Juliana Gorayeb Postal

AVALIAÇÃO DO USO DE QUANTIFICADORES DE TEORIA DA INFORMAÇÃO PARA IDENTIFICAÇÃO DE CONVERSAS ONLINE DE PEDOFILIA

Manaus

Junho de 2017

Juliana Gorayeb Postal

AVALIAÇÃO DO USO DE QUANTIFICADORES DE TEORIA DA INFORMAÇÃO PARA IDENTIFICAÇÃO DE CONVERSAS ONLINE DE PEDOFILIA

Dissertação apresentada ao Instituto de Computação da Universidade Federal do Amazonas, para a obtenção do Grau de Mestre em Informática.

Univesidade Federal do Amazonas

Instituto de Computação

Programa de Pós-graduação em Informática

Orientador: Eduardo Freire Nakamura

Manaus

Junho de 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Postal, Juliana Gorayeb
P857a Avaliação do Uso de Quantificadores de Teoria da Informação para Identificação de Conversas Online de Pedofilia / Juliana Gorayeb Postal. 2017
63 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. pedófilos. 2. chats. 3. teoria da informação. 4. redes sociais. I. Nakamura, Eduardo Freire II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Avaliando o Uso de Quantificadores de Teoria da Informação Para Identificação de Conversas Online de Pedofilia"

JULIANA GORAYEB POSTAL

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

Prof. Carlos Maurício Seródio Figueiredo - MEMBRO INTERNO

Prof. José Luiz de Souza Pio - MEMBRO EXTERNO

Manaus, 05 de Maio de 2017

Resumo

Redes sociais privativas de mensagens instantâneas, como Whatsapp, representam uma ameaça para crianças e adolecentes 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 para extração de características de texto baseado em dois quantificadores de teoria da informação, que utilizam histogramas individuais de palavras que representam as conversas e três histogramas médios que representam o padrão de discurso dos possíveis tipos de autores presentes na base de dados: Predador (pedófilo), vítima e regular (nem vítima e nem predador). O primeiro quantificador é a entropia de Shannon que indica repetição de assunto dos tipos de autor em conversas, o segundo é a divergência de Jensen-Shannon que mede a similaridade entre o discurso em uma conversa em relação ao padrão de discurso dos tipos de autor. O método proposto é capaz de resumir as conversas consideradas no estudo em três características de entropia e três características de divergência independente da quantidade de conversas consideradas nos experimentos. Este vetor de características compacto permite que um classificador seja 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.

Palavras-chave: pedófilos, chats, Teoria da Informação, redes sociais..

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. In this work, we present a method for extracting text features based on two information theory quantifiers, using individual histograms of words representing the conversations and three mean histograms that represent the discourse pattern of possible types of authors present on the basis of Data: Predator (pedophile), victim and regular (neither victim nor predator). The first quantifier is Shannon's entropy which indicates repetition of the subject's subject in conversations, the second is the Jensen-Shannon divergence that measures the similarity between speech in a conversation relative to the discourse pattern of author types. The proposed method is able to summarize the conversations considered in the study in three characteristics of entropy and three characteristics of divergence independent of the amount of conversations considered in the experiments. This compact feature vector allows a classifier to be able to identify pedophile conversations with a performance close to 90%, considering the measures F_1 and $F_0.5$, and that it becomes 72.8% faster than the state of the art.

Keywords: pedophiles, chats, Information Theory, social networks.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BoW	*Bag of Words*
KNN	*K-Nearest Neighbor*
LIWC	*Linguistic Inquiry and Word Count*
NCMEC	*National Center for Missing and Exploited Children*
NLP	*Natural Language Processing*
NPS	*Naval Postgraduate School*
RSM	*Redes Sociais Móveis*
RSO	*Redes Sociais Online*
SSS	*Small Sample Size*
SVM	*Support Vector Machines*
TF-IDF	*Term Frequency – Inverse Document Frequency*

Sumário

1 Introdução

Estamos em uma era onde a computação móvel e as redes sociais se combinaram em redes sociais móveis como um meio para as pessoas socializarem e se conectarem diretamente através de seus telefones celulares (CHIN; ZHANG, 2013). Crianças e adolecentes (aqui referenciados como **jovens**) são usuários frequentes de diferentes redes sociais online (RSO) (KONTOSTATHIS; EDWARDS; LEATHERMAN, 2010) e também de redes sociais móveis (RSM) devido a facilidade de acesso aos *smatphones*. 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%), além disso 33% dos jovens entrevistados afirmam acessar à Internet pelo uso de *smartphones* ou dispositivos portáteis como *tablets*. 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, principalmente pela utilização dos *smartphones* que são mais discretos que um computador.

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, RSOs, em especial as baseadas em mensagens instantâneas, representam uma ameaça real para crianças e adolecentes 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.

A disponibilidade generalizada da *Internet* e o anonimato que ela proporciona, trouxeram novas formas de crime. Por este motivo, muitos predadores sexuais criam perfis falsos, onde escondem sua identidade e idade (BOGDANOVA; ROSSO; SOLORIO, 2012). Outro resultado obtido pela *EU Kids Online Project* foi que uma pequena parte das crianças expostas a conteúdos sexuais ficaram de fato incomodadas, isto sugere que em muitos casos, elas não entendem o risco associado a estes conteúdos ou mensagens (LIVINGSTONE et al., 2010). Concluímos que jovens na faixa etária citada não possuem discernimento para evitar uma situação de perigo na *Internet*, e este é o fator explorado pelos molestadores de crianças. Eles procuram perfis de jovens em redes sociais e iniciam o contato por *chat*, tentando conquistar a confiança deles com o objetivo de marcar um encontro pessoalmente. Assim, este trabalho surge da necessidade de elaborar um modelo para detectar

[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; CHILDREN et al., 2010).

conversas online de pedofilia em RSO para *smartphones*, conversa online é um termo que utilizaremos para fazer referência a uma conversa na *Internet* e não para afirmar que a identificação das conversas é em tempo real.

