

Modelando Grooming como um Fenômeno Conversacional: Importância da Progressão dos Turnos na Detecção de Conversas Predatórias

João Marcos Tomáz S. Campos¹, Diogo T. Chaves¹, Rafael M. Gomes¹, Wagner Meira Jr.¹,
Marcos André Gonçalves¹, Leonardo Rocha²

¹Universidade Federal de Minas Gerais, ²Universidade Federal de São João del-Rei
joaomarcoscampos@dcc.ufmg.br;diogochaves@dcc.ufmg.br;rafaelgomes@dcc.ufmg.br
meira@dcc.ufmg.br;mgoncalv@dcc.ufmg.br;lrocha@ufs.j.edu.br

ABSTRACT

Online grooming detection is commonly formulated as a text classification task, although predatory behavior emerges through a progressive conversational process involving trust building, emotional bonding, and gradual escalation. This paper investigates whether grooming is better characterized as a conversational modeling problem than as a conventional document classification task. Using the PAN12 Sexual Predator Identification benchmark, we systematically compare lexical, distributed, contextual, and long-context representations combined with supervised classifiers, in-context learning, pre-trained moderation models, and hierarchical conversational architectures. To explicitly model conversational progression, we introduce the Contextual Grooming Pattern Network (CGPN), a turn-level architecture that encodes conversational turns with MiniLM and models their temporal evolution through a bidirectional GRU with attention pooling. CGPN achieves the best overall performance, outperforming flat-concatenation approaches and demonstrating that preserving conversational structure provides complementary information beyond aggregated textual content. We further show that discriminative signals emerge within the first 25% of the dialogue and that effective detection depends less on explicit toxicity than on relational patterns such as affection, connection, isolation, information seeking, and gradual escalation. Across linguistic, structural, early-detection, and explainability analyses, the results provide converging evidence that grooming is fundamentally a progressive conversational phenomenon and that explicitly modeling turn dynamics is critical for its detection.

Keywords

Online grooming detection; conversational modeling; text representation; contextual embeddings; explainable AI.

1 Introdução

A detecção automática de *grooming* online é um problema crítico para a segurança de crianças e adolescentes em plataformas digitais. Diferentemente de formas mais explícitas de

abuso ou assédio, o *grooming* manifesta-se como um processo relacional gradual, no qual o agressor constrói confiança, reduz barreiras emocionais e, progressivamente, conduz a interação a dinâmicas de isolamento, intimidade e sexualização [4, 24, 31, 49]. Muitas mensagens envolvidas nesse processo parecem amigáveis, ambíguas ou comuns quando analisadas isoladamente, permitindo que o comportamento predatório contorne mecanismos tradicionais baseados em palavras-chave, toxicidade ou conteúdo ofensivo explícito [15, 45].

Essa característica sugere que o desafio central da tarefa não é apenas identificar mensagens suspeitas, mas compreender como sinais de risco emergem e se acumulam ao longo da interação. **Em outras palavras, o *grooming* é inerentemente um fenômeno conversacional.** Seu significado não está concentrado em uma única mensagem, mas na progressão da conversa e na forma como padrões de aproximação afetiva, coleta de informações pessoais, construção de confiança e sexualização se combinam ao longo do tempo. Consequentemente, compreender essa progressão pode ser mais importante para a detecção de *grooming* do que analisar mensagens isoladas ou mesmo documentos resultantes da simples concatenação dos diálogos.

Do ponto de vista computacional, essa natureza progressiva impõe desafios que diferenciam a detecção de *grooming* de tarefas convencionais de classificação textual. O primeiro é o extremo desequilíbrio de classes, uma vez que conversas associadas a predadores representam apenas uma pequena fração das interações observadas [5, 21]. O segundo é a dependência de contexto conversacional de longo alcance, pois evidências relevantes podem surgir em momentos distintos da conversa. O terceiro é a elevada sobreposição semântica entre diálogos benignos e predatórios, especialmente em interações informais e afetivas. Consequentemente, modelos que se concentram apenas no conteúdo textual agregado podem ignorar aspectos fundamentais da dinâmica conversacional.

O corpus *PAN12 Sexual Predator Identification* [40] consolidou-se como o principal *benchmark* para essa tarefa e impulsionou uma ampla variedade de abordagens baseadas em características léxicas, estilísticas, *embeddings* distribuídos e modelos neurais. Entretanto, a maior parte desses trabalhos trata a conversa como um documento único, representado por meio da concatenação das mensagens. Embora essa estratégia preserve o conteúdo textual, ela tende a perder informações sobre ordem, estrutura e evolução dos turnos. Assim, permanece em aberto uma questão

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

fundamental: **até que ponto a detecção de *grooming* deve ser tratada como um problema de classificação textual ou como um problema de modelagem conversacional?**

Neste trabalho, investigamos essa questão a partir da hipótese de que a progressão conversacional constitui um sinal central para a detecção de *grooming*. Mais especificamente, argumentamos que preservar explicitamente a estrutura dos turnos e a evolução da interação pode ser mais importante do que simplesmente aumentar a quantidade de contexto textual disponível ao modelo. Diferentemente da maior parte dos trabalhos anteriores, nos quais o foco era melhorar a classificação de conversas predatórias e explorar partes não contíguas do documento para tal, buscamos compreender em que medida a preservação da dinâmica conversacional contribui para a identificação de interações predatórias. Para isso, realizamos uma comparação sistemática entre diferentes representações textuais, classificadores supervisionados, modelos de moderação e arquiteturas hierárquicas, analisando desempenho preditivo, detecção precoce, interpretabilidade e evidências comportamentais capturadas pelos modelos.

As seguintes questões de pesquisa orientam o estudo:

- **QP1:** Como diferentes estratégias de representação textual e modelagem conversacional impactam a detecção de *grooming*?
- **QP2:** Em que medida essas abordagens capturam padrões comportamentais compatíveis com a progressão relacional característica do fenômeno?

As principais contribuições deste trabalho são:

- **Avaliação comparativa abrangente:** realizamos um *benchmarking* sistemático envolvendo representações léxicas, distribuídas e contextuais com classificadores supervisionados, modelos de moderação e arquiteturas hierárquicas.
- **Modelagem da progressão conversacional:** propomos o **Contextual Grooming Pattern Network (CGPN)**, uma arquitetura hierárquica baseada em turnos que combina codificação contextual das mensagens, modelagem sequencial e agregação por atenção. Mais do que introduzir uma nova arquitetura, utilizamos o CGPN como instrumento para investigar o impacto da preservação explícita da progressão dos turnos.
- **Evidências da importância da estrutura conversacional:** por meio de análise de características estruturais, estudo de ablação e experimentos de detecção precoce, mostramos que sinais relevantes emergem antes da conclusão da conversa e que a preservação da dinâmica dos turnos contribui diretamente para o desempenho dos modelos.
- **Análise comportamental e interpretabilidade:** utilizamos SHAP, TCAV e modelos baseados em conceitos para investigar os sinais capturados pelos modelos, evidenciando a importância de fatores relacionais na detecção de *grooming*.

