Utilizando Teoria da Informação para Identificar Conversas de Pedofilia em Redes Sociais de Mensagens Instantâneas

Juliana G. Postal[1,2], **Eduardo F. Nakamura**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. General Rodrigo Octávio, 6200, CEP: 69080-900, Manaus – AM – Brasil

[2]Samsung Instituto de Desenvolvimento em Informática da Amazônia (SIDIA)
Av. Min. João G. de Souza, 880, CEP: 69075-830, Manaus – AM – Brasil

{juliana.postal,nakamura}@icomp.ufam.edu.br

Abstract. *Social networks of instant messaging, such as Whatsapp, represent a real threat for children and teenagers, who can easily become targets of sexual predators and pedophiles. Hence, the automatic identification of pedophile chats represent a key tool to protect the young users of social networks. However, these networks have two sensitive particularities: (1) messages are often stored only locally; (2) mobile devices of limited processing power are the major interfaces. In this context, the state-of-the-art has a prohibitive cost to run on mobile devices. On the other hand, the nature of the peer-to-peer communication of such networks make it inviable to process the chat on the cloud, without risking to expose the victims. In this work, we present a new method, based on the Shannon entropy and the Jensen-Shannon divergence, to identify pedophile chats, that achieves nearly* 90% *of* F_1 *and* $F_{0.5}$*, and can be up to 72.8% faster than the state-of-the-art.*

Resumo. *Redes sociais privativas de mensagens instantâneas, como Whatsapp, representam uma ameaça para crianças e adolescentes que podem se tornar alvos de pedófilos. Portanto, a identificação automática de conversas de pedofilia representa uma importante ferramenta para proteção de jovens usuários destas redes. Contudo, estas redes possuem como particularidades: (1) as mensagens são tipicamente armazenadas apenas localmente; e (2) dispositivos móveis de capacidade limitada de processamento são os principais veículos de utilização. Neste contexto, as soluções de estado-da-arte possuem um custo computacional proibitivo para execução em dispositivos móveis. Em contrapartida, a natureza da comunicação ponto-a-ponto destas redes torna, em muitos casos, inviável o processamento em nuvem sem correr o risco de expor as vítimas de pedofilia. Neste trabalho, apresentamos um método, baseado na entropia de Shannon e na divergência de Jensen-Shannon, capaz de identificar conversas de pedofilia com um desempenho próximo a* 90%*, considerando as medidas* F_1 *e* $F_{0,5}$*, e que chega a ser 72,8% mais rápido que o estado-da-arte.*

1. Introdução

Crianças e adolescentes (aqui referenciados como **jovens**) são usuários frequentes de diferentes redes sociais online (RSO) [Kontostathis et al. 2010]. Segundo Livingstone et al. [2010], 59% dos jovens[1] possuem um perfil em alguma rede social, e utilizam a Internet principalmente em casa (87%) e na escola (63%). Embora nesses locais os responsáveis estejam normalmente por perto, é impossível manter a supervisão por tempo integral dos jovens e de suas interações em ambientes online. O anonimato proporcionado pela Internet apresenta riscos [Reis et al. 2016], que esses jovens usuários podem não possuir maturidade para perceber. Além disso, os modelos de privacidade atuais em RSOs nem sempre oferecem proteção adequada para os diferentes perfis de usuários [Silva et al. 2016]. Como consequência, RSO, em especial as baseadas em mensagens instantâneas, representam uma ameaça real para crianças e adolescentes que podem ser assediados por pedófilos[2]. Portanto, a identificação automática de conversas de pedofilia representa uma importante ferramenta para proteção de jovens usuários de RSO.

Os principais desafios relacionados à identificação automática de conversas de pedofilia incluem: (a) bases de dados escassas e desbalanceadas e (b) erros de digitação, abreviações e gírias incomuns. As bases de dados são escassas devido à sensibilidade dos dados e à proteção das vítimas. Como resultado, as bases desbalanceadas podem levar a conclusões enganosas contaminadas pela alta acurácia das classes mais frequentes [Liu e Chawla 2011]. Os erros, abreviações e gírias (e.g., `pls` que significa `please` e `sexx` ao invés de `sex`) podem ser utilizados propositalmente para enganar observadores externos (algoritmos ou pessoas).

Os métodos atuais mais eficazes para a identificação de conversas de pedofilia são baseados na abordagem *Bag of Words* (BoW) que utiliza todas as palavras e suas ocorrências como características para alimentar um algoritmo de aprendizagem de máquina (e.g., *support vector machine* - SVM) [Villatoro-Tello et al. 2012]. Estas soluções são computacionalmente intensas e, frequentemente, trabalham com um vocabulário dinâmico e crescente. Portanto, estas soluções não são escaláveis em ambientes de redes de troca de mensagens instantâneas como Whatsapp, onde as mensagens não são processadas por servidores, mas trocadas ponto-a-ponto entre celulares ou dispositivos móveis. Nesse tipo de ambiente com restrições severas de privacidade e de processamento, os métodos tradicionais para identificação de mensagens de pedofilia apresentam um custo proibitivo para processamento local e, o processamento em nuvem não é, normalmente, uma opção viável, pois as mensagens são armazenadas apenas localmente.

A principal contribuição deste trabalho é um método baseado em quantificadores de Teoria da Informação para identificação de mensagens de pedofilia em RSO de troca de mensagens. O método proposto utiliza entropia de Shannon e divergência de Jensen-Shannon para reduzir a dimensão dos dados e aumentar a eficiência (menor custo computacional) mantendo uma eficácia próxima a 90%, considerando as métricas F_1 e $F_{0,5}$. Os resultados mostram que para conversas com mais de 2.500 palavras o método proposto chega a ser 72,8% mais rápido que o estado-da-arte.