1.1 Contexto e Motivação

A rotulagem de predadores em um *log* de *chat* é considerada uma tarefa de aprendizagem de máquina, especificamente uma tarefa de classificação de texto usando aprendizagem supervisionada (CHEONG et al., 2015). Na literatura de psicolinguística, existem fortes indícios que ligam o uso de linguagem natural à personalidade, as flutuações sociais e situacionais, e intervenções psicológicas. O interesse em particular está nas conclusões que apontam para o valor psicológico de estudar o uso das palavras para identificar o comportamento enganoso (PENNEBAKER; FRANCIS; BOOTH, 2001; MIHALCEA; STRAPPARAVA, 2009; HANCOCK et al., 2007; NEWMAN et al., 2003).

Abordagens que utilizam aprendizagem de máquina e processamento de linguagem natural são largamente utilizadas para realizar a tarefa de identificar pedófilos em *chats*. Porém existe uma abordagem baseada em teoria da informação que mostrou ser promissora em detectar estilos de linguagem em textos literários, o foco nesta abordagem está numa compreensão mais geral do padrão de frequência das palavras e sua distribuição em um *corpus* de textos literários (ROSSO; CRAIG; MOSCATO, 2009). Desta forma, acreditamos que o uso desta abordagem proporcionará um estudo interessante em extração de características, que é a forma de representar cada amostra de uma classe em aprendizagem de máquina, aplicada ao contexto de detecção de *chats* de aliciamento.

De acordo com uma pesquisa de vitimização feita nos Estados Unidos em 2008 pela NCMEC. Na Europa 1 em cada 12 crianças já encontraram pessoalmente alguém com quem conversaram *online*, e que 59% das crianças entrevistadas na pesquisa, disseram que acessam a *Internet* no seu próprio quarto ou em algum cômodo particular por meio de *tablets* ou *tablets* (LIVINGSTONE et al., 2010).

A organização *Perverted Justice*, foi fundada nos Estados Unidos em 2003 com o objetivo de catalogar pedófilos e evitar que eles obtenham sucesso em abusar de crianças no país. É composta de agentes policiais e voluntários que recebem treinamento para posarem como crianças em *chats online* e conduzir a conversa com o pedófilo a um rumo onde ele possa ser incriminado. Desde 2003 até hoje, a organização conseguiu 622 condenações utilizando esta estratégia, e todas as conversas *online* estão disponíveis no site http://www.perverted-justice.com/, que é à base de dados mais usada nos estudos relacionados à aliciamento na *Internet*.

Deste modo, desenvolver técnicas para apoiar o monitoramento e classificação de conversas online de pedofilia em RSO para *smartphones*, irá proporcionar uma maior segu-

rança para as crianças na *Internet*, colaboração com o trabalho da polícia e pode proporcionar importantes resultados para estudos sociológicos e psicológicos.

1.2 Desafios

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.

Dados públicos de *chats* de pedofilia são escassos, e por consequência, há desbalanceamento entre o número de conversas regulares e de pedofilia. Vários algorítimos de aprendizagem de máquina são sensíveis a desbalanceamento nos dados, tornando o problema mais desafiador. Para o caso do algorítimo KNN por exemplo, classes possuindo amostras muito frequentes tendem a dominar a vizinhança de uma instância de teste apesar das medidas de distância aplicadas, levando a um desempenho baixo de classificação na classe minoritária (LIU; CHAWLA, 2011). O SVM não é recomendado nestas condições por cair em problemas com o conhecido *Small Sample Size problem* (SSS), que ocorre quando o número de amostras é menor do que a dimensionalidade do seu vetor de características (CORTES; VAPNIK, 1995). Além disso, dependendo da proporção do desbalanceamento, uma escolha ruim de hiperplano de separação poderá produzir resultados enganosos de bom desempenho, por influência da classe com maior quantidade de amostras. Machado (2009) afirma que o uso de métodos de amostragem para balanceamento de dados visam mudar a distribuição dos dados de treinamento, de modo a aumentar a acurácia de seus modelos. Isto é alcançado com a eliminação de casos da classe majoritária (*undersampling*) ou replicação de casos da classe minoritária (*oversampling*), então por estes motivos iremos realizar o balanceamento dos dados utilizando a abordagem de *undersampling*. Existe também o desafio de implementar estes modelos de aprendizagem de máquina em aplicativos de celular para que executem com baixo custo computacional e boa capacidade de classificação das conversas devido à sua capacidade inferior de processamento em relação aos computadores. Sistemas que obtem conhecimento de dados normalmente possuem etapas de processamento intenso porque precisam tratar e extrair

características de muitos dados para obter resultados confiáveis.

1.3 Problema

A maneira habitual de incriminar esses predadores sexuais é quando agentes policiais treinados posam como crianças em salas de *chat online*. No entanto, o número de predadores sexuais *online* sempre superam os de policiais e voluntários. Encontrar pessoas dispostas a colaborar com a polícia e treiná-las, requer muito esforço e tempo. Outras dificuldades enfrentadas estão relacionadas à quantidade de pedófilos investigados, que está restrita aos suspeitos conhecidos pela polícia, a abrangência das investigações que depende diretamente de denúncias feitas pela população e a dificuldade da própria família em identificar o processo de aliciamento visto que conversas realizadas por RSO em *smartphones* são mais discretas.

Apesar de o problema da detecção de conversas *online* de pedofilia já possuir solução, nenhuma delas é direcionada para RSO que executam em *smartphones*, por este motivo em nosso trabalho, pretendemos contribuir com um método para identificação destas conversas online de pedofilia utilizando a Entropia de Shannon e a Divergência de Jensen-Shannon, dois quantificadores de teoria da informação que tem por objetivo resumir o vocabulário que representa uma conversa em apenas seis características descritivas.

1.4 Objetivos

O objetivo geral deste trabalho é propor um método e demonstrar sua eficácia por meio de avaliações de qualidade de classificação e tempo de execução para extrair características de texto reduzindo sua dimensionalidade para identificar de conversas online de pedofilia utilizando Entropia e JSD. Os objetivos específicos que orientam este trabalho são:

- Definir uma representação dos *chats* que utiliza uma quantidade fixa de características independente do tamanho da base de dados utilizada.
- Elaborar uma técnica que transforme o vocabulário de um *chat* em características de informação.
- Definir uma metodologia que possui melhor custo-benefício de qualidade de classificação e tempo de execução em *hardware* de baixa capacidade de processamento.