2 Trabalhos Relacionados

A detecção automática de *grooming* online dialoga com tarefas de moderação de conteúdo, como identificação de toxicidade, discurso ofensivo e *cyberbullying*. Entretanto, diferencia-se delas por sua natureza progressiva e dissimulada: o comportamento predatório frequentemente emerge por meio de

interações afetivas e ambíguas orientadas à construção de confiança, ao isolamento e à sexualização gradual [7, 14, 37, 39, 41, 43, 49]. Assim, a tarefa exige não só reconhecer conteúdos suspeitos, mas modelar a progressão da interação.

O corpus *PAN12 Sexual Predator Identification* [40] consolidou-se como um dos principais *benchmarks* da área, motivando abordagens baseadas em características léxicas, estatísticas e estilísticas combinadas a classificadores tradicionais, como SVM, Regressão Logística e Floresta Aleatória [22, 34, 40, 48, 53], ou até LLMs ajustadas para a tarefa [35]. Embora importantes, esses métodos normalmente representam a conversa como um documento plano, perdendo informações sobre ordem, estrutura e evolução dos turnos, ou removem grande parte das conversas, o que altera fortemente a prevalência da classe positiva.

Com o avanço dos modelos de linguagem, representações contextuais como BERT [13], MiniLM [1, 18, 42] e arquiteturas para textos longos, como Longformer [3] e BigBird [58], passaram a capturar melhor relações semânticas em nível de documento [33]. A exploração dessas relações semânticas também tem sido utilizada para enriquecer representações textuais e organizar coleções em estruturas de tópicos, evoluindo de representações baseadas em *word embeddings* para *contextual embeddings* [50–52]. Entretanto, no contexto de *grooming*, os sinais relevantes dependem tanto do volume de contexto quanto da progressão dos turnos e do acúmulo gradual de padrões relacionais.

Nesse sentido, trabalhos recentes têm avançado além da representação da conversa como um único documento. O framework Osprey [54] mostra que, para detecção de *grooming*, representar uma conversa como uma sequência de embeddings de suas mensagens, combiná-las com *features* conversacionais e processá-las com modelos recorrentes supera representações que agregam toda a conversa em um único vetor. Esses resultados demonstram a relevância de se preservar a estrutura das mensagens e os padrões de interação.

Além da representação textual, a escolha do classificador é relevante em cenários desbalanceados. Nesse sentido, classificadores tradicionais e modelos tabulares têm sido explorados em cenários textuais e desbalanceados nos quais são usadas representações densas e conjuntos com poucos dados rotulados [19–22]. Também são frequentes modelos genéricos de moderação e toxicidade, embora sua adequação ao *grooming* seja limitada, pois o comportamento predatório pode não apresentar agressividade ou toxicidade explícita [6, 12, 16, 17, 29, 36, 46, 47, 55, 57].

A detecção precoce, essencial para aplicações de segurança infantil, tem recebido atenção recentemente. Em Kim et al. [26], usando a mesma base e para a mesma tarefa, os autores propõem uma estratégia de memória que recupera sentenças anteriores similares à mensagem corrente. Os resultados mostram que a forma de selecionar o histórico conversacional influencia o *trade-off* entre desempenho e antecipação da detecção, criando um bom precedente a ser explorado.

A explicabilidade dos modelos, no entanto, aparece como aspecto ainda pouco explorado. Métodos como SHAP [28], LIME [44], TCAV [25] e Modelos de Gargalo de Conceito

(CBM) [27, 30] oferecem mecanismos complementares para investigar a importância de atributos, o alinhamento a conceitos humanos e a interpretabilidade por design.

Este trabalho contribui com uma avaliação sistemática de representações léxicas, distribuídas, contextuais e de texto longo, combinadas a classificadores supervisionados, aprendido no contexto e moderação pré-treinada. Por meio dessa, investigamos o quanto a preservação explícita da sequência de turnos acrescenta às representações planas, combinando essa comparação com análises estruturais e de interpretabilidade comportamental.

3 Metodologia

Este estudo foi estruturado como uma avaliação comparativa de representações textuais e estratégias de classificação para detecção automática de *grooming* online, modelada como uma tarefa binária no nível da conversa. Formalmente, dado um corpus $\mathcal{D} = \{\mathcal{C}_i, y_i\}_{i=1}^N$, em que $\mathcal{C}_i$ representa uma conversa e $y_i \in \{0, 1\}$ indica ausência ou presença de comportamento predatório, buscamos aprender uma função $f_{\mathcal{C}_i} \rightarrow y_i$. A classe positiva ($y_i = 1$) corresponde a conversas nas quais ao menos um participante está associado a um predador, conforme o *ground truth* do corpus. A metodologia compreende seis etapas: preparação do corpus, extração de representações textuais, treinamento de classificadores, proposição da arquitetura CGPN, avaliação de detecção precoce e análise de interpretabilidade comportamental.

3.1 Conjunto de Dados e Unidade de Análise

Utilizamos os conjuntos oficiais de treinamento e teste do *PAN12 Sexual Predator Identification* [40]. Após a consolidação dos arquivos XML brutos, o corpus resultou em 66.914 conversas de treinamento e 155.101 de teste. O cenário apresenta forte desequilíbrio de classes: conversas rotuladas como positivas correspondem a 3,0% do conjunto de treinamento e 2,4% do conjunto de teste.

A unidade de análise adotada é a conversa completa, pois o *grooming* tende a emergir da progressão da interação, e não de mensagens isoladas. Cada conversa $\mathcal{C}_i$ é composta por uma sequência cronológica de mensagens $\{m_{i,1}, m_{i,2}, \dots, m_{i,T_i}\}$. Para os modelos de representação plana, as mensagens foram concatenadas em um único documento textual D_i :

$$D_i = \sum_{t=1}^{T_i} m_{i,t} \quad (1)$$

O rótulo da conversa foi definido como $y_i = 1$ quando ao menos um participante consta na lista de predadores fornecida pelo PAN12, e $y_i = 0$ caso contrário. Essa formulação permite comparar abordagens que tratam a conversa como um único documento com aquelas que preservam explicitamente a sequência dos turnos.

3.2 Representações Textuais

Para avaliar o impacto da representação na detecção de *grooming*, consideramos quatro famílias de representações textuais. Todas as etapas de ajuste, vetorização e redução dimensional foram realizadas exclusivamente a partir do conjunto de treinamento, evitando vazamento de dados.