[1] Pesquisa realizada com 25.142 crianças e adolescentes entre nove e dezesseis anos.

[2] Pedófilo é um adulto cujas fantasias focam em jovens como parceiros sexuais [Lanning et al. 2010].

O restante deste artigo está organizado como descrito a seguir. A seção 2 apresenta os principais trabalhos relacionados com a identificação de conversas de pedofilia. A seção 3 detalha a abordagem proposta. Os resultados, comparando o método proposto com o estado-da-arte, são apresentados e discutidos na seção 4. A seção 5 apresenta as conclusões e trabalhos futuros. A título de referência rápida, o apêndice A apresenta alguns conceitos utilizados ao longo do artigo (TF-IDF, F_1 e $F_{0,5}$).

2. Trabalhos Relacionados

Pendar [2007] foi o primeiro autor a abordar o tema de detecção de pedofilia em conversas online, utilizando a base de dados da organização *Perverted Justice*[3]. Como estratégia de pré-processamento, para cada conversas as linhas escritas pelo pedófilo e pela vítima foram separadas em arquivos diferentes e as *stop words*[4] foram removidas. Para extrair características destas conversas, o autor testou N-Gram de grau 1, 2 e 3 e ponderando os termos utilizando TF-IDF (*term-frequency, inverted document frequency*). O algorítimo kNN foi utilizando, obtendo 94% de medida F_1, utilizando trigramas como características.

Villatoro-Tello et al. [2012] venceram a competição PAN 2012[5] para classificar conversas de pedofilia, altualmente representado o estado-da-arte em relação à eficácia na detecção desse tipo de conversa. Os autores utilizaram a base de dados do próprio evento, que inclui a base *Perverted Justice* em sua composição. Os autores filtraram as conversas com: (a) apenas um autor; (b) menos de seis linhas escritas por cada autor; e (c) longas cadeias de caracteres não-ascii; afirmando que estas conversas não possuem informação suficiente. As características utilizadas foram os valores de TF-IDF de todas as palavras do corpo de conversas. O resultado foi um F_1 de 95% utilizando o SVM.

Peersman et al. [2012] apresentaram uma abordagem em três etapas que combina previsões dos três níveis de uma conversa: o nível de mensagem individual, o nível de usuário, e a conversa inteira como combinação das duas anteriores. As características utilizadas são os valores de TF-IDF dos termos das conversas, e das linhas de cada autor, separadamente, para permitir a previsão em três etapas. Foram utilizados dois classificadores SVM: um para detectar uma linha de uma conversa predatória; outro para classificar um participante da conversa como um pedófilo ou não pedófilo. Os resultados desses dois classificadores foram combinados para nivelar o resultado final balanceando a precisão e revocação de cada classificador, obtendo 90% de medida F_1.

Cheong et al. [2015] identificaram pedófilos em uma base de conversas e mensagens de fóruns do jogo infantil *Movie Star Planet*, coletados durante 15 minutos de jogo. Como estratégia de pré-processamento, as linhas das conversas foram agrupadas em vítimas e pedófilos, por usuário/indivíduo, sendo que nesse último caso foram mantidas apenas as linhas onde há claramente um discurso predatório. Os autores utilizaram o TF-IDF para obter informações léxicas das conversas, análise de sentimento para detectar o comportamento de *rule-breaker* e a seleção manual dos trechos finais das conversas. Para a classificação, os autores utilizaram o *Naive Bayes*, obtendo 53% de medida F_1.

Morris e Hirst [2012] utilizaram características léxicas e comportamentais para identificar conversas de pedofilia. Os autores também utilizaram TF-IDF para representar

[3] `http://www.perverted-justice.com`.

[4] *Stop words* são palavras muito frequentes e pouco discriminantes.

[5] `http://pan.webis.de/clef12/pan12-web/`.

as características léxicas das conversas, acrescido de características comportamentais provenientes de informações que podem ser extraídas das conversas, tais como o número de mensagens enviadas por um indivíduo e o número total de conversas que este indivíduo participou. Para identificar predadores, os autores utilizaram o SVM e dois filtros para distinguir predadores de vítimas. A eficácia do método foi 77% de F_1.

Parapar et al. [2012] incorporaram *Linguistic Inquiry and Word Count* (LIWC) no processo de extração de características para a identificação de conversas de pedofilia. Como estratégia de pré-processamento, os autores criaram arquivos diferentes para cada indivíduo. Para a extração de características foram utilizadas as técnicas LIWC para obter a informação de até que ponto diferentes assuntos são usados por pessoas em conversas, e TF-IDF para as características léxicas do texto. O classificador usado foi o SVM que obteve 83% de medida F_1.

Bogdanova et al. [2012] utilizaram análise de sentimento para verificar se uma conversa é predatória ou não. Três bases foram utilizadas pelos autores: (1) *Perverted Justice* para formar o conjunto de dados positivos; (2) *logs* de conversas de *cybersex* adulto; e (3) o *corpus* de conversas chamado NPS de acesso pago. Os autores utilizaram a similaridade semântica de Leacock e Chodorow para encontrar trechos onde o discurso é focado em um único assunto. Estes cálculos consecutivos de similaridade semântica criam cadeias de termos semanticamente relacionados, chamados *Lexical Chains*. Para diferenciar conversas regulares e de pedofilia, os autores estabelecem dois valores, 0.5 e 0.7, como limiares para os tamanhos destas *Lexical Chains*, que se forem extrapolados a conversa é considerada de pedofilia. Não há uma medição de acertos e erros ou técnica de validação descrita nos resultados do trabalho.