1.5 Organização do Trabalho

Esta dissertação está organizada da seguinte forma: no capítulo 2 apresentamos os fundamentos teóricos necessários para o entendimento dos métodos adotados. No capítulo 3

fizemos uma síntese dos trabalhos relacionados, de autores que abordam o mesmo problema que estamos investigando e que utilizaram estes fundamentos teóricos. No capítulo 4 apresentamos a solução proposta pelo nosso trabalho. No capítulo 5 apresentamos os resultados dos experimentos feitos com o nosso método em comparação ao *baseline* e por fim no capítulo 6 fizemos uma conclusão geral do estudo realizado em nosso trabalho.

2 Referencial Teórico

Neste capítulo são apresentados os conceitos necessários para o entendimento e desenvolvimento do trabalho. O ferramental teórico é dividido em quatro seções. Na seção 2.1 será feita uma breve introdução ao Processamento de Linguagem Natural (*Natural Language Processing* – NLP) como ciência, na seção 2.2 é mostrada a técnica de *Lexical Chains* e sua estratégia de analisar as conexões de palavras em frases criando semântica, na seção 2.3 é mostrado os conceitos de análise de sentimento e como ele é aplicado para definir emoção expressa por usuários em RSO, na seção 2.4 é feito um resumo do *Bag of Words* e a ponderação de termos *Term Frequency – Inverse Document Frequency* explicando como ele realiza a extração de características de textos. Nas seções 2.5 e 2.6 é feito um resumo das duas técnicas de Teoria da Informação utilizadas neste trabalho: Entropia e JSD, e seu potencial para a extração de características dos *chats* para serem submetidos aos algorítimos de Aprendizagem de Máquina. As seções 2.7, 2.8, 2.9 e 2.10 apresentam um resumo das abordagens de aprendizagem de máquina utilizadas nos problemas de processamento de texto para detecção de *chats* de pedofilia. Na seção 2.11, descrevemos as métricas de classificação utilizadas em aprendizagem de máquina, e por fim na seção 2.12 encontram-se as considerações finais deste capítulo.

2.1 Processamento de Linguagem Natural

Segundo Manning, Schütze et al. (1999), a NLP é um campo da ciência da computação, inteligência artificial e linguística computacional interessado nas interações entre computadores e linguagens humanas naturais, e está relacionada com a área de interação humano-computador. Alguns desafios em NLP envolvem compreensão da linguagem natural, ou seja, permitir que computadores sejam capazes de extrair significado de uma linguagem humana como entrada, e também geração de linguagem natural. A NLP possui algumas tarefas de processamento do texto das quais utilizaremos em nosso trabalho apenas a *tokenização* e o *part-of-speech tagging*.

Pereira (2014) afirma que uma palavra é a menor partição possível de uma sentença no contexto de linguagem natural. Desta forma, dada uma sentença como entrada, a *tokenização* segmentará todo o seu texto em *tokens*(palavras). Este processo deve ser feito antes de qualquer análise sintática dos dados de entrada. Um exemplo de *tokenização* é ilustrado na figura 1.

Pereira (2014) também afirma que o *part-of-speech tagging* indica a classe gramatical de um *token* através dos rótulos (*tags*) definidos no trabalho de Santorini (1990) chamado

Figura 1 – Exemplo gráfico da *tokenização* de uma sentença – Adaptado de Manning, Raghavan e Schutze (2008).

Projeto *Penn Treebank*[1]. Isto é feito por meio da análise do comportamento sintático das palavras em uma sentença e do campo semântico em que a frase se insere. A figura 2 mostra o resultado do *part-of-speech tagging* aplicado aos *tokens* do exemplo ilustrado na figura 1, neste exemplo as *tags* são: PRP (pronome pessoal), NN (Substantivo), RB (Advérbio), VBZ (verbo na 3º pessoa do singular no presente), VBG (verbo no gerúndio ou particípio no presente).

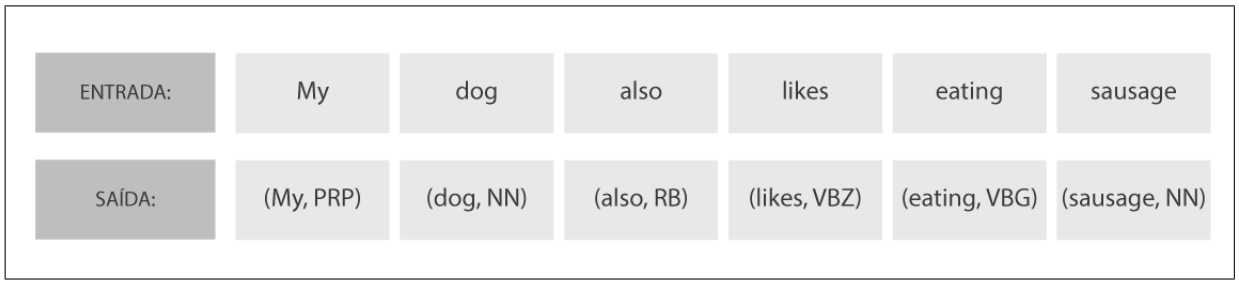


Figura 2 – Exemplo gráfico do *part-of-speech tagging* em um conjunto de *tokens* – Adaptado de Nivre et al. (2015).

Esta análise é feita na estrutura gramatical das sentenças, por exemplo, é feita uma detecção dos grupos de palavras que se combinam (como frases) e quais palavras são sujeito ou objeto de um verbo, isso faz com que o *part-of-speech tagging* seja capaz de fornecer a *tag* mais provável de uma palavra, mesmo se ela pertencer a mais de uma classe gramatical dependendo de onde ela está inserida em uma sentença.