- **Representação léxica esparsa:** aplicamos TF-IDF aos documentos concatenados, restringindo o vocabulário a 2.000 características mais informativas. O conjunto de teste foi representado apenas pela transformação do vocabulário aprendido no treinamento.
- **Representação distribuída estática:** treinamos um modelo Word2Vec [32] do tipo *Skip-gram*, com dimensionalidade $d = 100$ e janela de contexto $w = 5$, utilizando apenas as conversas de treinamento. A representação de cada conversa foi obtida pela média dos vetores dos *tokens* válidos.
- **Representações contextuais de sentença:** utilizamos o modelo all-MiniLM-L6-v2 [42], que produz *embeddings* densos de 384 dimensões. Essa representação foi avaliada tanto isoladamente quanto concatenada a características estruturais da conversa.
- **Representações contextuais profundas e de texto longo:** avaliamos BERT *bert-base-uncased* [13], com dimensão $d = 768$ e limite de 512 *tokens*, e Longformer *longformer-base-4096* [3], voltado a documentos extensos. Para o Longformer, adotamos janelas deslizantes de 1.024 *tokens*, com *stride* de 256. A representação final da conversa foi obtida pelo *mean pooling* dos vetores [CLS] das K janelas:

$$E_C = \frac{1}{K} \sum_{k=1}^K \text{CLS}_k \quad (2)$$

Além das representações textuais, extraímos características estruturais da conversa, incluindo o número de turnos, o tamanho médio das mensagens, a proporção de mensagens curtas e o número de autores únicos. Essas características foram concatenadas aos *embeddings* MiniLM em experimentos específicos, com o objetivo de avaliar se sinais estruturais simples contribuem para a detecção de *grooming*.

3.3 Classificadores Supervisionados

As representações extraídas foram avaliadas com diferentes famílias de classificadores, incluindo modelos lineares, métodos baseados em árvores, redes neurais e aprendizado no contexto. Para os modelos sensíveis ao desequilíbrio de classes, utilizamos ponderação inversamente proporcional às frequências das classes no treinamento.

Foram avaliados Regressão Logística, SVM, Floresta Aleatória, MLP, XGBoost [8] e LightGBM [23]. Nos modelos de *gradient boosting*, o peso da classe positiva foi ajustado por meio do parâmetro *scale_pos_weight*, de modo a compensar a raridade das conversas predatórias. Essa estratégia é particularmente relevante no PAN12, em que a classe positiva representa menos de 3% das instâncias. Também avaliamos o TabPFN [20], um classificador baseado em *in-context learning* para dados tabulares. A seleção de subconjuntos de treinamento tem sido explorada como estratégia para reduzir redundância e custo computacional preservando a efetividade [10, 11]. Devido à sua restrição computacional, o TabPFN foi treinado sobre um subconjunto balanceado de 1.000 instâncias e representações reduzidas para 100 dimensões por PCA. Esse experimento permite avaliar o comportamento de um paradigma alternativo em regime de poucos dados rotulados e forte desbalanceamento no teste.

3.4 CGPN: Modelagem Hierárquica da Progressão Conversacional

As abordagens planas avaliadas anteriormente representam a conversa como um documento concatenado. Embora essa formulação preserve o conteúdo textual, ela reduz a estrutura conversacional a uma representação única. Trabalhos recentes [54] já mostraram que sequências de mensagens processadas por redes recorrentes podem capturar padrões relevantes para a detecção de grooming. Tendo esse princípio de modelagem sequencial em vista e o estendendo por meio de uma arquitetura hierárquica em dois níveis, propomos o **Contextual Grooming Pattern Network (CGPN)**, visando capturar a hipótese central deste trabalho: sinais discriminativos emergem da progressão dos turnos.

No primeiro nível da arquitetura, cada mensagem $m_{i,t}$ da conversa C_i é codificada pelo MiniLM, produzindo uma sequência de vetores $\mathbf{e}_{i,t} \in \mathbb{R}^{384}$. Para limitar o custo computacional e padronizar as sequências, consideramos até $T_{\max} = 64$ turnos por conversa, preservando os mais recentes em conversas que excedem esse limite. Mesmo que essa estratégia descarte sinais iniciais de aproximação e construção de confiança, assumimos que os sinais de escalada tendem a se concentrar nas fases posteriores da interação.

Inspirados em redes de atenção hierárquica [56], no segundo nível, a sequência de vetores $\mathbf{e}_{i,1}, \dots, \mathbf{e}_{i,T}$ é processada por uma GRU bidirecional [9] com dimensão oculta $d_h = 256$. Em seguida, aplicamos uma camada de *attention pooling* [2] para produzir um vetor de contexto conversacional $\mathbf{c}_i$:

$$\mathbf{c}_i = \frac{1}{T} \sum_{t=1}^T \alpha_t \mathbf{h}_t, \quad \alpha_t = \frac{\exp \mathbf{w}^\top \mathbf{h}_t}{\sum_{t'} \exp \mathbf{w}^\top \mathbf{h}_{t'}} \quad (3)$$

em que $\mathbf{h}_t$ é o estado oculto bidirecional no turno t , α_t é o peso de atenção atribuído ao turno e $\mathbf{w}$ é um vetor aprendível. O vetor $\mathbf{c}_i$ é concatenado às características estruturais da conversa e classificado por um MLP com saída sigmoide.

O CGPN foi treinado com perda binária cruzada ponderada, usando peso da classe positiva $w = 32,2$, otimizador Adam com taxa de aprendizado $\eta = 10^{-3}$ e *weight decay* de 10^{-4} . O treinamento foi conduzido por até 10 épocas, com *early stopping* de paciência igual a 3. A melhor configuração foi selecionada pelo desempenho no conjunto de validação.

3.5 Protocolo de Avaliação

A avaliação foi conduzida no conjunto oficial de teste do PAN12. Considerando o forte desequilíbrio de classes, adotamos como métricas o F1 e o PR-AUC da classe positiva. Também reportamos precisão e revocação, pois representam diferentes compromissos operacionais em aplicações de segurança: modelos com maior precisão favorecem alertas de alta confiança, enquanto modelos com maior revocação reduzem falsos negativos e podem atuar como triagem inicial com revisão humana. Todas as comparações foram realizadas mantendo a mesma divisão oficial de treinamento e teste, garantindo consistência entre representações e classificadores.