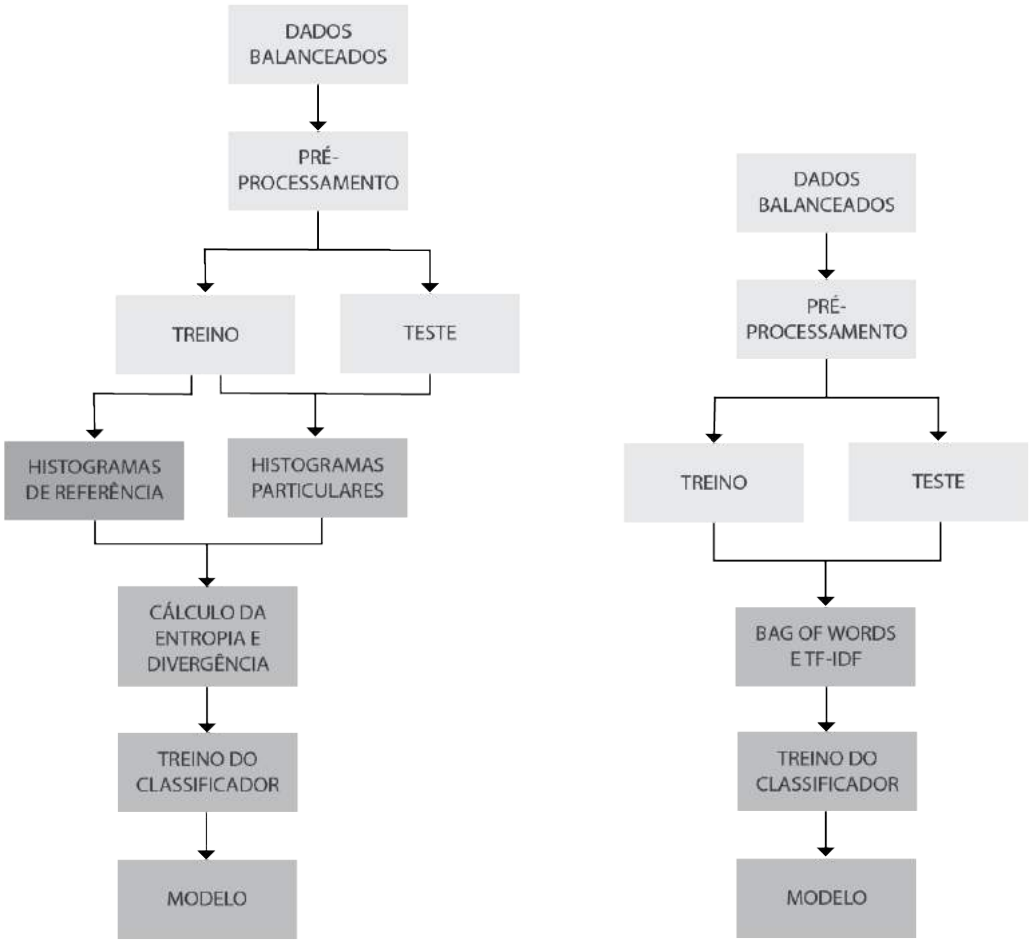
Em relação ao uso da entropia de Shannon e da divergência de Jensen-Shannon para caracterização de textos, Rosso et al. [2009] utilizaram estes quantificadores de informação para caracterizar obras de Shakespeare dentre outros autores renascentistas ingleses. Como estratégia de pré-processamento, os autores realizaram manualmente o trabalho de um *parser* gramatical para remover palavras funcionais e pontuações e aplicar *stemming*. Entretanto, os textos não são classificados, apenas caracterizados em relação usando quantificadores de Teoria da Informação.

3. Abordagem Proposta: H+JSD

A abordagem proposta, referenciada como **H+JSD**, utiliza entropia de Shannon (H) e divergência de Jensen-Shannon (JSD) como características descritivas das conversas que, então, são utilizadas para classificar se uma determinada conversa é de pedofilia ou não. A abordagem H+JSD é ilustrada na figura 1(a) e detalhada a seguir, e diverge da abordagem do estado-da-arte, conforme ilustra a figura 1(b).

3.1. Pré-Processamento

De forma semelhante a Rosso et al. [2009], o pré-processamento utilizado neste trabalho consiste no mapeamento das palavras em dois grandes grupos gramaticais: as palavras funcionais e as léxicas. O grupo de palavras funcionais é composto de palavras que possuem significado semântico fraco, útil apenas para estruturação das frases, raramente admite novas palavras. As classes gramaticais que compõem este grupo são verbos auxiliares, pronomes, conjunções, preposições, determinantes e modais. O grupo de palavras



(a) Método proposto (H+JSD). **(b)** Baseline (BoW).

Figura 1. Fluxos do método proposto (H+JSD) comparado ao baseline (BoW).

léxicas inclui palavras que fornecem significado semântico para as sentenças e é composto por substantivos, verbos principais, adjetivos, interjeições e advérbios.

Nesta etapa, utilizamos o *Stanford Core NLP* [Manning et al. 2008; 2014] para descobrir a classe gramatical de cada palavra e em seguida verificar se ela se encaixa no grupo de funcionais ou léxicas. O passo seguinte é a filtragem das palavras funcionais, consideradas *stop words*.

3.2. Extração de Características

O processo de extração de características inicia com a contabilização das frequências das palavras visando entender sua distribuição no *corpus*. A seguir, é necessário converter estas palavras e suas frequências em dois tipos de histogramas: o primeiro representa a frequência média das palavras de um *corpus*, chamado de histograma de referência; o segundo representa a frequência das palavras de uma única conversa, chamado histograma individual. Todos os histogramas são normalizados de maneira que a soma de todas as classes (probabilidades) é sempre igual a um (distribuição de probabilidades discreta).