Segundo Stubbs (2001), em estudos de linguagem as classes gramaticais das palavras estão divididas em dois grandes grupos: palavras funcionais e léxicas. No grupo das palavras funcionais estão classes como verbos auxiliares, pronomes, conjunções, preposições, determinantes e modais. As palavras deste grupo são utilizadas em estruturação de frases, raramente fazem parte de mais de uma classe gramatical e possuem significado semântico fraco, por isso a maioria destas palavras são removidas das bases de dados em problemas de processamento de texto e recebem o nome de *stopwords*. No grupo de palavras léxicas estão os substantivos, verbos principais, adjetivos, interjeições e advérbios, são as

[1] lista de tags e suas classes gramaticais correspondentes disponível em ⟨https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html⟩

palavras pertencentes a este grupo que fornecem o significado semântico das sentenças e comumente estão presentes em mais de uma classe gramatical. Desta forma, a tarefa de *part-of-speech tagging* é utilizada para filtrar as palavras funcionais e manter apenas as léxicas.

2.2 Lexical Chains

Segundo Barzilay e Elhadad (1999), *lexical chains* ou encadeamento léxico é uma técnica de encadeamento de palavras que relaciona frases através de substantivos relacionados com a sentença. No processamento de linguagem natural, muitos dos algoritmos de Lexalytics são baseados em *lexical chains*, Considere a seguinte sentença:

"I like beer. Miller just launched a new pilsner. But, because I'm a beer snob, I'm only going to drink pretentious Belgian ale."

Essas 3 frases são relacionadas através da beer->pilsner->ale. Mesmo que essas frases não sejam adjacentes um ao outro no texto, elas são lógicamente relacionadas entre si e, portanto, podem ser associadas entre si. Este é um conceito realmente importante - se os substantivos estão relacionados um com o outro, podemos encontrar essa cadeia conceitual (lexical) no conteúdo, mesmo quando essas frases são separadas por muitas outras frases não relacionadas. A "pontuação"de uma cadeia lexical está diretamente relacionada ao comprimento da cadeia e às relações entre os substantivos em cadeia (mesma palavra, antônimo, sinônimo, hiper). Dentro de implementações de Lexalytics, a extração temática usa *lexical chains* para pontuação de tema. A síntese usa *lexical chains* para escolher as frases mais representativas. A avaliação do sentimento da entidade usa *lexical chains* para associar o sentimento com frases com as próprias entidades.

2.3 Análise de Sentimento

Segundo Pang, Lee et al. (2008), a análise de sentimentos atua no tratamento computacional da opinião, do sentimento e da subjetividade no texto. Surgiu como resposta direta ao aumento do interesse em novos sistemas Que lidam diretamente com as opiniões sendo objeto de primeira classe. Tem como parte importante a coleta de informações para descobrir o que as outras pessoas pensam sobre assuntos. Isto é possível devido a crescente disponibilidade e popularidade de recursos ricos em opiniões, como sites de revisão on-line e blogs pessoais, já que as pessoas agora podem usar tecnologias de informação para buscar e entender as opiniões dos outros.

Desta forma, a análise de sentimento ou mineração de opinião está relacionada ao uso de processamento de linguagem natural, análise de texto e linguística computacional para identificar e extrair informações subjetivas em materiais de origem. É amplamente

utilizada em mídias sociais para uma variedade de aplicações, que vão desde *marketing* a serviço ao cliente. De um modo geral, a análise de sentimento tem como objetivo determinar a atitude de um autor em relação a algum tema, ou a polaridade contextual geral de um documento. A atitude pode ser o seu julgamento ou avaliação, seu estado emocional ou a comunicação emocional pretendida, que é o efeito emocional que o autor deseja despertar no leitor.

2.4 Bag of Words e Term Frequency – Inverse Document Frequency

Segundo Salton e Michael (1983), o modelo *Bag of Words* (BoW) no âmbito de recuperação da informação e de linguagem natural, é a representação desordenada de um documento através da frequência das palavras (termos) de seu dicionário, com o objetivo de transformar um *corpus* em dados discretos. No entanto, para alguns problemas é necessário conhecer a relação termo x documento de um *corpus*, e para isso aplica-se a ponderação *Term Frequency – Inverse Document Frequency* (TF-IDF). A figura 3 mostra a relação entre a frequencia da palavra e seu valor.

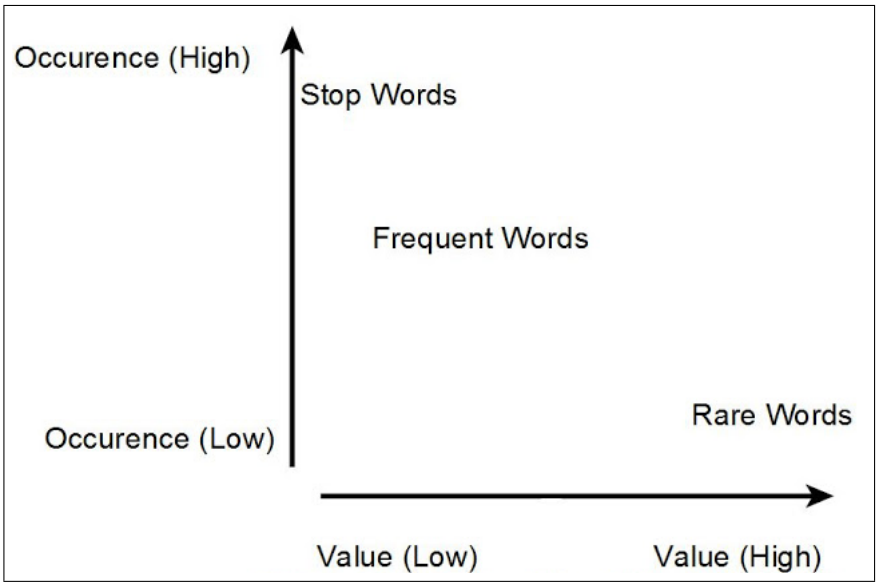


Figura 3 – Relação entre a ocorrência de um termo e sua relevância – Adaptado de Robertson (2004).

