

Master Thesis: Fusion of transformer-based language representations with psycholinguistic LIWC features for improved detection of cybergrooming in chat logs

Master thesis by Amelie Oberkirch
Date of submission: 13. October 2025

1. Review: Shiying Fan
2. Review: Prof. Dr. Martin Steinebach
Darmstadt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Informatik
Fraunhofer-Institut für
Sichere
Informationstechnologie
Media Security

Contents

1 Abstract	7
2 Introduction	8
3 Related Work	9
3.1 Understanding Cybergrooming	9
3.2 Stages of Cybergrooming	10
3.3 Towards Automated Detection of Online Grooming	11
3.3.1 Sexual Predator Identification	11
3.3.2 Grooming Process Modeling	12
3.3.3 Early Grooming Detection	12
3.3.4 Machine Learning Approaches	12
3.4 Transformer Models for Text Classification	14
3.4.1 Extension to cross-attention in transformer architectures.	14
3.5 BERT	15
3.6 Feature Fusion with BERT	16
3.6.1 Classic fusion approaches	16
3.6.2 Cross-attention as a key mechanism	17
3.7 Transformer-based Methods for Cybergrooming Detection	17
3.8 Linguistic Inquiry and Word Count (LIWC)	19
3.8.1 LIWC-22 Categories	19
3.8.2 Psychometric Profiling with LIWC	20
3.9 LIWC in Cybergrooming Research	20
3.9.1 Key psychometric LIWC Features in Grooming	21
3.10 Explainable AI with SHAP	21
3.11 Integrating Psycholinguistic Features and Explainable AI for Enhanced Grooming Detection . .	22
4 Dataset and Preprocessing	24
4.1 PAN12 Dataset	24
4.2 Perverted Justice (PJ) Dataset	25
4.3 Slang and Informal Language Handling	27
4.4 Raw Conversational Length Distributions	28
4.5 Generating Synthetic Non-Grooming Data	28
4.6 Final Dataset	29
5 Methodology	30
5.1 Initial Chunking Strategy for BERT Fine-tuning	30
5.1.1 Train-Test Split	30

5.2	BERT fine-tuning as Baseline	30
5.2.1	Evaluation Metrics	31
5.3	Limitations of Initial BERT Fine-tuning Approach	32
5.4	Chunking Strategy for BERT to reduce Domain and Length Leakage	32
5.4.1	Chunk-Length Distributions (For Train and Test) Across Sequence Lengths	32
5.4.2	Data Distributions after Chunking	34
5.5	Improved Training Pipeline	35
5.5.1	Improved Training Configuration	35
5.5.2	Evaluation Strategy with Additional Subsets	35
5.6	Choosing the final configuration for Feature Fusion and SHAP Explainability Analysis	36
5.7	Choosing the LIWC Feature Set for Feature Fusion and SHAP Explainability Analysis	36
5.8	LIWC Data Extraction	37
5.9	LIWC Data Analysis	37
5.9.1	Global LIWC Analysis on Complete Conversations	37
5.9.2	Chunk-based LIWC Analysis with 512 Tokens	39
5.10	Feature Fusion Strategy	39
5.10.1	Model Architecture	39
5.10.2	Training Pipeline and Evaluation	42
5.11	Explainability Analysis based on Feature Fusion Model	42
5.11.1	SHAP analysis of LIWC features	42
5.11.2	Confidence Analysis based on Label Flip and Confidence Shift	43
5.12	LIWC Analysis of Misclassifications	44
5.12.1	Descriptive per-group statistics	45
5.12.2	Per-feature group comparisons	45
5.12.3	Proximity hypothesis tests in LIWC feature space	45
5.12.4	SHAP-Based Proximity Analysis of Top-20 Misclassifications	46
6	Evaluation	48
6.1	Initial BERT Finetuning Results	48
6.2	Bert Finetuning Results on different Chunk Sizes and Data Setups	49
6.2.1	Fixed Chunk Size of 150 chunks	49
6.2.2	Fixed Chunk Size of 250 chunks	49
6.2.3	Fixed Chunk Size of 512 chunks	50
6.2.4	Confusion Matrices for BERT Baseline across Chunk Sizes and Data Setups	51
6.2.5	ROC Curves for BERT Baseline across Chunk Sizes and Data Setups	52
6.3	Comparing LIWC-2022 Macro Groups	53
6.4	Comparing LIWC Features between PJ and PAN12 on Full Conversations	54
6.5	Comparing LIWC Features between PJ and PAN12 and the Synthetic Dataset	55
6.6	Chunk-based LIWC Analysis	56
6.7	Feature Fusion Evaluation	58
6.7.1	Using all LIWC-2022 Features	58
6.7.2	Using a Subset of Psychometric LIWC-2022 Features	59
6.7.3	Confusion Matrices for Feature Fusion Models	60
6.7.4	ROC Curves for Feature Fusion Models	61
6.8	Ablation Studies based on SHAP	62
6.8.1	LIWC-Feature Importance Ranking	62
6.9	Confidence Analysis and Label Flip Analysis	68

6.10	Missclassification Analysis based on LIWC-Scores	72
6.10.1	Results Summary: Full LIWC feature set	72
6.10.2	Results Summary: Psychometric LIWC subset	75
6.11	Shap-based Analysis of Missclassification	78
6.11.1	Total LIWC Feature Set	78
6.11.2	Psychometric LIWC Subset	80
7	Discussion	82
7.1	Main Findings and Interpretation	82
7.1.1	Addressing Data Leakage and Shortcut Learning	82
7.1.2	Addressing Baseline Performance	83
7.1.3	Feature Fusion Architecture (BERT + LIWC via Late Fusion with Cross-Attention)	83
7.1.4	Performance Gains from LIWC Integration (Full Set vs. Psychometric Subset)	84
7.1.5	Model Confidence and Stability	84
7.1.6	LIWC-Features in Grooming and Non-Grooming Chats	85
7.2	Limitations	87
7.3	Future Work	87
8	Appendix	94

List of Figures

4.1	Comparison of LIWC macro-groups	28
5.1	Chunk Length Distributions	34
5.2	Feature Fusion Architecture	41
6.1	Confusion matrices of BERT baseline after epoch 3.	51
6.2	Zoomed ROC curves after epoch 3 for different chunk lengths.	52
6.3	Comparison of aggregated LIWC macro-groups	53
6.4	LIWC-Feature Comparison (PJ, PAN12) over Complete Conversations	54
6.5	Top 30 LIWC Differences with Synthetic Baseline	56
6.6	LIWC Feature Comparison at Chunk Level	56
6.7	Confusion matrices by epoch for feature-fusion models.	60
6.8	Zoomed ROC curves for feature-fusion models across epochs.	61
6.9	Cumulative importance of LIWC-2022 features.	63
6.10	Top 20 LIWC features ranked by percentages of the total significance.	65
6.11	Top 20 LIWC features ranked by mean signed SHAP value.	67
6.12	Confidence shift analysis with LIWC features.	69
6.13	Agreement matrices for predictions with vs. without LIWC features.	70
6.14	Volcano plots for all LIWC-2022 features.	72
6.15	Volcano plots for psychometric LIWC features.	75
6.16	Signed proximity plots for all LIWC-2022 features.	78
6.17	Signed proximity plots for the psychometric LIWC-2022 subset.	80

List of Tables

5.1	Initial data distribution after chunking	30
5.2	Token statistics per conversation/file across datasets	33
5.3	Split for Chunk-Length of 150 Tokens	34
5.4	Split for Chunk-Length of 250 Tokens	34
5.5	Split for Chunk-Length of 512 Tokens	34
6.1	Evaluation of initial BERT base model	48
6.2	Evaluation for Chunk Size 150, Separated Split	49
6.3	Evaluation for Chunk Size 150, Mixed Split	49
6.4	Evaluation for Chunk Size 250, Separated Split	49
6.5	Evaluation for Chunk Size 250, Mixed Split	50
6.6	Evaluation for Chunk Size 512, Separated Split	50
6.7	Evaluation for Chunk Size 512, Mixed Split	50
6.8	Evaluation: Feature Fusion with all LIWC-2022 Features	58
6.9	Evaluation: Feature Fusion with LIWC-2022 Subset	58
6.10	Confidence and label flip analysis for all LIWC-2022 features vs. psychometric subset.	70
6.11	Summary of misclassification analysis with all 118 LIWC features.	74
6.12	Summary of misclassification analysis with psychometric LIWC subset (49 features).	76
6.13	Key proximity results (Top-K of all LIWC features)	80
6.14	Key proximity results (Top-K psychometric LIWC features)	81
8.1	LIWC-22 categories with abbreviations and exemplar words.	95
8.2	Complete LIWC baseline results	97
8.3	LIWC baseline results (based on chunks)	99

1 Abstract

2 Introduction

In today's world, social media are an integral part of the lives of children and young people and offer extensive opportunities for global networking and the exchange of ideas. In addition to the great potential of the latest technologies, social media also pose many dangers, especially for children and young people. Cybergrooming represents a significant danger in this regard - whereby minors are often contacted and emotionally influenced by adults using communicative manipulation in order to initiate or prepare sexual contact.

3 Related Work

This chapter introduces the term “cybergrooming” and relevant research in the field of cybergrooming and cybergrooming detection. The first section provides an overview of the psychological and communicative dynamics of cybergrooming. The second section additionally focuses on technical approaches to detecting grooming behavior in online conversations, with an emphasis on machine learning methods and their integration with psycholinguistic features.

3.1 Understanding Cybergrooming

The terms “online grooming” and “cybergrooming” are often used synonymously to describe the process by which an adult establishes a digital relationship with a minor in order to achieve and facilitate subsequent sexual exploitation. In academic literature, “online grooming” is the internationally established term for this phenomenon, particularly in the fields of psychology, criminology, and child protection. However, the synonymous term “cybergrooming” is more commonly used in media education and German-language legal discourse, with an emphasis on the technological environment in which the grooming takes place [1, 2]. **To maintain terminological clarity, the term “cybergrooming” is used throughout this thesis.**

Cybergrooming can be seen as a psychological manipulation process in which a perpetrator uses digital communication technologies to befriend a minor, initiate sexual interactions online, and/or arrange a physical meeting with the aim of sexual abuse [3]. This process usually takes place in several stages, in which trust is gradually built up, emotional dependence is created, and the minor is desensitized to sexual content or requests [4]. Groomers often present themselves as friendly and understanding, using engaging and emotionally supportive language rather than openly deceiving [5]. This process is facilitated by digital media, which allow perpetrators to remain anonymous and gain unrestricted access to children without direct adult supervision [1].

It is important to note that online grooming does not refer to the sexual abuse itself, but rather to the preparatory phase and psychological manipulation. The perpetrator uses targeted strategies to influence the victim’s thoughts, emotions and decision-making processes in such a way that they serve his own intentions [2]. This grooming dynamic is not limited to online contexts. The broader term “grooming” refers to any deliberate strategy used by perpetrators to emotionally condition and exploit defenseless minors and is a significant mechanism in cases of sexual abuse [6].

While there is agreement on the manipulative nature of cybergrooming, researchers have developed various models to represent its dynamics, communication strategies and stages of development[7]. These models have been proposed to conceptualize the process of cybergrooming and often describe it as a multi-stage process involving various phases such as friendship formation, relationship building, exclusivity, risk assessment, sexual interaction and conclusion. Importantly, research suggests that these phases do not necessarily occur in

a fixed or linear order and that perpetrators may skip phases, return to earlier phases, or adapt their strategies depending on the victim and context. For example, Black et al. [8] conducted a qualitative content analysis of Facebook chat logs and identified linguistic patterns characteristic of different phases of grooming, providing empirical evidence for the nonlinear nature of the grooming process.

To illustrate different conceptual approaches and understand the grooming process, the following section focuses on two well-known models: O'Connell's stage-based framework and Gupta's linguistic profiling approach.

3.2 Stages of Cybergrooming

A key feature of grooming is its gradual and often strategically planned progression. Many studies describe grooming not as a one-time event, but as a multi-phase communication process in which perpetrators consciously build trust, establish emotional bonds and gradually introduce sexual content. The following study's illustrate this process using various phase models based on annotated chat logs. Despite their theoretical relevance, however, the empirical generalizability of these models remain controversial.

A frequently cited model comes from O'Connell [9], who identified six consecutive phases: *friendship building*, *relationship building*, *risk assessment*, *exclusivity*, *sexuality* and *conclusion*. These phases form a kind of "linear roadmap" for the grooming process. The communication strategies of the phases range from small talk and compliments to risk assessments and explicit sexual innuendo or appointments. Although O'Connell's model has greatly influenced the conceptual understanding of grooming dynamics, it has been criticized for its limited empirical basis. For example, Broome et al. [5] and Lorenzo-Dus and Kinzel [10] point to methodological ambiguities regarding sample size, perpetrator demographics and the structural characteristics of the conversations analyzed.

Furthermore, the assumption of a fixed sequence has been challenged by subsequent empirical findings suggesting that grooming often develops not linearly but with overlapping or cyclical phases [11]. Recent sequence analyses show that perpetrators often use trust-building measures, sexualization and risk assessment in parallel rather than in a strict sequence [12].

An alternative, linguistically based model was proposed by Lorenzo-Dus et al. [13]. Based on the analysis of 24 chat logs from the online archive Perverted-Justice.com, the authors divide grooming into three dynamic phases: *access*, *approach* and *entrapment*. In the entrapment phase in particular, linguistic strategies often overlap, making it difficult to draw clear boundaries between the phases. The authors identify four communication processes: deceptive trust development, sexual gratification, isolation and compliance testing. These are realized through strategic language use, such as praise, desensitization and personal disclosures. Politeness strategies in particular play a central role in manipulating the victim by simultaneously signaling control and trust. However, despite its depth, the study remains limited in its generalizability as it relies on a purely qualitative methodology and is based on lure chats with adult volunteers. More recent work shows that lure chats differ significantly from real victim interviews in key aspects such as threats, coercion and intensity of risk questioning [12].

Another influential model was developed by Beech and A [14], who analyzed convicted sex offenders in New Zealand. Their three-stage typology—*relationship building*, *sexual content* and *assessment*—emphasizes the dynamics of early manipulation. According to Beech and A [14], in the *relationship building* phase, offenders adapt their language to that of adolescents in order to create a feeling of closeness. The *sexual content* phase

often begins subtly, for example through games or advice and ends with explicit offers. The *assessment* phase serves to continuously evaluate risk and create an emotional profile of the victim.

A psychologically based extension is also provided by Olson's *Luring Communication Theory* (LCT), which was originally designed for offline grooming but can also be applied to online contexts [15]. This model comprises three phases: *Deceptive Trust Development*, *Grooming* and *Physical Approach*. In the first phase, an emotional connection is established through personal exchange (e.g., conversations about age or interests). This is followed by increasingly sexualized communication, which eventually transitions into a phase in which physical contact is initiated.

Taken together, these models show that grooming processes often proceed in strategically structured phases that are realized linguistically through specific communicative actions.

3.3 Towards Automated Detection of Online Grooming

The sheer volume of modern online communication makes manual monitoring of cybergrooming in online chats, social media and online games virtually impossible. Automated detection systems are therefore essential for filtering and categorizing suspicious dialogues in real time. [16] To support timely detection and reduce the burden on victims, current research is increasingly focused on developing automated systems capable of detecting grooming behavior in online conversations. These systems aim to flag suspicious interactions in real time or near real time, enabling efficient protective measures in digital environments.

Over the past decade, a variety of approaches have emerged that rely on classical machine learning, the extraction of linguistic and behavioral features and modern deep learning techniques. Broadly speaking, these approaches to cybergrooming can be divided into three main paradigms: (1) *Identification of sexual predators*, where the goal is to classify entire conversations as grooming or non-grooming; (2) *Modeling the grooming process*, where the focus is on identifying the phases and strategies of grooming; and (3) *Early detection* of grooming, with the goal of identifying harmful intentions in the earliest stages of communication. The following sections present studies from each of these areas, highlighting their methodology, datasets and key findings.

3.3.1 Sexual Predator Identification

Villatoro-Tello et al. proposed a two-stage framework specifically aimed at identifying sex offenders who engage in grooming behavior in online chat environments [17]. Their approach first filters suspicious conversations based on lexical, stylistic and emotional characteristics and then classifies individual messages according to different grooming phases. Based on the PAN12 dataset for identifying sexual offenders, their model captures the sequential and manipulative nature of offenders' communication. Broome et al. supported this approach and conducted a comprehensive review of machine learning methods for identifying sex offenders [5]. They confirmed that the most effective approaches are based on linguistic features and emphasized the growing importance of deep learning models. In particular, they highlight models that can capture contextual and temporal dependencies. Both studies underscore the value of stage-aware and linguistic modeling in the development of robust systems for detecting grooming, while also pointing to the ethical necessity of transparency and explainability in high-risk applications.

3.3.2 Grooming Process Modeling

The understanding of cybergrooming as a sequential and manipulative process has led to many efforts to model its phases. For example, Gupta et al. conducted one of the earliest qualitative studies, using the Perverted Justice dataset to analyze grooming behavior in chat logs [18]. They based their work on O'Connell's six-phase model and manually labeled the perpetrators' messages accordingly. Their findings underscored the structured nature of grooming strategies and laid important groundwork for subsequent modeling efforts.

Building on this foundation, Cano et al. operationalized Olson's Luring Communication Theory by defining three grooming phases themselves: Deceptive Trust Development, Grooming and Seeking Physical Approach [15]. They used linguistic, psycholinguistic and discourse-based features, including LIWC and sentiment analysis, to classify messages into the three phases using SVMs. Their results showed that discourse features were particularly effective in describing the progression of grooming behavior.

Furthermore, Black et al. emphasized the nonlinear and dynamic nature of grooming, highlighting that perpetrators often skip, repeat, or adapt phases depending on context and interpersonal factors [8]. This finding underscores the importance of process-based modeling over static classification.

3.3.3 Early Grooming Detection

Efforts to detect grooming at an early stage often focus more on classifying individual messages rather than entire conversations. One such approach is presented by Isaza et al., who developed a CNN-based model trained on the PAN12 dataset to classify short messages as grooming or non-grooming [19]. Their architecture utilizes semantic features through pre-trained word embeddings (Word2Vec) and applies convolutional filters of various sizes to detect n-gram-like patterns. Although the model achieved a high recall rate, it suffered from low precision due to class imbalance. Nevertheless, it proved useful for early intervention as it correctly marked a large number of true positives.

Another relevant contribution comes from Schläpfer et al. [20], which dealt with machine learning methods for the early detection of sex offenders in online chats. Using pre-trained language models and ensemble classifiers, their approach focused on identifying grooming behavior at the conversation level. Although their work did not involve manual annotation at the message level, it showed that linguistic signals embedded in early parts of conversations can be used to distinguish chats from sex offenders from harmless chats. Their findings underscore the potential of automated systems to detect grooming attempts in their early stages and contribute to the development of real-time safety measures.

3.3.4 Machine Learning Approaches

While early research focused on the classification of perpetrators and the modeling of phases, modern work reflects a broader methodological field, often motivated by the need for scalable and interpretable detection systems. In their systematic review, An et al. [21] emphasize that machine learning approaches play a central role in the automated detection of grooming, particularly due to their ability to process large amounts of online text. However, they caution that most systems remain purely reactive and lack integration with social science insights into the dynamics of grooming. To close this gap, hybrid models are needed that combine the statistical power of machine learning with insights from psychology and criminology.

From this perspective, Gunawan et al. [22] investigated behavioral indicators by training a classifier on the Perverted Justice dataset. Features such as message length, conversation turns and response time were used to capture conversation flow. Their results suggest that these features, combined with oversampling techniques such as SMOTE, can serve as strong predictors of predatory intent.

To capture the subtleties of psychological manipulation, Cook et al. [23] introduced a hybrid annotation approach involving psychologists who manually labeled over 6,000 chat messages according to eleven predefined grooming strategies. These annotations were then used to train deep learning models capable of detecting finely tuned manipulation tactics. The study not only demonstrated the effectiveness of such models, but also emphasized the importance of human expertise in capturing contextual meaning, further underscoring the need for hybrid human-AI systems in automated cybergrooming detection.

In addition, Leiva-Bianchi et al. [24] conducted a comparison of machine learning classifiers using the PAN12 dataset. They evaluated classical algorithms, ensemble methods and neural networks using lexical features such as bag-of-words, TF-IDF and N-grams. Ensemble approaches, in particular Random Forest and Gradient Boosting, proved to be the most robust across all evaluation metrics, highlighting the value of combining different weak learners for detecting grooming in chat logs.

Preuß et al. [25] went one step further and proposed a two-stage detection method that combines convolutional neural networks (CNNs) with a multilayer perceptron (MLP). Their system, trained on PAN12, analyzed lexical patterns using CNNs and supplemented them with behavioral signals such as mood, reaction time and message frequency. A manually created standard for line-level annotations enabled not only conversation-level classification but also the identification of semantically relevant messages, resulting in very good performance on several detection tasks. Similarly, Hamm and McKeever [16] investigate machine learning for grooming detection with a particular focus on conversational tone. Using the PAN12 dataset, they compare traditional models (SVM) with the large language model LLaMA 3.2 and show that positively toned, emotionally warm strategies are used more frequently by groomers and are also easier to detect than negatively phrased behaviors. By combining sentiment analysis (DistilBERT) with classifier comparisons, their results show that large language models not only achieve higher F1 scores but are also better able to capture finer linguistic patterns, further underscoring the value of psychological modeling approaches.

Together, these studies highlight the increasing complexity of machine learning methods in detecting grooming. However, as An et al. [21] point out, a key challenge remains the lack of interdisciplinary integration. Many models operate in isolation from psychological theories or social contexts, limiting their ability to generalize to real-world scenarios. The following sections therefore examine new transformer-based language models and explore how psycholinguistic profiling tools such as LIWC can be used to bridge this gap between behavioral knowledge and algorithmic accuracy.

3.4 Transformer Models for Text Classification

Transformer models were firstly introduced by Vaswani et al. [26] in their paper *Attention Is All You Need*. After being introduced, transformers became a key component in many state-of-the-art NLP models. Thanks to their design, Transformers are both highly effective for a wide range of NLP tasks and computationally efficient to train. Their success has led to a lot of variants, such as XLM [27], GPT [28], and XLNet[29]. Thereby, the central characteristic of Transformer Models lies on their reliance, unlike previous sequence models based on recurrence or convolution, entirely on an **attention mechanisms**, enabling greater parallelization and more effective modeling of long-range dependencies.

Therefore, the core component is called the **self-attention** mechanism, which computes a weighted representation of each input token by attending to all other tokens in the sequence. Specifically, the model uses **scaled dot-product attention**, where the attention weights are being calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q , K , and V represent query, key, and value matrices derived from the input. This allows the model to dynamically determine **contextual relevance** across the entire sequence [26].

To enhance the model's capacity, the Transformer applies **multi-head attention**, which projects the input into multiple subspaces and performs attention in parallel. This enables the model to capture different types of relationships simultaneously [26].

Transformers use a stack of encoder and decoder layers, each composed of multi-head self-attention, feed-forward networks, and residual connections with layer normalization. Since no recurrence is used, positional encodings based on sinusoidal functions are added to the input embeddings to encode word order vaswani2017attention.

The Transformer architecture has been widely adopted in a lot of NLP tasks, including machine translation, text classification and question answering. Its ability to model long-range dependencies and its parallelizable structure have made it a foundational model for many subsequent advancements in the field.

3.4.1 Extension to cross-attention in transformer architectures.

While self-attention enables a model to capture relationships within a single sequence, cross-attention extends the concept to the interaction between different sources of information [26]. The key change here lies in the origin of the query, key, and value matrices. In cross-attention, the queries (Q) originate from one modality or information source, while keys (K) and values (V) are created from another source [26]. This makes it possible, for example, to draw the attention of text representations to additional features. The mathematical formulation remains unchanged:

$$\text{CrossAttention}(\mathbf{Q}_{\text{text}}, \mathbf{K}_{\text{feature}}, \mathbf{V}_{\text{feature}}) = \text{softmax}\left(\frac{1}{\sqrt{d_k}} \mathbf{Q}_{\text{text}} \mathbf{K}_{\text{feature}}^\top\right) \mathbf{V}_{\text{feature}}.$$

Integration is often achieved via residual connections, allowing the model to dynamically decide when and to what extent the additional information should influence the original representation [30]. This flexibility

makes cross-attention a powerful tool for multimodal learning, where information from different modalities (e.g., text, images, audio) needs to be integrated effectively [31].

3.5 BERT

BERT (Bidirectional Encoder Representations from Transformers), was first introduced by Devlin et al. Devlin et al. [32] and is one of the most influential pre-trained language models which builds upon a multi-layer bidirectional Transformer encoder. Unlike unidirectional models such as GPT [28], which process text left-to-right, BERT jointly conditions on both left and right context, enabling it to capture deeper semantic relationships between words [33].

To achieve this, BERT was pre-trained using two unsupervised objectives: **Masked Language Modeling** (MLM) and **Next Sentence Prediction** (NSP). In MLM, random tokens in a sentence are masked, so that the model learns to predict them based on the surrounding context [32]. This prevents the model from only "seeing" the answer, enabling deep bidirectional context representations. NSP, in turn, helps the model understand inter-sentential relationships by training it to determine whether one sentence logically follows another. This has proven essential for a lot of downstream tasks such as Natural Language Inference (NLI) and Question Answering (QA) [34].

BERT was trained on a large scale, document level corpora including BooksCorpus (800M words) and English Wikipedia (2.5B words), avoiding shuffled sentence level datasets like the Billion Word Benchmark to preserve long range dependencies [34]. Unlike earlier approaches that entirely transferred word embeddings, BERT allows full parameter fine-tuning, making it highly adaptable for task specific applications [35]. This architectural design leads to better performance in several NLP tasks such as text classification, sentiment analysis and false news detection. Studies have shown that BERT outperforms prior models even with negligible task specific adjustments. For example Qasim et al. [36] found that BERT base and BERT large achieved modern accuracy across multiple real world datasets, highlighting its strength in classification based applications.

In addition to BERT, many language models such as RoBERTa [37] and DeBERTa [38] are based on the Transformer architecture and build methodologically on BERT. RoBERTa ("Robustly Optimized BERT Pre-training Approach") was trained on a larger training corpus than BERT and, unlike BERT, does not use the next sentence prediction task and optimizes the masking procedure. As a result, RoBERTa achieves better results on various benchmarks such as GLUE, RACE and SQuAD and is particularly preferred for demanding classification tasks [37]. DeBERTa, on the other hand, integrates further innovations, such as disentangled attention and a redesigned masked decoding procedure, which further improve the model's capacity and generalizability [38, 39].

Fine-tuning mechanism for classification tasks Fine-tuning BERT is a two-step process that adapts the pre-trained model to specific downstream tasks. In the first step, the entire BERT model is initialized with its pre-trained parameters, followed by an adjustment of all parameters using supervised learning on a labeled target dataset [32, 34]. This approach differs fundamentally from feature extraction, where the BERT parameters remain frozen and only an additional classification layer is trained [40]. Studies show that fine-tuning is better than feature extraction in most cases because it allows the model to perfectly tailor its internal representation to the target task [34, 40]. BERT performs particularly well in binary classification tasks with balanced datasets due to its bidirectional context modeling. [41].

The effectiveness of BERT in binary tasks is further enhanced by the fact that the model develops specific linguistic information in different layers during the fine-tuning process. Peters et al. [40] used mutual information analyses to show that in binary classification tasks, task-relevant knowledge is mainly concentrated in the upper layers of the model, while in more complex sentence pair tasks, information is gradually built up across middle and upper layers [40]. This finding explains why BERT performs particularly well in straightforward binary decisions. In addition, recent studies confirm that BERT-like models remain superior in many binary classification tasks, especially pattern-driven tasks, even in the age of large language models, while requiring fewer computational resources than modern LLMs [42]. However, it is important to note that the optimal configuration of the fine-tuning process, including consideration of the 512-token limit, appropriate learning rates, and batch sizes, is crucial for achieving very good results [34, 41].

3.6 Feature Fusion with BERT

Feature fusion refers to the combination of multiple information sources to train a model. The idea behind this is that more robust representations can be generated than would be possible with a single source of information, since combined features are more meaningful than individual ones. The principle of feature fusion is one of the central principles of modern deep learning models, especially in the context of multimodal architectures [31]. Transformer models such as BERT are particularly well suited for integrating additional information such as LIWC features, POS tags, or N-gram statistics due to their modular architecture and powerful context modeling [43].

3.6.1 Classic fusion approaches

A basic distinction is made between three types of feature fusion, which differ mainly in the timing of integration [31, 44]:

Early fusion: Features from different sources are merged at the input level (e.g., through concatenation or summation) and then processed together by a model. This allows interactions between modalities to be mapped early on, but is susceptible to noise in individual modalities [30].

Late Fusion: Here, the modalities are first processed separately and then combined at the decision level, for example, through weighted averaging, voting, or a special fusion layer. [45]. In multimodal settings, late fusion also enables interpretability using SHAP, for example the class-wise contribution of each uni-modal predictor can be quantified and contrasted even at the cost of an additional fusion stage [46].

Hybrid/Hierarchical Fusion: This approach combines early and late fusion, allowing iterative fusion at different levels of the network. This allows the advantages of both strategies to be exploited, but comes at the cost of higher costs and significantly greater architectural complexity, making interpretability more difficult [31].

Advanced work further distinguishes between “early-to-mid fusion” (integration in lower to middle layers), “mid-to-late fusion” (integration after the first stage but before the end of the model), or “deep hierarchical fusion” (multiple integration across different layers) depending on the chosen integration point [31].

3.6.2 Cross-attention as a key mechanism

A central trend in current research is the use of cross-attention to improve feature fusion. Cross-attention can be used flexible in all fusion approaches (early/mid/hybrid) and allows flexible interaction between modalities. [43, 47].

In the context of BERT, this means that the hidden states of the text representation can specifically pay attention to additional compressed features [30, 48]. This creates a context-dependent interaction that weights additional information without disrupting the proven language modeling [30]. In Cross-Attention, the integration is typically achieved with a gating system that adds the cross-attention output ($H' = H + g \odot \text{Attn}(H, t)$) in a way that additional information is only incorporated if it is actually informative. This ensures that the model can fall back on the original BERT representation if the additional features are not helpful [30].

3.7 Transformer-based Methods for Cybergrooming Detection

Transformer-based models have fundamentally transformed the field of automated cybergrooming detection, decidedly outperforming traditional machine learning approaches in recent years. The introduction of BERT (Bidirectional Encoder Representations from Transformers) and its subsequent improvement in RoBERTa (Robustly Optimized BERT Pretraining Approach) enabled robust semantic and syntactic modeling of chat conversations, even in the presence of informality or unstructured grammar, as frequently encountered in online communication. Research shows that BERT- and RoBERTa-based frameworks achieve high accuracy and robustness in detecting grooming behaviors, both in early and aggregate phases of chat dialogues [49].

In direct comparison, RoBERTa has been shown to outperform not only classical classifiers but also BERT itself in the context of online grooming detection [50]. A likely reason for this advantage is RoBERTa's more extensive and diverse pretraining corpus, coupled with the removal of the Next Sentence Prediction (NSP) objective, which allows for richer contextual representations [37].

At present, DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is considered among the leading architectures for nuanced text classification in social media contexts. Hassan et al. [51] demonstrate that ensembles of fine-tuned DeBERTa models achieve state-of-the-art results on complex, fine-grained classification tasks involving harmful online content, such as online sexism. Their approach, which systematically compares DeBERTa against other popular transformer models including BERT, RoBERTa, and XLM-RoBERTa, consistently shows that DeBERTa delivers the best performance in terms of precision, recall, and F1-score across all levels of granularity in the shared task.

A core limitation of transformer-based models such as BERT, RoBERTa, and DeBERTa is their maximum input length, typically 512 tokens. Chat conversations in the context of cybergrooming, however, often greatly exceed this limit. To address this, research such as Vogt et al. [49] propose a sliding window approach: long chats are divided into overlapping segments, each of which is analyzed separately by the model, and the results are then aggregated for a robust prediction over the entire conversation. This technique enables transformers to process arbitrarily long dialogues and is now standard for grooming detection on real data streams.

Despite this methodological progress, an important challenge remains. most transformer models excel at identifying *what* is said in chat conversations but base their decisions primarily on surface-level features such as token frequencies, syntactic patterns, and keywords. Their “black box” nature limits interpretability.