O histograma de referência é calculado utilizando a partição de treino. Existem três tipos de entidades em nossa base de dados: predador (predófilo), vítima e regular (nem vítima e nem predador), sendo assim teremos um histograma de referência para cada tipo entidade. Os valores associados às palavras nestes histogramas representam a sua probabilidade de aparecer em um texto daquele tipo de entidade. O histograma individual possui as mesmas palavras do histograma de referência, e seus valores expressam a frequência relativa das palavras presentes na conversa que o originou.

3.2.1. Entropia de Shannon Normalizada (H)

A entropia, no contexto de processamento de texto, descreve a riqueza no uso de um vocabulário. A entropia de Shannon [MacKay 2003] de uma conversa x é dada por

$$S(x) = -\sum_{w \in V} p(w) \log_2 p(w), \tag{1}$$

onde $p(w)$ é a probabilidade de ocorrência da palavra w do vocabulário V. Neste trabalho utilizamos a entropia normalizada, dada por

$$H(x) = \frac{S(x)}{S_{\max}(x)}, \tag{2}$$

onde $S_{\max}(x) = \log_2 |V|$ é a entropia máxima e $|V|$ é o tamanho do vocabulário. Valores próximos a zero indicam poucas palavras muito frequentes, enquanto valores próximos a um indicam uma distribuição mais uniforme entre as palavras.

Bogdanova et al. [2012] afirmam que pedófilos tendem a manter o assunto de cunho sexual com jovens em conversas. A entropia é capaz de detectar esta característica de vocabulário repetitivo em relação às conversas regulares. O processo de obtenção dos valores de entropia utiliza as palavras do histograma de referência como guia para sabermos quais palavras devemos contabilizar a frequência relativa. Portanto, computamos três valores de entropia como característica para cada conversa (um para cada vocabulário considerado: predador, vítima e regular).

No entanto, é possível que em conversas regulares haja repetição de assunto, uma vez que o estudo das palavras é feito do ponto de vista da frequência e não da semântica. Nesse caso, o comportamento da entropia será semelhante às conversas de pedofilia. Por este motivo, a utilizamos a entropia em conjunto com a divergência de Jensen-Shannon.

3.2.2. Divergência de Jensen-Shannon (JSD)

A divergência de Jensen-Shannon (JSD) mede a similaridade entre duas distribuições de probabilidade p e q [Lin 1991], e é dada por

$$\mathit{JSD}(p, q) = S\left(\frac{p+q}{2}\right) - \frac{S(p) + S(q)}{2}, \tag{3}$$

onde $S(\cdot)$ é a entropia de Shannon (equação 1). Quanto mais similares as distribuições, mais próximo de zero é a JSD, e quanto mais distintas, mais próximo de um é a JSD.

No contexto deste trabalho, cada um dos histogramas de referência representa o padrão de discurso de um determinado tipo de entidade (predador, vítima e regular). O cálculo de divergência é feito entre o histograma particular e cada um dos histogramas de referência, resultando em três valores de JSD. Portanto, cada conversa terá seis características, que serão as entropias e as divergências em relação ao padrão de discurso de cada tipo de entidade, indicando se cada entidade de cada conversa assemelha-se mais com um predador, uma vítima ou uma pessoa regular.

3.3. Classificação

Para a elaboração do modelo, utilizamos o SVM para realizar as predições, pois conforme mostram os trabalhos relacionados (seção 2), o SVM produz os resultados mais precisos. O ajuste de parâmetros e escolha do *kernel* foram feitos de forma empírica, com o objetivo de maximizar a eficácia da classificação segundo as medidas F_1 e $F_{0,5}$. A metodologia de ajustes dos parâmetros é detalhada a seguir.

4. Avaliação Experimental

Nesta seção, descrevemos a metodologia de avaliação utilizada, bem como os parâmetros escolhidos, e apresentamos uma discussão dos resultados referentes à: (1) eficácia da classificação; (2) análise de complexidade dos métodos; e (3) eficiência da execução.

4.1. Metodologia (Materiais e Métodos)

Neste trabalho, utilizamos a base de dados pública do *International Competition on Plagiarism Detection* (PAN) de 2012[6], que abordou o tema de aliciamento em conversas online. Esta é a base utilizada por Villatoro-Tello et al. [2012], cuja solução escolhemos como *baseline* por ser o estado-da-arte em desempenho de classificação (seção 2). O *baseline* será referenciado no restante deste documento como **BoW** (*Bag of Words*). A base original é desbalanceada, apresentando 208.248 conversas regulares e 3.677 de pedofilia. Para evitar viés, um subconjunto balanceado da base orinal, composto elas as 3.677 conversas de pedofilia, e 3.677 conversas do tipo regular escolhidas aleatoriamente.

Para uma comparação justa entre os métodos, utilizaremos a estratégia de seleção de palavras do BoW para que ambos os métodos utilizem as mesmas palavras nos experimentos, isolando apenas a forma de transformá-las em características. Este processo é ilustrado na figura 2.

Os métodos foram implementados usando Matlab R2015b, Weka 3.8.0, e Stanford Core NLP 3.7.0. Os experimentos com medições de tempo foram executados em um PC com 8GB de memória RAM, processador Intel core i5 $2,9$ GHz, HD SATA de 1TB.