O TF-IDF conta o número de ocorrências de cada palavra, após a normalização adequada, esta contagem de frequência de termos é comparada a uma contagem inversa de frequência de documentos, que mede o número de ocorrências de uma palavra em todo o corpus (geralmente em uma escala logarítmica e, novamente, adequadamente normalizada). Na prática, o cálculo do TF-IDF é uma multiplicação simples do TF, mostrado na

equação 2.1:

$$tf(t,d) = f_{t,d}, \tag{2.1}$$

onde t é o termo, d é o documento onde t ocorreu e f é a frequência absoluta de t em d. O cálculo do IDF é mostrado na equação 2.2:

$$idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}, \tag{2.2}$$

sendo $|D|$ a quantidade total de documentos e $|d \in D : t \in d|$ o número total de documentos onde t ocorre, ou seja, o idf é uma escala logarítmica de fração inversa dos documentos que contém o termo. O resultado final é uma matriz termo-por-documento cujas colunas contêm os valores TF-IDF para cada um dos documentos no corpus. Por consequência, os documentos de comprimento arbitrário se tornam listas de números fixos.

2.5 Entropia de Shannon

Segundo Rosso, Ospina e Frery (2016), A Entropia é uma quantidade básica com múltiplas interpretações específicas do ponto de vista de determinadas ciências: por exemplo, tem sido associada com desordem de sistemas, volume em espaço-estado e falta de informação. Ao lidar com o conteúdo da informação, a entropia de Shannon é muitas vezes considerada como o fundacional e mais natural.

De acordo com Rosso, Craig e Moscato (2009), a Entropia normalizada é útil para a análise da dispersão das palavras de textos sobre um espectro de palavras possíveis, este espectro pode ser ajustado conforme a necessidade do estudo a ser feito, por exemplo podem ser todas as palavras de um idioma, bem como podem ser apenas o grupo de palavras léxicas de um idioma. No ponto de vista da física, esta medida é interpretada como a homogeneidade entre os textos analisados. A equação é normalizada, então valores de entropia baixos (próximos a zero) indicam repetição ou redundância de palavras nos textos, ao passo que valores altos (próximos a um) indicam um vocabulário mais rico. A equação 2.3 se refere à Entropia normalizada:

$$H_{\mathsf{S}}[P] = S[P]/S_{max} = (-\sum_{j=1}^{N} p_j \ \log(p_j)) \ / \ \log(N), \tag{2.3}$$

onde $S[P]$ é a Entropia da distribuição de probabilidade de palavras P tal que $\{p_j;\ j = 1, \ldots, N\}$, sendo p_j é a probabilidade da palavra j e N é a quantidade de palavras únicas em um documento, também chamadas de vocabulário, e $S_{max} = \log(N)$ é a Entropia de uma distribuição de probabilidades uniforme, este é o termo que normaliza a equação.

2.6 Divergência de Jensen-Shannon

A JSD mede a distância, ou similaridade entre duas distribuições de probabilidade em função da entropia destas distribuições. Ainda no trabalho de Rosso, Craig e Moscato (2009), esta medida que revela o grau de variação de uma distribuição de probabilidade, que chamaremos de histograma particular, em relação a um referencial estabelecido, que é outra distribuição. Se utilizarmos como referencial uma distribuição de probabilidade média de palavras, que chamaremos de histograma de referência, composto pela probabilidade média das palavras do vocabulário obtido da maioria dos documentos de texto utilizados em um estudo, a JSD informa o valor de similaridade entre um texto em particular, em relação a este histograma de referência, indicando o quanto este texto está proximo do padrão representado pelo histograma médio. A equação 2.4 mostra o cálculo da JSD:

$$J_S[P_1, P_2] = S[(P_1 + P_2)/2] - S[P_1]/2 - S[P_2]/2, \tag{2.4}$$

onde $J_S[P_1, P_2]$ é a divergência entre um histograma particular P_1 e um histograma de referência P_2, $S[(P_1 + P_2)/2]$ é a entropia da média dos histogramas, $S[P_1]$ é a entropia do histograma particular e $S[P_2]$ é a entropia do histograma de referência. Se as entropias estiverem normalizadas, valores de JSD próximos a zero indicam similaridade entre os histogramas e valores próximos a um indicam divergência entre eles.

2.7 Support Vector Machines

Desenvolvido por Cortes e Vapnik (1995), o SVM é um dos mais populares algoritmos de classificação. Este método transforma os vetores de características em um espaço de dimensões maiores, em que as classes podem ser separadas linearmente por hiperplanos. A figura 4. mostra o processo para chegar ao hiperplano ótimo de separação, que maximiza a margem M, para um problema de duas classes.

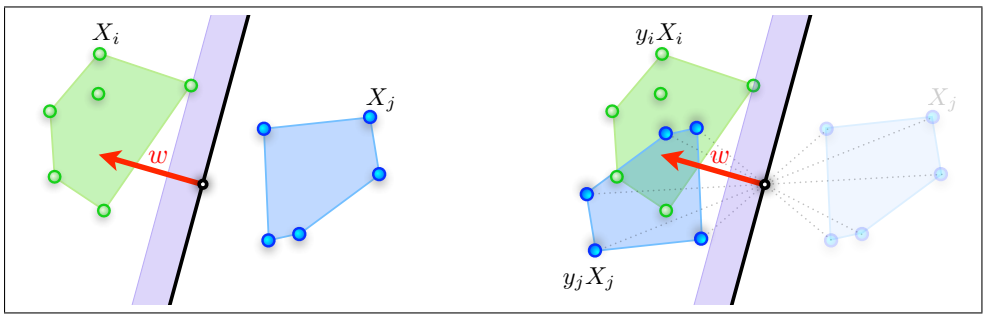


Figura 4 – Processo de obtenção do hiperplano ótimo para um problema de duas classes – Adaptado de Jaggi (2014).

As classes que as amostras desconhecidas x pertencem, são determinadas pela aplicação da equação correspondente a "separação" de classificação. Se o resultado é "positivo", a

amostra é classificada como pertencente a classe x_1 e se o resultado é "negativo", pertence a x_2. Para o problema de duas classes ilustrado na figura 4, o plano de separação é dado pela equação 2.5:

$$y = wx + b, \tag{2.5}$$

onde y é a classe da instância x, w é um *support vector* do hiperplano e b é uma constante de *offset*.

