate and reliable psychopathy detection model posed significant challenges, particularly in the initial stages of algorithm selection. It was unclear whether to begin with traditional ML algorithms or to employ deep learning techniques. Consequently, careful experimentation and systematic evaluation were required to determine the optimal modeling strategy, balancing predictive performance with computational efficiency and training feasibility.

Building A Contexual Vocabulary

Another challenge in this study was determining whether to use a predefined vocabulary consisting of words specifically associated with psychopathic and non-psychopathic speech, or to derive the vocabulary directly from the training dataset. Careful consideration was required to determine how to construct this vocabulary in a manner that maximizes its discriminative power, captures relevant linguistic patterns for each class, and avoids introducing bias or overfitting.

6.2. Commercial and Societal Value

The developed system has potential commercial applications in security, mental health, and criminology, providing tools for early risk assessment and decision support. Beyond commercial value, the research contributes to societal well-being by enabling more proactive identification of high-risk individuals and supporting preventive interventions.

6.3. Summary

This research set out to determine whether psychopathic linguistic patterns can be identified through NLP and machine learning, with the goal of creating a scalable, interpretable, and non-invasive detection framework. By constructing a novel dataset from verified psychopathic and non-psychopathic interviews, expanding vocabulary coverage using the DyVaT algorithm, and applying targeted preprocessing, semantic filtering, and data augmentation, the study achieved robust classification results.

The Linear SVM model emerged as the optimal solution, delivering an accuracy of 0.8031 and an

F1 score of 0.7957, outperforming alternative models such as Logistic Regression and Random Forest. These results underscore that psychopathy-related language patterns are consistent and detectable when captured through a well-engineered lexical feature space. Moreover, the interpretability of the Linear SVM provides valuable transparency-an essential quality in forensic, legal, and clinical applications.

The contributions of this study are threefold:

- **Data Contribution** Development of a real-world, verified dataset of psychopathic and non-psychopathic speech, providing a foundation for future research.
- **Methodological Innovation** Application of DyVaT for vocabulary expansion, enhancing feature quality while maintaining explainability.
- **Practical Relevance** Validation of an NLP-based detection tool that could support early risk assessment in mental health, security, and law enforcement contexts.

While the findings are promising, they should be interpreted in light of the study's limitations, including dataset size, English-only scope, and reliance on text transcripts without paralinguistic data. These constraints point toward several avenues for future research: expanding the dataset, incorporating multilingual sources, integrating multimodal features, and leveraging advanced deep learning architectures such as transformer-based models.

In essence, this project demonstrates that language serves not only as a medium for communication but also as a measurable indicator of underlying personality traits. Through careful design and rigorous validation, the framework developed here shows the potential of computational psycholinguistics to aid in the early identification of psychopathy-supporting, rather than replacing, professional judgment in high-stakes environments.

7. References

Jonathan Adkins, Ali Al Bataineh, and Anthos Khanal. 2025. A psycholinguistic nlp framework for forensic text analysis of deception and emotion. *Frontiers in Artificial Intelligence*, 8:1669542.

Saqib Alam and Nianmin Yao. 2019. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, 25(3):319–335.

Hesham Allam, Lisa Makubvure, Benjamin Gyamfi, Kwadwo Nyarko Graham, and Kehinde Akinwolere. 2025. Text classification: How machine learning is revolutionizing text categorization. *Information*, 16(2):130.

Yehia Ibrahim Alzoubi, Ahmet E Topcu, and Ahmed Enis Erkaya. 2023. Machine learning-based text classification comparison: Turkish language context. *Applied Sciences*, 13(16):9428.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Gregory Webster, and Damon Woodard. 2025. Psytex: A knowledge-guided approach to refining text for psychological analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 151–178.

Evan Bose, Chaitanya Anil Kumar, and N Meenakshi. 2025. Ai-driven psychological profiling on social media: Mechanisms, ethical breaches, and regulatory challenges in data inference. recent trends in social studies. 2025; 2 (1): 1–7p. *AI-Driven Psychological Profiling on Social Media Bose et al. STM Journals*, page 2.

A Bouguettaya, EM Stuart, and E Aboujaoude. Racial bias in ai-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. npj digit. med. 8, 332 (2025).

Alexis Carrillo, Simon Friedrich Roske, Rebeca Ianov-Vitanov, Enrico Perinelli, Alessandro Grecucci, and Massimo Stella. 2025. Textual forma mentis networks bridge language structure, emotional content and psychopathology levels in adolescents. *arXiv preprint arXiv:2505.06387*.

Vivek Chavan, Arsen Cenaj, Shuyuan Shen, Ariane Bar, Srishti Binwani, Tommaso Del Becaro, Marius Funk, Lynn Greschner, Roberto Hung, Stina Klein, et al. 2025. Feeling machines: Ethics, culture, and the rise of emotional ai. *arXiv preprint arXiv:2506.12437*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Mamata Das, PJA Alphonse, et al. 2023. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*.

Krishna Kant Dixit, Sumit Pundir, Anurag Shrivastava, C Praveen Kumar, Arun Pratap Srivastava, and Pankaj Singh. 2023. Analyzing textual data for mental health assessment: Natural language processing for depression and anxiety. In *2023 10th IEEE Uttar Pradesh Section International*

Conference on Electrical, Electronics and Computer Engineering (UPCON), volume 10, pages 1796–1802. IEEE.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.