3.6 Protocolo de Detecção Precoce

Avaliamos também a capacidade dos modelos de identificar sinais de risco antes do término da interação. Para isso, definimos um protocolo de detecção precoce em cinco estágios de progressão: 10%, 25%, 50%, 75% e 100% da conversa. Para cada conversa D_i , geramos versões truncadas D_i^τ , em que $\tau \in \{0,10,0,25,0,50,0,75,1,00\}$ representa a fração observada do diálogo. Os modelos foram treinados com conversas completas e avaliados com essas versões parciais. Em cada estágio, calculamos a precisão, a revocação e o F1 da classe positiva. Esse protocolo permite estimar quando os sinais discriminativos começam a emergir e qual perfil operacional cada modelo assume ao longo da progressão da conversa.

3.7 Análise de Explicabilidade e Interpretabilidade

Para investigar se os modelos capturam padrões compatíveis com a dinâmica do *grooming*, combinamos três estratégias: importância de características estruturais, explicabilidade baseada em conceitos e interpretabilidade por design.

Primeiro, aplicamos SHAP [28] aos modelos que combinam MiniLM e características estruturais, a fim de estimar a contribuição de atributos como o número de turnos, o tamanho médio das mensagens e o número de autores. Essa análise permite verificar se a estrutura da conversa fornece sinais relevantes além do conteúdo textual. Segundo, utilizamos Vetores de Ativação de Conceitos (TCAV) [25] para avaliar o alinhamento global dos modelos a conceitos comportamentais associados ao *grooming*. Consideramos conceitos derivados da literatura forense [37, 49], incluindo *Afeto*, *Conexão*, *Isolamento* e *Sexualização*. Cada conceito foi instanciado por conjuntos de palavras-semente usados como *probes* direcionais no espaço de representação. Para cada conceito C , treinamos um classificador linear para separar exemplos associados ao conceito de exemplos de controle. A relevância do conceito para a predição da classe predatória foi medida pela Sensibilidade Direcional Média:

$$S_{C,k,l} = \frac{|\{x \in \mathcal{X}_{teste} \mid \nabla h_{k,l} f_l x \cdot \mathbf{v}_C > 0\}|}{|\mathcal{X}_{teste}|} \quad (4)$$

em que $\mathbf{v}_C$ é o vetor de ativação do conceito, $f_l x$ é a representação do exemplo x na camada l e $h_{k,l}$ projeta essa representação para o *logit* da classe predatória.

Terceiro, implementamos um **Concept Bottleneck Model (CBM)** [27] para quantificar o custo de uma decisão totalmente auditável. O CBM opera em três etapas: (i) anotação heurística de conversas com base em seis conceitos psicológicos de *grooming* — *Afeto*, *Conexão*, *Isolamento*, *Sexualização*, *Busca de Informação* e *Solicitação de Mídia*; (ii) treinamento de classificadores lineares para prever esses conceitos a partir de *embeddings* MiniLM; e (iii) classificação final a partir dos conceitos preditos. A decisão final é dada por:

$$\hat{y} = \sigma \left(\sum_{j=1}^6 \beta_j \hat{c}_j + \beta_0 \right) \quad (5)$$

em que $\hat{c}_j$ representa o conceito j predito e β_j são coeficientes aprendidos. Essa formulação produz uma regra de decisão auditável, permitindo comparar interpretabilidade total e desempenho preditivo.

3.8 Modelos Genéricos de Moderação

Por fim, avaliamos modelos pré-treinados de moderação e toxicidade como referência externa à tarefa de *grooming*. O objetivo não é tratá-los como substitutos dos modelos especializados, mas verificar se marcadores genéricos de conteúdo ofensivo, sexual ou tóxico são suficientes para capturar interações predatórias. Foram considerados modelos como Toxic-BERT e KoalaAI-Moderation. Suas predições foram analisadas no nível de mensagem e comparadas entre mensagens de conversas predatórias e benignas. Essa análise permite verificar empiricamente a hipótese de que *grooming* não coincide necessariamente com toxicidade explícita, reforçando a necessidade de representações e classificadores especializados para capturar padrões relacionais e progressivos.

4 Resultados e Discussão

Esta seção apresenta os resultados à luz das duas questões de pesquisa. Primeiro, analisamos como diferentes representações textuais e estratégias de classificação impactam a detecção de *grooming* (QP1). Em seguida, investigamos em que medida os modelos capturam padrões compatíveis com a progressão relacional do fenômeno, considerando sinais linguísticos, estrutura conversacional, detecção precoce e explicabilidade comportamental (QP2).

4.1 Análise Linguística da Base

Antes da avaliação supervisionada, realizamos uma caracterização linguística no nível de mensagem. A análise de 225.008 mensagens indica que o *grooming* não é adequadamente capturado por detectores genéricos de abuso ou toxicidade. Modelos como KoalaAI-Offensive e DeHateBERT classificam apenas uma pequena fração de mensagens predatórias como ofensivas, reforçando que esse comportamento tende a se apoiar em persuasão, ambiguidade e construção gradual de intimidade, e não necessariamente em agressividade explícita.

O modelo KoalaAI-Moderation revelou-se mais informativo ao distribuir mensagens em categorias específicas. A Figura 1 mostra que a maior parte das mensagens foi classificada como **OK** em ambas as classes. Ainda assim, categorias associadas a conteúdo sexual aparecem com maior frequência relativa em mensagens predatórias (24% vs. 10%). Esse resultado sugere que modelos de moderação podem apoiar a triagem, mas são insuficientes como mecanismo principal de detecção, pois a predominância da classe OK confirma que o comportamento predatório frequentemente não coincide com marcadores explícitos de toxicidade.

A Tabela 1 apresenta padrões linguísticos frequentes extraídos por mineração sequencial utilizando o algoritmo PrefixSpan [38]. Além de termos comuns em conversas online, emergem padrões associados à aproximação afetiva, à busca de informações pessoais e à solicitação de imagens, como *love*, *baby*, *school*, *meet* e *pic*. Esses padrões são compatíveis com estágios descritos na literatura de *grooming*, nos quais

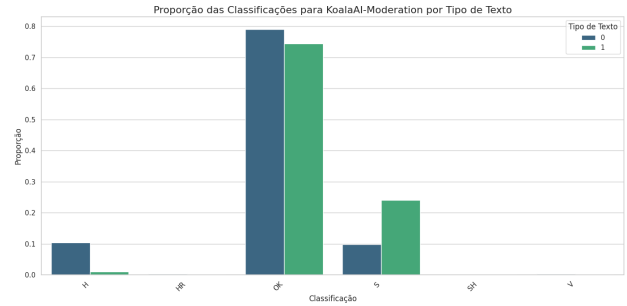


Figura 1: Distribuição de classes previstas pelo KoalaAI-Moderation para mensagens de predadores (1) e não predadores (0). A categoria S (sexual) é marcadamente mais frequente em mensagens predatórias.

o agressor busca construir vínculo, obter informações sobre a vítima e ampliar gradualmente a intimidade da interação.