Para todos os experimentos, adotamos uma avaliação baseada em validação cruzada de dez grupos (*ten-fold cross-validation*). A avalição de eficácia inclui as medidas F_1 e $F_{0,5}$. A medida F_1 serve como referência comparativa para os trabalhos relacionados e pondera igualmente a Precisão e Revocação dos métodos. A medida $F_{0,5}$ prioriza os valores de Precisão dos resultados e também foi utilizada por Villatoro-Tello et al. [2012].

Os intervalos de confiança referem-se a $\alpha = 0.05$ (95% de confiança).

[6]Disponível em: `http://pan.webis.de/clef12/pan12-web/`.

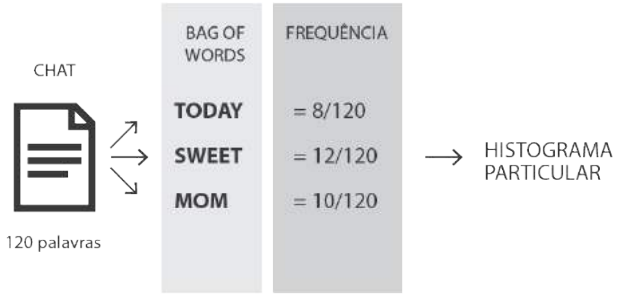
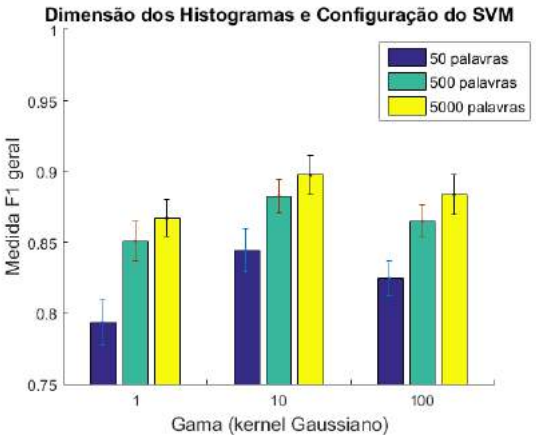


Figura 2. Processo de seleção de palavras.

Ajuste Empírico de Parâmetros

Diferente do BoW, no H+JSD, o histograma de referência não é o estado final das características, mas nessa etapa é importante descobrir um tamanho de vocabulário que não sacrifique os resultados de classificação e que reduza o custo computacional total. Sendo assim, para encontrar empiricamente os valores ideais de dimensão do vocabulário e configuração do SVM, realizamos uma série de experimentos onde testamos os tamanhos de vetor 50, 500 e 5.000 e os valores de γ iguais a 1, 10 e 100 para o *kernel* Gaussiano. Por fim, observamos os valores de medida F_1 geral obtidos. Os resultados, exibidos na figura 3, mostram que um vocabulário de tamanho 5.000 e $\gamma = 10$ é o suficiente para obtermos um valor próximo a 90% de F_1. Portanto, esta será a configuração utilizada no método proposto (H+JSD).



Palavras	γ 1	γ 10	γ 100
50	0,7935 $\pm$ 0,016	0,8443 $\pm$ 0,015	0,8247 $\pm$ 0,012
500	0,8511 $\pm$ 0,014	0,8827 $\pm$ 0,012	0,8651 $\pm$ 0,011
5.000	0,8672 $\pm$ 0,013	**0,8975 $\pm$ 0,014**	0,8838 $\pm$ 0,014

Figura 3. F_1 para diferentes tamanhos de histograma e diferentes valores de γ.

4.2. Resultados

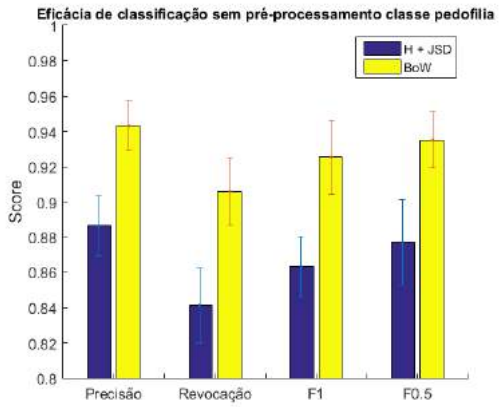
Esta seção apresenta uma avaliação quantitativa entre o H+JSD e o BoW, que inclui avaliações experimentais e analíticas (complexidade computacional). A análise de com-

plexidade e as medições de tempo de execuções são instrumentos complementares para demonstrar a maior eficiência do H+JSD em relação ao BoW.

4.2.1. Eficácia dos Métodos (Qualidade da Classificação)

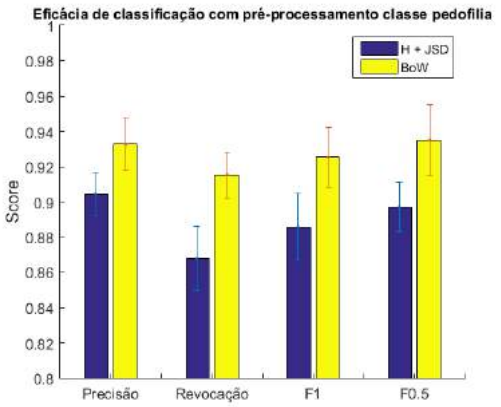
As figuras 4(a) e 4(b) apresentam os resultados de F_1 da classificação, com e sem pré-processamento, enquanto as figuras 5(a) e 5(b) apresentam os resultados de $F_{0,5}$. Os resultados mostram o impacto positivo do pré-processamento para ambos os métodos.