Barbara Gawda. 2022. The differentiation of narrative styles in individuals with high psychopathic deviate. *Journal of Psycholinguistic Research*, 51(1):75–92.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2025. Improving suicidal ideation detection in social media posts: Topic modeling and synthetic data augmentation approach. *JMIR Formative Research*, 9:e63272.

Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

J. T. Hancock, M. Woodworth, and R. Boochever. 2018. Psychopaths online: The linguistic traces of psychopathy in email, text messaging and facebook. *Media and Communication*, 6(3):83–92.

R. D. Hare. 2020. The pcl-r assessment of psychopathy. In *The Wiley International Handbook on Psychopathic Disorders and the Law*, pages 63–106. Wiley.

Jashanjot Kaur and P Kaur Buttar. 2018. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. 2022. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*, 10.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843.

Jingfang Liu, Peng Ding, and Jie Chen. 2025. Depglm: Depression degree recognition on social media based on large language models. *Digital Health*, 11:20552076251408281.

Huimin Mao and Qing Han. 2025. Enhancing textgcn for depression detection on social media with emotion representation. *Frontiers in Psychology*, 16:1612769.

Leberecht Maxim, Nedderhoff Andre, Zitzmann Steffen, and Hecht Martin. 2025. Comparing machine learning methods for predicting dark triad personality traits using social media text data. *Journal of Research in Personality*, page 104690.

Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Dena F Mujtaba and Nihar R Mahapatra. 2025. Behind the screens: Uncovering bias in ai-driven video interview assessments using counterfactuals. *arXiv preprint arXiv:2505.12114*.

Anam Naz, Hikmat Ullah Khan, Amal Bukhari, Bader Alshemaimri, Ali Daud, and Muhammad Ramzan. 2025. Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review*, 58(8):239.

NPR. Npr media dialog transcripts. Kaggle Dataset. Accessed: September 22, 2025.

Dhruv Patel and Anju Johnson. 2025. Detecting narcissistic personality disorder (npd): A hybrid regex and nlp based ai approach with phase-aware classification. *IEEE Access*.

Rilo Chandra Pradana and Derwin Suhartono. 2024. Synonym replacement augmentation for handling data imbalance in personality classification. In *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 1–6. IEEE.

Md Ashrafur Rahman, Evangelos Victoros, Rob Davis, Tariq Duaa, Yeasna Shanjana, and Md Rabiul Islam. 2025. Use of artificial intelligence in mental healthcare, health psychology, and related research: A narrative review to address challenges and opportunities. *Health Science Reports*, 8(12):e71595.

María E Raygoza-L, Jesús Heriberto Orduño-Osuna, Roxana Jimenez-Sanchez, and Fabian N Murrieta-Rico. 2025. Innovative artificial intelligence approaches for identifying and managing dsm cluster b personality disorders in mental health: A case study on the dark triad. In *Exploring Psychology, Social Innovation and Advanced Applications of Machine Learning*, pages 1–20. IGI Global Scientific Publishing.

Minhah Saleem and Jihie Kim. 2024. Intent aware data augmentation by leveraging generative ai

for stress detection in social media texts. *PeerJ Computer Science*, 10:e2156.

Bosubabu Sambana, Kondreddygari Archana, Suram Indhra Sena Reddy, Shaik Meethaigar Jameer Basha, and Shaik Karishma. 2025. Data augmentation for cognitive behavioral therapy: Leveraging ernie language models using artificial intelligence. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 204–209. IEEE.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Yuhan Wang, Imran Razzak, and Shoaib Jameel. 2025. Ll4g: Self-supervised dynamic optimization for graph-based personality detection. *arXiv preprint arXiv:2504.02146*.

Zhivar Sourati, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Nuan Wen, Ala Tak, Fred Morstatter, and Morteza Dehghani. 2024. Secret keepers: The impact of llms on linguistic markers of personal traits. *arXiv preprint arXiv:2404.00267*.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2025. Challenging the validity of personality tests for large language models. In *Proceedings of the 2025 Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 74–81.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133.

Avraham Treistman, Dror Mughaz, Ariel Stulman, and Amit Dvir. 2022. Word embedding dimensionality reduction using dynamic variance thresholding (DyVaT). *Expert Systems with Applications*, 208:118157.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and spaCy*. Packt Publishing. Accessed: December 6, 2025.

Michael Veale and Borgesius Frederik Zuiderveen. 2021. Demystifying the draft eu artificial intelligence act. *Computer Law Review International*, 22(4):97–112.

G Venkateshwarlu, S Akhila, V Kavyasree, S Vishnu, and VS Prasad. 2024. Enhanced text classification using random forest: Comparative analysis and insights on performance and efficiency. *Int. J. Comput. Eng. Res. Trends*, 11:1–8.

Sumona Yeasmin, Nazia Nowshin, and Tasnia Afrin Chowdhury. 2024. Identifying human dark triad from text data through machine learning models. *International Journal of Research and Innovation in Applied Science*, 9(6):89–104.

Jianlong Zhou and Fang Chen. 2023. Ai ethics: From principles to practice. *Ai & Society*, 38(6):2693–2703.