De forma particularmente relevante, os padrões mais discriminativos não são os explicitamente sexuais. Termos associados a afeto (*love*, *baby*), conexão (*meet*, *friend*) e busca de informações (*home*, *school*, *where*) apresentam diferenças de frequência superiores ou comparáveis às observadas para *sex*. Esse resultado antecipa uma das principais hipóteses investigadas neste trabalho: a detecção de *grooming* depende menos da identificação de linguagem sexual explícita e mais da modelagem de sinais relacionais que se acumulam ao longo da conversa.

Tabela 1: Subsequências frequentes extraídas de mensagens de predadores (PrefixSpan) e diferença de frequência relativa em relação a não predadores ($\Delta\%$).

Padrão	Categoria	$\Delta\%$	Padrão	Categoria	$\Delta\%$
home	Busca de info.	15,79	meet	Conexão	8,67
love	Afeto	15,01	friend	Conexão	8,54
baby	Afeto	13,26	show	Solicit. mídia	8,52
school	Busca de info.	10,58	where	Busca de info.	8,44
pic	Solicit. mídia	9,08	sex	Sexualização	7,94

4.2 QP1: Representações e Classificadores

A Tabela 2 apresenta a comparação entre representações e classificadores. Devido ao forte desbalanceamento de classes, utilizamos F1 e PR-AUC como métricas principais, complementadas por precisão e revocação.

Entre os modelos avaliados, o CGPN apresenta o melhor desempenho global, com F1 = 0,768 e PR-AUC = 0,871. Esse resultado responde diretamente à QP1 ao mostrar que a modelagem hierárquica da sequência de turnos supera as estratégias de concatenação plana. O ganho do CGPN indica que preservar a progressão conversacional é decisivo para a tarefa.

Considerando modelos baseados em concatenação plana, BERT + MLP e MiniLM + Estrutural + XGBoost alcançam os melhores resultados, embora apresentem perfis distintos. O BERT + MLP obtém maior precisão, enquanto MiniLM + Estrutural + XGBoost apresenta melhor equilíbrio entre precisão e revocação, além de menor custo computacional. A comparação com o Longformer é particularmente reveladora: apesar de ter sido projetado para processar documentos

extensos e capturar dependências de longo alcance, seu desempenho não supera os melhores modelos planos em termos de F1. Esse resultado sugere que simplesmente aumentar a quantidade de contexto textual disponível ao modelo não é suficiente para capturar os sinais característicos do *grooming*. Mais importante do que o volume de texto parece ser a capacidade de preservar e modelar explicitamente a progressão da interação ao longo dos turnos conversacionais, reforçando a hipótese central deste trabalho.

Tabela 2: Desempenho dos modelos na classe positiva (predador). †: 1.000 amostras + PCA 100 componentes. ‡: features estruturais concatenadas ao embedding.

Embedding	Classificador	Prec.	Rec.	F1	PR-AUC
MiniLM (turnos)	CGPN (BiGRU)	0,692	0,864	0,768	0,871
BERT	MLP	0,794	0,658	0,720	0,798
MiniLM+Est.‡	XGBoost	0,666	0,795	0,724	0,787
BERT	Deep Res. MLP	0,578	0,820	0,678	0,793
BERT	Deep MLP	0,666	0,715	0,690	0,786
MiniLM	XGBoost	0,834	0,531	0,649	0,763
MiniLM	LightGBM	0,838	0,513	0,637	0,758
Longformer	LogReg	0,366	0,949	0,528	0,768
BERT	SVM	0,382	0,931	0,542	0,705
BERT	LogReg	0,385	0,900	0,539	0,714
TF-IDF	LogReg	0,267	0,865	0,408	0,664
BERT	RF	0,600	0,326	0,423	0,461
MiniLM	TabPFN†	0,229	0,961	0,369	0,652
CBM	LogReg branco	0,074	0,669	0,133	0,135

O CGPN alcança o melhor desempenho geral, superando em aproximadamente 6% de F1 relativo o melhor modelo baseado em concatenação plana. Como ambos os métodos utilizam representações contextuais modernas e classificadores de alta efetividade, esse resultado sugere que preservar explicitamente a sequência dos turnos fornece informação complementar ao conteúdo textual agregado. Dessa forma, os resultados reforçam a hipótese de que a progressão conversacional constitui um sinal relevante para a detecção de *grooming*.

O TabPFN alcança alta revocação, mas baixa precisão, comportamento consistente com seu treinamento balanceado e avaliação em um teste com apenas 2,4% de positivos. Essa diferença entre as métricas indica importantes desdobramentos do desbalanceamento para esse modelo: um alto volume de falsos positivos pode ser detectado, o que, em aplicações reais, aumentaria a carga de trabalho sobre revisores humanos.

Os modelos genéricos e léxicos, por sua vez, apresentam desempenho inferior, indicando que sinais superficiais são insuficientes para captar o padrão relacional do *grooming*.

A Figura 2 complementa a análise ao mostrar o comportamento dos principais modelos em diferentes limiares de decisão. A curva do CGPN mantém vantagem consistente em precisão para níveis crescentes de revocação, o que é especialmente relevante em aplicações de segurança infantil, nas quais falsos negativos têm alto custo.

Assim, os modelos avaliados podem ser organizados em três perfis: (i) modelos de alta precisão, como XGBoost e

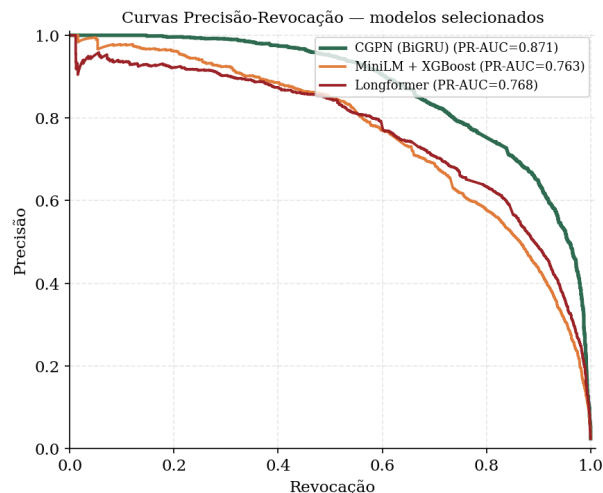


Figura 2: Curvas Precisão-Revocação dos principais modelos. O CGPN mantém precisão superior em todos os níveis de revocação (PR-AUC = 0,871), superando o MiniLM + XGBoost (0,763) e o Longformer (0,768).

LightGBM, úteis para alertas de maior confiança; (ii) modelos de alta revocação, como Longformer, SVM e TabPFN, adequados como primeira camada de triagem; e (iii) modelos equilibrados, como CGPN, BERT + MLP e MiniLM + Estrutural, mais adequados a cenários gerais de detecção.