A remoção de palavras funcionais melhora os resultados, pois evita que o classificador utilize palavras que não agregam significado às sentenças, ou que não possuem informação suficiente para distinguir o discurso das possíveis entidades presentes nas conversas. O método H+JSD mantém valores de F_1 e $F_{0,5}$ próximos a 90%, compadado a valores próximos a 94% para o BoW. Entretanto, ao observar as medidas de $F_{0,5}$ observamos que a diferença de desempenho entre o H+JSD e o BoW diminuem, isso ocorre porque a diferença de revocação é maior do que a de precisão.



Métricas	H+JSD	BoW
Precisão	0,884 ± 0,015	0,942 ± 0,014
Revocação	0,843 ± 0,023	0,905 ± 0,019
F_1	0,863 ± 0,017	0,925 ± 0,021
$F_{0.5}$	0,877 ± 0,024	0,937 ± 0,016

(a) Sem pré-processamento.



Métricas	H+JSD	BoW
Precisão	0,9018 ± 0,015	0,938 ± 0,015
Revocação	0,8721 ± 0,023	0,917 ± 0,012
F_1	0,886 ± 0,019	0,923 ± 0,017
$F_{0.5}$	0,897 ± 0,014	0,939 ± 0,020

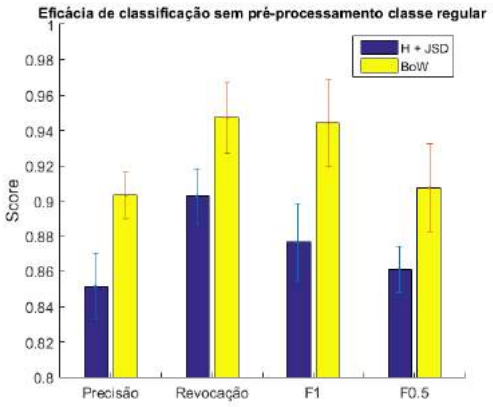
(b) Com pré-processamento.

Figura 4. Comparação das métricas da classe pedofilia com e sem pré-processamento.

4.2.2. Eficiência Analítica dos Métodos (Análise de Complexidade)

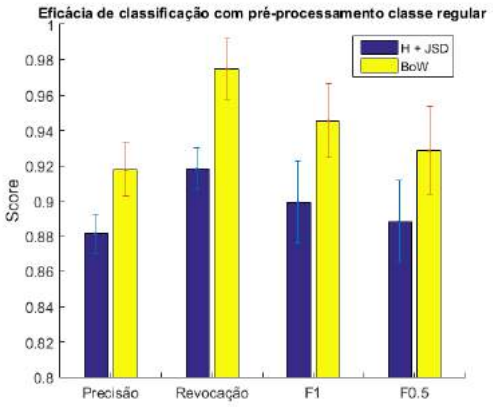
O H+JSD possui uma eficiência (velocidade) muito maior que o BoW. Isto ocorre porque o vetor de características compacto do H+JSD resulta um custo computacional muito menor que o BoW. Na etapa de extração de características ambos os métodos possuem o mesmo grau de complexidade $O(pm)$, onde p é a quantidade de palavras (termos) e m é a quantidade de conversas. A complexidade da etapa de classificação é composta pelo

custo do treino e teste do modelo. O treino do SVM tem custo $O(kt^2)$ onde t é o tamanho da coleção de conversas de treino e k é a quantidade de características que representa cada instância, este custo é referente às operações de produto escalar que o classificador realiza para encontrar o hiperplano ótimo de separação das instâncias. O teste tem custo $O(kv)$ onde v é o tamanho da coleção de teste. Nesse aspecto, o H+JSD leva vantagem significativa em relação ao BoW, pois independente da quantidade de conversas consideradas no estudo, ou da quantidade de palavras únicas encontradas na coleção, o H+JSD utiliza apenas seis características, tornando o custo de classificação no treino igual a $O(t^2)$ e o de teste igual a $O(v)$, enquanto que para o BoW, apesar de fixarmos o tamanho do vetor de características em 5.000 palavras por motivos de eficiência, esta quantidade é, a princípio, igual ao tamanho do vocabulário.



Métricas	H+JSD	BoW
Precisão	0,856 ± 0,020	0,901 ± 0,013
Revocação	0,901 ± 0,017	0,943 ± 0,020
F_1	0,876 ± 0,022	0,944 ± 0,025
$F_{0.5}$	0,861 ± 0,013	0,923 ± 0,023

(a) Sem pré-processamento.



Métricas	H+JSD	BoW
Precisão	0,881 ± 0,011	0,919 ± 0,015
Revocação	0,919 ± 0,013	0,978 ± 0,017
F_1	0,899 ± 0,023	0,945 ± 0,021
$F_{0.5}$	0,888 ± 0,022	0,928 ± 0,025

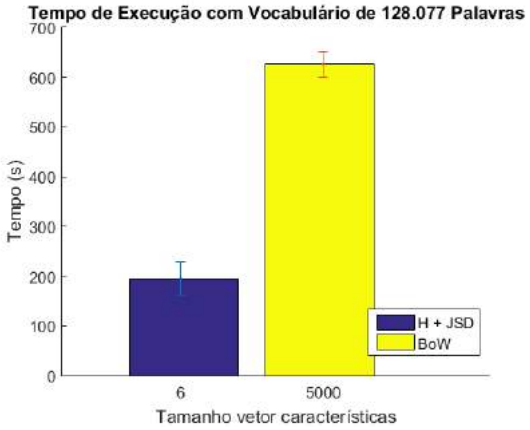
(b) Com pré-processamento.

Figura 5. Comparação de métricas da classe regular com e sem pré-processamento.