Psycholinguistic features, for example, cues related to trust building, deception, emotional manipulation, or the progression of grooming phases are rarely modeled explicitly. Broome *et al.* and Street *et al.* highlight that discursive and psychological dynamics central to grooming are often marginalized in favor of quantitative, corpus-driven features [5], [50]. Therefore, a crucial challenge remains in the field of cybergrooming detection. While models such as BERT, RoBERTa, and DeBERTa reliably classify what is said in chats, their decisions are largely grounded in superficial patterns like token frequencies, syntactic structures, and specific keywords. As highlighted by Mersha *et al.* [52] in their comprehensive analysis of explainability in neural NLP models, these transformers still lack genuine transparency—their inherent “black box” nature makes it difficult for users to understand or validate the basis of any given decision. This limitation is especially apparent in sensitive domains such as medical or social text, where even fine-tuned BERT models often remain hard to interpret [53].

In socially and psychologically complex tasks such as cybergrooming detection, where cues about trust-building, deception, manipulation, or progression across grooming phases are central for meaningful interpretation—a significant depth of psycholinguistic information remains underutilized. Broome *et al.* and Street *et al.* show that relevant discursive and psychological dynamics of grooming are often lost within purely quantitative, corpus-driven features [5], [50].

Recent explainability frameworks for language models specifically address this gap: Ribeiro *et al.* [54] present a methodology that combines techniques such as Integrated Gradients with psycholinguistic tools like LIWC. This hybrid approach not only clarifies *which* linguistic units drive model decisions, but also identifies the crucial psycholinguistic and emotional categories—such as trust markers and affectivity—that influenced those predictions. Methods such as SHAP or the explicit integration of LIWC attributions can thus substantially improve both model transparency and the psychological interpretability of transformer-based predictions.

This leads directly to the research focus of the present work. By strategically combining transformer architectures with psycholinguistic analysis tools like LIWC and explainability frameworks such as SHAP, there is a chance to strengthen both the predictive power and the interpretative depth of online grooming detection systems.

3.8 Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) was originally developed in the early 1990s by James W. Pennebaker and colleagues as a theory-driven tool for the analysis of written and spoken language [55]. The main goal of LIWC is to extract insights into psychological states, attitudes, and personality traits of the writer based on a predefined dictionary that maps words to psychologically meaningful linguistic categories. LIWC analyzes text by counting the frequency of words in these categories, providing a quantitative representation of the language used. This allows researchers to draw conclusions about the psychological and social dimensions of the text. The construction of the LIWC dictionary and category system followed an iterative process: initially, relevant words were selected based on psychological theory and expert evaluation. These were then revised, adapted, and psychometrically validated through continuous testing and updated over time in response to new research findings and computational methods. The LIWC dictionary has been continuously refined, with its most recent version released in 2022. [56].

3.8.1 LIWC-22 Categories

The main categories in LIWC-22 are structured as follows:

- **Linguistic dimensions:** Cover syntactic and functional aspects such as personal pronouns (e.g., first person singular/plural), articles, auxiliary verbs, verb tenses, negations, and other grammatical features.
- **Psychological processes:**
 - *Affective processes:* Positive and negative emotions, anger, anxiety, sadness, and swearing.
 - *Cognitive processes:* Insight, causation, discrepancies, tentativeness, certainty, differentiation, memory, and analytical thinking.
 - *Social processes:* Family, friends, references to males/females, group membership, communication terms, politeness, moralization, prosocial behavior.
- **Personal concerns:** Work, money, religion, home, leisure, sexuality, politics, ethnicity, and technology.
- **Biological/physical processes:** Body, health, illness, wellness, mental states, food, substances, death.
- **Cultural and digital phenomena:** Culture-related categories (such as politics, ethnicity), netspeak, conversation markers, and emojis.
- **Summary variables:** Analytic thinking, clout (social status), authenticity, and emotional tone—these provide higher-order, composite scores derived from subcategory frequencies.
- **Other metrics:** Additional features such as total word count, average words per sentence, dictionary word proportion, and the percentage of "big words" (long words) are also included.

Each word in the dictionary may belong to multiple categories simultaneously. For example, the word "cry" is counted within *negative emotion*, *sadness*, and *past tense* categories. LIWC's multidimensional approach enables detailed quantitative assessment of the psychological and social dimensions hidden within texts, making it a valuable tool across psychology, social science, computer science, and digital communication research [56].

3.8.2 Psychometric Profiling with LIWC

LIWC has become a cornerstone of psychological and behavioral language research due to its ability to extract psychologically meaningful dimensions from text. As Tausczik and Pennebaker [55] emphasize, LIWC bridges the gap between language and psychology by mapping linguistic usage onto cognitive, emotional, and social constructs. This enables researchers to quantify and interpret even subtle psychological processes on a large scale. This capability makes LIWC indispensable across a variety of fields. For example, in forensic psychology, different LIWC categories facilitate the detection of deceptive patterns in communication. Fornaciari and Poesio [57] successfully applied LIWC to distinguish between truthful and fabricated statements in forensic transcripts, identifying linguistic markers of deception.

A recent study by Glasauer and Alexandrowicz [58] illustrates LIWC's utility for objective personality assessment. Using expressive writing samples from 124 participants, they modeled Big Five traits as latent constructs predicted by LIWC categories. While correlations with traditional self-reports were moderate—particularly for Neuroticism, Conscientiousness, and Openness—the approach showed promising model fits and validity. This study underscores LIWC's ability to capture personality-related language patterns and highlights its usefulness for unobtrusive, scalable personality profiling in psychological and forensic contexts.

The relevance of LIWC has further increased through its integration into modern machine learning pipelines. Recent work, such as that by Kilic et al. [59], demonstrates that LIWC-based features not only enhance the predictive performance of neural networks but also provide interpretable insights into the linguistic markers underlying complex behavioral predictions. Thus, LIWC serves as both a theoretical and practical bridge between classical psycholinguistic analysis and the current state of computational modeling.

3.9 LIWC in Cybergrooming Research

As highlighted in the comprehensive interdisciplinary review by An et al. [21], LIWC is among the most frequently utilized psycholinguistic tools in computational cybergrooming detection pipelines. Its ability to map word usage onto psychologically meaningful constructs has greatly improved the interpretability and effectiveness of machine learning models, supporting both risk analysis and offender profiling.

A central application of LIWC (Linguistic Inquiry and Word Count) lies in the linguistic characterization of grooming stages in chat-based conversations between adults and minors. Gupta et al. [18] were among the first to segment annotated pedophile chat logs based on grooming theory and to use LIWC for creating psycholinguistic profiles of individual grooming phases. Their results showed that certain LIWC categories such as *social processes*, *family*, or *sexual* terms exhibit distinctive frequency patterns along the grooming timeline. Cano et al. [15] confirmed these findings by demonstrating that integrating LIWC-derived features into machine learning models significantly improves the detection of grooming activities and phases, particularly in comparison to simple lexical baselines.

More recent studies have extended the application of LIWC beyond offender profiling. Guo et al. [60], for example, examined psychological vulnerability markers of potential victims and quantified them using LIWC categories. They found that personality traits and social support dimensions derived from LIWC robustly distinguish between more and less vulnerable minors. Another empirical study by Broome et al. [5] involved focus groups with police and correctional officers to identify relevant LIWC categories (e.g., affective, social, cognitive, biological processes), whose salience was then validated through the analysis of authentic grooming

chats. The findings emphasized that grooming discourse is often not solely characterized by overtly sexual content but also by frequent social bonding signals and a focus on the present.

The review by An et al. [21] further corroborate the high relevance of LIWC in grooming research. The authors highlight three main advantages:

- LIWC provides interpretable and theory-driven language features that can be effectively integrated into scalable machine learning models,
- It supports phase detection and offender profiling through the quantitative mapping of conversational structure,
- It enables cross-platform and cross-linguistic analysis due to its validated and standardized category framework.

An et al. [21] further emphasize that LIWC categories—such as affective states, social processes, biological cues, and cognitive markers—are not arbitrarily selected but have been empirically validated through repeated application in leading computational and social science studies.

3.9.1 Key psychometric LIWC Features in Grooming

The following LIWC dimensions have consistently proven to be most informative in the context of cybergrooming detection and linguistic analysis[5], [18], [21]):

- **Affective processes:** positive/negative emotion, sadness, anxiety
- **Social processes:** family, friends, communication, group references
- **Cognitive processes:** insight, certainty, tentative language, causation
- **Biological processes:** body, sexual keywords
- **Drives and informal language:** risk, reward, netspeak, swear words

In summary, LIWC offers a validated, scalable, and interpretable set of psycholinguistic features that has proven effective both for descriptive analyses and as input variables for machine learning models in cybergrooming research [5], [15], [18], [21], [60]. Its widespread adoption and continuous methodological refinement make it an important resource for understanding, detecting, and ultimately preventing online grooming.

3.10 Explainable AI with SHAP

While deep learning models achieve remarkable performance in natural language processing tasks, their decision-making processes often remain obscure. This is particularly relevant for deep neural networks such as transformer-based language models, as such models act as *black boxes*: due to millions of parameters and complex, non-linear relationships, it is difficult to directly interpret the decision logic [61], [62]. The field of Explainable AI (XAI) seeks to address this by developing methods that make model behavior interpretable and comprehensible to human users. One of the key techniques in XAI is the use of feature attribution methods, which assign importance scores to individual input features based on their contribution to the model's output. This allows users to understand which aspects of the input data most strongly influence the model's predictions. Two widely used methods for this purpose are SHAP (Shapley Additive explanations) and

LIME (Local Interpretable Model-agnostic Explanations). Both approaches aim to explain individual model predictions by quantifying the importance of input features, yet they differ significantly in their underlying methodology and scope.

SHAP is grounded in cooperative game theory and builds on the concept of *Shapley values* [63]. It treats the prediction task as a game in which each input feature contributes to the final outcome. For a given model and instance, SHAP calculates how the model's prediction changes when a specific feature is added or removed from all possible subsets of features. The resulting Shapley value of a feature is its average marginal contribution across all possible feature combinations. This makes SHAP a mathematically sound and globally consistent attribution method.

One of SHAP's core strengths is its satisfaction of desirable properties such as:

- **Fairness:** Features that have no impact on the model output receive a SHAP value of zero.
- **Consistency:** If a model is changed such that a feature contributes more to the output, its SHAP value does not decrease.
- **Additivity:** The sum of all SHAP values plus a base value (often the average model output) exactly equals the model prediction for the instance.

SHAP can be applied to any machine learning model, including complex deep learning architectures, making it a versatile tool for interpreting model behavior across different domains. However, calculating exact Shapley values can be computationally expensive, especially for models with many features [64], [65]. Also, it is important to note that SHAP explanations can be affected by class imbalance. In strongly skewed datasets, the attributions may be dominated by the majority class, potentially obscuring signals from the minority class. Therefore, balancing the data or carefully designing the background distribution is recommended to ensure faithful and interpretable explanations [66], [67].

3.11 Integrating Psycholinguistic Features and Explainable AI for Enhanced Grooming Detection

Previous approaches to detecting cyber grooming have primarily focused on linguistic features for the early identification of problematic communication [59], but have systematically neglected the integration of validated psychological frameworks into explainable AI approaches [5]. Recent research demonstrates that combining LIWC features with Transformer-based models and SHAP can significantly enhance both the interpretability and predictive performance of models.

Preiss and Chen [68] show that integrating LIWC-22 into a Mental-RoBERTa classifier, combined with SHAP analyses, enables reliable extraction of psycholinguistically relevant text segments. Similar findings are reported by Wewelwala et al. [69] in clinical emotion analysis: combining ClinicalBERT with LIWC/NRC and SHAP yielded complementary contributions, with the Transformer providing around 68% of the predictive signal and LIWC/NRC contributing the remaining 32%. These features offered additional psychological insights and improved interpretability.

Miah et. al.[70] and Salminen et. al. [71] extend these findings in a related, but for this work particularly relevant, context: the detection of toxic and harmful online communication. Both studies show that combining contextual embeddings (BERT variants) with LIWC-based psycholinguistic features leads to more robust predictions and clearer, psychologically grounded explanations, even in complex, multimodal chat data.

Despite these advances, a significant research gap remains in the domain of cyber grooming detection: the fusion of psycholinguistic features with modern language models, and their transparent evaluation using SHAP, has not yet been systematically implemented.

This work addresses this gap by combining Transformer-based text representations (BERT) with LIWC features and applying SHAP to quantify the contribution of both feature sets to model predictions. The aim is to significantly improve the detection performance of cyber grooming while enhancing explainability through psychologically grounded, transparent model interpretations, thereby establishing a methodological foundation for future applications.

4 Dataset and Preprocessing

For this thesis, data from two complementary sources were collected and processed: the *Perverted Justice* (PJ) archive and the PAN12 Sexual Predator Identification dataset.

4.1 PAN12 Dataset

The PAN12 Sexual Predator Identification Corpus was introduced by Inches and Crestani as part of CLEF 2012 and represents the first large-scale benchmark dataset for the automatic detection of sexual predators in chat conversations [72]. The PAN12 dataset was designed for two specific tasks: First, identifying sexual predators among all users, and second, detecting specific chat lines that are most characteristic of predatory behavior. It remains a reference benchmark in cyber grooming detection and has been employed in a lot of follow-up studies. [72]

The collection combines four distinct sources:

1. Perverted Justice logs containing confirmed grooming conversations
2. Omegle chats between consenting adults
3. IRC logs from *irclog.org*
4. IRC logs from *krijnhoetmer.nl*

This composition was chosen to include both genuine positive cases and potentially misleading negative examples containing sexual explicit but non-grooming messages.

The dataset comprises a total of **357,622 conversations**. Among them:

- **11,350 conversations** ($\approx 3.17\%$) originate from *Perverted Justice (PJ)* logs and constitute the **positive class (grooming)**.
- **346,272 conversations** represent the **negative class**.

The data was split in the following way.

- **Training set:**
 - 66,927 conversations
 - 903,607 messages
 - 97,689 unique authors, including 142 predators
- **Test set:**

- 155,128 conversations
- 2,058,781 messages
- 218,702 unique authors, including 254 predators

And can be seen as a classification task of either a binary classification between predator and non-predator chats. Also, the conversation can be classified in three different ways, with a distinction of:

- (P) Grooming conversations
- (A) Sexual conversations between adults
- (N) Non-sexual chats

A big challenge of the Pan12 dataset is constituted by the **class imbalance**, as only a small fraction of all conversations are grooming-related, making the detection of sexual predators and relevant text passages particularly difficult.

4.2 Perverted Justice (PJ) Dataset

The Perverted Justice (PJ) dataset is based on chat logs collected by the U.S.-based Perverted Justice Foundation (PJFI), which undertook a proactive approach in the detection of online sexual predators. They trained adult volunteers posed as young people (decoy victims) in chat rooms and engaged in conversations with adults who have sexual intentions towards minors and to record these interactions. The goal was to identify and report potential predators to law enforcement agencies. Once the adult was convicted in an American Court of Law, the Foundation made all the material related to them available online for public access. The PJ dataset consists of chat logs from over 600 confirmed grooming cases, with each conversation containing multiple messages. The conversations are typically structured as dialogues between a decoy victim and an adult groomer. PJ data have been used in nearly all major peer-reviewed studies on automatic grooming detection and form the basis for the positive class in the PAN12 dataset. The organisation ceased its operations in 2019 following public criticism because of the controversial nature of its methods and the potential risks to decoy victims. The original PJ dataset is no longer publicly available; however, archived chat logs remain available upon request for research purposes. The dataset has been used in various studies to analyze psycholinguistic features and develop machine learning models for grooming detection.

Collecting Perverted Justice (PJ) Data. The original PJ dataset is no longer publicly available. Following a formal request, academic access to an archived version of the site via the Internet Archive (Wayback Machine) was granted. Based on this permission, the full archive was retrieved and processed for research purposes.

A custom scraping tool was developed to extract all available chat logs from a stable 2023 snapshot. Chat dialogues were extracted from the HTML structure by identifying `<div class="chatLog">` sections and splitting them into lines. Regular expressions were used to clean the data, removing timestamps, HTML comments, and formatting artifacts. Conversations were split into individual sessions based on paragraph breaks.

After scraping, the pipeline proceeded as follows:

- **Scraping:** All chat logs were downloaded and extracted. This yielded **605** raw PJ chat sessions which consisted of several partial conversations over a long period of time.

- **Parsing:** Speaker-message pairs were extracted using regex rules and normalized. Conversations were split into multiple sub-dialogues whenever empty lines indicated a topic or session change. This yielded **10,811** parsed PJ chat sessions across all extracted files.
- **Anonymization:** A batch-wise anonymization step was applied to ensure privacy and remove personally identifiable information. This included: Replacing emails with *[EMAIL]*, URLs with *[URL]*, IP addresses with *[IP]*, and phone numbers with *[PHONE]*.
- **Filtering:** Conversation Sessions with fewer than 8 messages were removed, resulting in **6,175** grooming-labeled conversations.
- **Final preprocessing:** The remaining chats were standardized into a uniform JSON format with labeled roles (groomer/decoy), consistent dialogue structure, and metadata fields.

The final PJ grooming dataset consists of **6,175** labeled conversations.

Adding PAN12 Data (Non-Grooming Subset). To create a reliable negative class, a carefully filtered subset of the PAN12 dataset was used. PAN12 contains over 100,000 conversations, including more than 5,000 involving known sexual predators. The preprocessing steps were as follows:

- **Raw corpus loading:** The full dataset consists of **66 927** training and **155 128** test conversations, totaling **222 055** chats across both splits.
- **Initial filtering:** All conversations involving identified predators or labeled as containing grooming behavior (according to the official PAN12 ground truth) were excluded.
- **Participant & length constraints:** Only conversations with exactly two participants and a minimum of eight messages were retained.
- **Spam removal & quality filtering:** An additional manual filtering step excluded problematic cases:
 - Extremely short, non-linguistic messages (e.g., “m”, “e”, “?_?”).
 - Symbolic or spam-like ASCII sequences (e.g., “yearyearyear...” or repeated “WARNING” messages).
 - Excessively long sequences that could not be meaningfully tokenized or chunked (e.g., single messages with over 512 tokens).

For the PAN12 dataset, anonymization was already conducted during dataset creation. As stated by Inches and Giacomo [72], all personal identifiers (e.g., names, emails, phone numbers, URLs) were removed or replaced by synthetic placeholders prior to release. After all preprocessing and filtering steps, the resulting PAN12 non-grooming dataset contained **27,755** conversations.

All further preprocessing steps were applied consistently across both datasets and integrated into the pipeline prior to modeling.

4.3 Slang and Informal Language Handling

For preprocessing the chat data, a custom script was developed that uses the OpenAI API (*gpt-4.1-mini*) to perform targeted normalization. The employed prompt specifies clear rules:

1. **Replace slang, internet abbreviations, and leetspeak** (e.g., *wanna* → “want to”, *ur* → “you”) with their standard English equivalents.
2. **Resolve predefined placeholders** according to a mapping table (e.g., *_laugh_* → “haha”, *_thinking_* → “hmm”).
3. **Correct obvious spelling and grammar mistakes.**
4. **Do not modify** emojis, emoticons, or stylistic punctuation sequences (e.g., *:P*, *XD*, *??*).
5. **Preserve** the informal style and emotional tone, including explicit or manipulative language.

This normalization was performed *prior* to fine-tuning BERT for three key reasons:

Avoiding source effects. The dataset combines conversations from different sources (PAN12 and Perverted Justice). Without normalization, the model could exploit differences in slang usage as a shortcut, instead of learning actual grooming-related patterns. Therefore, it might be the case that it is too easy for the model to learn the decision boundary for these two classes, not because the model is good at detecting sexual predators, but because these two datasets were generated in different ways. [20]

Synchronizing BERT and LIWC input. The planned model fusion relies on BERT embeddings and LIWC features extracted from the *same* text chunks. BERT processed the raw text while LIWC analyzed the cleaned text, a semantic mismatch would occur. [38] Consistency ensures that, for example: “*This chunk had high LIWC values in the sexuality category and was also classified as grooming by BERT.*”

Improving model efficiency and robustness. Uniformly cleaned text results in more stable tokenization, reduces rare tokens, avoids model misinterpretations, and can improve performance by removing noise. [73]

Additionally, LIWC relies on a fixed dictionary and does not recognize misspelled words [56]. Without correction, such words would not be assigned to any category, leading to incomplete feature counts. Normalization ensures that relevant terms are recognized and correctly mapped to their LIWC categories.

4.4 Raw Conversational Length Distributions

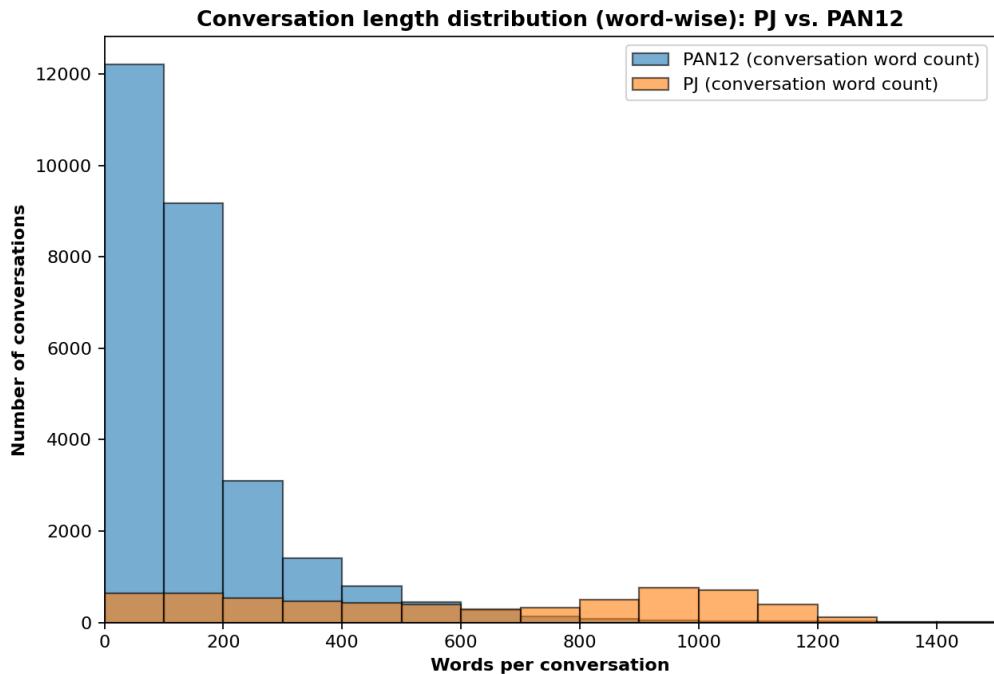


Figure 4.1: Comparison of aggregated LIWC macro-groups between PAN12 (non-grooming) and PJ (grooming) over global conversations.

After initial preprocessing, the length distributions of the raw conversations in both datasets were analyzed. Figure 5.4.1 shows the number of words per conversation for both PJ and PAN12 datasets. It is evident that PJ conversations are clearly longer than those from PAN12, with a **mean length of 609 words (PJ) versus 159 words (PAN12)**. This difference in length constitutes a potential confounder, as the model could exploit it as a shortcut for classification instead of learning real grooming-related, semantic patterns. Therefore, the preprocessing pipeline was designed to mitigate this effect by enforcing fixed-length chunking and incorporating synthetic non-grooming data in PJ style (see Section ??).

4.5 Generating Synthetic Non-Grooming Data

To further increase model robustness and mitigate dataset-specific artifacts, a custom script was developed that uses the OpenAI API (*gpt-4.1-mini*) to generate synthetic negative examples in the style of the Perverted Justice (PJ) data. The employed prompt specifies clear rules:

1. **Replicate the linguistic style of PJ chats** by providing the model with 20 already normalized PJ conversations as few-shot examples.
2. **Enforce strict formatting**, requiring every line to follow the pattern *speaker_1: <text>* or *speaker_2: <text>*.

-
3. **Generate long dialogues** with 90–120 turns, in order to reduce the risk of length-based shortcuts. This choice was made because PJ conversations are usually much longer than those from PAN12 (see Figure 4.1 for a comparison of conversation length distributions).
 4. **Restrict content** to everyday, harmless topics without any references to minors, schools, or sexual activity.
 5. **Include small imperfections** (e.g., occasional double dots, lowercase starts, or short fragments) to match the PJ normalization profile more closely, while keeping the conversations coherent.

Avoiding length leakage. Another confounder is the systematic difference in conversation length: PJ logs are generally much longer than PAN12 conversations. Without adjustment, the model could exploit length alone as a predictive cue. By explicitly generating long synthetic non-grooming conversations in PJ style, the data balance is improved and the reliance on length as a shortcut is mitigated.

Avoiding domain leakage. Since all PJ chats are labeled as class 0 and all PAN12 chats as class 1, the model could otherwise learn to separate classes merely based on dataset-specific style differences. The synthetic PJ-style non-grooming data helps reduce this effect by providing negative examples that mirror the PJ distribution but lack grooming semantics.

Ethical considerations. Due to usage restrictions, it was not possible to generate synthetic grooming conversations in the PAN12 style. Therefore, only non-grooming chats were created, ensuring that no artificial grooming data was introduced.

4.6 Final Dataset

The final dataset used for model training and evaluation consists of the following combination of grooming and non-grooming conversations from the two sources, along with synthetically generated non-grooming dialogues. The final conversation counts are the following:

- **PJ: 6175 Dialogues**
- **PAN12: 27755 Dialogues**
- **Synthetic: 600 Dialogues with mean lenght of 150 messages**

5 Methodology

5.1 Initial Chunking Strategy for BERT Fine-tuning

The first preprocessing pipeline was designed to provide a straightforward baseline for fine-tuning. Each conversation was reconstructed from the *dialogue* field into a single string with explicit speaker tags. Tokenization was then applied to the full conversation, with a maximum sequence length of 512 to provide the maximum information context in each chunk, meaning that chunk boundaries could occur in the middle of a message. Conversations exceeding this limit were split into non-overlapping chunks, which could occur at arbitrary positions within the text. Therefore, a single conversation could have multiple chunks, each treated as an independent training example. Finally, dynamic padding was applied at batch time and the tokenized dataset was then passed to the Hugging Face *Trainer* for BERT fine-tuning without additional preprocessing steps such as fixed-length padding or message-boundary control.

5.1.1 Train-Test Split

For the finetuning, a group-level train-test split was applied using *GroupShuffleSplit* from *sklearn*, ensuring that all conversations from a single predator were contained entirely within either the training or test set. This approach prevents data leakage across splits, which could otherwise lead to overoptimistic performance estimates. The split ratio was set to 70% for training and 30% for testing. Also, the same ratio of grooming to non-grooming conversations was maintained in both sets to ensure balanced class distributions in the training and train and test data.

The resulting data distribution after chunking is shown in the following table.

Table 5.1: Initial data distribution after Chunking (max length 512, no message boundary control).

Split	Grooming	Non-Grooming PAN12	Total
Train	14997	30781	45778
Test	1330	3429	4759

5.2 BERT fine-tuning as Baseline

Based on the initial chunking pipeline, BERT was fine-tuned for binary classification.

The normal baseline followed a classic fine-tuning pipeline for binary classification with *bert-base-uncased*. The training configuration was as follows:

- **Backbone:** *bert-base-uncased* (standard model dropouts)
- **Chunk length:** 512
- **Chunking:** tokenizer-based with *return_overflowing_tokens=True*, no overlap, no alignment to message boundaries; dynamic padding in batch
- **Trainer/Optimization:** Hugging Face *Trainer*
- **Epochs:** 3
- **Batch Size:** 8
- **Learning Rate:** $2e-5$
- **Weight Decay:** 0.01
- **Warmup:** none
- **Gradient Clipping:** not explicitly set

5.2.1 Evaluation Metrics

To evaluate the performance of the binary classifier, the most common metrics for a binary classification task were used in **all the following model evaluations**:

Let TP , FP , TN , and FN denote true positives, false positives, true negatives, and false negatives. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5.1)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (5.2)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (5.3)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5.4)$$

Therefore, Accuracy measures the overall proportion of correctly classified instances, while precision measures the fraction of predicted positive instances that are true positives. Recall measures the fraction of actual positive instances that are correctly identified, emphasizing the model's ability to minimize false negatives. Finally, the F_1 -score is reported as the harmonic mean of precision and recall, providing a balanced measure that accounts for both.

Since the dataset is slightly imbalanced, the F_1 -score for the positive class is used as the primary metric, ensuring that both precision and recall are equally considered to make the evaluation robust against class imbalance.

The evaluation was performed every 3000 steps during training, using the test dataset.

5.3 Limitations of Initial BERT Fine-tuning Approach

While the initial fine-tuning approach enabled a fast and effective baseline, it also raised methodological concerns:

1. **Domain Leakage:** Since all PJ conversations carry label 0 and all PAN12 conversations label 1, the model could rely on dataset-specific artifacts (domain leakage) instead of semantic cues. As a result, the model might learn to distinguish between datasets rather than genuine grooming patterns.
2. **Length Leakage:** In addition, PJ conversations are generally longer than the ones from PAN12 (Figure 4.1), making conversation length itself a potential shortcut (length leakage). As a result, the model could exploit systematic length differences rather than learning meaningful grooming-related features.

As shown later in the evaluation (Table 6.1), the initial model indeed achieved very high performance, which motivated the design of a stricter preprocessing pipeline. Therefore, a second preprocessing script was developed, which introduced fixed-length padding strategies, enforced chunking at message boundaries, and incorporated synthetic non-grooming data in PJ style to reduce both domain and length leakage. The following sections describe this improved pipeline in detail.

5.4 Chunking Strategy for BERT to reduce Domain and Length Leakage

To mitigate leakage effects and evaluate robustness, a stricter preprocessing pipeline was implemented. Instead of splitting conversations at arbitrary positions, chunks were created only at message boundaries, ensuring that single utterances remained intact.

During development, additional options such as random left/right padding and random message shuffling were tested. While these techniques can increase robustness by preventing the model from overfitting to position-specific cues, they were not retained in the final pipeline. There were two reasons for this decision: first, later feature fusion with LIWC required that padding positions remain consistent, otherwise token-level interpretability would be undermined. Second, SHAP was planned as an explainability method, and artificially shuffled or randomly padded inputs would have introduced noise, making it harder to attribute predictions to relevant linguistic or psychometric features.

To further mitigate domain leakage and length leakage, synthetic non-grooming chats in PJ style were added (Section 4.5). Models were then trained and tested under two complementary setups:

1. **Separated Split:** With synthetic data only in the test set, to probe generalization.
2. **Mixed Split:** With synthetic data included in both train and test sets, to strengthen robustness.

5.4.1 Chunk-Length Distributions (For Train and Test) Across Sequence Lengths

To determine the optimal chunk length for a robust BERT baseline the mean, standard deviation, median, minimum, and maximum of tokens per conversation across the different datasets were calculated and are shown in table 5.2.

Table 5.2: Token statistics per conversation/file across datasets

Dataset	Files	Mean	Std (pop)	Std (sample)	Median	Min	Max
PJ (grooming)	6 175	723.83	446.54	446.58	719	12	1 850
PAN12	27 751	210.12	202.20	202.20	144	1	3 656
Synthetic	621	969.47	161.24	161.37	972	488	1 435
PAN12 + Synthetic	28 372	226.74	230.01	230.01	147	1	3 656
ALL (PJ + PAN12 + Synthetic)	34 547	315.59	339.65	339.65	171	1	3 656

The statistics show that PJ conversations are generally much longer (mean: 724 tokens, median: 719 tokens) than PAN12 conversations (mean: 210 tokens, median: 144 tokens). The synthetic conversations are even longer (mean: 969 tokens, median: 972 tokens) and thus closer to the PJ distribution. The combined PAN12 + Synthetic dataset has a mean of 227 tokens and a median of 147 tokens, which is still significantly shorter than PJ. Overall, the complete dataset (PJ + PAN12 + Synthetic) has a mean of 316 tokens and a median of 171 tokens.

To fairly evaluate the base model across these regimes, it was decided to test the baseline training with the following three fixed chunk sizes to evaluate if the chunk size has an impact on performance and model robustness:

- **150 tokens:** closely matches the PAN12 median, minimizing fragmentation for PAN12 while providing a compact, compute-efficient context budget.
- **250 tokens:** offers additional headroom beyond the PAN12 median (covering a larger share of its distribution) and tests whether modestly more context improves performance without a large compute increase.
- **512 tokens:** targets full-length context for longer PJ/Synthetic conversations, reducing truncation and assessing the payoff of richer context windows.

Therefore, multiple target sequence lengths (150, 250, and 512) were generated, and padding on these fixed lengths was applied to test the impact of chunk size on performance, preventing the model from relying on systematic length differences between datasets.