2.8 Árvores de Decisão

De acordo com Kamiński, Jakubczyk e Szufel (2017) árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Modelos de árvore são feitos utilizando a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema, por exemplo, a tarefa e construir a árvore consiste em etapas como escolher características da amostra nos nós, para isso utiliza como critério o valor produzido pelo cálculo de ganho de Informação, que é um cálculo baseado na entropia, então cada etapa se torna uma subtarefa e assim constrói-se otimizadamente o modelo. As árvores de decisão estão entre os mais populares algoritmos de inferência e tem sido aplicado em várias áreas como, por exemplo, diagnóstico médico e risco de crédito, e é possível entender o modelo produzido pelo algorítimo deles pode-se extrair regras do tipo "se-então" que são facilmente compreendidas. A capacidade de discriminação de uma árvore vem da divisão do espaço definido pelos atributos em sub-espaços e a cada sub-espaço é associada uma classe.

2.9 Naive Bayes

Lewis (1998) afirma que o Naive Bayes é um classificador probabilístico simples baseado na aplicação de teorema de Bayes. Possui premissas de independência entre os atributos, isto é, assume que a presença (ou ausência) de uma característica particular de uma classe não está relacionado com a presença (ou ausência) de qualquer outro recurso. Por exemplo, uma fruta pode ser considerada uma maçã se for vermelha, redonda, e tem aproximadamente 10 cm de diâmetro. Um classificador Naive Bayes considera que cada característica contribui de forma independente para a probabilidade de que esta fruta é uma maçã, independentemente de quaisquer possíveis correlações entre as características de cor, circularidade e diâmetro. Abstratamente, o modelo de probabilidade para um classificador é um modelo condicional sobre uma variável de classe C dependente com um pequeno número de resultados ou classes, representada por uma determinada quantidade

de características que vão de F_1 à F_n, mostrado na equação 2.6:

$$p(C|F_1, \ldots, F_n), \tag{2.6}$$

o problema é que, se o número de características é grande ou quando uma característica pode assumir uma grande variedade de valores, então o modelo se torna inviável. Para tornar o modelo tratável, utiliza-se o teorema de Bayes, descrito na equação 2.7:

$$p(C|F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n)}{p(F_1, \ldots, F_n)}, \tag{2.7}$$

assim o denominador no teorema não tem dependência em C, tratando-se de uma constante, por este motivo, geralmente se considera apenas o numerador do teorema.

2.10 K-Nearest Neighbor

De acordo com Zhang e Zhou (2007), *K-Nearest Neighbor* (KNN) é um algoritmo simples que armazena todos os casos disponíveis e classifica novos casos com base em uma medida de similaridade, por exemplo, funções de distância. As funções de distância mais comumente utilizadas são a Euclidiana, Manhathan e Minkowski, descritas respectivamente nas equações 2.8, 2.9 e 2.10 para $K > 1$ e variáveis numéricas:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}, \tag{2.8}$$

$$\sum_{i=1}^{k}|x_i - y_i|, \tag{2.9}$$

$$(\sum_{i=1}^{k}(|x_i - yi|)^q)^{\frac{1}{q}}, \tag{2.10}$$

se $K = 1$, então o caso é simplesmente atribuído à classe de seu vizinho mais próximo. O KNN foi usado na estimativa e padrão de reconhecimento estatístico já no início da década de 1970, como uma técnica não-paramétrica. Um caso é classificado pelo voto da maioria de seus vizinhos, assim, ele é atribuído à classe mais comum entre os seus K vizinhos mais próximos medidos por uma função de distância.

2.11 Métricas de Classificação

2.11.1 Medida F

Na análise estatística da classificação, Powers (2011) afirma que a medida F é uma métrica para validação de resultados de classificação. Considera tanto a precisão como a revocação

dos experimentos para calcular a pontuação. A precisão é o número de verdadeiros positivos (tp) dividido por todos os resultados positivos (tp e fp), como mostra a equação 2.11:

$$\text{Precisão} = \frac{tp}{tp + fp}, \tag{2.11}$$

a revocação é o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos com os falsos negativos (fn), como mostra a equação 2.12:

$$\text{Revocação} = \frac{tp}{tp + fn}, \tag{2.12}$$

na equação 2.13 é descrita a fórmula para encontrar a medida F para qualquer β desejado, que regula a relevância dada à precisão ou revocação.

$$MedidaF_\beta = (1 + \beta^2) \cdot \frac{\text{precisão} \cdot \text{revocação}}{\beta^2 \cdot \text{precisão} + \text{revocação}}, \tag{2.13}$$

A medida F1 é um caso especial da equação geral de medida F e pode ser interpretada como uma média harmônica da precisão e revocação, descrita na equação 2.14:

$$F1 = 2 \cdot \frac{\text{precisão} \cdot \text{revocação}}{\text{precisão} + \text{revocação}}, \tag{2.14}$$

onde atinge seu melhor valor em 1 e pior em 0.

A medida $F_{0.5}$ é outra métrica de validação utilizada para dar maior ênfase à Precisão ao invés da Revocação. É calculada através da equação geral de F_β mostrada na equação 2.13, utilizada para customizar quanta relevância desejamos dar à precisão ou revocação.

2.12 Considerações Finais

Neste capítulo foi apresentado um resumo dos conceitos teóricos utilizados em problemas de processamento de texto e técnicas de validação dos resultados gerados por algorítimos de classificação, necessários para entender o desenvolvimento deste trabalho, mostrando suas origens, utilidades e escopos de aplicação. O objetivo foi mostrar algumas das técnicas de pré-processamento, extração de características e classificação para a criação de modelos que extraem conhecimento dos dados. As escolhas feitas para definir o nosso método baseiam-se nestes conceitos. No próximo capítulo veremos a aplicação destes conceitos na prática descritos em trabalhos realizados por outros autores que abordam o mesmo problema que estudamos, para compreendermos suas metodologias, e obtermos direcionamento na construção da nossa solução com base em seus resultados e experiências.