4.2.3. Eficiência Medida dos Métodos (Tempo de Execução)

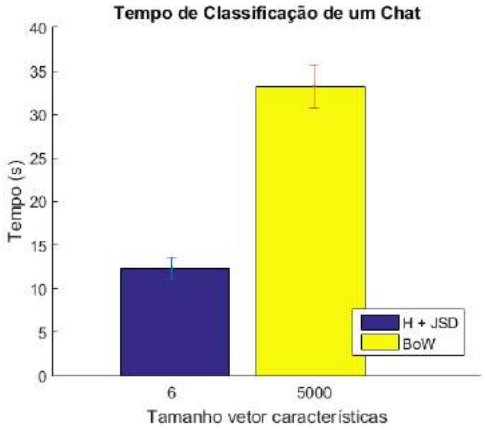
O custo computacional e o tempo de execução dos métodos são métricas fundamentais, principalmente considerando redes de mensagens instantâneas ponto-a-ponto baseadas em dispositivos móveis, tais como o Whatsapp. A figura 6(a) exibe tempo de execução total (128.077 conversas, ten-fold cross-validation), enquanto a figura 6(b) apresenta o tempo médio de classificação de uma conversa. O impacto do custo computacional reduzido do H+JSD, evidenciado na análise de complexidade apresentada, se traduz em um tempo menor de processamento. Em particular, para a tarefa de classificação, o tempo médio do H+JSD é $62,85\%$ menor que o BoW (12,35s contra 33,24s).

As figuras 7(a) e 7(b) apresentam o tempo de execução de ambos os métodos para



H+JSD	BoW
194,58s ± 33,16s	626,36s ± 25,43s

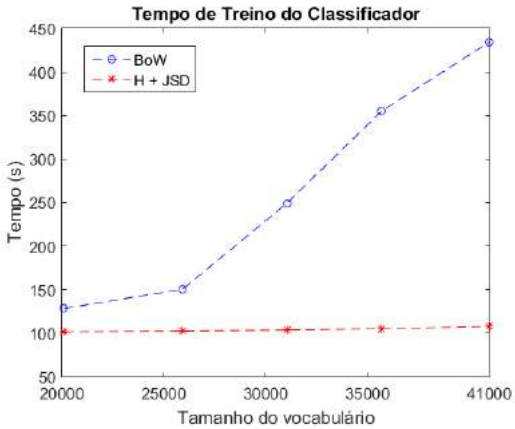
(a) Treino.



H+JSD	BoW
12,35s ± 1,27s	33,24s ± 2,46s

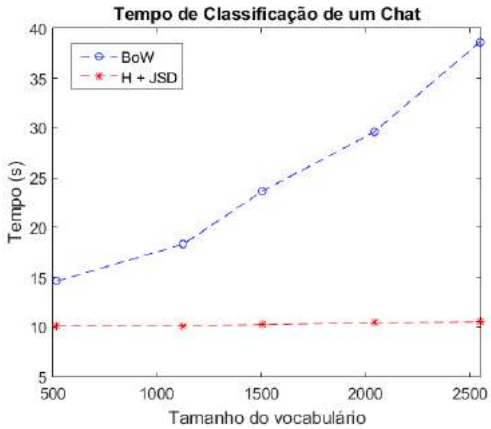
(b) Classificação de uma conversa.

Figura 6. Avaliação do tempo de treino e de classificação de uma conversa.



# Palavras	H+JSD	BoW
20.092	101,71s	128,34s
25.914	102,49s	150,27s
31.036	103,48s	249,12s
35.711	104,81s	355,43s
40.996	107,62s	434,51s

(a) Treino.



# Palavras	H+JSD	BoW
517	10,08s	14,59s
1.124	10,12s	18,31s
1.503	10,23s	23,63s
2.039	10,45s	29,58s
2.548	10,51s	38,67s

(b) Classificação de uma conversa.

Figura 7. Tempo de treino e de classificação de uma conversa pelo vocabulário.

o treino do modelo e para a classificação de uma conversa. Neste experimento, variamos o tamanho do vocabulário utilizado pelo BoW e nos histogramas de referência do H+JSD. Os resultados mostram que o método proposto H+JSD é muito mais escalável que o BoW. Para conversas com mais de 2.500 palavras o H+JSD chega a ser 72,8% mais rápido que o BoW (10,51s contra 38,67s) e, para conversas maiores, esta diferença será ainda maior.

5. Conclusão

Neste trabalho, apresentamos um método para identificação de conversas de pedofilia em redes sociais de mensagens instantâneas. O método proposto (H+JSD) alcança valores de F_1 e de $F_{0,5}$ próximos a 90%, comparados a 94% do estado-da-arte (BoW). Contudo, conforme demonstramos, o estado-da-arte não é escalável e seu custo computacional é proibitivo para dispositivos móveis. Em contraste, a solução proposta é escalável, com um custo computacional reduzido comparado ao BoW. Esta escalabilidade é um requisito chave para implementação de filtros de pedofilia em aplicativos móveis como Whatsapp, pois as mensagens são armazenadas apenas localmente não sendo desejável (ou possível) que as conversas sejam processadas na nuvem sem a prévia autorização de todos os participantes da conversa.

Embora a solução proposta seja significativamente mais eficiente (até 72,8% mais rápida que o estado-da-arte) ainda há espaço para redução do custo computacional e aumento da eficácia (qualidade da classificação). Como trabalhos futuros, estamos avaliando outras métricas de dissimilaridade comumente usadas em Teoria da Informação, Recuperação de Informação e Análise de Dados, tais como as distâncias de Hellinger, Jaccard, Cosseno e Bray-Curtis.