Figure 5.1 shows the resulting chunk length distributions for the three target lengths (150, 250, 512) after applying the improved chunking strategy, to show the impact of different chunk sizes on the dataset. Each subplot displays the distribution of chunk lengths in both training and test sets combined. The orange bars represent grooming chunks collected from PJ, while the blue bars represent non-grooming chunks collected from PAN12. Additionally the green bars show the synthetic non-grooming chunks in PJ style. The Chunk lengths were measured before padding on a fixed size of 512, 250 or 150 chunks was applied.

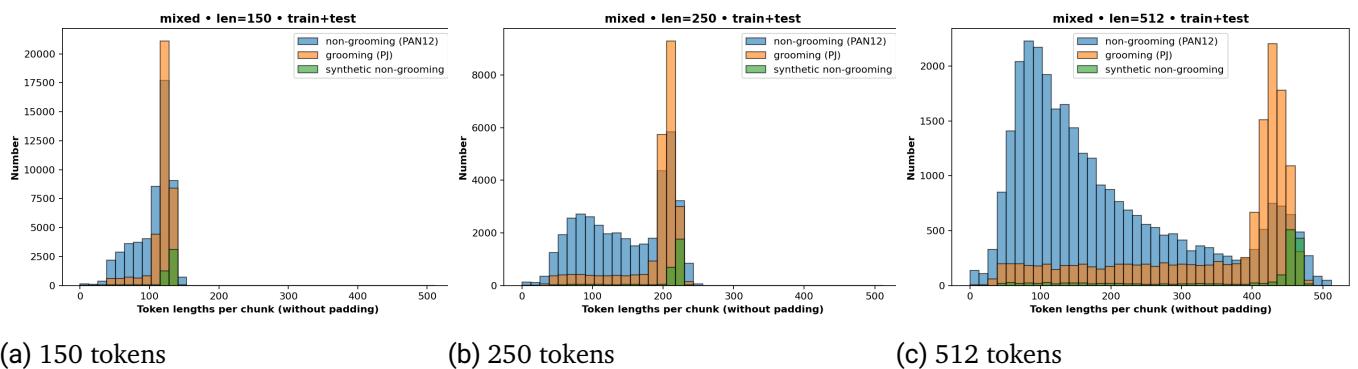


Figure 5.1: Chunk length distributions (train+test) for different sequence lengths (before padding).

5.4.2 Data Distributions after Chunking

The following tables show the resulting data distributions after chunking for the three target lengths (150, 250, 512) and each setup (synthetic data only in test set and synthetic data in test and train set). Each table lists the number of chunks per class in the training and test sets for, together with the total number of chunks.

Split	Grooming	Non-Grooming PAN12	Non-Grooming Synthetic	Total
Separated Train	27883	36954	0	64837
Separated Test	9625	16211	4749	30585
Mixed Train	26583	37356	3254	67193
Mixed Test	10925	15809	1495	28229

Table 5.3: Split for Chunk-Length of 150 Tokens.

Split	Grooming	Non-Grooming PAN12	Non-Grooming Synthetic	Total
Separated Train	17487	27057	0	44544
Separated Test	6003	11848	2925	20776
Mixed Train	16662	27321	2008	45991
Mixed Test	6828	11584	917	19329

Table 5.4: Split for Chunk-Length of 250 Tokens.

Split	Grooming	Non-Grooming PAN12	Non-Grooming Synthetic	Total
Separated Train	9699	21259	0	30958
Separated Test	3267	9195	1583	14045
Mixed Train	9197	21352	1087	31636
Mixed Test	3769	9102	496	13367

Table 5.5: Split for Chunk-Length of 512 Tokens.

5.5 Improved Training Pipeline

For training on this improved dataset, the same BERT backbone was used but with a slightly more robust training configuration. Dropout rates in the classification head were increased (0.3), and label smoothing (0.1) was applied to reduce overconfidence. Gradient checkpointing was enabled for memory efficiency.

5.5.1 Improved Training Configuration

The training configuration for the improved pipeline was the following:

- **Backbone:** *bert-base-uncased* with increased dropouts
- **Sequence length:** Variants $\in \{150, 250, 512\}$
- **Dropout (BERT):** $hidden_dropout_prob = 0.3$, $attention_probs_dropout_prob = 0.3$
- **Loss function:** Cross-Entropy
- **Label smoothing:** 0.1
- **Epochs:** 3
- **Batch size:** 8
- **Learning rate:** 2×10^{-5}
- **Weight Decay:** 0.01
- **Warmup Ratio:** 0.06
- **Max Grad Norm:** 1.0
- **Gradient Checkpointing:** enabled

5.5.2 Evaluation Strategy with Additional Subsets

The model was trained and tested across all three chunking target lengths (150, 250, 512) under two different setups

1. **Synthetic data only in the test set.** This setup was designed to evaluate whether the model can generalize to PJ-style negatives without having seen them during training.
2. **Synthetic data in both train and test sets.** This setup was designed to test whether including synthetic negatives in the training phase improves robustness and reduces reliance on dataset-specific artifacts.

The evaluation was again performed every 3000 steps on the test set, using accuracy, precision, recall, and F1 score as metrics.

To further test robustness, an evaluation callback ran three subsets after each evaluation step:

- **Real-only test set (consisting of only PAN12 + PJ data).**

- **Synthetic-only test set.**: Here the model was only evaluated based on the accuracy since precision, recall, and F1 are not defined for a single class.

Additionally the confusion matrix was calculated for the complete test set after each evaluation step to analyze the types of errors made by the model. Also, the ROC-Curve and the AUC score were calculated after each evaluation step to analyze the model's performance across different classification thresholds.

The detailed outcomes the training across alle different chunk size and train/test split configurations are shown in the evaluation section (Section 6.2).

5.6 Choosing the final configuration for Feature Fusion and SHAP Explainability Analysis

Based on the evaluation across different chunk sizes and dataset configurations (Section 6.2), **it was decided to continue the feature fusion and explainability analysis with a chunk size of 512 tokens in combination with the mixed split strategy**. Although shorter chunks were hypothesized to reduce length leakage, the results show no significant performance improvement for 150 or 250 tokens. Instead, 512-token chunks consistently achieved the best F1 scores (up to 0.97) while at the same time providing richer conversational context, which is crucial for a clean LIWC analysis. Moreover, the separated split revealed poor generalization, indicating overfitting on training data. By contrast, the mixed split setup led to more robust performance across evaluation subsets, making it the more reliable and realistic configuration for further experiments. Therefore, this final choice balances model accuracy with interpretability and robustness, setting a solid foundation for the following feature fusion experiments and shap analyses.

5.7 Choosing the LIWC Feature Set for Feature Fusion and SHAP Explainability Analysis

The Feature-Fusion and following SHAP explainability analysis was performed with two variants of feature sets in each case:

1. **All-Features-Fusion**: Use of all 118 LIWC features.
2. **Psychometric-Fusion**: Use of a subset of **49 psychometric LIWC-features** coverings all psycholinguistic LIWC-Features that are highlighted in the literature as relevant for manipulative communication in Grooming Chats (Section 3.9.1). The subset contains the following categories and dimensions:

Psychometric LIWC Subset (49 features).

- **Drives**: *Drives, affiliation, achieve, power, reward, risk, curiosity, allure*
- **Cognition**: *Cognition, allnone, cogproc, insight, cause, discrep, tentat, certitude, differ, memory*
- **Affect & Emotion**: *Affect, tone_pos, tone_neg, emotion, emo_pos, emo_neg, emo_anx, emo_anger, emo_sad, swear*
- **Social**: *Social, socbehav, prosocial, polite, conflict, moral, comm, socrefs, family, friend, female, male*

- **Physical/Biological:** *Physical, health, illness, wellness, mental, substances, sexual, food, death*

The idea for this two-staged approach was to evaluate whether a set of only psycholinguistically relevant features could provide similar or better interpretability and performance compared to using the full LIWC-2022 feature set.

5.8 LIWC Data Extraction

The LIWC features were extracted using the official LIWC-22 software. The extraction was done once over complete PJ and PAN12 conversations as well as the synthetic non-grooming conversations in PJ style. Furthermore, the LIWC features were also extracted for each chunk in the improved chunked dataset with a chunk size of 512 tokens(Section 5).

Also, the LIWC Features were extracted in two variants: once with all 118 features and once with the psychometric subset of 49 features (Section 5.7). The extracted LIWC features were then stored in a separate sidecar file.

5.9 LIWC Data Analysis

In order to analyze the effect size of LIWC features prior to fine-tuning, a comprehensive analysis of LIWC features across all conversations was performed. The LIWC-based analysis was carried out in two steps. First, a global comparison was performed on complete conversations, followed by a chunk-based comparison using 512 chunks, which was tailored to the subsequent feature fusion. Both analyses were performed once for the complete set of LIWC-2022 features and once restricted to the psychometric subset of LIWC features to determine whether psychometric variables alone are sufficient to distinguish grooming conversations from non-grooming conversations.

5.9.1 Global LIWC Analysis on Complete Conversations

For the global analysis, all grooming conversations from the PJ were first aggregated into one conversation per groomer to represent their general language style, while non-grooming conversations from PAN12 were used in their original form. For each conversation, the complete set of LIWC-2022 features was calculated. To improve interpretability, the features were grouped into macro categories according to the LIWC-2022 manual:

- **Linguistic:** Function, pronoun, ppron, i, we, you, shehe, they, ipron, det, article, numeral, preposition, auxiliary verb, adverb, conjunction, negation, verb, adjective, quantity
- **Punctuation:** Period, comma, question mark, exclamation mark, apostrophe, other
- **Emoji:** Emoji
- **Drives:** Belonging, Achievement, Power
- **Motivation:** Reward, Risk, Curiosity, Enticement

- **Cognition:** Cognition, allnone, cogproc, Insight, Cause, Discrepancy, Attempt, Certainty, Difference, Memory
- **Affect:** Positive tone of voice, negative tone of voice, emotion, positive emotion, negative emotion, fear, anger, sadness, swearing
- **Social:** Social behavior, prosocial, polite, conflict, morality, communication, social references, family, friend, female, male, social
- **Physical:** Health, illness, well-being, mental, substances, sexual, food, death, physical
- **Perception:** Attention, movement, space, visual, auditory, feeling
- **Culture:** Politics, ethnicity, technology
- **States:** Need, desire, acquisition, lack, fulfillment, fatigue
- **Time:** Time, focus on the past, focus on the present, focus on the future
- **Conversation:** Internet slang, agreement, non-fluency, filler words

For each conversation, the value of a macro group was defined as the arithmetic mean of all available member characteristics. Subsequently, the macro group means were averaged across all conversations for each class (PJ vs. PAN12). The results were visualized as grouped bar charts (PJ vs. PAN12) sorted by the overall mean $\frac{1}{2}(\bar{y}_{PJ} + \bar{y}_{PAN})$.

In addition to macro group aggregation, each individual LIWC feature was statistically analyzed. For each feature f , the mean and standard deviation across all conversations were calculated per class,

$$\bar{x}_{PJ}, s_{PJ}, \quad \bar{x}_{PAN}, s_{PAN},$$

followed by the mean difference $\Delta = \bar{x}_{PJ} - \bar{x}_{PAN}$ and Cohen's d [74],

$$d = \frac{\Delta}{s_p}, \quad s_p = \sqrt{\frac{(n_{PJ} - 1)s_{PJ}^2 + (n_{PAN} - 1)s_{PAN}^2}{n_{PJ} + n_{PAN} - 2}}.$$

The descriptive statistics and effect sizes were compiled into a summary table for all LIWC features which is provided in the appendix (table 8.1).

Furthermore, the 30 LIWC-2022 features with the largest absolute effect sizes $|d|$ were selected to identify the most discriminative variables. This procedure was performed twice. Once for the complete set of LIWC features and once restricted to the psychometric LIWC subset to analyze if psychometric features alone capture the major differences between the two classes.

To further assess the quality and representativeness of the synthetically generated non-grooming data, an additional global LIWC analysis was conducted comparing PJ Grooming and PAN12 Non-Grooming to the synthetic dataset. For each comparison, Cohen's d was again computed for all LIWC features, and the top 30 features by absolute effect size were visualized as horizontal bar plots. This step was intended to identify potential linguistic deviations between the synthetic and real-world data and to evaluate whether the synthetic data exhibits similar psycholinguistic patterns as PAN12.

The results of the global LIWC analysis are presented in the evaluation section (Section 6.3).

5.9.2 Chunk-based LIWC Analysis with 512 Tokens

To examine patterns within conversations, a chunk-based analysis was performed with a fixed chunk size of 512 tokens (identical to the size later used for model training). For each chunk, a LIWC-2022 feature vector was calculated and the average values per class were determined across all chunks. also based on the data level , which For each feature, the range

$$\text{range}(f) = \max_s \bar{f}_s - \min_s \bar{f}_s$$

was determined for each feature across all sources ($s \in \{\text{PJ, PAN12, SYN}\}$), and the $K = 15$ features with the largest ranges were selected for visualization. Two heatmaps were created: one limited to these 15 most important distinguishing features and one containing all available features (in alphabetical order).The rows represent PJ Grooming, Synthetic Non-Grooming, and PAN12 Non-Grooming, while the columns represent the LIWC features. This process was repeated once for all LIWC features and once for the psychometric features to determine whether psychometric variables alone exhibit comparable distinction at the chunk level.

The visualization and evaluation of the chunk-based LIWC analysis is presented in section 6.6.

5.10 Feature Fusion Strategy

To extend the BERT baseline with the additional input of LIWC features, the LIWC features were first extracted for all chunks from the dataset with 512 tokens and stored as numerical vectors in a separate sidecar file. For this purpose, the sidecar contained a key consisting of *conv_id* and *chunk_index* for each chunk, as well as the corresponding chunk-specific LIWC features. When loading the dataset, these sidecar files were assigned to the respective chunks based on the *conv_id* and the *chunk_index* (for dialogues consisting of several chunks), so that the training and test data sets then contained an additional column for *liwc* in addition to *input_ids*, *attention_mask*, and *labels*.

To integrate the LIWC features into the BERT architecture, a **late fusion approach** was implemented using cross-attention and a gating mechanism. This method allowed the model to select information from LIWC into the contextualized token representations learned by BERT.

5.10.1 Model Architecture

The proposed feature fusion model extends *bert-base-uncased* (12 transformer layers, hidden size $d_h = 768$, 12 attention heads) with a late-fusion mechanism that integrates psycholinguistic LIWC features into the transformer encoder in the upper encoder layers at layer 6 and 12. The design was used to keep the original BERT architecture and its pre-trained weights intact while allowing the model to leverage additional psycholinguistic information from LIWC. This keeps the tokenization and main contextual processing of the text unchanged, while the LIWC information is only introduced after language modeling, which holds comparability with the BERT baseline and allows a later downstream explainability using SHAP. The overall process can be summarized as followed:

- **Input and Feature Projection:**

- For each text chunk, a LIWC feature vector $x \in \mathbb{R}^{d_{liwc}}$ is extracted ($d_{liwc} = 118$ for all features or $d_{liwc} = 49$ for the psychometric subset).
 - x is normalized using *LayerNorm* and mapped through a linear projection with *GELU* activation to a compact fusion dimension $z \in \mathbb{R}^{d_p}$ (*proj_dim*, e.g., 128/768).
 - From z , a single *LIWC token* $t \in \mathbb{R}^{d_h}$ is generated via a linear layer. This token summarizes the psychometric information for the entire chunk.
- **Fusion via Cross-Attention:**
 - Text hidden states $H \in \mathbb{R}^{T \times d_h}$ from intermediate layers serve as **queries**.
 - The LIWC token t is used as **key/value** in a multi-head cross-attention mechanism ($n_heads=4$, mask-aware).
 - This enables each token to attend to the psycholinguistic context selectively.
 - **Gating and Residual Update:**
 - A channel-wise gate $g \in \mathbb{R}^{d_h}$ (*gate_type=channel*) is computed from the *[CLS]* representation and z .
 - The final representation is updated via

$$H' = H + g \odot \text{Attn}(H, t)$$

followed by a 10% dropout.

- A post-fusion *LayerNorm* is applied to stabilize training and normalize the residual update.

- **Integration in Encoder:**

- The fusion block is applied in the upper encoder layers (layers 6 and 12; *fusion_at_layers* with *depth=1*).
- After the final fusion layer, the *[CLS]* vector is pooled as in the baseline model and passed to a linear classification head.
- The pooled *[CLS]* representations are combined either by averaging (*mean*). The resulting fused representation is then passed to the linear classifier.

The overall feature-fusion architecture is illustrated in the following figure 5.2.

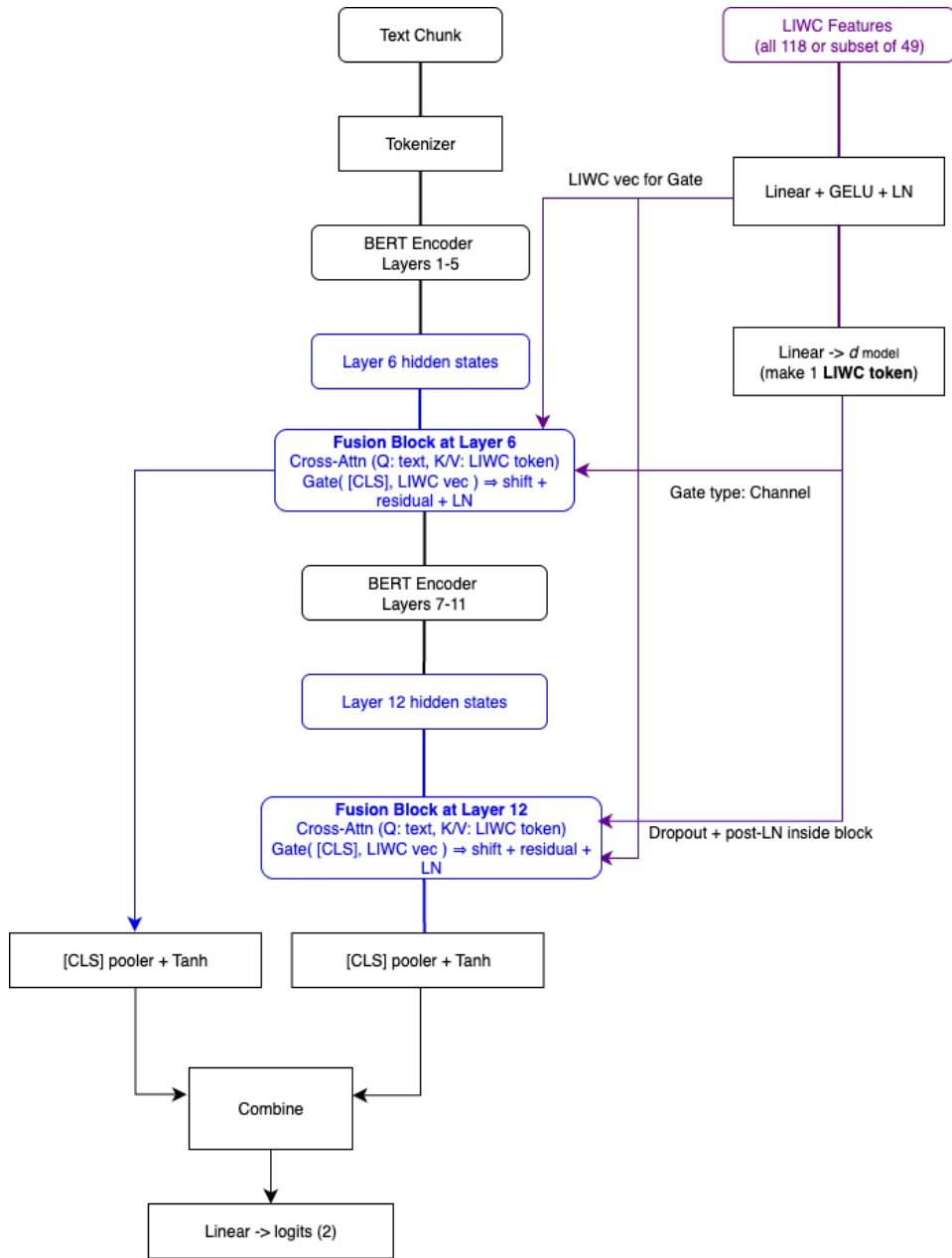


Figure 5.2: Late fusion of LIWC with BERT: LIWC features (either all 118 or the 49 psychometric subset) are projected, transformed into a single *LIWC token*, and fused with hidden states from layers 6 and 12 via cross-attention plus a gating mechanism (Multimodal Shifting). The resulting fusion blocks are applied in parallel to the BERT encoder and do not feed back into subsequent layers. The [CLS] representations from both fusion points are pooled, combined, and classified. The **blue blocks** highlight where fusion takes place, while the **purple line** represents the LIWC stream (projection and token).

5.10.2 Training Pipeline and Evaluation

The training and evaluation of the feature fusion model followed a similar pipeline as the improved BERT-Baseline (Section 5.5). With the following hyperparameters:

- **Number of Epochs:** 3
- **Saving Checkpoint at Epoch:** 1
- **Evaluation Steps:** 3000
- **Learning Rate:** 2×10^{-5}
- **Weight Decay:** 0.01
- **Warmup Ratio:** 0.06
- **Max Grad Norm:** 1.0
- **Batch Size:** 8
- **Loss Function:** Cross-Entropy with Label Smoothing of 0.1
- **Dropout inside Fusion:** 0.1
- **Number of Heads in Cross-Attention:** 4
- **Fusion Depth:** 1

The Evaluation was performed with the same metrics as in the BERT-Baseline (accuracy, precision, recall, F1) on the same splits (test-all, synthetic-only), which have been logged every 3000 steps during training.

5.11 Explainability Analysis based on Feature Fusion Model

The following Analysis was done based on the Feature Fusion Model with a Chunk size of 512 tokens and Synthetic Data in both Train and Test set after **three epochs of training**. This decision was made to ensure that the model had sufficient exposure to the data and to capture more complex patterns in the feature interactions.

To analyze the model's decision-making process, SHAP was used to show insights into how both token-level text features and LIWC psychometric features contributed to the model's predictions. The following sections describe the two complementary explainability analyses that were conducted.

5.11.1 SHAP analysis of LIWC features

To understand which psycholinguistic features are relevant for the classification decision, a SHAP analysis of the LIWC features was performed. Both the complete LIWC feature set with 118 dimensions and a psychometric subset with 49 selected categories were considered. The fixed-length LIWC vector calculated for each chunk served as input for the analysis.

Reducing Computational Costs with KernelSHAP

The explanations were generated using KernelSHAP, whereby the text input of the model was capped to a fixed sequence length and kept constant for all explanations. To keep the computational costs of the SHAP analysis manageable, a fixed configuration was used for both the complete LIWC feature set (118 dimensions) and the reduced psychometric subset (49 dimensions). Specifically, 64 random samples from the test set were selected for explanation (`--samples 64`). For the background distribution, 32 LIWC vectors were chosen and reduced to at most 20 representative centroids using k-means clustering (`--background 32`). Each instance was then perturbed 256 times (`--nsamples 256`) in order to approximate the marginal feature contributions.

The model was executed in mini-batches of 4 explanations (`--batch_size 4`), with a maximum of 16 perturbations processed in parallel (`--predict_max_batch 16`). In addition, the text input was capped to a constant sequence length of 64 tokens (`--seq_cap 64`), and inference was carried out in mixed precision with disabled gradients. These measures substantially reduced GPU memory consumption and made it feasible to perform KernelSHAP for both feature sets without exceeding hardware limits.

Determining Feature Importance and Direction of Effect

For each instance in the test dataset, the prediction for the positive grooming class was explained. SHAP determines how much the model prediction changes when a single LIWC feature is neutralized. A high SHAP value indicates that this feature contributed significantly to the decision. In this way, the **mean absolute** SHAP value ($\text{mean } |SHAP|$) was calculated for each LIWC feature, describing the global importance of the feature. To make the results easier to interpret, these values were then converted into percentages of the total significance:

$$\text{FeatureImportance}_i = \frac{\text{mean } |SHAP|_i}{\sum_{j=1}^F \text{mean } |SHAP|_j} \times 100\%$$

where F is the number of LIWC features. This shows the percentage of the total model decision that can be explained by a single feature (Figure 6.10).

In addition, the **mean signed** SHAP value was determined to indicate the direction of the effect. Therefore a positive SHAP value means the feature pushes the prediction toward the positive class (grooming), while a negative SHAP value means it pushes toward the negative class (non-grooming). Thus, higher projected values indicate more grooming-like language, lower values indicate more non-grooming-like language. The results were summarized in several steps and can be found in section 6.8.

5.11.2 Confidence Analysis based on Label Flip and Confidence Shift

In order to further investigate the impact of LIWC features on model decisions, model confidence and class assignment were calculated once with the regular LIWC feature vectors (*LIWC on*) and once with LIWC vectors set to zero (*LIWC off*). This allowed the influence of the additional LIWC features on model confidence and class assignment to be analyzed.

For both conditions, the class probability was determined by applying the softmax function to the logits. The highest probability p_{\max} represents the model confidence for the predicted class:

$$p_{\max}^{(\text{on})} = \max_k p_k^{(\text{LIWC on})}, \quad p_{\max}^{(\text{off})} = \max_k p_k^{(\text{LIWC off})}.$$

The difference between these two values defines the so-called **confidence shift**:

$$\Delta p_{\max} = p_{\max}^{(\text{on})} - p_{\max}^{(\text{off})}.$$

A positive value indicates that the LIWC features increased the certainty of the model decision, while a negative value describes a decrease in confidence.

In addition, it was analyzed whether the predicted class of an input changed as a result of LIWC fusion (for example, model decision = grooming before and model decision = non-grooming after). If such a change occurs, it is referred to as a **label flip**:

$$\text{flip} = \mathbb{1} \left[\hat{y}^{(\text{on})} \neq \hat{y}^{(\text{off})} \right], \quad \hat{y} = \arg \max_k p_k.$$

The number and rate of label flips show the extent to which the LIWC features lead to a changed classification decision.

For quantitative analysis, the following metrics were calculated:

- $\Delta\mu$, $\Delta\tilde{x}$, $\Delta\sigma$: Mean, median, and standard deviation of Δp_{\max} .
- Δp_{10} , Δp_{90} : 10th and 90th percentiles of the distribution of Δp_{\max} .
- $n_{\text{class } 0}$, $n_{\text{class } 1}$: Frequency distribution of the predicted classes with LIWC.
- n_{flips} , flip rate: Number and relative frequency of prediction changes.

The results are presented in Section 6.9.

5.12 LIWC Analysis of Misclassifications

To identify potential patterns in the misclassifications made by the feature fusion model, a LIWC analysis of false positives and false negatives was conducted. It was tested whether misclassified samples resemble the opposite correct class in LIWC space, i.e., whether false negatives are closer to true negatives and false positives are closer to true positives according to their LIWC Feature values. Furthermore, it was analyzed which LIWC features differ significantly between misclassified and correctly classified samples. This analysis was again conducted with both the full LIWC feature set and the psychometric subset. The same fusion model, tokenizer, and LIWC sidecar vectors as in the previous sections were used to generate predictions and per-sample LIWC vectors. The analysis was performed on the complete test set.

5.12.1 Descriptive per-group statistics

For each LIWC feature and each confusion group $g \in \{\text{TN}, \text{TP}, \text{FP}, \text{FN}\}$, the following descriptive statistics were computed over the per-sample values:

count, mean, std, median, q1 (25th), q3 (75th), min, max.

These summaries describe central tendency and spread within groups and are used as context for the inferential tests below. The complete summary statistics for all LIWC features are provided in the appendix (table ??).

5.12.2 Per-feature group comparisons

Per-feature comparisons were run for four group pairs: FP vs. TN, FN vs. TP, TN vs. FN, and TP vs. FP. For each LIWC feature f , let x denote values in the first named group and y in the second. The following statistics were computed:

- **Mean difference** $\Delta\mu = \bar{x} - \bar{y}$.
- Again, Cohen's d was used to quantify effect size
- **Significance tests.** The Mann–Whitney U test (two-sided) was used to assess the significance of differences in distributions of FP/TN and FN/TP.
- **Multiple testing control.** The Benjamini–Hochberg procedure was applied across features to obtain FDR-adjusted q -values. For plotting, a significance score $-\log_{10}(q)$ was used.

5.12.3 Proximity hypothesis tests in LIWC feature space

To test whether misclassifications resemble the opposite correct group in LIWC space, distances were computed in standardized feature space (z-scored per LIWC dimension). Let \mathbf{c}_{TN} and \mathbf{c}_{TP} be the centroids of TN and TP, respectively. For each FN sample with standardized vector \mathbf{z} , Euclidean distances $d_{\text{TN}} = \|\mathbf{z} - \mathbf{c}_{\text{TN}}\|_2$ and $d_{\text{TP}} = \|\mathbf{z} - \mathbf{c}_{\text{TP}}\|_2$ were computed and the following one-sided hypotheses were tested:

$$H_1^{\text{FN}} : \Pr[d_{\text{TN}} < d_{\text{TP}}] > 0.5 \quad \text{and} \quad H_1^{\text{FP}} : \Pr[d_{\text{TP}} < d_{\text{TN}}] > 0.5.$$

For each hypothesis, the analysis reports:

- **Proportion of samples** closer to the hypothesized centroid:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[d_{\text{closer}} < d_{\text{farther}}].$$

- **One-sided binomial test** against 0.5 to assess whether the observed proportion \hat{p} is significantly greater than chance.
- **Paired one-sided tests** on the distance differences $d_{\text{TN}} - d_{\text{TP}}$ for FN and $d_{\text{TP}} - d_{\text{TN}}$ for FP. Both the Wilcoxon signed-rank test and the one-sample t -test were used to assess whether the mean/median difference is significantly greater than zero.

The results of the LIWC analysis of misclassifications are presented in section 6.10.

5.12.4 SHAP-Based Proximity Analysis of Top-20 Misclassifications

For a deeper analysis of misclassifications, an evaluation was performed based on the **top 20 LIWC features**, which were identified in Section 5.11.1 as the most relevant using global SHAP values (*mean_abs*) and explain the majority of model decisions. This allows examining whether misclassifications can be explained by their proximity to the opposite correct class along these LIWC dimensions. Also, focusing on the 20 most important features reduces noise from less informative variables and enables a more targeted analysis of group differences.

Top-K selection and SHAP direction. All LIWC-2022 features were globally ranked by their *mean_abs* SHAP values, and the top 20 were selected. In addition, the SHAP sign (*mean_signed*) was used for each feature to define the “TP-like” versus “TN-like” direction.

Z-scaling and SHAP-oriented projection. To ensure comparability, feature columns were z-scaled. The values were then multiplied *per sample* by the SHAP sign (+1 or -1). On this aligned axis, values *to the right of* $x = 0$ correspond to greater TP similarity, while values *to the left of* $x = 0$ indicate greater TN similarity. The dashed vertical line at $x = 0$ in the plots marks this separation threshold.

Group statistics and confidence intervals. For each group (<{TP, TN, FP, FN}) and each top-K feature, group means \bar{x} , standard deviations s , and standard errors

$$SE = \frac{s}{\sqrt{n}}$$

with n as the group size were calculated. For the misclassification groups (FP and FN), a 95% confidence interval was then computed:

$$CI_{95} = \bar{x} \pm 1.96 \cdot SE.$$

The position of this CI relative to $x = 0$ was classified as:

- *TP-side*: CI_{95} lies completely to the right of 0 (clearly TP-like),
- *TN-side*: CI_{95} lies completely to the left of 0 (clearly TN-like),
- *crosses-0*: CI_{95} includes 0 (ambiguous),
- *unknown*: no CI could be computed (e.g., missing SE).

Proximity rate (distance logic). Additionally, for each feature it was assessed which correct class the misclassifications were closer to. This was determined using the distances between the misclassification mean \bar{x}_{MC} and the TP or TN means:

$$d_{TP} = |\bar{x}_{MC} - \bar{x}_{TP}|, \quad d_{TN} = |\bar{x}_{MC} - \bar{x}_{TN}|.$$

For **FP**, features with $d_{TP} < d_{TN}$ (closer to TP) were counted. For **FN**, features with $d_{TN} < d_{TP}$ (closer to TN) were counted. The proportion across all top-K features yields the *proximity rate* (e.g., 0.75 = 75% of features support the expected proximity).

Summary measures. The following aggregated values were derived:

- **Proximity rate:** proportion of features supporting the expected proximity ($FP \rightarrow TP$ or $FN \rightarrow TN$),
- **CI TP-side (n):** number of features with 95% CI fully to the right of 0,
- **CI TN-side (n):** number of features with 95% CI fully to the left of 0,
- **CI crosses 0 (n):** number of features with 95% CI intersecting 0 (ambiguous).