3 Trabalhos Relacionados

Neste capítulo apresentamos uma seleção de trabalhos de autores que abordam o problema de identificação de *chats* de pedofilia na *Internet*, para estudar suas metodologias, entender suas escolhas e obter embasamento para a construção da nossa solução. Separamos em seções diferentes os trabalhos que fazem pré-processamento dos que não o fazem, e em cada trabalho identificamos as etapas de escolha da base de dados, extração de características, método de classificação e métricas utilizadas para avaliar os resultados para fazer um resumo das estratégias adotadas. No final do capítulo organizamos todas estas informações dos trabalhos em uma tabela e descrevemos as escolhas para a nossa abordagem.

3.1 Trabalhos com Pré-Processamento

Pendar (2007) foi o primeiro autor a abordar o tema de detecção de pedofilia em *chats*. Utilizou apenas a base de dados da organização *Perverted Justice*, que é uma base onde os pedófilos são reais mas as vítimas são policiais ou voluntários disfarçados de crianças, apesar desta limitação na base de dados ainda é possível realizar um estudo válido, onde para cada pedófilo encontrado em um *chat*, coloca-se um rótulo de "conversa suspeita". O autor ressalta a importância de se ter bases de dados onde os pedófilos e vítimas são reais, e bases com conversas erotizadas entre dois adultos para um modelo de aprendizagem de máquina com maior capacidade de generalização, porém afirmou ter dificuldades em encontrar estes tipos de dados. Como estratégia de pré-processamento, o autor separou em arquivos diferentes as linhas escritas pelo pedófilo e pela vítima, isto é, cada *chat* é segmentado em dois. Foram utilizados 701 *chats* no trabalho e após o pré-processamento, a base ficou com 1.402 *chats*. Além disso, criou sua própria lista de *stopwords*, composta das 79 palavras mais frequentes da coleção de *chats* e aplicou nos arquivos. O autor chama a atenção para problemas como gírias e palavras escritas incorretamente e também para o cuidado em tentar tratá-los para não introduzir mais erros nos dados. Para extrair características destes *chats*, o autor testou unigramas, bigramas e trigramas associados aos *chats* com as linhas escritas pelas vítimas e predadores individualmente e ponderando os termos utilizando TF-IDF. Na classificação o autor utilizou o algorítimo KNN e seu melhor resultado foi 94% de medida F1 utilizando trigramas como características, validando os resultados com *cross-validation* de *two-folds*.

Cheong et al. (2015) criou um método que é capaz de identificar pedófilos ou *rule-breakers* em um jogo *online* infantil. Para isso, utilizou como base de dados textos de *chats* e fóruns do jogo infantil *Movie Star Planet* minerados durante 15 minutos de um servidor no Reino Unido, resultando em 8.707 autores não pedófilos e 62.704 linhas de

texto, mais 59 autores pedófilos e 40.413 linhas de texto. Os próprios moderadores do jogo rotularam os conteúdos de pedofilia de acordo com os seguintes critérios:

- Usuários que iniciam uma conversa de cunho sexual no jogo.
- Usuários que aceitam este tipo de conversa no jogo e que respondem de forma similar.
- Usuários que tentam obter acesso a outro usuário, por exemplo, endereço, telefone ou contato em redes sociais.

Como estratégia de pré-processamento, as linhas de *chat* foram agrupadas em vítimas e pedófilos por autor, sendo que para este último caso foram mantidas apenas as linhas onde há claramente um discurso predatório, afirmando que deseja construir um modelo especializado em detectar intenções de aliciamento. Para a extração de características, utilizou TF-IDF para obter informações léxicas dos *chats*, análise de sentimento para detectar o comportamento de *rule-breaker* se houver e selecionou manualmente os trechos finais dos *chats* imediatamente antes de serem rotulados como pedofilia pelos moderadores do jogo. Para a classificação utilizou o *Naive Bayes* e obteve 53% de medida F1 utilizando *cross-validation* de *two-folds*, contribuindo com uma funcionalidade que apoia o trabalho dos moderadores do jogo em detectar pedófilos ou *rule-breakers* para um ambiente de jogo mais seguro. O autor deixa em aberto como trabalho futuro a elaboração de um método para detectar estes usuários em tempo real.

Rosso, Craig e Moscato (2009) utiliza técnicas de teoria da informação para identificar obras de *Shakespeare* dentre outros autores renascentistas ingleses, no entanto, seu trabalho não se encaixa no padrão de abordagens de aprendizagem de máquina. Utiliza como base de dados uma coleção de 185 obras literárias eruditas renascentistas inglesas digitalizadas de uma fonte chamada Literatura Online, onde 30 destes textos foram escritos por Shakespeare. Como estratégia de pré-processamento, realiza manualmente a identificação das classes gramaticais das palavras como funcionais ou léxicas, remoção pontuações, *tokenização* e *stemming*. Após as modificações, as palavras são agrupadas em 3 conjuntos distintos, o primeiro contendo apenas as palavras funcionais, o segundo apenas as léxicas, e o terceiro é uma união dos dois conjuntos anteriores. Realiza uma espécie de extração de características que consiste na obtenção dos quantificadores de teoria da informação Entropia, JSD e Complexidade Estatística dos textos. A Entropia de cada texto é calculada do seu histograma de frequência relativa de palavras, a JSD é obtida através do cálculo de similaridade entre o histograma do texto em relação a um histograma médio obtido de todos os textos juntos, e a Complexidade Estatística é o produto entre a Entropia e a JSD. Não há um processo de classificação, e nem métricas de comparação como em problemas de aprendizagem de máquina. O método utiliza os quantificadores dos textos

para criar um plano de agrupamento Entropia x Complexidade Estatística, onde estes valores são capazes de separar as obras de *Shakespeare* dos outros autores. O resultado do estudo mostra que *Shakespeare* tinha o estilo de escrita que mais obedecia os padrões renascentistas ingleses, como se o método como um todo fosse capaz de simular um crítico literário.

Parapar, Losada e Barreiro (2012) incorporam *Linguistic Inquiry and Word Count* (LIWC) no processo de extração de características para a identificação de *chats* de pedofilia. Utilizam a base de dados PAN 2012 nos experimentos e como estratégia de pré-processamento, criaram arquivos diferentes para cada autor, resultando em 97689 *chats*. 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 *chats*, e TF-IDF para extrair as características léxicas do texto. Utilizou o classificador SVM devido a alta dimensionalidade do vetor de características e obteve 83% de medida F1 utilizando *cross-validation* de *four-folds*. O autor conclui que características baseadas em *chats* representam a atividade dos participantes da conversa, o número de assuntos mencionados pelo participante, o percentual de conversas iniciadas por ele e outras características da conversa. Um aspecto interessante da pesquisa foi que o uso do LWIC não melhorou os resultados.