A. Apêndice: Conceitos Fundamentais

A título de suporte, a seguir são apresentadas as definições de alguns conceitos comuns aos trabalhos relacionados e que são utilizados neste artigo. Mais detalhes sobre estes conceitos são fornecidos por Manning et al. [2008].

A.1. Características Usadas por Bag of Words

Definição 1 (Frequência de termo (TF)) *O TF contabiliza a frequência de ocorrência de um dado termo (palavra ou conjunto de palavras) em um conjunto de documentos. Intuitivamente, esta frequência é proporcional a importância do termo para o universo de discurso. O valor de TF de um termo t em um documento d é dado por*

$$\text{tf}_{t,d} = \textit{número de vezes que } t \textit{ ocorre em } d. \tag{4}$$

Definição 2 (Frequência invertida de documento (IDF)) *O IDF estima a raridade de um termo em uma coleção de documentos, de forma que se um dado termo ocorre em todos os documentos, seu IDF é zero. O valor de IDF de um termo t, sobre o corpus C, é dado por*

$$\text{idf}_{t,C} = \log \frac{|C|}{|\{d \in C : t \in d\}|}, \tag{5}$$

onde $|C|$ é o tamanho do corpus C e $|\{d \in C : t \in d\}|$ é o número de documentos de C em que t aparece.

Definição 3 (Frequência de termo-frequência invertida de documento (TF-IDF)) *O TF-IDF é dado por*

$$\text{tf-idf}_{t,d,C} = \text{tf}_{t,d} \cdot \text{idf}_{t,C}. \tag{6}$$

A.2. Métricas de Avaliação de Classificadores

Neste artigo usamos a medida F_β para avaliar os métodos, em particular considerando os valores $\beta = 0,5$ e $\beta = 1$.

Definição 4 (Medida F_β) *A medida F_β, para $\beta \geq 0$ é dada por*

$$F_\beta = (1 + \beta^2) \frac{\text{Precisão} \cdot \text{Revocação}}{\beta^2 \cdot \text{Precisão} + \text{Revocação}}, \tag{7}$$

onde Precisão e Revocação são dados por

$$\text{Precisão} = \frac{VP}{VP + FP}, \qquad \text{Revocação} = \frac{VP}{VP + FN},$$

e VP são os verdadeiros positivos, FP são os falsos positivos e FN são os falsos negativos para a classe avaliada.

Referências

Bogdanova, D., Rosso, P., e Solorio, T. (2012). On the impact of sentiment and emotion based features in detecting online sexual predators. In *In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 110 – 118, Jeju, Korea. Association for Computational Linguistics.

Cheong, Y., Jensen, A. K., Gudnadottir, E. R., Bae, B., e Togelius, J. (2015). Detecting predatory behavior in game chats. *Transactions on Computational Intelligence and AI in Games*, 7(3):220 – 232.

Kontostathis, A., Edwards, L., e Leatherman, A. (2010). *Text Mining and Cybercrime*, pages 149 – 164. John Wiley & Sons, Ltd, West Sussex, United Kingdom.

Lanning, K. V., for Missing & Exploited Children, N. C., et al. (2010). *Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children*. National Center for Missing & Exploited Children with Office of Juvenile Justice and Delinquency Prevention, Virginia, USA.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Liu, W. e Chawla, S. (2011). *Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets*, pages 345 – 356. Springer Berlin Heidelberg, Berlin, Heidelberg.

Livingstone, S., Haddon, L., Görzig, A., e Ólafsson, K. (2010). *Risks and safety on the Internet: the perspective of European children*. LSE: EU Kids Online, London, United Kingdom.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Manning, C. D., Raghavan, P., e Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., e McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Morris, C. e Hirst, G. (2012). Identifying sexual predators by svm classification with lexical and behavioral features. In *Working notes of the 3rd Conference and Labs of the Evaluation Forum, Evaluation Labs and Workshop*, volume 12, pages 1 – 29, Rome, Italy. The CLEF Initiative.

Parapar, J., Losada, D., e Barreiro, A. (2012). A learning-based approach for the identification of sexual predators in chat logs. In *Working notes of the 3rd Conference and Labs of the Evaluation Forum, Evaluation Labs and Workshop*, volume 12, pages 1 – 12, Rome, Italy. The CLEF Initiative.

Peersman, C., Vaassen, F., Van Asch, V., e Daelemans, W. (2012). Conversation level constraints on pedophile detection in chat rooms. In *Working notes of the 3rd Conference and Labs of the Evaluation Forum, Evaluation Labs and Workshop*, volume 12, pages 1 – 13, Rome, Italy. The CLEF Initiative.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *In Proceedings of the International Conference on Semantic Computing (ICSC)*, volume 1, pages 235 – 241, California, USA. IEEE.

Reis, J., Miranda, M., Bastos, L., Prates, R., e Benevenuto, F. (2016). Uma análise do impacto do anonimato em comentários de notícias online. In *Anais do 13o. Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*, pages 1290–1304. SBC.

Rosso, O. A., Craig, H., e Moscato, P. (2009). Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*, 388(6):916 – 926.

Silva, C., Barbosa, G., Silva, I., Silva, T., e Mourão, F. (2016). Privacidade para crianças e adolescentes em redes sociais online sob a lente da usabilidade: Um estudo de caso no facebook. In *Anais do 13o. Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*, pages 1245–1259. SBC.

Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., Montes-y Gómez, M., e Pineda, L. V. (2012). Two-step approach for effective detection of misbehaving users in chats. In *Working notes of the 3rd Conference and Labs of the Evaluation Forum, Evaluation Labs and Workshop*, volume 12, pages 1 – 12, Rome, Italy. The CLEF Initiative.