The results of this analysis, including visualization and summary tables, are presented in Section 6.11.

6 Evaluation

This chapter presents the results of these evaluations, focusing on the performance improvements achieved through feature fusion compared to baseline models using only transformer embeddings or LIWC features alone.

6.1 Initial BERT Finetuning Results

Table 6.1: Evaluation of BERT base model

Step (Epoch)	Loss	Accuracy	Precision	Recall	F1
3000 (0.52)	0.007	0.998	0.999	0.998	0.999
9000 (1.57)	0.003	0.999	0.100	1.000	0.100
15000 (2.61)	0.004	0.999	0.999	1.000	0.100
17169 (3.00)	0.002	1.000	0.999	1.000	0.100

As shown in table 6.1, the initial BERT fine-tuning on the unmodified PJ and PAN12 datasets achieved **near-perfect performance, with an F1 score of 0.999 after only 3000 train steps**. This extremely high accuracy suggested that the model was likely exploiting dataset-specific artifacts rather than learning genuine grooming-related patterns. For example, all PJ conversations were labeled as grooming (1) and all PAN12 conversations as non-grooming (0), allowing the model to rely rather on superficial cues such as conversation length or stylistic differences between datasets than on semantic content. This motivated the development of a stricter preprocessing pipeline to mitigate domain leakage and length leakage. In the following sections the improved preprocessing, training and evaluation strategies will be evaluated in detail.

6.2 Bert Finetuning Results on different Chunk Sizes and Data Setups

To evaluate the impact of chunk size and dataset splitting strategy on model performance, a series of experiments were conducted using three different chunk sizes (150, 250 and 512 tokens) and two dataset configurations: a *separated split*, where synthetic non-grooming data was included only in the training set and a *mixed split*, where synthetic data was present in both training and test sets. The chunk distribution and results for each configuration are summarized in the following tables.

6.2.1 Fixed Chunk Size of 150 chunks

Table 6.2: Evaluation for Chunk size 150, separated split (synthetic data only in train).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.37)	0.611	0.837	0.665	0.975	0.790	0.231
9000 (1.11)	0.555	0.859	0.696	0.981	0.814	0.303
15000 (1.85)	0.529	0.872	0.719	0.976	0.828	0.314
21000 (2.59)	0.498	0.888	0.747	0.973	0.845	0.408
24315 (3.00)	0.553	0.869	0.710	0.985	0.825	0.299

Table 6.3: Evaluation for chunk size 150, mixed split (synthetic data in train + test).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.36)	0.352	0.939	0.870	0.992	0.927	0.940
9000 (1.07)	0.257	0.975	0.958	0.979	0.969	0.993
15000 (1.79)	0.253	0.977	0.968	0.974	0.971	0.987
21000 (2.50)	0.248	0.980	0.965	0.986	0.975	0.993
25200 (3.00)	0.255	0.979	0.957	0.989	0.973	0.991

6.2.2 Fixed Chunk Size of 250 chunks

Table 6.4: Evaluation for chunk size 250, separated split (synthetic data only in train).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.54)	0.567	0.854	0.672	0.969	0.793	0.158
9000 (1.62)	0.647	0.845	0.653	0.988	0.786	0.082
15000 (2.69)	0.608	0.858	0.673	0.991	0.802	0.145
16704 (3.00)	0.561	0.871	0.695	0.989	0.816	0.200

Table 6.5: Evaluation for chunk size 250, mixed split (synthetic data in test + train).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.52)	0.298	0.958	0.903	0.988	0.944	0.980
9000 (1.57)	0.255	0.976	0.941	0.996	0.967	0.974
15000 (2.61)	0.242	0.984	0.964	0.993	0.978	0.991
17247 (3.00)	0.242	0.984	0.962	0.994	0.978	0.991

6.2.3 Fixed Chunk Size of 512 chunks

Table 6.6: Evaluation for chunk size 512, separated split (synthetic data only in train).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.78)	0.472	0.894	0.690	0.989	0.813	0.186
6000 (1.55)	0.517	0.883	0.668	0.987	0.797	0.092
9000 (2.33)	0.518	0.885	0.670	0.994	0.800	0.116
11610 (3.00)	0.534	0.885	0.670	0.994	0.801	0.093

Table 6.7: Evaluation for chunk size 512, mixed split (synthetic data in train + test).

Step (Epoch)	All					Synth-only
	Loss	Accuracy	Precision	Recall	F1	Accuracy
3000 (0.76)	0.251	0.978	0.935	0.990	0.962	0.970
6000 (1.52)	0.282	0.970	0.907	0.995	0.949	0.954
9000 (2.28)	0.248	0.981	0.941	0.996	0.968	0.982
11865 (3.00)	0.245	0.982	0.945	0.996	0.970	0.982

The results of the different chunk sizes and split strategies show that **BERT generally achieves very high performance on the balanced dataset**, regardless of the chosen chunk size. In addition, the recall is significantly higher than the precision in all runs, indicating that the model correctly identifies almost all grooming cases, but produces some false positives.

Another notable feature is that the model does not generalize well on synthetic data when it occurs exclusively in the test set. The accuracy in the “synth-only” column remains low in this setting, indicating a distribution shift between real and synthetic data. However, when the proportion of synthetic data is integrated into the training, the effect disappears almost completely and the model learns to successfully incorporate the synthetic data into its decision boundaries.

In addition, it can be seen that the marginal label baseline score increases slightly with increasing chunk size, which is probably due to greater stability of the classifications in longer contexts. Overall, the results confirm that BERT is able to recognize the grooming class almost perfectly, even with a larger context and in a balanced dataset, which is reflected in F1 scores of up to 0.97.

Finally, it should be emphasized that the dataset is way more balanced than the original PAN12 dataset (33% grooming cases vs. only 5% in PAN12), which explains the excellent metrics. The high data balance leads to clearly separable classes, which further benefits the performance of the model.

6.2.4 Confusion Matrices for BERT Baseline across Chunk Sizes and Data Setups

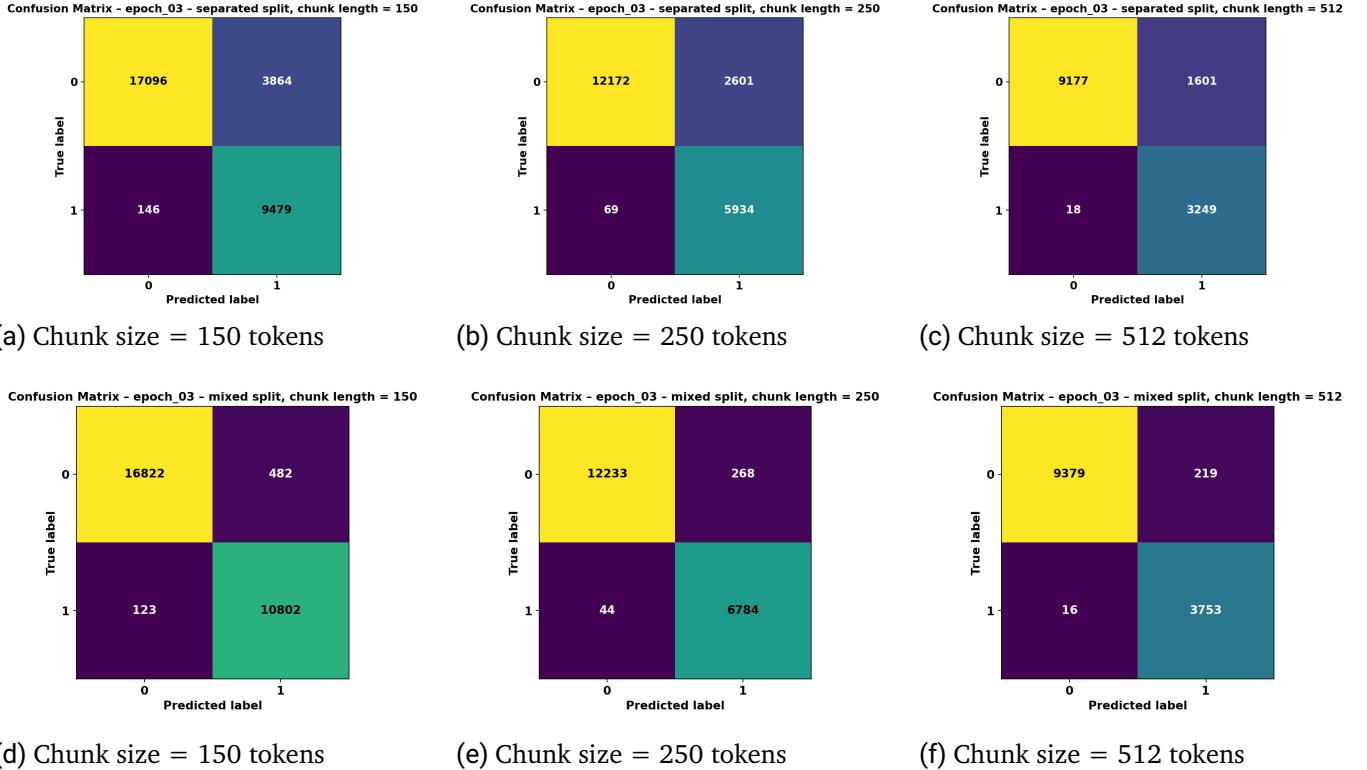


Figure 6.1: Confusion matrices of the BERT baseline after epoch 3. Top row: models trained with synthetic data included only in the test set. Bottom row: models trained with synthetic data included in both train and test set. Each column corresponds to a different chunk length (150, 250, 512).

Figure 6.1 shows the confusion matrices for the BERT baseline across different chunk sizes and dataset configurations after epoch 3. When looking at the chunk lengths, the false positives and false negatives decrease steadily with an increasing chunk size for both, the false positive and false negative rate and the mixed and separated data setup. This indicates that the model benefits from a larger context and higher word count, which is consistent with the improved metrics observed in table 6.1. It is evident, that the model with the synthetic data in both train and test set achieves a clearly better performance across all chunk sizes with a distinctly lower false positive and also a slightly lower false negative rate across all chunk sizes. This confirms the earlier observation that including synthetic data in training helps the model generalize better to this distribution, leading to fewer misclassifications.

What is the most striking, is the general higher false positive rate compared to the false negative rate across all configurations. Especially in the setup with the synthetic data only in the test set, the false positive rate is notably higher than the false negative rate (e.g. 1601 FP vs only 18 FN for chunk length 512 tokens), particularly for smaller chunk sizes (e.g. 3864 FP vs 146 FN for chunk length = 150 tokens). Even in the

mixed setup, where the performance is generally better, the false positive rate remains slightly higher than the false negative rate. This indicates that the model is generally more conservative in its predictions, preferring to classify uncertain cases as grooming rather than risking missing actual grooming conversations. This behavior is often desirable in practical applications where false negatives (missed grooming cases) can have serious consequences. As already mentioned, the dataset is very balanced, which also contributes to the low false negative rates. Overall, the confusion matrices confirm that BERT performs very well across all configurations, with performance improving with larger chunk sizes and when synthetic data is included in training and with the best performance for chunk size 512 and mixed data setup. Note that the values for accuracy, precision and f1 scores are still lower for the mixed setup with chunk size of 512 tokens compared to the mixed setup with chunk size of 250 tokens in table 6.1. This happens because the dataset for 512-token chunks contains fewer samples overall. While the absolute number of false positives and false negatives is lower, their relative share among predictions is higher, which reduces precision. At the same time, recall slightly improves since almost no true positives are missed.

6.2.5 ROC Curves for BERT Baseline across Chunk Sizes and Data Setups

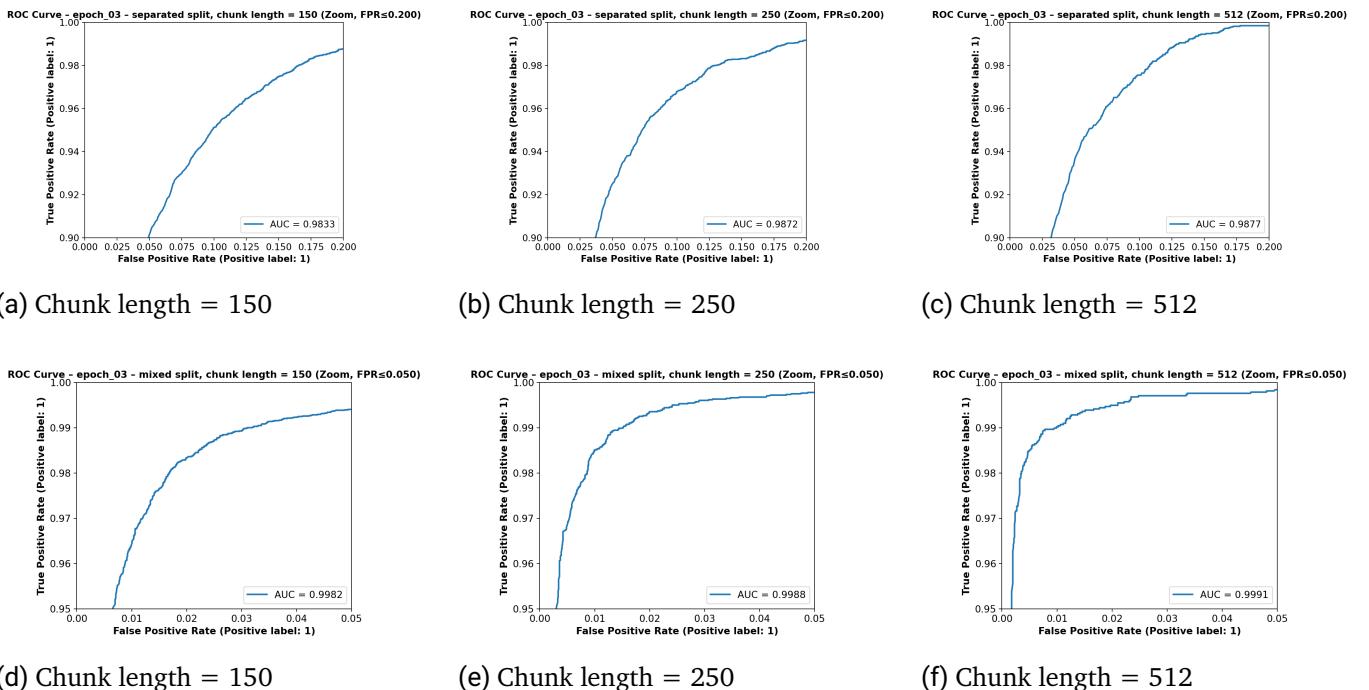


Figure 6.2: Zoomed ROC curves after epoch 3 for different chunk lengths (BERT baseline). Top row: models evaluated with synthetic data included *only in the test set* ($\text{FalsePositiveRate} \leq 0.20$ and $\text{TPR} \geq 0.90$). Bottom row: models trained and evaluated with synthetic data in *train & test* and therefore shown with a *tighter zoom* to expose finer differences ($\text{FalsePositiveRate} \leq 0.05$ and $\text{TPR} \geq 0.95$). Each panel shows the zoomed range of the ROC curve to better reveal separation at low false-positive rates; the legend reports the pAUC for the full ROC.

Figure 6.2 shows the zoomed ROC curves for the BERT baseline across different chunk sizes and dataset configurations after epoch 3. The ROC curves were plotted zoomed for both configurations to highlight the

differences more clearly, as the model already performed very well, especially for the mixed dataset, with small differences in the F1 score of < 1% after epoch 3. Note that the zoom levels differ between the two setups. For the setup with synthetic data only in the test set, the ROC curves are shown with a zoom on false positive rates ≤ 0.20 and true positive rates ≥ 0.90 , highlighting the overall performance at moderately low FPRs where differences between chunk sizes are still relatively small. For the mixed data setup, a stronger zoom was applied ($FPR \leq 0.05$, $TPR \geq 0.95$) to emphasize finer distinctions between the models in the critical region, where even small errors become relevant.

When looking at the chunk lengths, it is evident that the ROC curves improve steadily with increasing chunk size for both configurations. This indicates that the model benefits from a larger context and higher word count, which is consistent with the improved metrics observed in table 6.1 and the confusion matrices in figure 6.1. The pAUC values also increase with chunk size, confirming that larger chunks lead to better overall discrimination between grooming and non-grooming conversations. When comparing the pAUC scores between the two configurations, it is clear that the models trained with synthetic data in both train and test sets achieve significantly higher pAUC values across all chunk sizes. This confirms the earlier observation that including synthetic data in training helps the model generalize better to this distribution, leading to improved performance across the entire ROC curve. Overall, the zoomed ROC curves confirm that BERT performs very well across all configurations, with performance improving with larger chunk sizes and when synthetic data is included in training, consistent with previous analyses.

6.3 Comparing LIWC-2022 Macro Groups

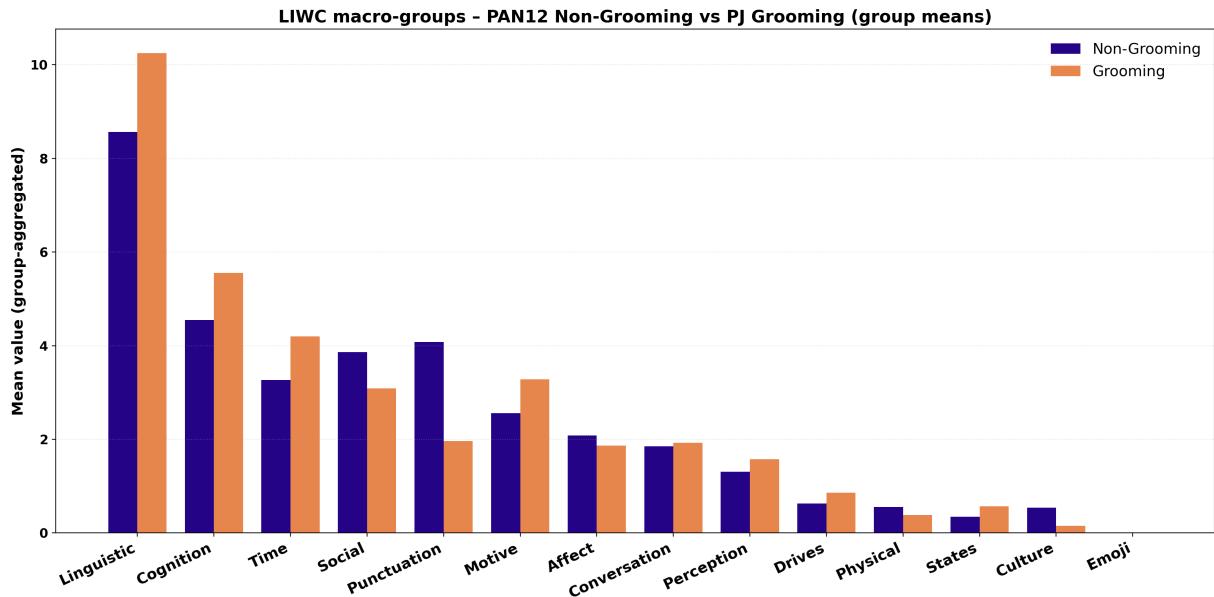


Figure 6.3: Comparison of aggregated LIWC macro-groups between PAN12 (non-grooming) and PJ (grooming) over global conversations.

Figure 6.3 compares the mean values between the LIWC scores of PAN12 (non-grooming) and PJ (grooming) conversations, aggregated into their Macro Groups.

The results indicate that linguistic features dominate both corpora, with PJ showing significant higher values. Note that the collected PJ conversations are generally longer and therefore containing more linguistic markers overall. More interesting is, that the groups *Time*, *Cognition* and *Social* stand out, where PJ conversations show a stronger presence of temporal references (often linked to future planning of meetings), cognitive processes and social markers. Also, the Category “Emoji” shows no Liwc-Values for both Datasets as a result of the slang handling and data preprocessing steps and different kind of emoji usage in the time, the datasets were collected.

Overall, the group-level aggregation highlights the major shifts across all LIWC-dimensions, providing a perspective to the category-wise analysis. This confirms that grooming communication is characterized not only by increased length and density of language, but also by a distinct emphasis on cognitive, temporal and social processes.

6.4 Comparing LIWC Features between PJ and PAN12 on Full Conversations

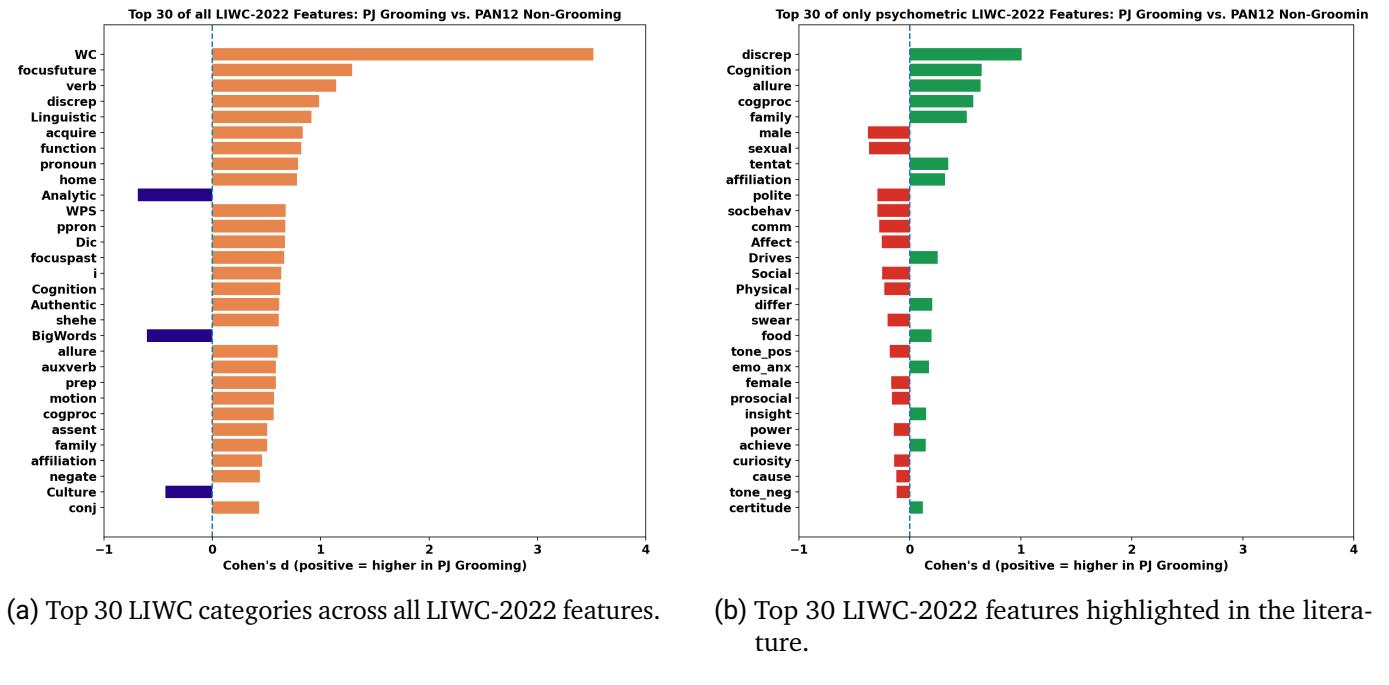


Figure 6.4: LIWC feature comparison between grooming (PJ) and non-grooming (PAN12) dialogues based on cohens d [74]. Positive values indicate higher feature usage in grooming.

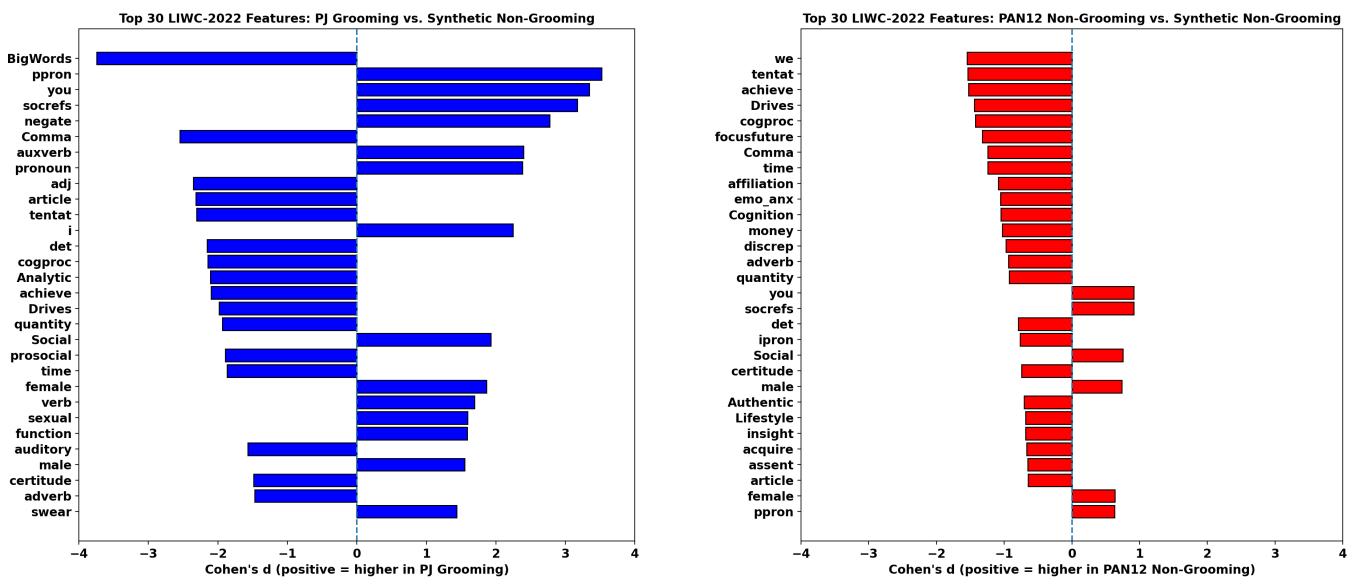
In addition, the effect size **Cohen's d** [74] was computed to quantify group differences. Figure 6.4 (left) shows the top 30 LIWC features with the largest absolute effect sizes across all LIWC-2022 features, while figure 6.4 (right) focuses on the top 30 features from the psychometric subset highlighted in prior literature (Section 5.7). For both plots, values over 0 indicate higher feature values in PJ-Grooming conversations, while values below 0 indicate higher values in PAN12 non-grooming conversations. Note, that effect sizes appear lower in the psychometric subset, since it is fully contained within the complete feature set and does not include the strongest differentiating features as shown in the left figure.

Again, as shown in figure 6.4 (left), PJ conversations are overall much longer (higher word counts), have longer sentences, and have overall more linguistically features like pronouns, verbs and function words. Since that is a strong confounder, for more content-related interpretations, these features should be ignored. When looking at the dialogical style, many functions words seem to be more prevalent in grooming conversations, which could be related to the more complex sentence structures and higher word counts. Also, the higher word count makes the pj conversations capture a broader range of topics leading to more diverse linguistic markers. What is noticeable is the strong presence of the feature *focusfuture*, which reflects the grooming strategy of planning future meetings. Also, next to the linguistic features, the thematic references like *family*, *home* and *affiliation* stand out, being higher present in grooming conversations from PJ than in PAN12, which could also be more consistent with the typical grooming narratives (e.g. asking, if the parents are home). Furthermore, the complexity in grooming conversations tends to be lower than in non-grooming conversations, as indicated by the lower *big words*, *Analytic* and *Culture* score, which could be due to the sources that PAN12 was collected from (e.g. forums, chatrooms) which often containing computer-related and more complex language.

When focusing on the psychometric features (right), the differences between grooming and non-grooming conversations become more pronounced. It is noticeable, that grooming conversations show a higher values in features like *Cognition*, *cognitive processes* and *Allure* which could be caused by a content related reference to seduction and manipulation. Therefore grooming conversations are clearly distinguishable based on psychological strategies like building closeness (social/affiliation), attraction (allure) and cognitive engagement (cognitive processes). Also, the strongest signal of grooming conversations lies in the feature *discrep* (would, should, could), showing a higher presence of words which might be used in boundary testing, conditioning and suggestions. This is accompanied by slight positive effects for *tentant* (hedging) and *polite* (courtesy or relationship building). Additionally, the features *Drives*, *insight*, *achieve* and *emotion anxiety* are more prevalent in grooming conversations, which could be related to the manipulative strategies used by groomers to build trust and emotional connection. It is striking, that the PAN12 Conversations contain higher values in the features *sexual*, *male/female*, *physical*, *swear* and *Social/social behavior/prosocial*. This is likely due to the fact that the PAN12 dataset contains a considerable amount of sexually explicit but non-grooming conversations for which were included to improve model-robustness. **Because LIWC computes scores as relative proportions, the shorter PAN12 conversations, which often consist almost entirely of sexual content, produce inflated values in sexual-related categories. In comparison, the longer and more diverse PJ logs dilute these terms within a broader linguistic context, leading to lower proportional scores.** Therefore it should be considered, that the PAN12 dataset is not a perfect representation of non-grooming conversations, but rather a challenging counterbalance to the grooming data. Overall, the LIWC analysis confirms that grooming conversations are characterized not only by increased length and density of language, but also by a distinct emphasis on cognitive, temporal and social processes, as well as specific psychological strategies related to manipulation and relationship building. These insights provide a deeper understanding of the linguistic and psychological markers of grooming behavior, which can inform the development of more effective detection models.

6.5 Comparing LIWC Features between PJ and PAN12 and the Synthetic Dataset

Additionally, Figure 6.5 (left) shows the comparison between PJ and synthetic non-grooming data. Large discrepancies are visible across several LIWC dimensions, including strong positive shifts for all LIWC-categories such as *Big Words*, *ppron* and *you*, suggesting that the synthetic samples over-represent certain linguistic markers. Figure 6.5 (right) shows the comparison between PAN12 and the synthetic data. Here, the synthetic



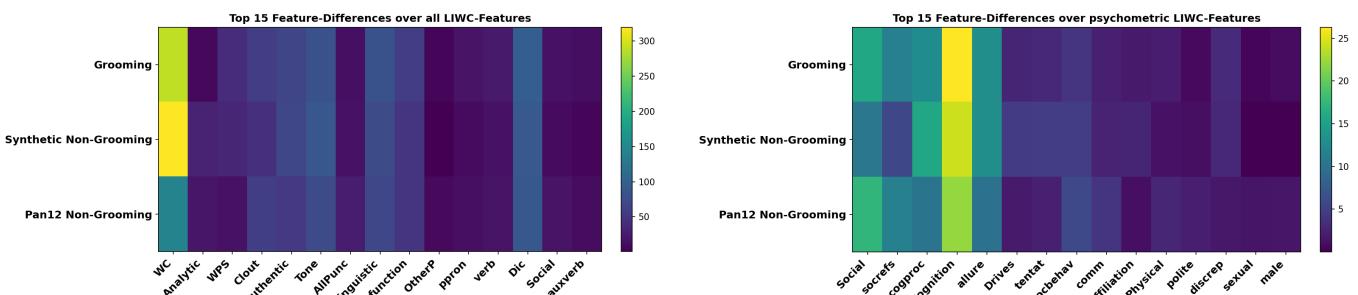
(a) PJ vs. synthetic

(b) PAN12 vs. synthetic

Figure 6.5: Top LIWC differences with synthetic baseline based on cohens d [74].

data again diverges notably across all LIWC-Categories. This confirms that the synthetic data differs clearly from both real corpora. Including such data in training therefore **challenges the model to generalize beyond the specific statistical patterns of PJ and PAN12 alone**. While this reduces predictive performance on real data in isolation only marginally, it improves the model's ability to handle out-of-distribution examples, making the overall model more robust. Still, when interpreting the LIWC-Features of the predicted grooming and non-grooming class a clear distinction between real and synthetic data should be done, since the synthetic data postpones the real distribution in many LIWC-Categories.

6.6 Chunk-based LIWC Analysis



(a) Top 15 LIWC categories on 512 chunk-level across all LIWC-2022 features.

(b) Top 15 LIWC-2022 features on 512 chunk-level highlighted in the literature.

Figure 6.6: LIWC feature comparison between grooming and non-grooming dialogues based on LIWC-2022 Features at chunk level.

In addition to an analysis of the overall conversations, an analysis of the LIWC features at the chunk level of 512 chunks (as used in the later feature fusion) was performed to analyze local variations across chats.

Figure 6.6 shows a heat map of the top 30 mean features of grooming and non-grooming conversations (left) with the largest difference across all LIWC features and (right) with the largest difference across only the psychometric features highlighted in the literature. Rows represent PJ Grooming, Synthetic Non-Grooming and PAN12 Non-Grooming, while columns represent LIWC features. Color intensity encodes the mean feature value per source. The heatmaps show, that the LIWC-Features differ significantly throughout all features and three sources at the chunk level, making it easier to understand how a model can already achieve high performance at this level.