4.3 QP1 e QP2: Características Estruturais

A Tabela 3 apresenta a ablação de características estruturais e temporais concatenadas ao MiniLM, mostrando que atributos simples da conversa, como número de turnos e tamanho médio das mensagens, agregam informação relevante ao conteúdo textual.

Tabela 3: Ablação de features estruturais e temporais (MiniLM + XGBoost). ΔF1 relativo ao baseline MiniLM puro.

Configuração	Prec.	Rec.	F1	PR-AUC	ΔF1
MiniLM (baseline)	0,613	0,749	0,675	0,742	—
+ Estrutural (8 feat.)	0,666	0,795	0,724	0,787	0,049
+ Estrutural + Temporal	0,668	0,792	0,725	0,787	0,050

A adição de características estruturais gera ganho de +4,9 pontos de F1 em relação ao MiniLM puro. A análise SHAP da configuração completa (Figura 3) mostra que `n_turns_norm` é o preditor estrutural dominante, com importância média muito superior às demais variáveis. Esse resultado conecta QP1 e QP2: além de melhorar o desempenho, a estrutura da conversa mostra que características estruturais da interação fornecem sinal discriminativo adicional.

A inclusão de características temporais acrescenta apenas +0,1 ponto de F1, pois os *timestamps* não estavam disponíveis no *dataset* utilizado neste estudo; assim, variáveis como `delta_t_mean`, `delta_t_std`, `night_ratio` e `burst_ratio` assumem valor zero. Esse resultado não invalida a relevância do tempo na detecção de *grooming*, mas impediu que as características temporais fossem exploradas de forma informativa: bases futuras com metadados temporais podem permitir avaliar melhor o ritmo e a cadência da interação predatória.

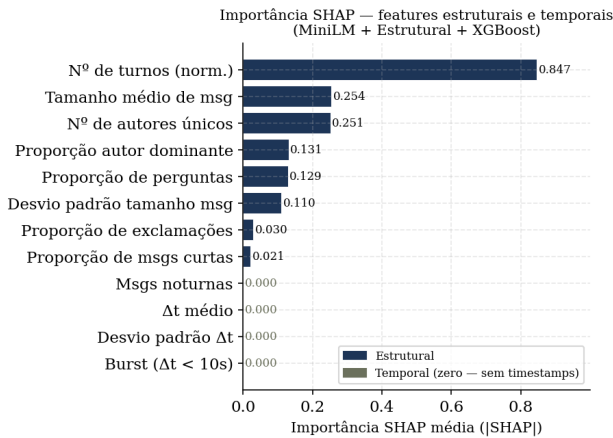


Figura 3: Importância média absoluta SHAP das features estruturais e temporais no modelo MiniLM + Estrutural + XGBoost. O número de turnos domina amplamente com SHAP = 0,847; features temporais assumem valor zero por ausência de timestamps no PAN12.

4.4 QP2: Detecção Precoce ao Longo da Conversa

A Figura 4 apresenta as curvas de detecção precoce para MiniLM + XGBoost e MiniLM + TabPFN ao longo dos cinco estágios de progressão. Esses modelos foram selecionados por representarem perfis operacionais distintos: o XGBoost prioriza maior precisão, enquanto o TabPFN apresenta alta revocação desde os estágios iniciais.

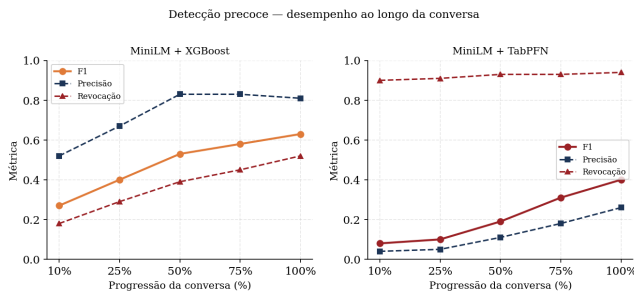


Figura 4: Detecção precoce: MiniLM + XGBoost (esquerda) e MiniLM + TabPFN (direita) ao longo da progressão da conversa. Os dois modelos representam pontos operacionais opostos no compromisso entre precisão e revocação.

Os resultados mostram que sinais discriminativos emergem já nos primeiros 25% da conversa. No MiniLM + XGBoost, esse estágio atinge precisão de 0,67 e F1 = 0,40, indicando que parte relevante dos casos identificados no início da interação já corresponde a conversas predatórias. O aumento progressivo do F1 até 100% confirma que os padrões de *grooming* se acumulam ao longo do diálogo, reforçando a adequação de estratégias de monitoramento contínuo. O TabPFN apresenta comportamento complementar: desde 10% da conversa, alcança revocação próxima de 0,90, mas com precisão muito baixa. Esse perfil é útil para compreender o compromisso operacional entre detectar o maior número possível de casos suspeitos e controlar falsos positivos. Em contextos com revisão humana, modelos de alta revocação podem atuar como

primeira camada de triagem; em contextos com menor tolerância a falsos positivos, modelos mais precisos são preferíveis.

4.5 QP2: Explicabilidade Mecanicista e Comportamental

Para investigar se os modelos aprendem sinais compatíveis com a dinâmica relacional do *grooming*, aplicamos TCAV ao espaço latente dos modelos. Os escores indicam em que medida o deslocamento de uma conversa em direção a um conceito psicológico aumenta a probabilidade de predição da classe predatória. A Figura 5 mostra que conceitos relacionais exercem maior influência do que sexualização explícita. Para a Random Forest, Isolamento e Conexão apresentam escores superiores aos de Sexualização. No TabPFN, o efeito é ainda mais pronunciado: Afeto e Conexão dominam a decisão, enquanto Sexualização exerce influência marginal. Essa diferença evidencia que arquiteturas distintas constroem representações funcionalmente diversas para a mesma tarefa.

Esses resultados respondem à QP2 ao indicar que os modelos não dependem exclusivamente de termos explícitos ou ofensivos. Ao contrário, os sinais mais informativos estão associados a padrões de aproximação, vínculo e progressão relacional. Essa evidência complementa os resultados do CBM, da análise linguística e da detecção precoce, consolidando a interpretação de que o *grooming* deve ser tratado como um fenômeno conversacional e comportamental.