When looking at the full feature set (left), it is noticeable, that the main differences between grooming and (synthetic) non-grooming chunks mainly lie on linguistic features such as *word count*, *WPS*, *function words* and *pronouns*, which are all more prevalent in grooming chunks. This confirms the earlier observation that grooming conversations tend to be longer and more linguistically rich, even at the chunk level. It can be seen, that the synthetic data have a very high word count balance out possible length leakage effects on chunk level. The differences in all other categories are less pronounced, but still visible with the real data seeming to be more similar across all LIWC-Categories than the synthetic data.

When focusing on the psychometric features (right), the differences between grooming and non-grooming chunks become more pronounced. This could be due to the fact, that the scale of the psychometric features is smaller, making differences more visible since there is no feature like *word count* dominating the scale. It is noticeable, that grooming chunks show values in features like *Cognition*, *cognitive processes* and *Allure* which could be caused by a content related reference to seduction and manipulation. Therefore grooming conversations are clearly distinguishable based on psychological strategies like building closeness (social/affiliation), attraction (allure) and cognitive engagement (cognitive processes). It is striking that at the chunk level, the grooming conversations have a higher proportion of sexual words, which can be explained by the higher overall relations of sexual terms in smaller conversation segments. Still, the difference is not as strong as expected from prior work [5], [21], which could be due to the fact that the PAN12 dataset contains a considerable amount of sexually explicit but non-grooming conversations. When comparing the synthetic non-grooming data to the real non-grooming data, it can be seen, that the synthetic data again shows a very different pattern across all psychometric features, confirming that the synthetic data differs clearly from both real corpora.

This chunk-level perspective complements the global analysis by showing that distinguishing features are not only visible when analyzing entire conversations, but also occur within short text chunks. While the bar charts of the global analysis in Figure 6.4 highlight markers such as *word count*, *pronoun usage* and *function words*, the heat maps show that even locally, grooming dialogues have higher values in many categories. In particular, the dominance of sexual terms becomes apparent at the chunk level, as shorter segments amplify their relative frequency, whereas these signals are diluted in complete conversations due to more diverse linguistic content. At the same time, psychometric dimensions such as cognitive and social processes remain dominant in both analyses, underscoring their theoretical relevance.

Taken together, these results suggest that grooming behavior can be detected in the dataset through both global stylistic patterns and local lexical cues, which explains why BERT already shows very strong performance at the finetuning at chunk level.

6.7 Feature Fusion Evaluation

Table 6.8: Evaluation: Feature Fusion with all LIWC-2022 Features

Step (Epoch)	Test Set Metrics				
	Loss	Accuracy	Precision	Recall	F1
3000 (0.76)	0.234	0.985	0.959	0.989	0.974
6000 (1.52)	0.231	0.987	0.962	0.994	0.978
9000 (2.28)	0.218	0.993	0.983	0.992	0.987
11865 (3.00)	0.218	0.993	0.989	0.985	0.987

Table 6.9: Evaluation: Feature Fusion with Subset of LIWC-2022 Features

Step (Epoch)	Test Set Metrics				
	Loss	Accuracy	Precision	Recall	F1
3000 (0.76)	0.249	0.976	0.925	0.996	0.959
6000 (1.52)	0.221	0.991	0.983	0.985	0.984
9000 (2.28)	0.219	0.992	0.990	0.983	0.986
11865 (3.00)	0.218	0.993	0.989	0.986	0.987

The presented tables 6.8 and 6.9 summarize the performance of the feature fusion model that combines BERT embeddings with LIWC-2022 features. Both configurations were evaluated at four training steps (3000, 6000, 9000 and 11865), corresponding to epochs of 0.76, 1.52, 2.28 and 3.00.

6.7.1 Using all LIWC-2022 Features

When using all LIWC-2022 features (Table 6.8), the fusion model achieved an improvement of the F1 score in all training steps of the feature-fusion model, starting from 0.974 at 3000 steps and reaching up to 0.987 at 11865 steps. This indicates that the additional psycholinguistic information provided by the LIWC features helps the model better distinguish between grooming and non-grooming conversations. It is striking, that the precision increased in all training steps from 0.935 to 0.959 at 3000 steps and up to a total difference of 4.4 % to 0.989 at 11865 steps. This suggests that the fusion model is more effective at reducing false positives, which is crucial in practical applications where misclassifying non-grooming conversations as grooming can have serious consequences. On the other hand, there was a small decrease in recall from 0.990 to 0.989 at 3000 steps and from 0.996 to 0.985 at 11865 steps, indicating that the model is slightly less sensitive in identifying all grooming cases. This leads to the conclusion, that the fusion model is more conservative in its predictions using additional LIWC features, which may be beneficial in reducing false alarms while still maintaining high overall accuracy.

6.7.2 Using a Subset of Psychometric LIWC-2022 Features

When using only a subset of psychometric LIWC-2022 features (table 6.9), the same trend of improvement in F1 score was observed, starting from 0.959 at 3000 steps and reaching up to 0.987 at 11865 steps. This indicates that even a targeted selection of psycholinguistically relevant features can significantly enhance model performance. The precision also showed a substantial increase from 0.925 to 0.959 at 3000 steps and up to 0.989 at 11865 steps, similar to the results when using all features. This reinforces the idea that psychometric features are particularly valuable in improving the model's ability to correctly identify grooming conversations while minimizing false positives. However, the recall showed a slight decrease from 0.996 to 0.985 at 6000 steps and from 0.983 to 0.986 at 11865 steps, indicating a minor trade-off in sensitivity. Overall, the results suggest that even a focused set of psychometric LIWC features can provide benefits when fused with BERT embeddings, enhancing the model's robustness and reliability in detecting grooming behavior.

6.7.3 Confusion Matrices for Feature Fusion Models

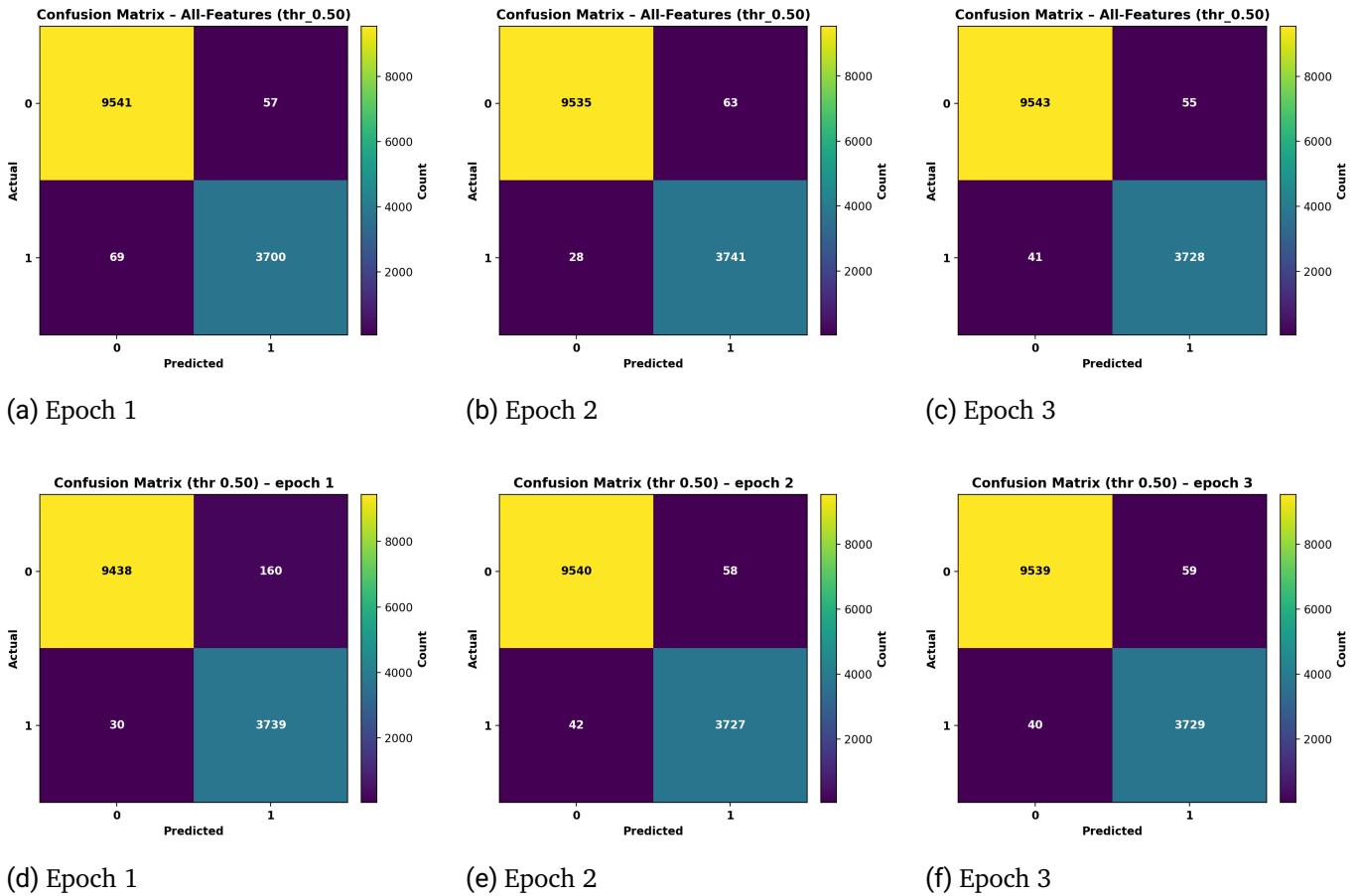


Figure 6.7: Confusion matrices by epoch for feature-fusion models (BERT baseline). Top row: fusion with all LIWC-2022 features. Bottom row: fusion with the psychometric LIWC-2022 subset. Each panel shows true labels (rows) vs. predicted labels (columns); diagonal cells indicate correct predictions, off-diagonal cells are errors. The confusion matrices were generated using a classification threshold of 0.5. For fair visual comparison, keep the same color scale across all panels.

Figure 6.7 shows the confusion matrices for both feature-fusion configurations across three epochs. It is notable that both configurations achieve a very balanced performance, having very high true positive rates (TPR) and true negative rates (TNR), with stable values across all epochs, indicating excellent overall classification performance. The number of false positives (FP) and false negatives (FN) is very low in all cases, demonstrating that the models are effective at minimizing both types of errors.

When comparing the two configurations, it is evident that the model using all LIWC-2022 features shows a slightly better performance in epoch 1, with fewer false positives and false negatives compared to the model using only the psychometric subset. However, as training progresses to epochs 2 and 3, the differences between the two configurations become negligible, with both achieving nearly identical performance in epoch 3. This suggests that while the additional LIWC features may provide some initial benefit, the psychometric subset is sufficient for achieving high accuracy once the model has been adequately trained. Consequently, the

confusion matrices reveal that both models maintain high sensitivity and specificity across all three epochs of training and for both configurations (all LIWC-Features and only psychometric subset of features). This suggests that the feature-fusion approach effectively balances the trade-off between detecting grooming conversations and avoiding false alarms. Also, when comparing the feature fusion models to the BERT-Baseline with 512 Chunks and a mixed split 6.1, it is evident, that the feature fusion models achieve a smaller false positive rate, indicating a more "conservative" classification behavior, which is particularly important in practical applications where false alarms can have serious consequences. However, as already shown in table 6.8 and 6.9, there is a small increase in the false positive rate, indicating a minor trade-off in sensitivity. Still, the differences are relatively small, suggesting that while LIWC features provide additional value, the BERT embeddings already capture much of the relevant information. Overall, the results suggest that the feature-fusion approach is better in classifying non-grooming conversations directly but with a small increase in the false negative rate but an overall more balanced performance than the BERT baseline.

6.7.4 ROC Curves for Feature Fusion Models

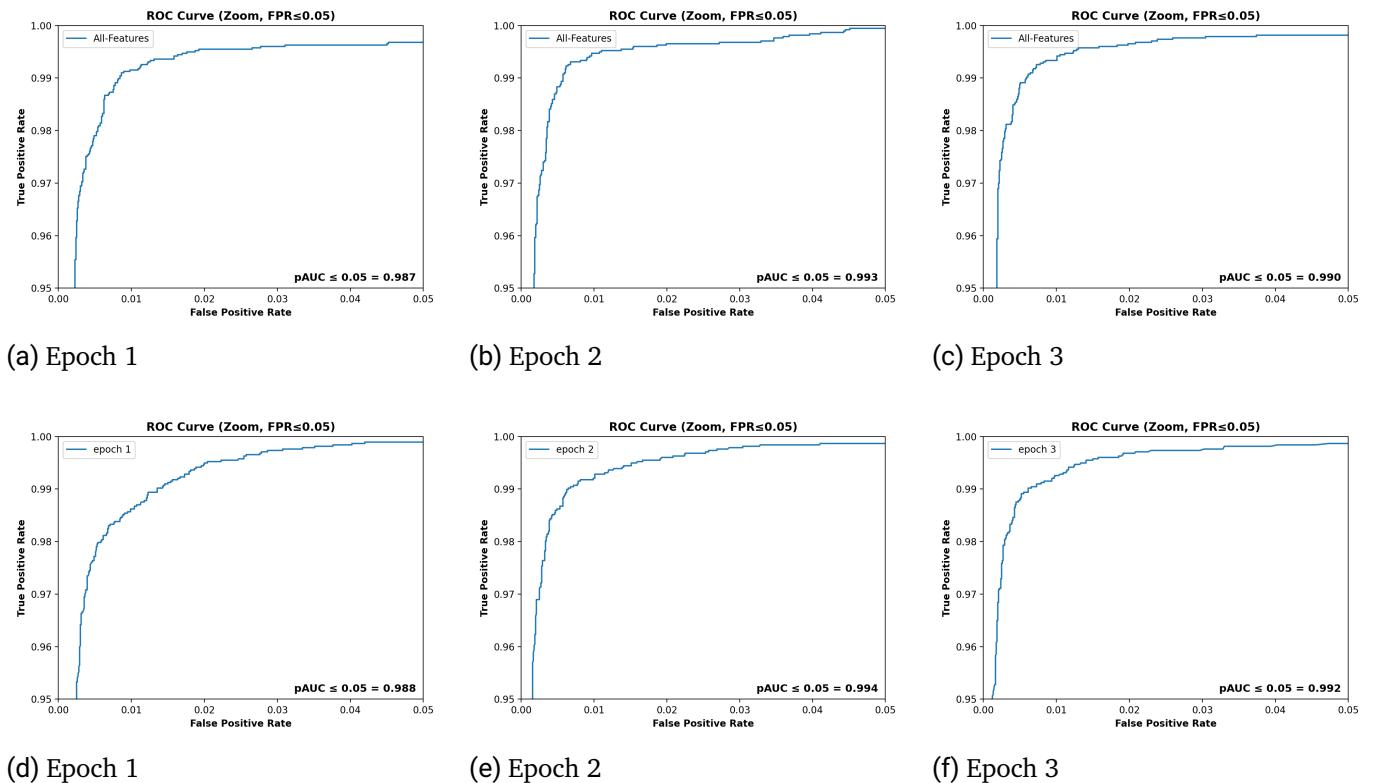


Figure 6.8: Zoomed ROC curves for feature-fusion models across all three epochs. Top row: fusion with all LIWC-2022 features. Bottom row: fusion with the psychometric LIWC-2022 subset. Columns show epochs 1–3. All panels display the same zoom window to highlight differences at low false-positive rates ($FPR \leq 0.05$, $TPR \geq 0.95$); the legend in each plot reports the pAUC of the full ROC curve.

Figure 6.8 shows the zoomed ROC curves for both feature-fusion configurations across three epochs. It is notable that both configurations achieve very high pAUC values, indicating excellent overall discrimination

between grooming and non-grooming conversations. Also, the curves for both configurations are very steep at a very low false positive rate, indicating that the models can achieve high true positive rates while keeping false positives very low. Especially in epoch 1, there is a slight difference between the two configurations, with the model using all LIWC-2022 features showing a slightly better performance. However, as training progresses to epochs 2 and 3, the differences between the two configurations become negligible, with both achieving nearly identical, almost perfect performance in epoch 3. Consequently, the zoomed-in view reveals that both models maintain high true positive rates even at very low false positive rates, across all three epochs of training and for both configurations (all LIWC-Features and only psychometric subset of features). This suggests that the feature-fusion balances sensitivity and specificity effectively. Also, the pAUC lies between 0.987 and 0.994, showing stable performance across all epochs and configurations.

When comparing the ROC curves of the feature-fusion models to the BERT baseline in figure 6.2, it is evident that the fusion models achieve a better performance, particularly at very low false positive rates for both configurations with more stable and steep ROC-Curves. This is also true when using only a subset of psychometric LIWC-Features, indicating that the addition of LIWC features helps the model maintain high sensitivity while reducing false positives as already shown in the confusion matrices at Figure 6.7. However, the differences are relatively small since the differences are only visible at a very small FPR, suggesting that while LIWC features provide additional value, the BERT embeddings already capture much of the relevant information. **Overall, the ROC analysis confirms that feature fusion enhances model robustness and reliability in detecting grooming behavior, particularly in scenarios where minimizing false positives is critical.**

6.8 Ablation Studies based on SHAP

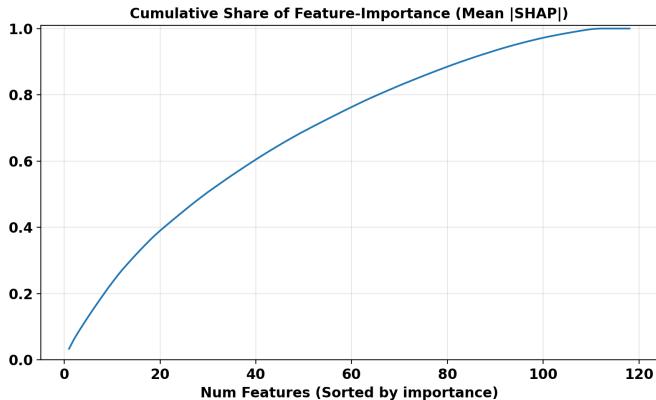
The results of the ablation studies based on SHAP are presented in the following sections. The analysis focuses on understanding the contribution of individual LIWC features to the model's predictions and identifying the most influential features for distinguishing between grooming and non-grooming conversations. The results of the SHAP Analysis were once computed for only the "real" data (PJ + PAN12) to show how they can be distinguished based on LIWC-Features and once for the complete dataset (PJ + PAN12 + synthetic) to show how the synthetic data influences the feature importance since the synthetic data was used during training to improve model robustness.

6.8.1 LIWC-Feature Importance Ranking

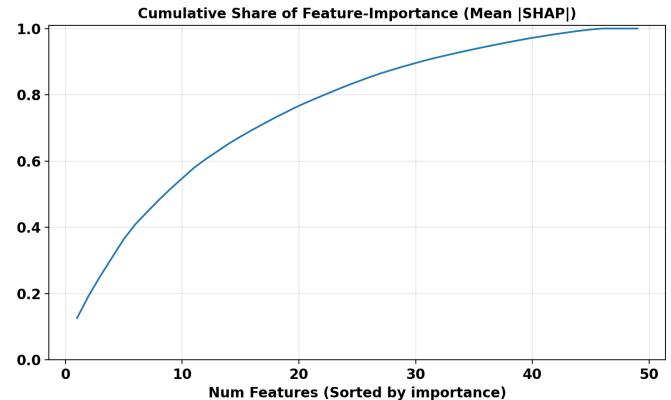
To show the results of the LIWC-based SHAP Analysis, several steps were taken. First, the cumulative curves of all features were plotted to show how many features are needed to explain a certain percentage of the model decision (Figure 6.9). Next, a global ranking of the top 20 features was created based on their percentages of the total significance (Figure 6.10). The results were then evaluated by class and visualized according to the direction of the effect (Figure 6.11). This makes it possible to understand which LIWC features are particularly relevant and whether they shift the model prediction toward grooming or non-grooming.

Cumulative Feature Importance

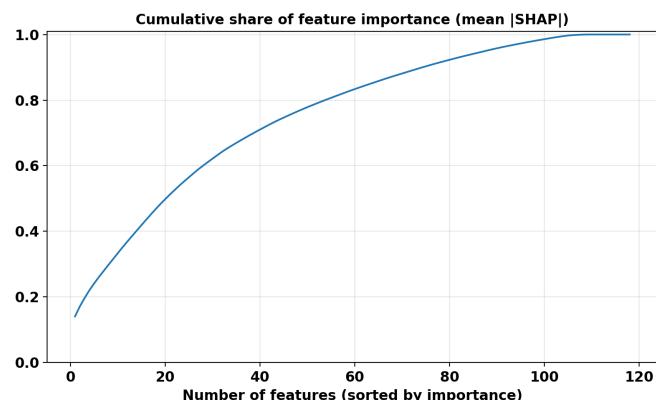
To gain a better understanding of the cumulative importance of features, the cumulative feature importance plots for all the used LIWC-Features were created. These plots show how many features are needed to explain a certain percentage of the model decision.



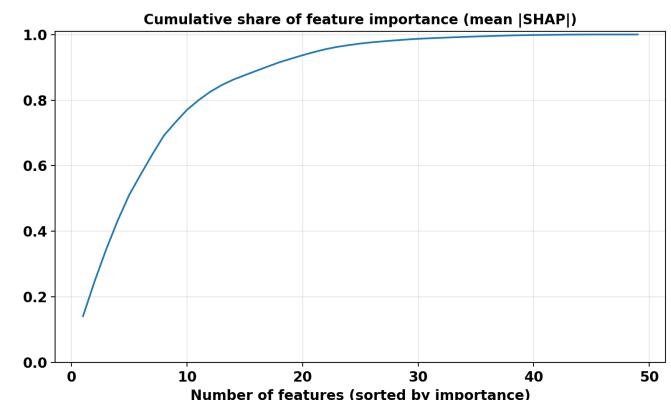
(a) All LIWC-2022 features (with synthetic data).



(b) Psychometric subset (with synthetic data).



(c) All LIWC-2022 features (no synthetic data).



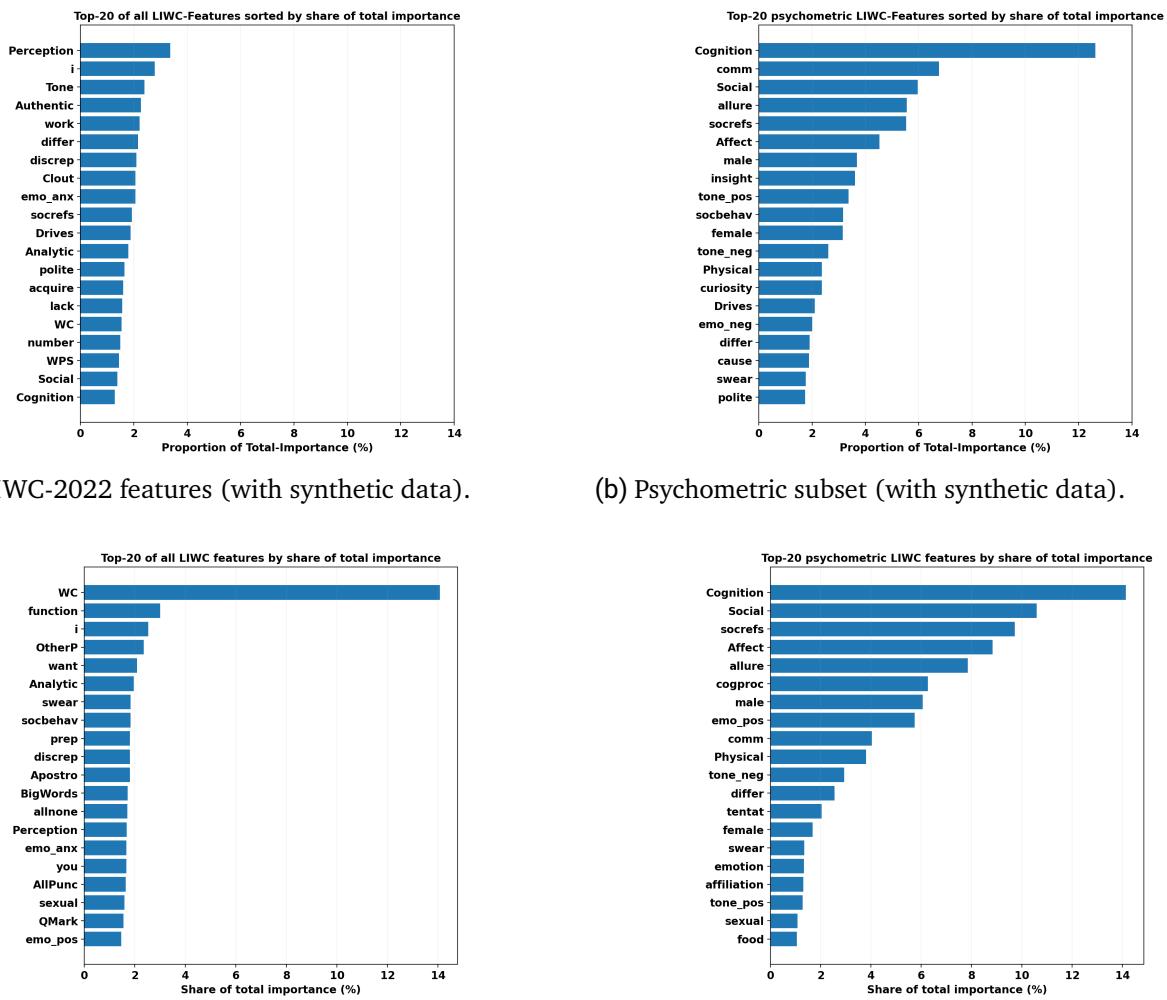
(d) Psychometric subset (no synthetic data).

Figure 6.9: **Cumulative importance of LIWC-2022 features.** Top row: with synthetic data. Bottom row: no synthetic data.

When looking at the curves of the cumulative feature importance (Figure 6.9), especially on the psychometric subset with no synthetic data, it is noticeable that the top 10 features already explain about 80% of the model decision and the top 20 features explain almost 100%. This indicates that the model relies strongly on a few key features to make its predictions when only real data is used. Furthermore, when looking at the cumulative share at the psychometric subset including synthetic data, the top 10 features explain only about 60% of the model decision, showing that the synthetic data leads to a more even distribution of feature importance across all features, which can be interpreted as a more robust model that does not rely too heavily on individual features. Still, the top 20 features explain almost 80% of the model decision, indicating the reliance of the model on less than a half of the used psychometric features.

Note, that the smaller size of the psychometric subset feature size leads in an overall higher importance per feature compared to the full feature set, where the importance is distributed over a larger number of features. In comparison to the curves based on the psychometric subset, the curves based on all LIWC-2022 features show a more gradual increase, indicating that the model relies on a broader range of features to make its predictions. This is particularly evident when synthetic data is included, where the top 20 features explain only about 40% of the model decision and the top 40 features explain about 60%. In contrast, when no synthetic data is included, the top 20 features explain about 50% of the model decision and the top 40 features explain about 70% which is still a significant portion regarding the total amount of 118 features. This suggests that the inclusion of synthetic data encourages the model to consider a wider array of features, potentially leading to improved generalization and robustness. Also, it is noticeable that the last 20 feature with the least importance in the full feature set explain only about 5% of the model decision, indicating that these features contribute very little to the model's predictions and could potentially be excluded without significantly impacting the model performance.

To gain a deeper understanding of which specific LIWC features are the most influential in the model's decision-making process, a global ranking of the top 20 features was created based on their percentages of the total significance since the top 20 features hold the majority of the model's importance for both fusion configurations. The results are presented in the following Figure 6.10.



(a) All LIWC-2022 features (with synthetic data). (b) Psychometric subset (with synthetic data).
(c) All LIWC-2022 features (no synthetic data). (d) Psychometric subset (no synthetic data).

Figure 6.10: Top 20 LIWC features ranked by percentages of the total significance. Top row: with synthetic data. Bottom row: no synthetic data.

The global rankings of the top 20 features (Figure 6.10) confirm the observations from the cumulative curves. A comparatively small core of features contributes most to the model decision, with the type of features used differing between the full LIWC set and the psychometric subset, as well as between training with and without synthetic data.

All LIWC features (with synthetic data). What is striking is a stronger weighting of *content/psychological categories like: Perception, i, Tone, Authentic, Analytic, Social/Cognition* and relationship and status markers such as *Clout* and *polite*. Formal variables such as *WC, WPS* and *number* also appear in the top 20, but are significantly less pronounced. This suggests that synthetic data demonstrably reduces dependence on pure length signals and thus emphasizes semantic-psychological signals more.

Note that, as already shown in the cumulative curves, the percentage feature importance per feature is generally lower for all LIWC-Features than in the psychometric subset. As already mentioned, this is probably

explained by the generally larger feature set, which distributes the importance over a larger number of features that the model can rely on.

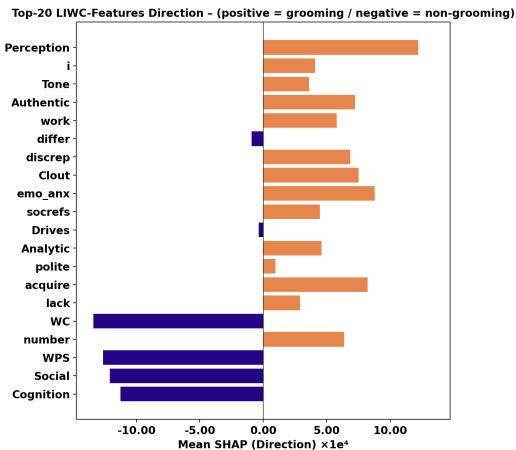
All LIWC features (without synthetic data). However, without synthetic data, *formal* features dominate: WC clearly ranks at the top, followed by *function*, pronouns (*i, you, OtherP*) and punctuation and word form features (*AllPunc, QMark, Apostro, BigWords*). Psychological categories (*Analytic, Perception, emo_anx, sexual*) are present, but are clearly overshadowed by style and length indicators.

Psychometric subset (with synthetic data). Here, the focus is clearly on cognitive, social and affective processes such as Cognition (dominant), Social including socrefs/socbehav, Affect, *allure, comm, insight, tone_pos/tone_neg* and *Drives*. Gender markers (*male/female*) and *Physical* appear as contextual signals. The distribution is more balanced than without synthetic data.

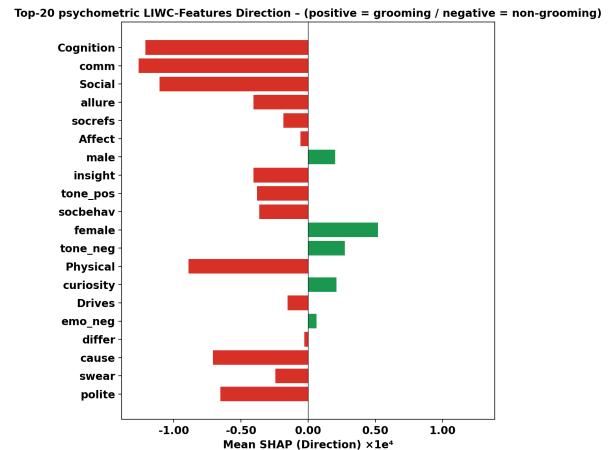
Psychometric subset (without synthetic data). The ranking remains stable at its core (*Cognition, Social, Affect, allure*), but the top is even more concentrated. In addition, thematic markers (*sexual, food*), uncertainty and deliberation signals (*tentat, differ*) and affiliation (*affiliation*) are more prominent. This indicates that the model relies particularly heavily on fewer psychological features and more individual thematic cues in real data. Generally, it can be said, that the more even distribution of feature importance when using synthetic data suggests that the model is less dependent on individual features and thus more robust. Also, the importance of the features varies depending on the used configuration, which indicates that the model relies on different features to make its predictions. Still, in the psychometric subset, features like *Cognition, Social, allure* and *Affect* are consistently important across configurations, highlighting their relevance in distinguishing grooming from non-grooming conversations while in the full feature set, the importance of features varies more widely, indicating that the model may be leveraging a broader range of linguistic cues when all LIWC-2022 features are available.

To further gain insights into the direction of the effects of the individual features, the mean signed SHAP values were visualized in the following section.