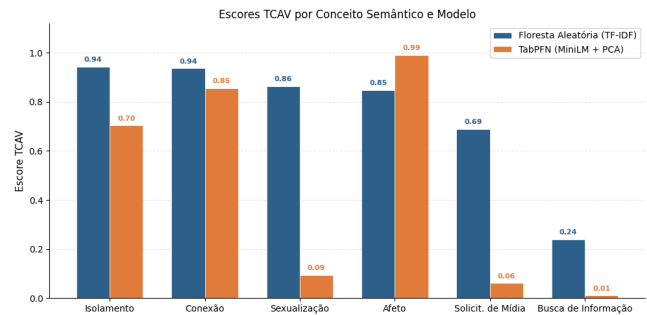


Figura 5: Escores TCAV (Sensibilidade Direcional Média) comparando a dependência conceitual da Random Forest (baseada em TF-IDF) e do TabPFN (baseado em MiniLM + PCA). A análise revela que conceitos de dinâmica relacional superam a linguagem puramente explícita.

4.6 Análise de Erros e Limitações dos Modelos de Moderação

A análise de erros reforça a insuficiência de modelos genéricos de moderação. Mensagens de predadores não detectados por nenhum classificador são predominantemente classificadas como **OK** pelo KoalaAI-Moderation, sugerindo interações mais camufladas. O Toxic-BERT ainda identifica sinais de obscenidade em alguns casos, mas com menor intensidade e consistência do que nas conversas corretamente detectadas.

Esses resultados mostram que modelos de moderação são úteis como suporte complementar, mas não substituem modelos especializados de detecção de *grooming*. As conversas mais difíceis são justamente aquelas em que o comportamento predatório se manifesta de forma sutil,

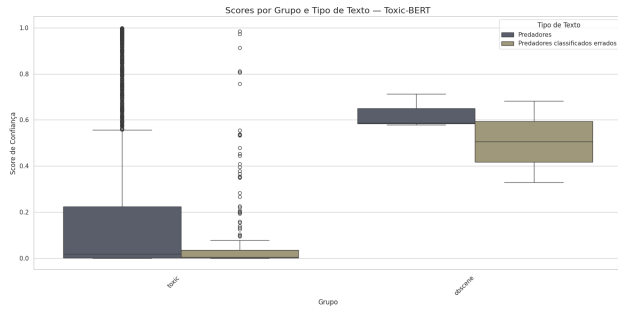


Figura 6: Distribuição dos escores Toxic-BERT para predadores detectados e não detectados. Casos não detectados apresentam menor toxicidade.

gradual e relacional. Assim, a análise de erros reforça a tese central do artigo: detectar *grooming* exige representações capazes de capturar a progressão conversacional, e não apenas marcadores explícitos de toxicidade ou abuso.

4.7 Estudo de Ablação do CGPN

A Tabela 4 avalia a contribuição dos principais componentes do CGPN por meio da remoção ou substituição da projeção das mensagens, bidirecionalidade do GRU, mecanismo de atenção e características estruturais auxiliares.

Tabela 4: Estudo de ablação do CGPN. Cada linha remove um componente mantendo os demais fixos. $\Delta F1$ relativo ao modelo completo (FULL). Melhor resultado em negrito.

Configuração	Prec.	Rec.	F1	PR-AUC	$\Delta F1$
FULL (baseline)	0,692	0,864	0,768	0,871	—
NO_AUX (sem aux.)	0,691	0,862	0,767	0,857	-0,001
UNI-GRU (forward)	0,691	0,861	0,767	0,867	-0,001
MEAN_POOL	0,675	0,862	0,757	0,869	-0,011
LAST_POOL	0,685	0,843	0,756	0,852	-0,012
NO_PROJ (sem proj.)	0,679	0,831	0,747	0,853	-0,021
NO_AUX+NO_PROJ	0,620	0,865	0,723	0,841	-0,046

A **projeção das mensagens** apresenta a maior contribuição individual: sua remoção reduz o F1 em 0,021, indicando que a transformação supervisionada dos embeddings MiniLM auxilia sua adaptação à modelagem sequencial. A substituição da atenção por *mean* ou *last pooling* também reduz o desempenho ($\Delta F1 = -0,011$ e $-0,012$), sugerindo que os turnos contribuem de forma desigual para a decisão e que sinais relevantes não se concentram apenas no final da interação.

A remoção isolada da **bidirecionalidade** ou das **features auxiliares** produz impacto mínimo ($\Delta F1 = -0,001$), enquanto a remoção conjunta das features auxiliares e da projeção resulta na maior queda ($\Delta F1 = -0,046$), indicando complementaridade entre os componentes. Em conjunto com a ablação estrutural da Tabela 3, esses resultados mostram que projeção e atenção respondem pelas maiores perdas individuais, enquanto a integração entre representação semântica, estrutura conversacional e modelagem sequencial produz o melhor desempenho.

5 Conclusão e Trabalhos Futuros

Este trabalho investigou se o *grooming* é melhor modelado como um problema de classificação textual ou como um fenômeno conversacional progressivo. Os resultados fornecem evidências consistentes em favor da segunda hipótese. Em resposta à QP1, observamos que modelos que preservam explicitamente a estrutura da conversa superam abordagens baseadas apenas na concatenação de mensagens. Em particular, o **Contextual Grooming Pattern Network (CGPN)** alcançou o melhor desempenho geral, demonstrando que a modelagem da sequência de turnos fornece informação discriminativa adicional além daquela capturada por representações textuais convencionais. Em resposta à QP2, verificamos que os sinais mais relevantes para a detecção de *grooming* estão associados principalmente a padrões de aproximação, afeto, conexão e progressão relacional, e não apenas à sexualização explícita.

Os diferentes experimentos convergem para essa interpretação. A análise linguística, as características estruturais, a detecção precoce, as análises por SHAP e TCAV e o estudo de ablação indicam que a progressão conversacional constitui um componente central para a tarefa. Em conjunto, os resultados sugerem que avanços futuros na detecção de *grooming* dependem menos de modelos cada vez maiores para classificação textual e mais de abordagens capazes de representar explicitamente a evolução da interação ao longo do tempo.

Assim, este trabalho contribui ao integrar avaliação preditiva, detecção precoce e explicabilidade comportamental em uma análise sistemática da detecção de *grooming*. Mais do que propor uma nova arquitetura, os resultados sustentam a hipótese de que a preservação explícita da dinâmica dos turnos constitui um fator fundamental para a identificação de interações predatórias. Dessa forma, o estudo reforça a caracterização do *grooming* como um fenômeno conversacional progressivo, e não apenas como um problema convencional de classificação textual.

Algumas limitações devem ser consideradas. O estudo foi conduzido exclusivamente no corpus PAN12, coletado em um contexto tecnológico distinto do atual, e não dispõe de metadados temporais reais que permitam modelar explicitamente a cadência das interações. Embora a hipótese investigada esteja associada principalmente à progressão conversacional, e não a vocabulários específicos, a generalização dos resultados para plataformas contemporâneas deve ser avaliada em trabalhos futuros. Além disso, o *ground truth* do PAN12 rotula uma conversa como positiva pela presença de um participante identificado como predador, não por anotação de *grooming* naquela interação específica. Assim, podem existir divergências entre o conteúdo observável de algumas conversas e seus rótulos.