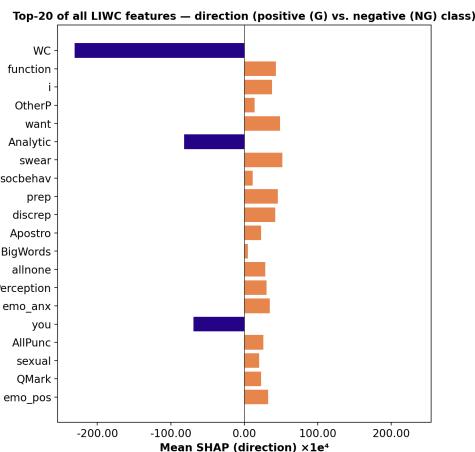
Singed Feature Importance by Class



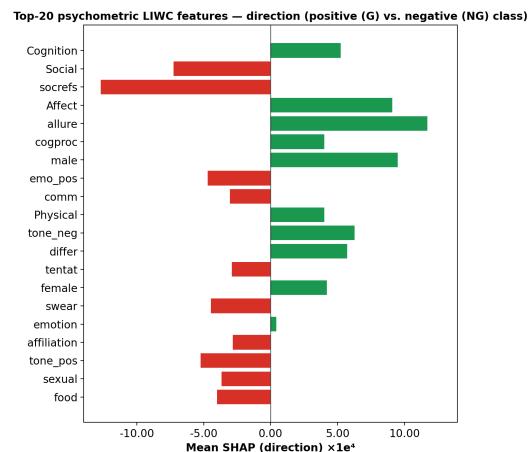
(a) All LIWC-2022 features (with synthetic data).



(b) Psychometric subset (with synthetic data).



(c) All LIWC-2022 features (no synthetic data).



(d) Psychometric subset (no synthetic data).

Figure 6.11: **Top 20 LIWC features ranked by mean signed SHAP value.** Top row: with synthetic data. Bottom row: no synthetic data. Positive values indicate a shift toward the grooming class, negative values toward the non-grooming class. Note: the scales differ between the plots. The left-hand plots (all features) have a larger range than the right-hand plots (psychometric subset).

Figure 6.11 shows the top 20 LIWC features ranked by their mean signed SHAP values for both configurations (all LIWC-2022 features and psychometric subset) with and without synthetic data. Note, that the mean signed Feature plots show the same features as the global importance plots, but now colored by their direction of effect. Positive values indicate a shift toward the grooming class, negative values toward the non-grooming class.

Note on axis scaling (effect magnitudes). The x-axis (mean signed SHAP value) is scaled very differently across panels. In particular, *All LIWC-2022 features (with synthetic data)* spans roughly $[-200, +200] \times 10^4$, whereas *Psychometric subset (no synthetic data)* spans about $[-10, +10] \times 10^4$ and *Psychometric subset (with synthetic data)* spans about $[-1.00, +1.00] \times 10^4$.

synthetic data) about $[-1, +1] \times 10^4$. Thus, absolute effects are *much larger* in the "All LIWC + synthetic" panel, even when bars look visually similar, followed by the psychometric/no-synthetic panel and, lastly, the psychometric/with-synthetic panel. Consequently, bar lengths should be compared primarily *within* a panel and not in between. For cross-panel comparisons, only focus on *direction* (sign) and *ranking*.

Interpretation of directional effects. Instead of listing individual features, Figure 6.11 reveals a pattern according to which dialogues are classified as grooming or non-grooming:

Language that is structured in a *cognitive-reflexive and cooperative* manner (planning/explaining, referring to conversation partners, polite and predominantly positive tone) shifts the decision toward **non-grooming**. In contrast, *emotionally charged, boundary-testing and closeness-creating* signals (negative/tense tone, heightened affect, uncertainty/consideration, targeted information gathering or "luring," and explicit sexual references) indicate **grooming**. This picture complements the global and cumulative analyses of the previous sections. The content-related psychological processes drive the separation, while formal text features (length, punctuation) act as stronger indicators, especially without synthetic data.

All features with vs. without synthetic data With synthetic data, the *semantic-psychological* contrast becomes clearly apparent. Here, cognitively and socially ordered and positive conversation tends to lean toward non-grooming, while tonal/affective tensions, authenticity/status signals and typical questioning movements (e.g., about age/contact) shift the decision toward grooming. Without synthetic data, however, the decision shifts more frequently toward non-grooming via *formal proxies* such as text length or punctuation patterns, while grooming components are fed by many smaller, stylistic-affective cues. This shift toward formal characteristics explains the steeper, "sharper" cumulative importance without synthetic data and underscores the risk of more unnatural correlations.

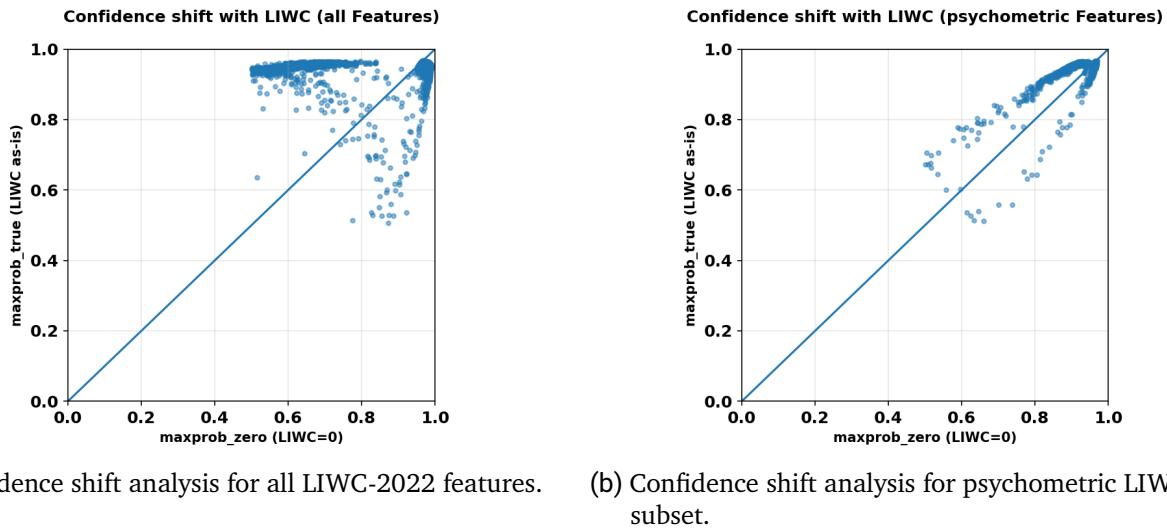
Psychometric subset with vs. without synthetic data. In the psychometric space, the decision direction is particularly plausible. *Cognition/cooperation/positive tone* mark non-grooming, while *negative affect, tone intensification, boundary tests/uncertainty and intimate context signals* mark grooming. With synthetic data, however, ambivalent categories (such as "compliments/allure" or vulgar language markers) are smoothed out and lose their one-sided grooming tendency. Without synthetic data, the profile becomes sharper and grooming and boundary testing patterns come to the fore. The cumulative curves 6.9 explain this difference. Synthetic data distributes importance more evenly and reduces the dominance of individual "sharp" indicators.

Note that the SHAP results indicate that several features change their direction when synthetic data are included, suggesting that the feature attribution is sensitive to the underlying data composition. This is particularly evident in the psychometric subset, where top features like *allure* and *Cognition* shift the direction of their effect when synthetic data is included. This sensitivity highlights the importance of considering data composition when interpreting feature importance and suggests that the model's reliance on certain features may vary depending on the training data used.

6.9 Confidence Analysis and Label Flip Analysis

In addition to the performance metrics and feature importance analysis, further evaluations were conducted to analyze the model's confidence in its predictions and to analyze instances where the model's predictions

changed between evaluations after using LIWC-Features als additional model Inputs. This analysis was conducted on the dataset including synthetic data to assess the overall behavior of the model with and without LIWC features. For the analysis, a total of 6849 samples were evaluated to reduce computational effort.



(a) Confidence shift analysis for all LIWC-2022 features. (b) Confidence shift analysis for psychometric LIWC-2022 subset.

Figure 6.12: **Confidence shift analysis with LIWC features.** Left: all LIWC-2022 features. Right: psychometric subset of LIWC-2022 features. Both plots show model confidence with LIWC set to zero (x-axis) versus LIWC as-is (y-axis). The diagonal indicates no effect. Points above the diagonal correspond to increased confidence due to LIWC features, while points below indicate reduced confidence. The farther a point lies from the diagonal, the stronger the influence of LIWC features on model confidence.

When looking on figure 6.12 it can been seen that for all LIWC-2022 features (left), a lot of samples lie above the diagonal, indicating that the model's confidence in its predictions increased when LIWC features were included. Also, most of the samples lie close to the upper border (0.9 - 1.0) indicating an overall really high confidence of the model in its predictions using all LIWC-2022 features. This suggests that the additional information provided by the full LIWC feature set helps the model make more confident predictions in most of the cases. Note that there are also some points underneath the diagonal showing that in some cases the model confidence decreased when LIWC features were included.

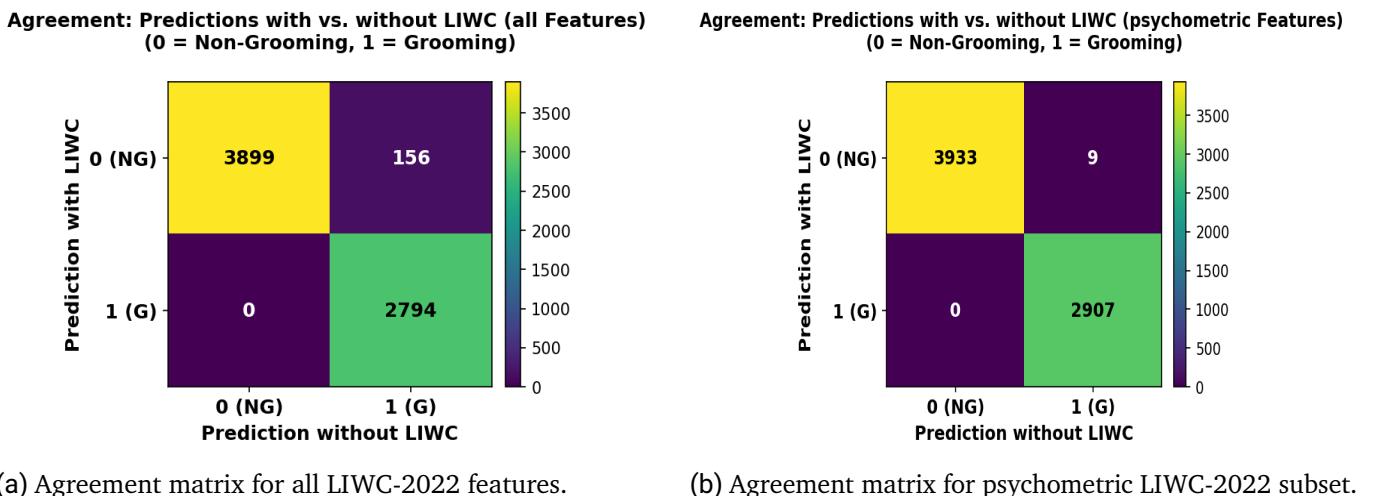
When looking on the psychometric subset of LIWC-2022 features (right), a similar pattern can be observed, but the effect is less pronounced. While there are still many points above the diagonal indicating increased confidence, the points are more widely spread and lie closer to the diagonal, suggesting that the impact of the psychometric subset on model confidence is more moderate compared to using all LIWC features. This indicates that while the psychometric features are still valuable, they may not provide as much additional information as the full feature set.

The following table 6.10 additionally shows the results of the confidence and label flip analysis for both the complete LIWC-2022 feature set and the psychometric subset after analyzing 6849 samples.

Table 6.10: Confidence and label flip analysis for all LIWC-2022 features vs. psychometric subset.

Metric	All Features	Psychometric Subset
Number of samples (n)	6849	6849
$\Delta\mu$ (mean)	0.1543	0.0219
$\Delta\tilde{x}$ (median)	0.2363	0.0313
$\Delta\sigma$ (std)	0.1481	0.0256
Δp_{10}	-0.0166	-0.0020
Δp_{90}	0.3081	0.0454
Class 0 count	4055	3942
Class 1 count	2794	2907
Flipped predictions	156	9
Flip rate	0.0228	0.0013

Additionally the Agreement matrices for the model with all LIWC-2022 features and the psychometric subset are shown in the following figure:



(a) Agreement matrix for all LIWC-2022 features.

(b) Agreement matrix for psychometric LIWC-2022 subset.

Figure 6.13: **Agreement matrices for predictions with vs. without LIWC features.** Left: all LIWC-2022 features. Right: psychometric subset of LIWC-2022 features. Each matrix compares the predicted class with LIWC features (y-axis) against the prediction without LIWC features (x-axis). Values on the diagonal indicate stable predictions, while off-diagonal values represent *label flips*.

When looking at table 6.10, it can be seen that the model with all LIWC-2022 features shows a much stronger confidence shift (mean increase of 0.1543) compared to the psychometric subset (mean increase of 0.0219). This indicates again, that adding the full LIWC-feature set has a more pronounced effect on the models confidence in its predictions. The median values also reflect this trend, with a larger increase for the full feature set (0.2363) compared to the subset (0.0313). The standard deviation is higher for the full feature set (0.1481) than for the subset (0.0256), suggesting greater variability in confidence shifts when using all features which is also reflected when looking at the samples in figure 6.12.

Also, the model with the complete LIWC-2022 Feature set seems to have a higher amount of label flip, flipping

156 prediction labels in total from Label Grooming to label Non-Grooming. In comparison, the model using only the psychometric subset of LIWC-Features has this kind of label flip. This indicates that the inclusion of all LIWC-2022 features leads to more changes in the model's predictions, which could be due to the additional information provided by a higher amount of LWIC-Features.

Note, that there is only a label flip direction of Grooming to Non-Grooming, but not the other way around. This suggests that the LIWC features help the model to be more conservative in its predictions, reducing false positives by reclassifying some grooming predictions as non-grooming. Also, when looking at the total amount of samples which were analyzed (6849 in total), the number of label flips is relatively small (2.28% for all features and 0.13% for the psychometric subset), indicating that the model's predictions are generally stable even when LIWC features are not included.

Overall, these results suggest that while both configurations enhance model confidence, the full LIWG-2022 feature set has a more substantial impact on both confidence shifts and prediction stability.

6.10 Missclassification Analysis based on LIWC-Scores

6.10.1 Results Summary: Full LIWC feature set

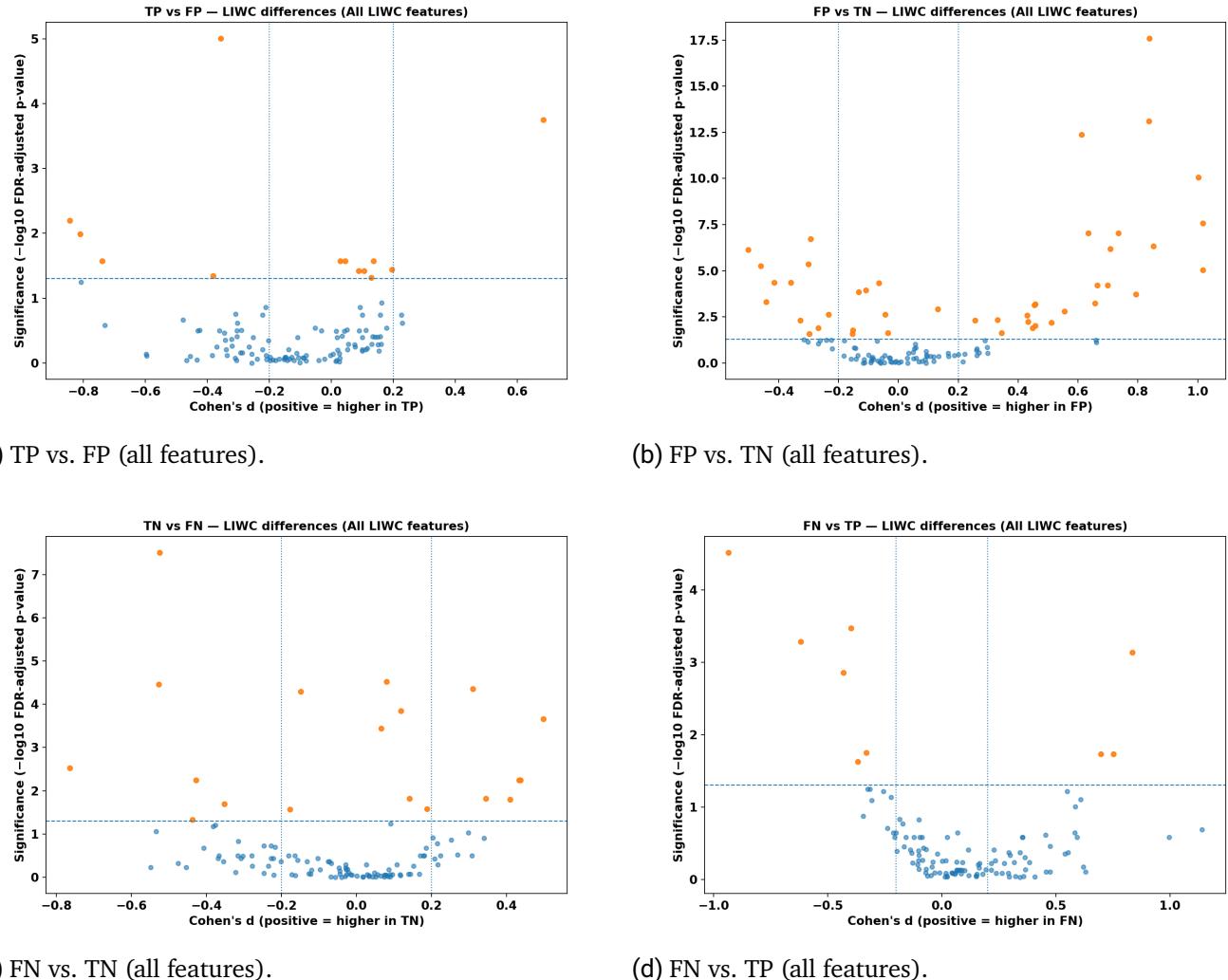


Figure 6.14: **Volcano plots for all LIWC-2022 features.** Top row: TP vs. FP and TP vs. TN. Bottom row: FN vs. TN and FN vs. TP.

Interpretation of Volcano Plots (All LIWC-2022 Features)

Figure 6.14 shows volcano plots for all LIWC-2022 features, comparing misclassified with correctly classified samples across all four groups (TP/TN/FP/FN). Each dot corresponds to a LIWC feature, with Cohen's d [74] on the x -axis (direction and magnitude of effect size) and significance on the y -axis ($-\log_{10}$ of the FDR-adjusted p -value). The vertical dashed line at $x = 0$ indicates no mean difference between the groups, with features plotted to the right being more frequent in the first group and those to the left in the second group. The horizontal dashed line marks the significance threshold of $FDR = 0.05$; points above this line represent features

with statistically significant group differences. Therefore Orange points indicate significant differences between the features of the groups with FDR<0.05. Since the complete LIWC-2022 feature set is used, the number of points in each plot is the same (118) with varying numbers of significant features showing differences between comparisons groups.

Top row (TP vs. FP and TP vs. TN). The comparison of TP vs. FP (Figure 6.14a) shows a moderate number of significant LIWC-features, suggesting that false positives differ from true positives on selected LIWC dimensions, but overall remain relatively close. In contrast, the FP vs. TN comparison (Figure 6.14b) displays clearly more significant differences, indicating that false positives and true negatives have more distinct positions in LIWC space. Taken together, this supports the interpretation that FP samples are linguistically closer to TP than to TN, leading to their missclassification based on LIWC features.

Bottom row (FN vs. TN and FN vs. TP). For FN vs. TN (Figure 6.14c), some more significant features appear in comparison to the top row, suggesting measurable differences between these groups. FN vs. TP (Figure 6.14d) shows fewer significant points, indicating that false negatives are more similar to TP than to TN, although both comparisons reveal noticeable divergence. This suggests that FN samples have an intermediate position, but lean slightly closer to TP.

Taken together, the volcano plot patterns suggest that FP samples are more strongly aligned with TP, while FN samples occupy a more ambiguous space between TP and TN. Note, that the model trained on the total LIWC feature subset set achieves very high classification performance ($F1=0.987$), so the number of misclassifications is very low (only 96 out of 13,367 samples). This limits the statistical power of the comparisons, especially for FN (only 41 samples) and may explain the relatively small number of significant features in some contrasts.

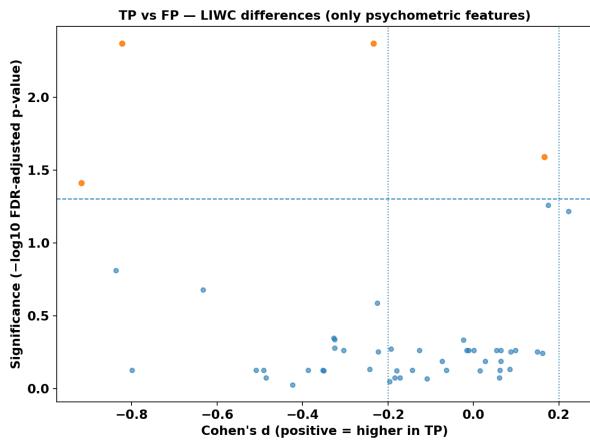
In addition table 6.11 summarizes key statistics from the misclassification analysis using all 118 LIWC-2022 features. It includes counts of true/false positives/negatives, centroid distances, numbers of significant features in each pairwise comparison, effect size statistics and proximity test results as an overview of the findings.

Table 6.11: Summary of misclassification analysis with all 118 LIWC features.

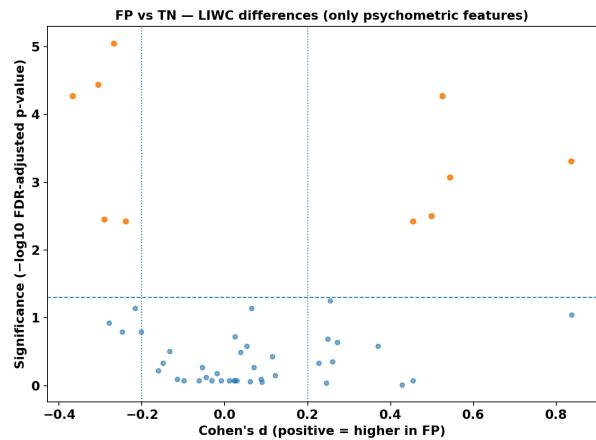
Metric	Value
# Samples (total)	13,367
TN / TP / FP / FN	9,543 / 3,728 / 55 / 41
Centroid dist. TN vs. TP (z-space)	3.04
FP vs. TN #sig (FDR<0.05)	44
Median $ d $ of sig. features	0.445
Max $ d $ of sig. features	1.018
Top-5 features (by $ d $ & FDR)	focusfuture, discrep, verb, i, Linguistic
FN vs. TP #sig (FDR<0.05)	9
Median $ d $ of sig. features	0.618
Max $ d $ of sig. features	0.936
Top-5 features	WC, politic, adj, BigWords, acquire
TN vs. FN #sig (FDR<0.05)	19
Median $ d $ of sig. features	0.352
Max $ d $ of sig. features	0.764
Top-5 features	focusfuture, function, Linguistic, male, ppron
TP vs. FP #sig (FDR<0.05)	13
Median $ d $ of sig. features	0.195
Max $ d $ of sig. features	0.844
Top-5 features	sexual, Comma, BigWords, WC, function
FN closer to TN than TP	Prop. 0.463, $p_{\text{bin}} = 0.73$
FP closer to TP than TN	Prop. 0.873, $p_{\text{bin}} < 10^{-8}$

Summary False positives are strongly aligned with true positives in LIWC space: nearly 87% of FP samples are closer to the TP centroid than to the TN centroid, a result that is highly significant across all proximity tests. This pattern is reflected in the per-feature comparisons, where 44 LIWC features differ significantly between FP and TN with medium-to-large effect sizes. In contrast, false negatives do not consistently resemble true negatives: only 46% are closer to the TN centroid and no statistical evidence supports a systematic shift toward TN. Per-feature differences between FN and TP exist for a small subset of features (9 significant), some with relatively strong effect sizes, but overall less consistent than for FP vs. TN. FN vs. TN contrasts also reveal some differences (19 features), suggesting partial but not decisive separation. Overall, misclassification patterns in the full LIWC space are dominated by the FP–TP alignment.

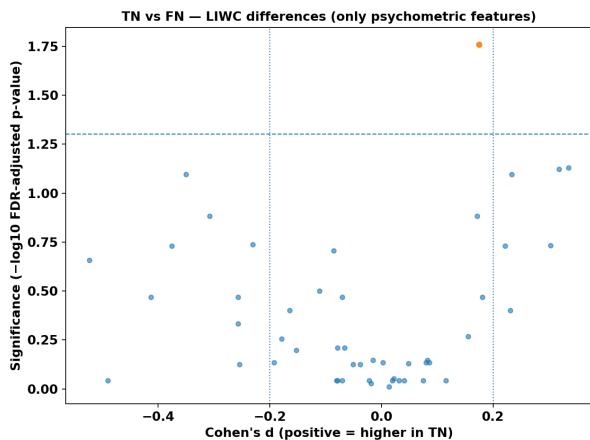
6.10.2 Results Summary: Psychometric LIWC subset



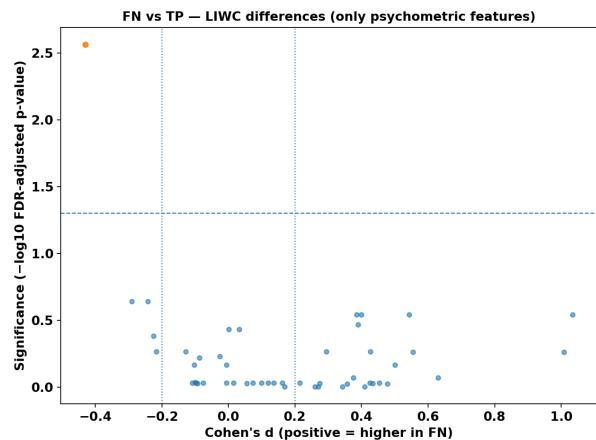
(a) TP vs. FP (psychometric subset).



(b) FP vs. TN (psychometric subset).



(c) FN vs. TN (psychometric subset).



(d) FN vs. TP (psychometric subset).

Figure 6.15: Volcano plots for psychometric LIWC features. Top row: TP vs. FP and TP vs. TN. Bottom row: FN vs. TN and FN vs. TP.

Interpretation of Volcano Plots (Psychometric LIWC Subset)

Figure 6.15 shows volcano plots restricted to the 49-dimensional psychometric LIWC subset. The plot layout and axes are the same as in the previous section and can be interpreted in the same way. The only difference is the number of points, which is now 49 in each plot, corresponding to the total number of psychometric LIWC features. Also, compared to the full feature set, the number of significant effects is much smaller, which indicates that the psychometric dimensions have fewer differences between confusion groups.

Top row (TP vs. FP and FP vs. TN). The TP vs. FP comparison (Figure 6.15a) shows only a very small number of significant features, suggesting that false positives resemble true positives fairly closely in terms of the psychometric LIWC categories. In contrast, FP vs. TN (Figure 6.15b) produces the highest concentration

of significant differences, highlighting that false positives are linguistically much closer to true positives than to true negatives in this reduced feature space.

Bottom row (FN vs. TN and FN vs. TP). The FN vs. TN comparison (Figure 6.15c) shows only one significant feature, while FN vs. TP (Figure 6.15d) reveals a similarly sparse pattern. This suggests that false negatives are difficult to distinguish from both true positives and true negatives using only the psychometric subset, although the relative number of significant effects still points to slightly more overlap with TPs than TNs.

Summary. Overall, the psychometric subset results in weaker separability compared to the full LIWC feature set. Again it has to be said, that the model trained on the psychometric LIWC feature subset also achieved a very high classification performance ($F1=0.985$), so that the number of misclassifications is very low (only 99 out of 13,367 samples). This limits the statistical power of the comparisons, especially for FN (only 40 samples) and may explain the relatively small number of significant features in some contrasts. The patterns still suggest that false positives align more closely with true positives than with true negatives, while false negatives occupy a more ambiguous position with limited distinctiveness in this reduced feature space. The following table 6.12 summarizes key statistics from the misclassification analysis using the psychometric LIWC-2022 subset including the same metrics as table 6.11 above.

Table 6.12: Summary of misclassification analysis with psychometric LIWC subset (49 features).

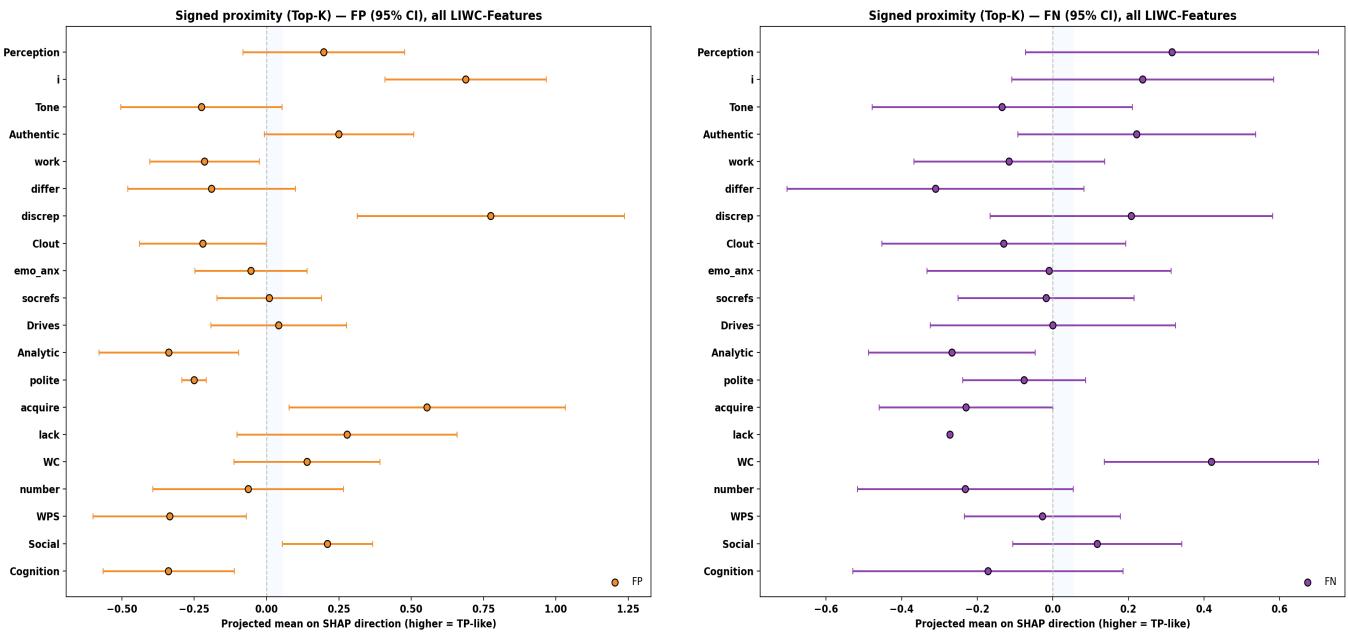
Metric	Value
# Samples (total)	13,367
TN / TP / FP / FN	9,539 / 3,729 / 59 / 40
Centroid dist. TN vs. TP (z-space)	1.54
FP vs. TN #sig (FDR<0.05)	10
Median $ d $ of sig. features	0.411
Max $ d $ of sig. features	0.836
Top-5 features (by $ d $ & FDR)	discrep, allnone, allure, cogproc, Cognition
FN vs. TP #sig (FDR<0.05)	1
Median $ d $ of sig. features	0.430
Max $ d $ of sig. features	0.430
Top-5 features	family
TN vs. FN #sig (FDR<0.05)	1
Median $ d $ of sig. features	0.175
Max $ d $ of sig. features	0.175
Top-5 features	polite
TP vs. FP #sig (FDR<0.05)	4
Median $ d $ of sig. features	0.528
Max $ d $ of sig. features	0.918
Top-5 features	swear, allnone, mental, family
FN closer to TN than TP	Prop. 0.55, $p_{bin} = 0.32$
FP closer to TP than TN	Prop. 0.661, $p_{bin} = 0.009$

Summary (psychometric subset). Within the psychometric subset, the alignment of false positives with true positives remains visible but weaker than with the full LIWC feature set: about two thirds of FP samples are closer to the TP centroid than to TN, a result that is still statistically significant. The number of significant per-feature differences between FP and TN is smaller (10 features), though several show medium effect sizes. FN samples again show no consistent resemblance to TN, with proportions close to chance and non-significant tests. Only a single feature differentiates FN from TP and TN, indicating limited explanatory power of psychometric LIWC categories for FN errors. Overall, the psychometric subset captures some meaningful FP–TP alignment, but FN patterns remain largely unresolved.

While this full-feature analysis provides an overview of differences between confusion groups, it also introduces substantial noise from dimensions that carry little explanatory value. To refine the evaluation, a focused analysis was therefore conducted on the Top-20 LIWC features, identified as most influential according to global SHAP values (Section 6.11). This allows for a clearer examination of whether misclassifications align with the patterns of the classes they were mistaken for.

6.11 Shap-based Analysis of Missclassification

6.11.1 Total LIWC Feature Set



(a) False Positives vs. True Positives/Negatives. The orange dots indicate the mean positions of the false positives, with horizontal bars showing 95% confidence intervals.

(b) False Negatives vs. True Positives/Negatives. The purple points indicate the mean positions of the false negatives, with horizontal bars showing 95% confidence intervals.

Figure 6.16: **Signed proximity plots for all LIWC-2022 features.** Left: False Positives relative to True Positives and True Negatives. Right: False Negatives relative to True Positives and True Negatives. Features are ordered by global SHAP importance. The x-axis is aligned with the global SHAP direction. Values to the right of zero indicate greater similarity to true positives, values to the left indicate greater similarity to true negatives. The purple/orange dots show the mean projected position of the misclassified samples.

Figure 6.16 shows the signed proximity plots for the Top-K LIWC features based on global SHAP importance when using all LIWC-2022 features. Note, that the sample size for FP (55) and for FN (41) is very small, which automatically makes the confidence intervals broader and the results less reliable. Therefore, the results have to be interpreted with caution since broader confidence intervals increase the probability of crossing the zero-line. Still, the direction of the projected means, where values greater than zero indicate greater similarity to true positives, while values less than zero indicate greater similarity to true negatives, as well as the aggregated proximity statistics in table 6.13 can provide dominant trends in the top-20 LIWC features according to missclassification. Therefore, when interpreting the results, the focus should rather lie on the consistent pattern across all features than only the individual features itself. There are two visible effects that can be captured when looking at the plots and the table:

- the position of the projected group means relative to zero (indicating similarity to TP vs. TN)

- the proportion proximity rate, showing the proportion of the top 20 features for which FP is closer to TP or FN is closer to TN

Even with broad confidence intervals, the general direction stays visible, while the confidence intervals counts make the uncertainty per feature transparent.

When looking at the False Positives (left plot) at Figure 6.16, it can be seen, that especially the features *Perception*, *i*, *Authentic* and *discrep* lie more on the TP side (right side of zero), while the features *Tone*, *work* and *differ* lie more on the TN side (left side of zero). Four features have 95% CIs entirely on the TP side: *i* (rank 2), *discrep* (rank 7), *acquire* (rank 14) and *Social* (rank 19). Five features have confidence intervals fully on the TN side: *work* (rank 5), *Analytic* (rank 12), *polite* (rank 13), *WPS* (rank 18) and *Cognition* (rank 20). All remaining top-20 features, including highly ranked ones such as *Perception* (rank 1), *Tone* (rank 3) and *Authentic* (rank 4), cross zero and therefore their alignment is ambiguous. Notably, confidence intervals on the positive (TP-like) side tend to be broader than on the negative side, indicating greater uncertainty. Despite the 50:50 split in mean positions, the fact that the **clearly TP-aligned features include high-ranked items, while several clearly TN-aligned features sit in the mid-to-lower ranks supports the proximity result**. Overall, FP appear more TP-like on the most influential LIWC dimensions, which is also reflected by the proximity rate in Table 6.13).

When looking at the False Negatives in figure 6.16 (right), a more ambiguous pattern compared to the False Positives can be seen. The projected means span a balanced range between -0.7 and +0.7, with most of the confidence intervals crossing the zero line. Only a few features show stable tendencies, such as *WC* (rank 15), which lies fully on the TP side, while *Analytic* (rank 12) and *lack* (rank 15) are completely TN-oriented but on a lower rank than the ambiguous top features. The generally wider confidence intervals reflect the smaller FN sample size ($n=41$) and further underscore the uncertainty. **Note, that the proximity rate in table 6.13 is only 0.30 even if more of the projected group means lie on the TN side (supporting the hypothesis that FN should be closer to TN)**. This is due to the fact that the proximity rate is based on a **distance comparison** to the TP and TN group means. Since many FN samples lie close to zero on the SHAP-oriented axes and TP are also positioned close to zero, the distance to TP (d_{TP}) is often smaller than the distance to TN (d_{TN}) – even if the FN mean value is negative. Wide confidence intervals (caused by the small FN sample size) further amplify this effect. The CI categories (TP-side, TN-side, or crosses-0) and the proximity rate thus capture different aspects. The former reflects the direction of each feature, while the latter quantifies the relative closeness to the class means. Taken together, both provide a picture of a TN tendency with many borderline cases. This highlights that the confidence interval based directionality and the proximity rate should not be seen as contradictory, but rather as complementary. The former indicates whether features clearly align with TP or TN, while the latter quantifies relative distances. **Especially for higher-ranked features, consistent tendencies across both measures offer stronger evidence than individual features alone.**

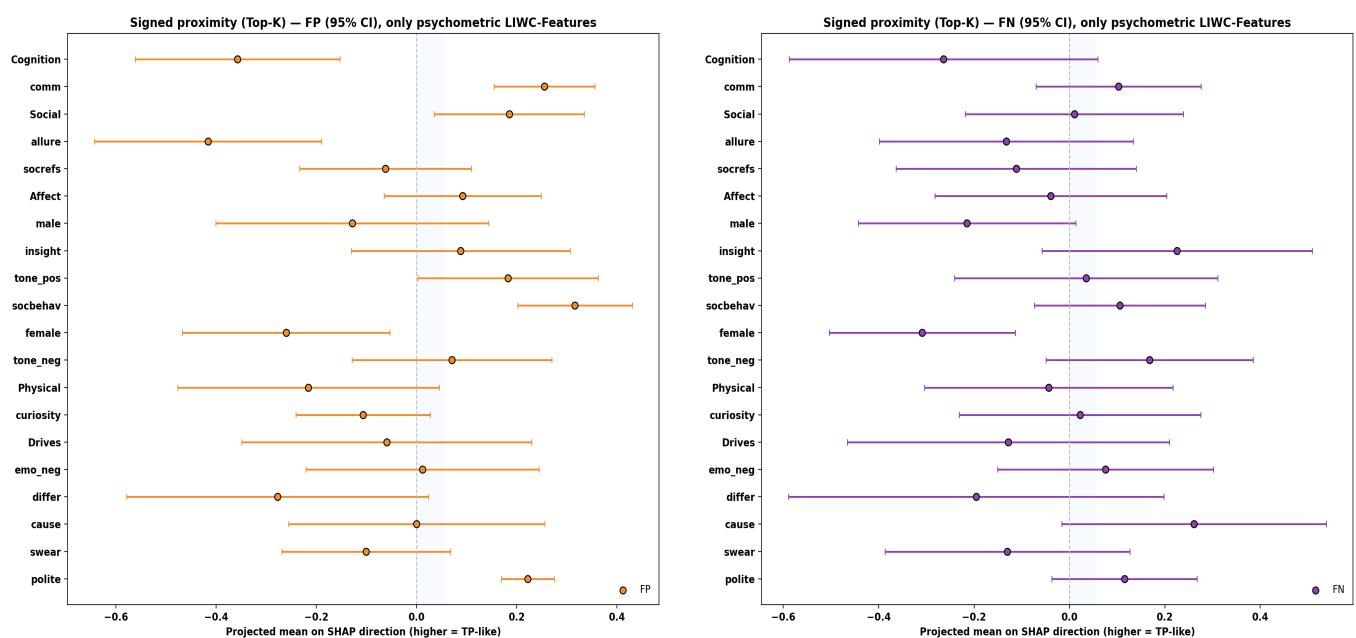
When considering the SHAP ranking of features, it becomes clear that alignment in the higher-ranked dimensions is more informative than in the lower-ranked ones. However, it remains difficult to draw a clear conclusion for the FN group. While the two highest-ranked features (*Perception* and *i*) show a tendency towards the TP side, most of the remaining features are highly ambiguous with confidence intervals crossing zero. This overall uncertainty makes it challenging to determine whether FNs generally align more with TNs, despite some individual features suggesting such a trend.

Table 6.13 summarizes the key proximity statistics for the Top-K LIWC features based on global SHAP importance when using all LIWC-2022 features.

Table 6.13: Key proximity results for Top-K of all LIWC features. The proximity rate was calculated by comparing the mean position of misclassified samples (FP or FN) to the means of the correctly classified groups (TP and TN) along the SHAP-oriented feature axis. The columns CI TP-side/TN-side/crosses-0 indicate the number of features (out of the Top-K) for which the 95% confidence interval of the projected mean lies entirely on the TP side (greater than zero), entirely on the TN side (less than zero), or crosses zero (indicating uncertainty).

Group	Proximity rate	CI TP-side (n)	CI TN-side (n)	CI crosses 0 (n)
FP	0.75	4	6	10
FN	0.30	1	2	17

6.11.2 Psychometric LIWC Subset



(a) False Positives vs. True Positives/Negatives. The orange dots indicate the mean positions of the false positives, with horizontal bars showing 95% confidence intervals.

(b) False Negatives vs. True Positives/Negatives. The purple points indicate the mean positions of the false negatives, with horizontal bars showing 95% confidence intervals.

Figure 6.17: Signed proximity plots for the psychometric LIWC-2022 subset. Left: False Positives relative to True Positives and True Negatives. Right: False Negatives relative to True Positives and True Negatives. As before, features are ordered by global SHAP importance and the x-axis shows projected means on the SHAP direction where values greater than zero indicate greater similarity to true positives, while values less than zero indicate greater similarity to true negatives. Again, the purple/orange dots show the mean projected position of the misclassified samples.

Figure 6.17 shows the signed proximity plots for the Top-K LIWC features based on global SHAP importance when using the psychometric LIWC-2022 subset. Again, the sample size for FP (59) and for FN (40) is

very small, which automatically makes the confidence intervals broader and the results in the plots and table less reliable.

When looking at the False Positives (left plot) in Figure 6.17, it can be seen that especially the feature *Cognition* lies on the TN side, while the features *comm* and *Social* lie completely on the TP side. Generally, for the False Positives, fewer confidence intervals cross the zero line compared to the False Negatives, indicating a somewhat clearer psychometric tendency. This suggests that False Positives more often display stable TP-like psychometric signals, which explains why they were misclassified despite belonging to the non-grooming class. At the same time, the presence of individual TN-oriented features such as *Cognition* highlights that their profiles are not entirely homogeneous, but rather contain a mix of grooming-like and neutral tendencies. This is also supported by the proximity rate in table 6.14, which shows that 70% of the False Positive means lie closer to the True Positive means. Taken together, it can be said that the False Positives show a moderately clear pattern with several features indicating stable TP-like psychometric tendencies, while others remain ambiguous. This suggests that the misclassified samples often contain psychometric signals resembling grooming-like communication, but also include some neutral elements. The presence of multiple confidence intervals crossing zero further underlines that psychometric features alone are insufficient to clearly separate those cases. Consequently, the observed False Positives likely result from a complex interaction between psychometric cues and linguistic triggers, highlighting the necessity of integrating psychometric features with richer linguistic or sequential indicators to improve robustness.

When looking at the False Negatives in figure 6.17 (right), a similar pattern compared to the False Negatives at figure 6.16 can be seen. The projected means again span a balanced range between -0.7 and $+0.7$, with most of the confidence intervals crossing the zero line. Only a few features show stable tendencies, such as female (rank 11), which lies fully on the TN side, while no confidence interval is completely TN oriented. When looking at the proximity rate in table 6.14, it can be seen that the proportion is 0.45, indicating that 45% of the False Negative means lie closer to the True Negative means. Taken together with the results of the confidence intervals, it can be said that the False Negatives show again a strongly ambiguous pattern with many borderline cases. In other words, their psychometric profile does not consistently resemble that of True Negatives, but rather fluctuates around the decision boundary. This suggests that the misclassified samples are not simply “harmless” cases that the model failed to detect, but instead contain psychometric signals partly overlapping with grooming-like communication. The high number of confidence intervals crossing zero further underlines that the psychometric features alone are insufficient to clearly separate those cases. Consequently, the observed False Negatives likely result from a complex interaction between weak psychometric cues and missing or subtle linguistic triggers, highlighting the necessity of integrating psychometric features with richer linguistic or sequential indicators to improve robustness.

Table 6.14: Key proximity results for Top-K psychometric LIWC features. The proximity rate was calculated by comparing the mean position of misclassified samples (FP or FN) to the means of the correctly classified groups (TP and TN) along the SHAP-oriented feature axis. The columns CI TP-side/TN-side/crosses-0 indicate the number of features (out of the Top-K) for which the 95% confidence interval of the projected mean lies entirely on the TP side (greater than zero), entirely on the TN side (less than zero), or crosses zero (indicating uncertainty).

Group	Proximity rate	CI TP-side (n)	CI TN-side (n)	CI crosses 0 (n)
FP	0.70	5	3	12
FN	0.45	0	1	19

7 Discussion

In this chapter, the presented results will be discussed and interpreted in a larger context. The implications of the findings for real-world applications will be considered, as well as the limitations of the study and potential avenues for future research.

7.1 Main Findings and Interpretation

The main findings of this thesis will be summarized and interpreted in the following sections. The focus will be on the challenges of data leakage and shortcut learning, the strong baseline performance, and the effectiveness of the proposed feature fusion architecture that integrates psycholinguistic features from the LIWC lexicon into a BERT-based model for online grooming detection.

7.1.1 Addressing Data Leakage and Shortcut Learning

The near-perfect baseline performance ($F1 \approx 0.999$) observed in initial BERT experiments was identified as a potential symptom of **data leakage, primarily due to domain and length artifacts between the PJ and PAN12 datasets**. To overcome this challenge, the generation of synthetic non-grooming chats in PJ style with a higher length than typical PAN12 chats was employed. By introducing synthetic non-grooming PJ chats, the model could no longer rely on source domain as a label proxy, and by enforcing stricter train/test splits, chunk-based leakage across datasets was prevented. Also, label smoothing was applied to mitigate overconfidence in its predictions, as well as an increased dropout rate to reduce overfitting on the training data and therefore improve generalization to unseen data. To further analyze the impact of chunk lengths, the improved baseline was tested with a fixed padding on different chunk sizes, revealing that shorter chunks led to a **slight decrease in performance**. Still, the performance was really high across all three different kind of chunk sizes (512, 250, 150 tokens), indicating that the model could still learn relevant patterns even in smaller text segments.

Also, it was striking, that including synthetic non-grooming Chats only in the test set, a strong drop in performance was visible, highlighting the lack of generalization, as well as a strong domain specific shortcut learning. This highlighted the risks associated with shortcuts in state-of-the-art models, particularly in security applications where robustness is critical. However, questions regarding the generalizability of synthetic negative examples and potential residual artifacts remain open for further investigation. When integrating the synthetic data in the train set, the performance drop disappeared, showing that the model could learn to generalize better with the synthetic data. It should be highlighted, that the synthetic data especially affected the precision, since the model achieved a much higher recall already without synthetic data, but the precision was much lower, indicating that the model made more false positive errors. To further improve the model's robustness, it could be considered to additionally integrate **data augmentation techniques** such as

paraphrasing or backtranslation in the training data, to increase the diversity of the training data and help the model learn more robust features that are less sensitive to specific wording or phrasing. However, in the present work such methods were deliberately avoided, since **LIWC features are lexically defined and therefore highly sensitive to textual modifications** [55]. Artificial augmentation (for example synonym replacement, backtranslation) risks shifting the distribution of key categories (for example pronouns, affective terms, sexual language), which would reduce the validity of subsequent psychometric analyses. Such distortions would likely also affect SHAP explanations, as the method would attribute importance scores based on artificially altered inputs rather than on authentic linguistic patterns, thereby undermining interpretability.

7.1.2 Addressing Baseline Performance

Given that all three chunk sizes (150, 250, and 512 tokens) had a consistently strong performance, especially when synthetic data was integrated into the training set, the subsequent feature fusion was conducted with the 512-token mixed configuration. This setup provided the richest conversational context and thereby maximized the coverage of LIWC categories within each chunk, offering the most informative basis for integration with transformer representations and allowing for a more comprehensive analysis of psychometric features. The overall strong performance across all settings can be partly attributed to the use of balanced training data, as prior work has shown that transformer-based models such as BERT are highly sensitive to class imbalance and achieve more stable and reliable results under balanced conditions [75]. Importantly, such balancing was required in this work to enable stable and meaningful SHAP analyses, ensuring that the derived explanations were not dominated by class imbalance effects [66], [67]. Still, overfitting cannot be completely ruled out as an explanation for the overall high performance after three epochs of training, even if the model was trained with increased dropout and label smoothing to mitigate overconfidence. What limits this concern is, that the relative gains from LIWC fusion over the BERT baseline were consistent across epochs 1–3. Nevertheless, **a stricter control would include a small held-out dev set with early stopping, reporting train/dev learning curves, and complementing single-split results with group-stratified k -fold cross-validation and out-of-domain evaluation.**

7.1.3 Feature Fusion Architecture (BERT + LIWC via Late Fusion with Cross-Attention)

The integration of psycholinguistic features from the LIWC lexicon into a BERT-based model for online grooming detection has demonstrated clear improvements in model performance and decision stability. The integration of psycholinguistic features from the LIWC lexicon into a BERT-based model for online grooming detection demonstrated clear improvements in performance and decision stability. The proposed feature-fusion architecture integrated LIWC features via late fusion at layer 6/12 using cross-attention with gating. This design aligns with findings from multimodal fusion research, which suggest that late fusion is particularly effective for heterogeneous modalities and avoids the risks of overloading early linguistic representations [76]. Furthermore, attribution studies show that late fusion improves interpretability, as SHAP attributions can more clearly disentangle the contributions of linguistic tokens and auxiliary features [77]. The performance of the feature-fusion model indicates that integration at layer 6/12 was especially effective in this setting, although the optimal integration depth is likely backbone-specific and requires further investigation. Also, the modular nature of the proposed mechanism makes it transferable to other transformer architectures such as RoBERTa or DeBERTa v3, which are known to be more powerful than BERT [37], [39]. However, it should be noted that these models were trained on a much broader corpus, which may include the PAN12 dataset,

potentially leading to data leakage if used as a baseline. Therefore, future work should carefully evaluate the use of these models in this context.

7.1.4 Performance Gains from LIWC Integration (Full Set vs. Psychometric Subset)

The integration of LIWC features consistently improved the F1 score to approximately 0.987 for both the full feature set and the psychometric subset after three epochs, with a **particularly notable increase in precision, leading to fewer false positives**. As the confusion matrices in figure 6.7 illustrated, the feature-fusion model led to a more balanced trade-off between precision and recall, with a significant reduction in false positive errors. This balance of precision and recall is especially relevant in practical applications as it reduces the risk of false alarms, which can be disruptive and lead to unnecessary interventions while keeping the false negative rate low. The improved precision is particularly valuable in clinical and forensic contexts, where the consequences of false positives can be significant, including reputational damage and increased workload for human reviewers.

It was shown, that the full LIWC feature set provided slightly stronger effects in the feature-fusion performance and later explainability analysis, while the psychometric subset offered a more streamlined approach with nearly equivalent performance. This has implications for deployment and complexity, suggesting that a reduced feature set may be preferable in resource-constrained environments without significant loss in effectiveness. Still, the psychometric subset allowed an analysis of the most relevant psycholinguistic categories for online grooming detection, especially when combined with SHAP explanations deepening the focus of the analysis beyond only linguistic features. Still, SHAP analyses across both the full LIWC set and the psychometric subset reveal a substantial overlap in the most influential psychometric categories, especially including cognition and social words. These features consistently emerge as strong indicators to distinguish between grooming and non-grooming behavior regardless of the feature space, suggesting a core set of psychometric markers that carry much of the discriminative signal. At the same time, the full LIWC configuration highlights additional linguistic proxies such as function words, analytic style, and word count, which indirectly reflect psychological processes. This indicates that a compact psychometric core could be sufficient for interpretability-focused applications, while the extended feature set offers complementary cues that may further strengthen classification in practice.

7.1.5 Model Confidence and Stability

The integration of LIWC features has also been shown to enhance model confidence and stability. By providing additional context and grounding for the model's predictions, LIWC features help to reduce uncertainty and variability in the decision-making process. This was evidenced by a significant increase in the median confidence shift (

$\Delta \mu \approx 0.1543$) when LIWC features were included, compared to a much smaller shift ($\Delta \mu \approx 0.0219$) when only the psychometric subset was used. The low rate of label flips (2.28% for the full set and 0.13% for the subset) further underscores the stabilizing effect of LIWC integration, with most changes being conservative reclassifications from grooming to non-grooming. This suggests that LIWC features help to clarify positive borderline cases by providing additional context and reducing ambiguity. Still, the number of label flips was very low, indicating that the model's decisions were generally stable and robust. This stability is particularly important in high-stakes applications where consistent and reliable predictions are crucial.

7.1.6 LIWC-Features in Grooming and Non-Grooming Chats

The analysis of the LIWC-Features in grooming and non-grooming on a global, chunk level and shap basis revealed several key insights into the psycholinguistic patterns associated with online grooming behavior and draw a direct line with the existing literature based on a psycholinguistic analysis of cybergrooming. Most striking is the high proportion of *future*, *home*, *family*, and *affiliation* in the PJ grooming conversations (Figure 6.4). These findings reflect typical grooming narratives, as described by Cano et al. [15]. In the trust development phase, categories such as *future* and *home* frequently occur, while the approach phase is marked by planning and frame semantics (“arriving”, “visiting”). Gupta et al. [18] confirm these patterns, identifying *family* as a marker of risk assessment (queries about parents and environment) and *time/motion* as characteristic for the conclusion/approach stage. Thus, the analysis of complete conversations (Figure 6.4) replicate established stage models of grooming.

A particularly strong signal lies in the category *discrep* (*would/should/could*), which is also elevated in the global PJ grooming conversations analysis of the psychometric subset (Figure 6.4, right). This aligns with the top predictors of the approach phase reported by Cano et al. [15], where *discrep* and *funct* serve as tools for boundary testing and conditioning. Slightly increased values for *tentat* further indicate strategies of relationship maintenance and polite, indirect approaches, as also observed by Broome et al. [5] in LIWC profiles of convicted offenders.

Beyond these psychometric dimensions, the results also highlight differences in thematic scope. Grooming conversations contain more social, relationship-oriented markers (*pronouns*, *function words*, *affiliation*), while the PAN12 non-grooming data are more strongly characterized by sexualized but non-grooming conversations (Figure 6.4, Figure 6.6). This is consistent with Gupta et al. [18], who list *social* as a marker of friendship forming and *sexual* as an indicator of the sexual phase. Broome et al. [5] also found that groomers primarily act in a relationship- and present-focused manner, while explicit sexual language is comparatively rare. This explains why sexualized markers are diluted in the global measures due to thematic diversity, but become more salient on the chunk level (Figure 6.6).

The chunk based analysis underlines, that these markers are not only detectable globally, but also already apparent in short text segments of 512 tokens (Figure 6.6). This complements the work of Cano et al. [15], who classify grooming processes at the line level according to stage-specific markers, and explains why models such as BERT can achieve high detection performance even at the chunk level. While global analyses emphasize length and diversity, local analyses highlight specific psycholinguistic patterns (sexuality, allure, cognition, social processes) that become especially evident in LIWC-based profiles.

Methodologically, it must be noted that formal features such as word count and sentence length represent strong confounders, as also emphasized by Broome et al. [5]. Without synthetic data, the model’s decision shifts more strongly towards such formal proxies, whereas with synthetic data, semantic and psycholinguistic contrasts become more pronounced (Figure 6.11). This supports the argument of An et al. [21], who state that data quality and balance are crucial to avoid artificial correlations and to make the underlying psycholinguistic processes visible.

When looking at the SHAP explanations (Figure 6.11), it is evident that the direction of effect of several features changes when synthetic data is included. This could be explained by the fact that SHAP values are **always model- and dataset-specific**. Through the integration of synthetic data, the underlying feature distributions and their correlations with the target variable shift, leading to cases where features associated with grooming in real data appear more frequently in non-grooming contexts when synthetic data are included. Moreover, the sample-based calculation of SHAP values contributes to this effect. Since the analysis was

performed on a randomly selected subset of 256 instances, **sampling variance can result in changes in the direction of individual features**. These issues are amplified when synthetic data are used, as their distributions may differ from real data, causing further adjustments of model parameters and the relative attribution of features. Overall, this highlights the sensitivity of explainable AI methods to data augmentation and sampling, and demonstrates that interpretations of SHAP directions should **always be made cautiously and in the context of the specific training and evaluation data**.

7.2 Limitations

There are several limitations within the study that need to be addressed. Firstly, the limitations of using decoy victim data from Perverted Justice should be considered. Such interactions are designed to result in an arrest, meaning a decoy may be more compliant and open to explicit talk than an adolescent victim. However, studies have shown that the adult leads these interactions, using higher frequencies of words (Drouin, Ryan et al., 2017) and directing the conversation to offline contact (Winters et al., 2017). Research has also shown that individuals often have previous offences for child pornography and similar grooming offences against actual children (Mitchell, Wolak, & Finkelhor, 2005) [5]

7.3 Future Work

Bibliography

- [1] M. Mladenović, V. Ošmjanski, and S. Vujičić Stanković, “Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges,” *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–42, 2021. doi: [10.1145/3424246](https://doi.org/10.1145/3424246).
- [2] S. Wachs, A. Schittenhelm, A. Cebulla, and M. Gámez-Guadix, “Cybergrooming victimization among young people: A systematic review of prevalence rates, risk factors, and protective factors,” *Journal of Child Online Safety*, vol. 10, no. 2, pp. 1–15, 2024. doi: [10.1007/s40894-024-00248-w](https://doi.org/10.1007/s40894-024-00248-w).
- [3] S. Webster et al., “European online grooming project - final report,” Mar. 2012.
- [4] H. Whittle, C. E. Hamilton-Giachritsis, A. R. Beech, and G. Collings, “A review of young people’s vulnerabilities to online grooming,” *Aggression and Violent Behavior*, vol. 18, no. 1, pp. 62–70, 2013.
- [5] L. J. Broome, C. Izura, and J. Davies, “A psycho-linguistic profile of online grooming conversations: A comparative study of prison and police staff considerations,” *Child Abuse & Neglect*, vol. 109, p. 104647, 2020. doi: [10.1016/j.chabu.2020.104647](https://doi.org/10.1016/j.chabu.2020.104647).
- [6] S. Craven, S. Brown, and E. Gilchrist, “Sexual grooming of children: Review of literature and theoretical considerations,” *Journal of Sexual Aggression*, vol. 12, no. 3, pp. 287–299, 2006, Conceptual discussion of grooming as a process. doi: [10.1080/13552600601069414](https://doi.org/10.1080/13552600601069414).
- [7] J. A. Kloess, A. R. Beech, and L. Harkins, “Online child sexual exploitation: Prevalence, process, and offender characteristics,” *Trauma, Violence, & Abuse*, vol. 15, no. 2, pp. 126–139, 2014. doi: [10.1177/1524838013511543](https://doi.org/10.1177/1524838013511543).
- [8] P. J. Black, M. Wollis, M. Woodworth, and J. T. Hancock, “A linguistic analysis of grooming strategies of online child sex offenders: Implications for prevention and detection,” *Child Abuse and Neglect*, vol. 44, pp. 140–149, 2015.
- [9] R. O’Connell, “A typology of cybersexploitation for youth,” *Cyberspace Research Unit, University of Central Lancashire*, 2003.
- [10] N. Lorenzo-Dus and A. Kinzel, “So is your mom as cute as you?” examining patterns of language use by online sexual groomers,” *Journal of Corpora and Discourse Studies*, vol. 2, pp. 14–39, 2019.
- [11] A. Joleby, J. A. Kloess, and A. R. Beech, “Offender strategies for engaging children in online sexual activity,” *Child Abuse & Neglect*, vol. 120, p. 105214, 2021, issn: 0145-2134. doi: <https://doi.org/10.1016/j.chabu.2021.105214>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0145213421002878>.
- [12] T. R. Ringenberg, K. Seigfried-Spellman, and J. Rayz, “Assessing differences in grooming stages and strategies in decoy, victim, and law enforcement conversations,” *Computers in Human Behavior*, vol. 152, p. 108071, 2024, issn: 0747-5632. doi: <https://doi.org/10.1016/j.chb.2023.108071>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563223004223>.

- [13] N. Lorenzo-Dus, C. Izura, and R. Pérez-Tattam, “Understanding grooming discourse in computer-mediated environments,” *Discourse, Context and Media*, vol. 12, pp. 40–50, 2016, ISSN: 2211-6958. doi: 10.1016/j.dcm.2016.02.004. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2211695816300095>.
- [14] A. Beech and E. A, “Identifying sexual grooming themes used by internet sex offenders,” *Deviant Behavior*, vol. 34, Feb. 2013. doi: 10.1080/01639625.2012.707550.
- [15] A. E. Cano, M. Fernandez, and H. Alani, “Detecting child grooming behaviour patterns on social media,” in *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, L. M. Aiello and D. McFarland, Eds. Cham: Springer International Publishing, 2014, pp. 412–427, ISBN: 978-3-319-13734-6. doi: 10.1007/978-3-319-13734-6_30. [Online]. Available: https://doi.org/10.1007/978-3-319-13734-6_30.
- [16] L. Hamm and S. McKeever, “Comparing machine learning models with a focus on tone in grooming chat logs,” *Frontiers in Pediatrics*, vol. Volume 13 - 2025, 2025, ISSN: 2296-2360. doi: 10.3389/fped.2025.1591828. [Online]. Available: <https://www.frontiersin.org/journals/pediatrics/articles/10.3389/fped.2025.1591828>.
- [17] E. Villatoro-Tello, M. Montes-y-Gómez, L. Villaseñor-Pineda, and P. Rosso, “A two-step approach for effective detection of online grooming behavior in chat rooms,” in *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2012, pp. 252–263.
- [18] A. Gupta, P. Kumaraguru, and A. Sureka, *Characterizing pedophile conversations on the internet using online grooming*, 2012. arXiv: 1208.4324 [cs.CY]. [Online]. Available: <https://arxiv.org/abs/1208.4324>.
- [19] G. Isaza, F. Muñoz, L. Castillo, and F. Buitrago, “Classifying cybergrooming for child online protection using hybrid machine learning model,” *Neurocomputing*, vol. 484, May 2022. doi: 10.1016/j.neucom.2021.08.148.
- [20] P. Schläpfer, S. Prayaga, and S. I. Perisanidis, *Online grooming detection*, https://www.researchgate.net/publication/353491539_Early_Detection_of_Sexual_Predators_in_Chats, Unpublished manuscript, 2022.
- [21] H. An et al., “Toward integrated solutions: A systematic interdisciplinary review of cybergrooming research,” *arXiv preprint arXiv:2503.05727*, 2025. [Online]. Available: <https://arxiv.org/abs/2503.05727>.
- [22] F. Gunawan, L. Ashianti, S. Candra, and B. Soewito, “Detecting online child grooming conversation,” pp. 1–6, Nov. 2016. doi: 10.1109/KICSS.2016.7951413.
- [23] D. Cook, M. Zilka, H. DeSandre, S. Giles, and S. Maskell, “Protecting children from online exploitation: Can a trained model detect harmful communication strategies?,” Aug. 2023, pp. 5–14. doi: 10.1145/3600211.3604696.
- [24] M. Leiva-Bianchi, N. Castillo, C. Astudillo, and F. Ahumada, “Effectiveness of machine learning methods in detecting grooming: A systematic meta-analytic review,” *Scientific Reports*, vol. 15, Mar. 2025. doi: 10.1038/s41598-024-83003-4.
- [25] S. Preuß et al., “Automatically identifying online grooming chats using CNN-based feature extraction,” in *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, K. Evang, L. Kallmeyer, R. Osswald, J. Waszczuk, and T. Zesch, Eds., Düsseldorf, Germany: KONVENS 2021 Organizers, Sep. 2021, pp. 137–146. [Online]. Available: <https://aclanthology.org/2021.konvens-1.12/>.