Como trabalhos futuros, pretende-se validar as arquiteturas em bases mais recentes e com metadados temporais, permitindo investigar a cadência das interações, além de explorar estratégias de *ensemble*, arquiteturas conversacionais baseadas em modelos de linguagem e conceitos comportamentais derivados da psicologia forense.

6 Cuidados Éticos

Utilizamos um *dataset* público e pseudoanonimizado, sem possibilidade de reidentificação direta dos usuários. O estudo não envolveu nova coleta de dados com seres humanos e está em conformidade com a LGPD (Lei n.º 13.709/2018) e com a Resolução CNS n.º 510/2016.

Acknowledgments

Este trabalho foi apoiado por CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILDIAR, processo n.º 408490/2024-1), CIA-Saúde (processo n.º PPE-00030-21), UNIMED-BH, FAPEMIG, AWS, NVIDIA e FAPESP.

Ferramentas de IA Generativa foram utilizadas apenas como suporte à revisão gramatical de todo o conteúdo. Os autores realizaram uma revisão completa do texto e assumem responsabilidade pela integridade das informações apresentadas.

Referências

- [1] S Alur et al. 2020. Contextual Representation for Cyberbullying Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4741–4752.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. In *arXiv preprint arXiv:2004.05150*.
- [4] Roxane Bois et al. 2023. Early Detection of Online Grooming with Language Models. *Medium Engineering Blog* (2023).
- [5] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.
- [6] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7, 2 (2015), 223–242.
- [7] Despoina Chatzakou et al. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. 13–22.
- [8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.
- [9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [10] Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 665–674. <https://doi.org/10.1145/3539618.3591638>
- [11] Washington Cunha, Felipe Viegas, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2025. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM Transactions on Information Systems* 43, 1 (2025). <https://doi.org/10.1145/3701769>
- [12] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*.
- [14] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of text-based cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5.
- [15] David Finkelhor and Lisa M Jones. 2006. Youth exposure to violence: Past, current, and future trends. *The Journal of Pediatrics* 149, 1 (2006), S116–S119.
- [16] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic corpus-based machine learning for hate speech detection. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–29.
- [17] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- [18] Victor Machado Gonzaga, Nils Murrugarra-Llerena, and Ricardo Marcacini. 2021. Multimodal intent classification with incomplete modalities using text embedding propagation. In *Proceedings of the 27th Brazilian Symposium on Multimedia and the Web (WebMedia)*. 1–8.
- [19] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data?. In *Advances in Neural Information Processing Systems*, Vol. 35. 507–520.
- [20] Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. 2023. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*.
- [21] Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (2019), 1–54.
- [22] Erin Jones and Sarah Thompson. 2018. Identifying Online Predators Using Machine Learning. In *Proceedings of the 2018 ACM Workshop on Privacy in the Electronic Society*.
- [23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30.
- [24] A Khalil et al. 2021. Early detection of online grooming using natural language processing. *Computer Speech & Language* 68 (2021), 101188.
- [25] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Classification Output: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*.
- [26] Dohyeon Kim, Taehoon Kim, and Jihoon Yang. 2025. Early Detection of Online Grooming with Language Models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. 963–970. <https://doi.org/10.1145/3672608.3707796>
- [27] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 5338–5348.
- [28] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017).
- [29] Sean MacAvaney et al. 2019. Hate speech detection: Challenges and solutions. *PLoS one* 14, 8 (2019), e0221152.
- [30] Binny Mathew, Punyajoy Saha, Seid Muhie Yerdon, Tirthankar Sandhan, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.
- [31] J McGhee et al. 2011. Risk factors for online grooming. *Journal of Forensic Psychology* (2011).
- [32] Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, Vol. 26.
- [33] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)* 54, 3 (2021), 1–40.
- [34] Gabriel Nascimento, Flavio Carvalho, Alexandre M da Cunha, Carlos Roberto Viana, and Gustavo P Guedes. 2019. Hate Speech Detection Using Brazilian Imageboards. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web (WebMedia)*. 325–328.

- [35] Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2024. Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 613–618. <https://doi.org/10.14428/esann/2024.ES2024-222>
- [36] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.
- [37] Rachel O'Connell, J. Price, and C. Barrow. 2004. Cyber stalking, abusive cyber sex and online grooming: A programmatic intervention. *University of Central Lancashire, Centre for Cyber Victim Counselling* (2004).
- [38] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. 2001. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proceedings of the 17th International Conference on Data Engineering*.
- [39] R. Pelle, C. Alcântara, and VP Moreira. 2018. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web (WebMedia)*. 237–243.
- [40] João Pires and Jorge Villar. 2012. Sexual Predator Identification in Chat Conversations. In *CLEF (Online Working Notes/Labs/Workshop)*.
- [41] Flor Miriam Plaza-del Arco et al. 2021. Early detection of cyberbullying on social media networks. *Future Generation Computer Systems* 125 (2021), 706–717.
- [42] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- [43] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*. IEEE, 241–244.
- [44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [45] Julian Risch and Ralf Krestel. 2020. Toxic comment detection in online discussions. *Deep Learning-Based Approaches for Sentiment Analysis* (2020), 85–109.
- [46] Joni Salminen et al. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *ICWSM* (2018).
- [47] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media* (2017), 1–10.
- [48] M Schulze et al. 2020. Online grooming detection with advanced machine learning architectures. *Expert Systems with Applications* (2020).
- [49] Peter K. Smith and Georges Steffgen. 2020. Cyberbullying Through the New Media: Findings from an International Network. *Psychology Press* (2020).
- [50] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Cunha, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 753–761. <https://doi.org/10.1145/3289600.3291032>
- [51] Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Gonçalves. 2020. CluHTM: Semantic Hierarchical Topic Modeling Based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8138–8150. <https://doi.org/10.18653/v1/2020.acl-main.724>
- [52] Felipe Viegas, Washington Cunha, Christian Gomes, Leonardo Rocha, and Marcos Gonçalves. 2025. Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics. *Computational Linguistics* (2025).
- [53] Inna Vogel et al. 2020. Detecting online grooming using deep learning. *Neurocomputing* 404 (2020), 310–321.
- [54] Hamed Waezi, Reza Barzegar, and Hossein Fani. 2024. Osprey: A Reference Framework for Online Grooming Detection via Neural Models and Conversation Features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4101–4106. <https://doi.org/10.1145/3627673.3679974>
- [55] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
- [56] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.
- [57] Dawei Yin et al. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB 2.0*. 1–7.
- [58] Manzil Zaheer, Guru Guruganesh, Kumar Avinash Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, Vol. 33. 17283–17297.