- [26] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fb053c1c4a845aa-Abstract.html.
- [27] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” *arXiv preprint arXiv:1901.07291*, 2019. [Online]. Available: <https://arxiv.org/abs/1901.07291>.
- [28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” OpenAI, Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *arXiv preprint arXiv:1906.08237*, 2019. [Online]. Available: <https://arxiv.org/abs/1906.08237>.
- [30] Y. Cai, X. Li, Y. Zhang, et al., “Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning,” *Scientific Reports*, vol. 15, no. 1, p. 2126, 2025. doi: [10.1038/s41598-025-85859-6](https://doi.org/10.1038/s41598-025-85859-6). [Online]. Available: <https://doi.org/10.1038/s41598-025-85859-6>.
- [31] S. Li and H. Tang, *Multimodal alignment and fusion: A survey*, 2024. arXiv: 2411.17040 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2411.17040>.
- [32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>.
- [33] A. Rogers, O. Kovaleva, and A. Rumshisky, *A primer in bertology: What we know about how bert works*, 2020. arXiv: 2002.12327 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2002.12327>.
- [34] C. Sun, X. Qiu, Y. Xu, and X. Huang, *How to fine-tune bert for text classification?* 2020. arXiv: 1905.05583 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/1905.05583>.
- [35] M. Koroteev, “Bert: A review of applications in natural language processing and understanding,” Mar. 2021. doi: [10.48550/arXiv.2103.11943](https://doi.org/10.48550/arXiv.2103.11943).
- [36] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, “A fine-tuned bert-based transfer learning approach for text classification,” *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 3498123, 2022. doi: <https://doi.org/10.1155/2022/3498123>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/3498123>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/3498123>.
- [37] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: <https://arxiv.org/abs/1907.11692>.
- [38] P. He, X. Liu, J. Gao, and W. Chen, *Deberta: Decoding-enhanced bert with disentangled attention*, 2021. arXiv: 2006.03654 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2006.03654>.
- [39] P. He, J. Gao, and W. Chen, *Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing*, 2023. arXiv: 2111.09543 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2111.09543>.
- [40] M. E. Peters, S. Ruder, and N. A. Smith, “To tune or not to tune? adapting pretrained representations to diverse tasks,” in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 7–14. doi: [10.18653/v1/W19-4302](https://doi.org/10.18653/v1/W19-4302). [Online]. Available: <https://aclanthology.org/W19-4302/>.

- [41] M. Bilal and A. Almazroi, “Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews,” *Electronic Commerce Research*, vol. 23, pp. 2737–2757, Apr. 2022. doi: 10.1007/s10660-022-09560-w.
- [42] J. Zhang, Y. Huang, S. Liu, Y. Gao, and X. Hu, *Do bert-like bidirectional models still perform better on text classification in the era of llms?* 2025. arXiv: 2505.18215 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2505.18215>.
- [43] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, *Attention bottlenecks for multimodal fusion*, 2022. arXiv: 2107.00135 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2107.00135>.
- [44] Y. Li et al., “A review of deep learning-based information fusion techniques for multimodal medical image classification,” *Computers in Biology and Medicine*, vol. 177, p. 108635, 2024, ISSN: 0010-4825. doi: <https://doi.org/10.1016/j.combiomed.2024.108635>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482524007200>.
- [45] G. Sharma, R. Chinmay, and R. Sharma, “Late fusion of transformers for sentiment analysis of code-switched data,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6485–6490. doi: 10.18653/v1/2023.findings-emnlp.430. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.430/>.
- [46] F. Abdullakutty and U. Naseem, “Decoding memes: A comprehensive analysis of late and early fusion models for explainable meme analysis,” May 2024, pp. 1681–1689. doi: 10.1145/3589335.3652504.
- [47] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani - Hiring Postdocs, “Memocmt: Multimodal emotion recognition using cross-modal transformer-based feature fusion,” *Scientific Reports*, vol. 15, Feb. 2025. doi: 10.1038/s41598-025-89202-x.
- [48] J. Biggiogera et al., *Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions*, 2021. arXiv: 2106.01536 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2106.01536>.
- [49] M. Vogt, U. Leser, and A. Akbik, “Early detection of sexual predators in chats,” pp. 4985–4999, 2021. doi: 10.18653/v1/2021.acl-long.386. [Online]. Available: <https://aclanthology.org/2021.acl-long.386>.
- [50] J. Street, L. Chen, and A. Patel, “Enhanced online grooming detection using context-aware transformers,” *arXiv preprint arXiv:2401.12345*, Sep. 2024. [Online]. Available: <https://arxiv.org/abs/2409.07958>.
- [51] F. Hassan, A. Bouchekif, and W. Aransa, “FiRC at SemEval-2023 task 10: Fine-grained classification of online sexism content using DeBERTa,” in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1824–1832. doi: 10.18653/v1/2023.semeval-1.252. [Online]. Available: <https://aclanthology.org/2023.semeval-1.252/>.
- [52] M. Mersha, M. Bitewa, T. Abay, and J. Kalita, *Explainability in neural networks for natural language processing tasks*, 2025. arXiv: 2412.18036 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2412.18036>.

- [53] S. Talebi, E. Tong, A. Li, G. Yamin, G. Zaharchuk, and M. Mofrad, “Exploring the performance and explainability of fine-tuned bert models for neuroradiology protocol assignment,” *BMC Medical Informatics and Decision Making*, vol. 24, Feb. 2024. doi: 10.1186/s12911-024-02444-z.
- [54] M. Ribeiro et al., *A methodology for explainable large language models with integrated gradients and linguistic analysis in text classification*, 2024. arXiv: 2410 . 00250 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2410.00250>.
- [55] Y. Tausczik and J. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, Mar. 2010. doi: 10.1177/0261927X09351676.
- [56] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, *The development and psychometric properties of liwc-22*. Austin, TX, 2022. [Online]. Available: www.liwc.app.
- [57] T. Fornaciari and M. Poesio, “Automatic deception detection in italian court cases,” *Artificial Intelligence and Law*, vol. 21(3), Sep. 2013. doi: 10.1007/s10506-013-9140-4.
- [58] C. Glasauer and R. W. Alexandrowicz, “The big-five between the lines: Approaching quantitative operationalization of text analytical personality assessment using linguistic markers,” *Psychological Test and Assessment Modeling*, vol. 64, no. 2, pp. 186–209, 2022.
- [59] I. Yakut Kilic and S. Pan, “Incorporating LIWC in neural networks to improve human trait and behavior analysis in low resource scenarios,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari et al., Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 4532–4539. [Online]. Available: <https://aclanthology.org/2022.lrec-1.482/>.
- [60] Z. Guo, P. Wang, J.-H. Cho, and L. Huang, “Text mining-based social-psychological vulnerability analysis of potential victims to cybergrooming: Insights and lessons learned,” in *Companion Proceedings of the ACM Web Conference 2023*, Austin, TX, USA: Association for Computing Machinery, 2023, pp. 1381–1388, ISBN: 9781450394192. doi: 10.1145/3543873.3587636. [Online]. Available: <https://doi.org/10.1145/3543873.3587636>.
- [61] A. Bilal, D. Ebert, and B. Lin, *Llms for explainable ai: A comprehensive survey*, 2025. arXiv: 2504 . 00125 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2504.00125>.
- [62] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, *Xai for transformers: Better explanations through conservative propagation*, 2022. arXiv: 2202 . 07304 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2202.07304>.
- [63] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705 . 07874 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/1705.07874>.
- [64] B. Rozemberczki et al., *The shapley value in machine learning*, 2022. arXiv: 2202 . 05594 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2202.05594>.
- [65] K. Aas, M. Jullum, and A. Løland, “Explaining individual predictions when features are dependent: More accurate approximations to shapley values,” *Artificial Intelligence*, vol. 298, p. 103 502, 2021, issn: 0004-3702. doi: <https://doi.org/10.1016/j.artint.2021.103502>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0004370221000539>.
- [66] M. Liu, Y. Ning, H. Yuan, M. E. H. Ong, and N. Liu, *Balanced background and explanation data are needed in explaining deep learning models with shap: An empirical study on clinical decision making*, 2022. eprint: 2206 . 04050. [Online]. Available: <https://arxiv.org/abs/2206.04050>.

- [67] Y. Chen, R. Calabrese, and B. Martin-Barragan, “Interpretable machine learning for imbalanced credit scoring datasets,” *European Journal of Operational Research*, vol. 312, no. 1, pp. 357–372, 2024, issn: 0377-2217. doi: <https://doi.org/10.1016/j.ejor.2023.06.036>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0377221723005088>.
- [68] J. Preiss and Z. Chen, “Incorporating word count information into depression risk summary generation: INF@UoS CLPsych 2024 submission,” in *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 211–217. [Online]. Available: <https://aclanthology.org/2024.clpsych-1.19/>.
- [69] S. Wewelwala et al., “Hybrid clinical emotion recognition: Integrating bert representations and psycholinguistic features for enhanced accuracy and explainability,” *Forum for Linguistic Studies*, vol. 7, no. 5, pp. 9660–9682, 2025. [Online]. Available: <https://journals.bilpubgroup.com/index.php/f1s/article/download/9660/6386/47180>.
- [70] M. W. R. Miah, J. Yearwood, and S. Kulkarni, “Detection of child exploiting chats from a mixed chat dataset as a text classification task,” in *Proceedings of the Australasian Language Technology Association Workshop 2011*, D. Molla and D. Martinez, Eds., Canberra, Australia, Dec. 2011, pp. 157–165. [Online]. Available: <https://aclanthology.org/U11-1020/>.
- [71] J. Salminen, M. Mustak, S.-G. Jung, H. Makkonen, and J. Jansen, “Decoding deception in the online marketplace: Enhancing fake review detection with psycholinguistics and transformer models,” *Journal of Marketing Analytics*, pp. 1–18, Mar. 2025. doi: <10.1057/s41270-025-00393-8>.
- [72] G. Inches and F. Crestani, “Overview of the international sexual predator identification competition at pan-2012,” Jan. 2012.
- [73] K. Sun, P. Qi, Y. Zhang, L. Liu, W. Wang, and Z. Huang, “Tokenization consistency matters for generative models on extractive NLP tasks,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 300–13 310. doi: <10.18653/v1/2023.findings-emnlp.887>. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.887/>.
- [74] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd. Hillsdale, NJ: Lawrence Erlbaum, 1988.
- [75] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, “A survey of methods for addressing class imbalance in deep-learning based natural language processing,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 523–540. doi: <10.18653/v1/2023.eacl-main.38>. [Online]. Available: <https://aclanthology.org/2023.eacl-main.38/>.
- [76] S. Shankar, L. Thompson, and M. Fiterau, *Progressive fusion for multimodal integration*, 2022. arXiv: <2209.00302> [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2209.00302>.
- [77] J. Wang, Y. Mao, N. Guan, and C. J. Xue, *Shap-cat: A interpretable multi-modal framework enhancing wsi classification via virtual staining and shapley-value-based multimodal fusion*, 2024. arXiv: <2410.01408> [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2410.01408>.

8 Appendix

Table 8.1: LIWC-22 categories with abbreviations and exemplar words. Adapted from Boyd et al. [56].

Category	Abbrev.	Word Examples
Linguistic	linguistic	General linguistic markers
Total function words	function	the, to, and, I
Total pronouns	pronoun	I, you, that, it
Personal pronouns	ppron	I, you, my, me
1st person singular	i	I, me, my, myself
1st person plural	we	we, our, us, lets
2nd person	you	you, your, u, yourself
3rd person singular	shehe	he, she, her, his
3rd person plural	they	they, their, them, themsel*
Impersonal pronouns	ipron	that, it, this, what
Determiners	det	the, at, that, my
Articles	article	a, an, the, a lot
Numbers	number	one, two, first, once
Prepositions	prep	to, of, in, for
Auxiliary verbs	auxverb	is, was, be, have
Adverbs	adverb	so, just, about, there
Conjunctions	conj	and, but, so, as
Negations	negate	not, no, never, nothing
Common verbs	verb	is, was, be, have
Common adjectives	adj	more, very, other, new
Quantities	quantity	all, one, more, some
Drives	Drives	we, our, work, us
Affiliation	affiliation	we, our, us, help
Achievement	achieve	work, better, best, working
Power	power	own, order, allow, power
Cognition	Cognition	is, was, but, are
All-or-none	allnone	all, no, never, always
Cognitive processes	cogproc	but, not, if, or, know
Insight	insight	know, how, think, feel
Causation	cause	how, because, make, why
Discrepancy	discrep	would, can, want, could
Tentative	tentat	if, or, any, something
Certitude	certitude	really, actually, of course, real
Differentiation	differ	but, not, if, or
Memory	memory	remember, forget, remind, forgot
Affect	Affect	good, well, new, love
Positive tone	tone_pos	good, well, new, love
Negative tone	tone_neg	bad, wrong, too much, hate
Emotion	emotion	good, love, happy, hope
Positive emotion	emo_pos	good, love, happy, hope
Negative emotion	emo_neg	bad, hate, hurt, tired
Anxiety	emo_anx	worry, fear, afraid, nervous
Anger	emo_anger	hate, mad, angry, frustr*
Sadness	emo_sad	:(. sad, disappoint*, cry
Swear words	swear	shit, fuckin*, fuck, damn
Social processes	Social	you, we, he, she
Social behavior	socbehav	said, love, say, care
Prosocial behavior	prosocial	care, help, thank, please
Politeness	polite	thank, please, thanks, good morning
Interpersonal conflict	conflict	fight, kill, killed, attack
Moralization	moral	wrong, honor*, deserv*, judge
Communication	comm	said, say, tell, thank*
Social referents	socrefs	you, we, he, she
Family	family	parent*, mother*, father*, baby
Friends	friend	friend*, boyfriend*, girlfriend*, dude
Female references	female	she, her, girl, woman
Male references	male	he, his, him, man
Culture	Culture	car, united states, govern*, phone
Politics	politic	united states, govern*, congress*, senat*
Ethnicity	ethnicity	american, french, chinese, indian
Technology	tech	car, phone, comput*, email*

Continued on next page

Category	Abbrev.	Word Examples
Lifestyle	lifestyle	work, home, school, working
Leisure	leisure	game*, fun, play, party*
Home	home	home, house, room, bed
Work	work	work, school, working, class
Money	money	business*, pay*, price*, market*
Religion	relig	god, hell, christmas*, church
Physical	physical	medic*, food*, patients, eye*
Health	health	medic*, patients, physician*, health
Illness	illness	hospital*, cancer*, sick, pain
Wellness	wellness	healthy, gym*, supported, diet
Mental health	mental	mental health, depressed, suicid*, trauma*
Substances	substances	beer*, wine, drunk, cigar*
Sexual	sexual	sex, gay, pregnan*, dick
Food	food	food*, drink*, eat, dinner*
Death	death	death*, dead, die, kill
States	states	short-term internal states that drive or inhibit behavior
Need	need	have to, need, had to, must
Want	want	want, hope, wanted, wish
Acquire	acquire	get, got, take, getting
Lack	lack	don't have, didn't have, *less, hungry
Fulfilled	fulfill	enough, full, complete, extra
Fatigue	fatigue	tired, bored, don't care, boring
Motives	motives	broad motivational drives
Reward	reward	opportun*, win, gain*, benefit*
Risk	risk	secur*, protect*, pain, risk*
Curiosity	curiosity	scien*, look* for, research*, wonder
Allure	allure	have, like, out, know
Perception	Perception	in, out, up, there
Attention	attention	look, look* for, watch, check
Motion	motion	go, come, went, came
Space	space	in, out, up, there
Visual	visual	see, look, eye*, saw
Auditory	auditory	sound*, heard, hear, music
Feeling (perceptual)	feeling	feel, hard, cool, felt
Time orientation	time_orient	temporal focus variables
Time	time	when, now, then, day
Past focus	focuspast	was, had, were, been
Present focus	focuspresent	is, are, I'm, can
Future focus	focusfuture	will, going to, have to, may
Conversational	Conversation	yeah, oh, yes, okay
Netspeak	netspeak	:), u, lol, haha*
Assent	assent	yeah, yes, okay, ok
Nonfluencies	nonflu	oh, um, uh, i i
Fillers	filler	rr*, wow, sooo*, youknow

Table 8.2: Complete LIWC baseline results across PJ grooming (G) and PAN12 non-grooming (NG) groups over complete conversations.

Feature	Mean (G)	Std (G)	Mean (NG)	Std (NG)	Cohen's <i>d</i>	Mean Diff
WC	6722.89	8040.96	288.70	1461.97	3.52	6434.19
focusfuture	2.71	0.88	0.97	1.35	1.29	1.73
verb	23.05	1.99	17.19	5.17	1.14	5.86
discrep	3.33	0.83	1.72	1.64	0.99	1.61
Linguistic	81.10	5.30	69.40	12.92	0.91	11.70
acquire	1.25	0.41	0.53	0.87	0.84	0.72
function	58.87	4.57	49.61	11.38	0.82	9.25
pronoun	22.77	2.34	18.22	5.79	0.79	4.54
home	0.50	0.24	0.15	0.45	0.78	0.35
Analytic	5.01	3.98	19.61	21.43	-0.69	-14.60
WPS	77.28	412.95	18.76	65.42	0.68	58.52
ppron	17.43	2.15	13.74	5.51	0.67	3.69
Dic	95.80	1.82	89.03	10.20	0.67	6.77
focuspast	2.65	0.90	1.47	1.78	0.67	1.17
i	8.20	1.33	6.14	3.26	0.64	2.06
Cognition	26.13	2.49	21.94	6.72	0.63	4.19
Authentic	70.06	14.99	52.31	29.02	0.62	17.75
shehe	0.83	0.54	0.27	0.90	0.61	0.55
BigWords	6.41	1.24	10.25	6.40	-0.60	-3.83
allure	12.77	1.78	9.77	5.01	0.60	3.00
auxverb	12.74	1.34	10.51	3.84	0.59	2.23
prep	9.42	1.19	7.38	3.52	0.59	2.05
motion	1.54	0.48	0.86	1.21	0.57	0.68
cogproc	12.76	2.00	9.93	5.06	0.57	2.84
assent	3.24	1.16	2.13	2.20	0.51	1.11
family	0.70	0.53	0.29	0.83	0.50	0.41
affiliation	1.78	0.56	1.09	1.53	0.46	0.70
negate	3.30	0.81	2.42	2.04	0.44	0.89
Culture	0.46	0.25	1.62	2.72	-0.43	-1.16
conj	5.17	1.08	4.11	2.50	0.43	1.07
Apostro	3.15	1.43	1.76	3.37	0.42	1.39
adverb	7.28	1.06	6.07	3.17	0.38	1.21
male	0.85	0.57	1.60	2.04	-0.37	-0.75
want	1.26	0.53	0.77	1.34	0.37	0.49
sexual	0.55	0.49	1.55	2.78	-0.36	-1.00
time	4.43	0.90	3.44	2.81	0.36	0.99
Perception	8.66	1.47	7.29	3.92	0.35	1.37
politic	0.01	0.04	0.80	2.27	-0.35	-0.79
Drives	2.57	0.67	1.86	2.04	0.35	0.71
tentat	2.97	0.79	2.28	2.03	0.34	0.69
ipron	5.33	0.97	4.48	2.71	0.32	0.85
need	0.49	0.25	0.28	0.67	0.31	0.21
Tone	85.82	15.39	76.82	29.17	0.31	9.00
socbehav	3.84	0.91	5.89	7.12	-0.29	-2.05
polite	0.56	0.31	2.30	6.04	-0.29	-1.73
we	0.52	0.28	0.32	0.70	0.29	0.20
comm	2.27	0.66	4.14	6.87	-0.28	-1.87
feeling	1.11	0.53	0.77	1.24	0.27	0.34
Comma	0.78	1.13	1.91	4.21	-0.27	-1.13
Affect	8.34	2.25	10.27	7.72	-0.25	-1.93
Social	15.27	2.39	17.60	9.50	-0.25	-2.33
ethnicity	0.02	0.06	0.29	1.13	-0.24	-0.26
Physical	2.03	0.84	2.84	3.59	-0.23	-0.81
article	2.41	0.68	2.99	2.62	-0.22	-0.58
Emoji	0.00	0.02	0.00	0.00	0.21	0.00
differ	3.58	0.82	3.12	2.25	0.21	0.46
swear	0.46	0.46	0.98	2.64	-0.20	-0.52
they	0.37	0.25	0.25	0.63	0.20	0.12
food	0.50	0.43	0.30	1.07	0.19	0.21
adj	6.80	1.24	7.55	3.93	-0.19	-0.75

Continued on next page

Feature	Mean (G)	Std (G)	Mean (NG)	Std (NG)	Cohen's <i>d</i>	Mean Diff
you	7.44	1.56	6.67	4.18	0.19	0.77
Exclam	0.31	0.53	0.82	2.78	-0.19	-0.51
number	2.17	4.26	2.86	3.74	-0.18	-0.69
tone_pos	6.55	2.14	7.82	7.10	-0.18	-1.26
space	4.98	1.12	4.46	3.03	0.17	0.52
emo_anx	0.11	0.09	0.05	0.31	0.17	0.05
filler	0.27	0.25	0.18	0.55	0.17	0.09
female	1.09	0.61	1.43	2.07	-0.17	-0.34
netspeak	3.09	1.79	3.86	4.74	-0.16	-0.77
prosocial	0.55	0.27	0.76	1.29	-0.16	-0.20
AllPunc	11.77	5.33	24.49	80.31	-0.16	-12.72
Lifestyle	2.15	0.68	1.81	2.22	0.16	0.34
OtherP	3.13	3.49	7.30	28.20	-0.15	-4.17
money	0.22	0.19	0.15	0.49	0.15	0.07
insight	2.70	0.69	2.42	1.96	0.14	0.28
power	0.22	0.26	0.35	0.93	-0.14	-0.13
achieve	0.57	0.25	0.44	0.90	0.14	0.13
curiosity	0.20	0.17	0.34	1.01	-0.14	-0.14
cause	1.35	0.39	1.55	1.52	-0.13	-0.20
QMark	3.13	1.84	8.85	48.47	-0.12	-5.72
tone_neg	1.24	0.44	1.53	2.44	-0.12	-0.29
certitude	0.73	0.33	0.62	0.98	0.11	0.11
conflict	0.08	0.10	0.21	1.19	-0.11	-0.13
memory	0.06	0.06	0.04	0.22	0.11	0.02
risk	0.14	0.11	0.10	0.42	0.11	0.04
health	0.20	0.14	0.15	0.54	0.09	0.05
tech	0.43	0.24	0.54	1.18	-0.09	-0.11
work	0.68	0.35	0.82	1.50	-0.09	-0.14
quantity	2.63	0.68	2.42	2.31	0.09	0.21
visual	1.21	0.47	1.05	1.75	0.09	0.15
moral	0.12	0.12	0.18	0.76	-0.09	-0.07
friend	0.37	0.22	0.29	0.91	0.08	0.08
Period	1.27	2.09	3.85	31.95	-0.08	-2.58
fulfill	0.04	0.05	0.06	0.27	-0.07	-0.02
death	0.05	0.10	0.08	0.48	-0.07	-0.03
fatigue	0.16	0.13	0.21	0.83	-0.07	-0.06
leisure	0.65	0.36	0.56	1.28	0.07	0.09
allnone	1.94	0.62	1.81	1.90	0.07	0.13
relig	0.14	0.16	0.19	0.79	-0.06	-0.05
Conversation	7.31	2.37	6.95	5.73	0.06	0.36
mental	0.00	0.01	0.01	0.11	-0.06	-0.01
det	8.28	1.34	8.05	4.18	0.06	0.23
nonflu	1.11	0.61	1.22	2.20	-0.05	-0.12
socrefs	11.34	1.91	11.65	6.37	-0.05	-0.31
illness	0.05	0.07	0.04	0.30	0.05	0.01
focuspresent	7.02	1.03	7.17	3.35	-0.04	-0.15
reward	0.01	0.03	0.03	0.39	-0.04	-0.02
attention	0.37	0.20	0.41	1.05	-0.04	-0.04
emo_sad	0.15	0.15	0.17	0.62	-0.04	-0.02
auditory	0.25	0.18	0.28	1.04	-0.03	-0.03
emo_neg	0.67	0.32	0.71	1.50	-0.02	-0.04
emo_pos	3.35	1.68	3.27	3.48	0.02	0.08
lack	0.21	0.17	0.20	0.74	0.02	0.02
emotion	4.19	1.83	4.13	3.93	0.02	0.06
Clout	60.57	18.54	60.98	33.57	-0.01	-0.41
substances	0.03	0.07	0.03	0.25	0.01	0.00
emo_anger	0.09	0.09	0.09	0.48	0.00	0.00
wellness	0.01	0.04	0.01	0.11	0.00	0.00

Table 8.3: Complete LIWC baseline results across PJ grooming (G), PAN12 non-grooming (NG), and Synthetic non-grooming (SYN) groups over chunked conversation

Feature	Mean (PJ)	Std (PJ)	Mean (PAN12)	Std (PAN12)	Mean (SYN)	Std (SYN)
WC	289.01	108.41	143.04	95.76	319.74	112.40
Analytic	7.02	8.56	20.06	22.12	31.67	12.97
Clout	58.74	27.23	59.42	34.18	43.96	19.22
Authentic	67.54	25.11	53.43	29.32	68.80	19.81
Tone	80.64	23.58	73.49	30.08	87.57	17.78
WPS	40.70	57.83	16.57	20.59	34.16	45.71
BigWords	6.46	2.27	10.45	6.60	12.22	2.57
Dic	95.67	2.92	88.45	10.61	89.70	3.10
Linguistic	81.01	6.01	68.67	13.32	73.22	3.69
function	58.89	6.03	50.07	11.62	50.28	3.76
pronoun	22.83	3.75	18.27	5.99	16.08	2.91
ppron	17.55	3.53	13.73	5.69	9.94	2.30
i	8.32	2.47	6.21	3.37	5.15	1.57
we	0.52	0.67	0.32	0.73	1.30	1.10
you	7.37	2.57	6.55	4.28	2.91	1.58
shehe	0.89	1.39	0.29	0.98	0.15	0.50
they	0.39	0.63	0.26	0.67	0.39	0.58
ipron	5.28	2.13	4.54	2.79	6.14	1.79
det	8.21	2.65	8.26	4.26	10.78	2.13
article	2.37	1.36	3.09	2.67	4.45	1.55
number	1.65	3.59	2.94	4.03	0.93	0.85
prep	9.45	2.38	7.55	3.64	9.10	1.81
auxverb	12.54	2.62	10.47	3.94	5.67	1.89
adverb	7.39	2.25	6.13	3.26	8.87	1.99
conj	5.13	1.96	4.22	2.59	5.28	1.67
negate	3.20	1.64	2.42	2.10	0.98	0.66
verb	23.03	3.49	17.26	5.29	15.67	2.58
adj	6.94	2.41	7.57	4.04	9.48	1.93
quantity	2.65	1.52	2.49	2.38	4.24	1.68
Drives	2.65	1.58	1.91	2.11	4.61	1.79
affiliation	1.83	1.35	1.10	1.57	2.65	1.43
achieve	0.62	0.69	0.46	0.93	1.76	1.06
power	0.21	0.54	0.37	1.01	0.23	0.36
Cognition	26.37	4.65	22.11	6.88	24.31	3.40
allnone	1.91	1.21	1.82	1.93	1.33	0.87
cogproc	12.73	3.99	10.12	5.17	15.66	3.38
insight	2.82	1.57	2.46	2.01	3.55	1.34
cause	1.35	0.95	1.56	1.54	0.88	0.68
discrep	3.24	1.77	1.74	1.69	2.91	1.13
tentat	2.98	1.66	2.32	2.10	4.85	1.86
certitude	0.73	0.79	0.64	1.01	1.33	0.77
differ	3.50	1.76	3.18	2.34	3.09	1.40
memory	0.08	0.23	0.04	0.23	0.15	0.28
Affect	8.25	3.36	9.19	7.45	8.63	2.86
tone_pos	6.61	3.19	6.85	6.78	7.49	2.96
tone_neg	1.17	1.08	1.40	2.31	1.05	0.87
emotion	3.99	2.43	3.09	3.32	3.12	1.47
emo_pos	3.20	2.25	2.38	2.93	2.35	1.37
emo_neg	0.62	0.77	0.56	1.21	0.67	0.57
emo_anx	0.11	0.29	0.06	0.32	0.37	0.43
emo_anger	0.13	0.40	0.09	0.48	0.09	0.21
emo_sad	0.05	0.25	0.06	0.42	0.03	0.11
swear	0.38	0.75	0.99	2.66	0.00	0.02
Social	15.54	4.42	17.37	9.59	10.42	3.28
socbehav	4.14	2.56	5.83	7.07	4.71	2.37
prosocial	0.59	0.91	0.77	1.31	1.74	1.28
polite	0.68	1.20	2.21	5.96	1.06	1.30
conflict	0.07	0.26	0.21	1.20	0.08	0.21
moral	0.12	0.31	0.19	0.80	0.10	0.22
comm	2.38	1.85	4.05	6.80	2.47	1.49

Continued on next page

Feature	Mean (PJ)	Std (PJ)	Mean (PAN12)	Std (PAN12)	Mean (SYN)	Std (SYN)
socrefs	11.28	3.22	11.48	6.46	5.63	1.90
family	0.70	0.96	0.29	0.85	0.15	0.48
friend	0.36	0.58	0.29	0.91	0.11	0.29
female	1.11	1.40	1.39	2.06	0.11	0.42
male	0.78	1.16	1.57	2.05	0.09	0.39
Culture	0.43	0.67	1.63	2.74	0.41	0.62
politic	0.01	0.09	0.77	2.24	0.01	0.07
ethnicity	0.02	0.16	0.29	1.14	0.00	0.04
tech	0.40	0.65	0.57	1.29	0.39	0.62
Lifestyle	2.23	1.65	1.86	2.30	3.04	1.61
leisure	0.66	0.93	0.57	1.32	0.95	1.04
home	0.51	0.67	0.15	0.47	0.36	0.63
work	0.72	0.93	0.85	1.55	1.30	1.04
money	0.23	0.49	0.15	0.53	0.58	0.84
relig	0.14	0.34	0.19	0.81	0.00	0.03
Physical	2.11	1.82	2.84	3.69	1.27	1.29
health	0.22	0.53	0.15	0.56	0.29	0.52
illness	0.08	0.36	0.04	0.30	0.16	0.37
wellness	0.01	0.07	0.01	0.12	0.04	0.17
mental	0.00	0.05	0.01	0.12	0.01	0.09
substances	0.03	0.16	0.03	0.26	0.01	0.06
sexual	0.44	0.79	1.52	2.87	0.00	0.04
food	0.62	1.12	0.30	1.09	0.66	1.06
death	0.04	0.17	0.08	0.48	0.01	0.07
need	0.55	0.84	0.30	0.74	0.47	0.47
want	1.22	1.01	0.76	1.39	0.79	0.60
acquire	1.29	0.99	0.54	0.89	1.11	0.75
lack	0.20	0.50	0.20	0.76	0.05	0.20
fulfill	0.04	0.15	0.07	0.29	0.12	0.25
fatigue	0.17	0.40	0.22	0.83	0.05	0.14
reward	0.03	0.16	0.03	0.39	0.08	0.21
risk	0.14	0.31	0.11	0.44	0.28	0.41
curiosity	0.21	0.45	0.34	1.03	0.26	0.42
allure	12.94	3.38	9.78	5.05	12.96	2.84
Perception	8.44	3.12	7.36	4.04	8.75	2.34
attention	0.37	0.57	0.41	1.07	0.55	0.57
motion	1.55	1.16	0.89	1.27	1.57	0.98
space	4.79	2.46	4.49	3.12	4.01	1.68
visual	1.16	1.13	1.06	1.78	1.01	0.95
auditory	0.27	0.54	0.29	1.00	0.86	0.72
feeling	1.08	1.02	0.77	1.26	1.12	0.68
time	4.76	2.41	3.47	2.92	7.17	2.82
focuspast	2.97	2.08	1.52	1.85	1.56	1.08
focuspresent	6.78	2.33	7.13	3.44	3.72	1.23
focusfuture	2.91	2.14	0.99	1.46	2.11	1.48
netspeak	2.64	2.27	2.57	4.02	1.18	1.23
assent	3.33	1.89	2.09	2.21	3.57	1.26
nonflu	1.11	1.03	1.23	2.25	0.87	0.61
filler	0.32	0.51	0.18	0.55	0.03	0.15
AllPunc	12.98	8.42	26.39	176.15	16.64	5.53
Period	1.45	2.66	4.38	53.03	2.59	3.96
Comma	0.81	1.38	2.19	19.45	7.18	3.47
QMark	2.99	2.29	8.51	46.55	2.27	1.23
Exclam	0.40	1.00	0.86	3.22	0.46	0.87
Apostro	3.08	2.36	1.95	17.48	3.66	1.47
OtherP	4.25	6.73	8.50	94.45	0.49	0.92