3.2 Trabalhos sem Pré-Processamento

Bogdanova, Rosso e Solorio (2012) utilizam análise de sentimento para verificar se uma conversa de *chat* é predatória ou não. Utilizam três bases de dados, a primeira é da organização *Perverted Justice* para formar o conjunto de dados positivos, a segunda são os *logs* de *chats* de cunho sexual entre adultos, conhecido como *cybersex*, que possui aproximadamente 30 *chats*, e a última é o *corpus* de *chats* chamado *Naval Postgraduate School* (NPS) de acesso pago, para formar o conjunto de dados negativos. Extrai características utilizando os valores resultantes do cálculo da similaridade de Leacock e Chodorow que estima a similaridade semântica de termos nos *chats*, para encontrar os trechos de conversa onde o discurso se torna fixo, ou insistente em algum assunto. os autores acreditam que os pedófilos se comportam de maneira distinta, isto é, são emocionalmente instáveis e sofrem de problemas psicológicos. Desta forma, tentam detectar o texto predatório utilizando as características de marcadores emocionais, buscando no texto palavras que expressem sentimentos como alegria, tristeza, raiva, surpresa, aversão e medo. Palavras positivas e negativas, *emoticons*, e frases imperativas também foram consideradas. Outras características incluem o uso das palavras de classes gramaticais específicas para detectar o nível de neuroticismo do autor no *chat* (por exemplo, porcentagens de pronomes pessoais e reflexivos e verbos modais de imposição). Estes cálculos consecutivos de similaridade semântica criam cadeias de termos semanticamente relacionados, chamados *Lexical*

Chains, e os autores realizam as medições destas *Lexical Chains* para identificar os *chats* de pedofilia, sem algorítimos convencionais de aprendizagem de máquina.

Villatoro-Tello et al. (2012) foi o vencedor da competição PAN 2012 que consistia na elaboração do método que obtivesse a melhor métrica de classificação para *chats* de pedofilia. O método possui uma arquitetura de duas camadas, a primeira identifica o *chat* suspeito, e a segunda identifica o pedófilo na conversa. Utilizou apenas a base de dados do próprio evento. Não realizou nenhum tipo de pré-processamento, no entanto, fez uma filtragem onde *chats* com apenas um autor, com menos de 6 linhas escritas por cada autor e *chats* com longas cadeias de caracteres não-ascii são desconsiderados dos experimentos. Ele afirma que estes *chats* possuem informação insuficiente para a realização dos estudos. Desta forma, considerou 5790 *chats* regulares e 798 de pedofilia nos experimentos. Para a extração de características foi utilizado o TF-IDF, e o processamento da base de dados resultou em 16.709 termos no vetor de características, e obteve 95% de medida F1 utilizando o classificador SVM com validação por *cross-validation* de *two-folds*.

Morris e Hirst (2012) utilizam características léxicas e comportamentais para identificar *chats* de pedofilia. Utilizaram a base de dados PAN 2012 em seus estudos. Utilizaram o TF-IDF para representar as características léxicas dos *chats*, e também foram utilizadas características comportamentais provenientes de informações que podem ser extraídas das conversas, como o número de mensagens enviadas por um autor e o número total de conversas que o autor participou. Para identificar predadores os autores utilizaram um classificador SVM com *kernel* gaussiano e dois filtros para distinguir predadores de vítimas, uma vez que grande parte dos falsos positivos foram vítimas. Usando apenas recursos léxicos os autores conseguiram obter uma pontuação de 77% para a medida $F1$. Uma descoberta foi que as características comportamentais não melhoraram os resultados quando usadas com as características léxicas, mas ao ser usada sozinha resultou em uma classificação de 56% para a medida $F1$, todos os resultados foram validados com *cross-validation* de *two-folds*.

Peersman et al. (2012) apresentam 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, para a classificação de *chats* de pedofilia. Utilizam a base de dados PAN 2012 em seu estudo. Em sua abordagem utilizam o TF-IDF para extrair características léxicas do *chat* completo, e das linhas de cada autor separadamente para permitir a previsão em três etapas. Os autores usam dois classificadores SVM, um para detectar uma linha de *chat* predatória, e outra para classificar um participante da conversa como um pedófilo ou não pedófilo. Os resultados destes dois classificadores foram combinados para nivelar o resultado final balanceando a precisão e revocação de cada classificador, obtendo 90% de medida F1 com *cross-validation* de *two-folds*.

3.3 Síntese dos trabalhos relacionados

Os trabalhos atuais de detecção de pedófilos incorporam métodos automáticos de classificação, porém, nem todos realizam o pré-processamento dos dados. Os que não realizam, compensam a exclusão dessa etapa com alguma outra estratégia no momento da classificação. A tabela 1 resume os aspectos principais a serem considerados nos trabalhos relacionados neste capítulo apresentando a base de dados utilizada, a estratégia de pré-processamento, a extração de características, o método de classificação, e o resultado de medida F1 obtido.

Autor	Database	Pre-Process	Extração Carac.	Método de Class.	F1
(PENDAR, 2007)	Perverted Justice	Chats por Autor e filtro de Stopwords	N-Gram / TF-IDF	KNN	94%
(BOGDANOVA; ROSSO; SOLORIO, 2012)	Perverted Justice, cybersex, NPS	Nenhum	Similaridade de Leacock and Chodorow	Lexical Chains	N/A
(VILLATORO-TELLO et al., 2012)	PAN 2012	Nenhum	TF-IDF	SVM	95%
(MORRIS; HIRST, 2012)	PAN 2012	Nenhum	TF-IDF	SVM	83%
(PEERSMAN et al., 2012)	PAN 2012	Nenhum	TF-IDF	SVM	90%
(CHEONG et al., 2015)	logs Movie Star Planet	Seleção de texto por tempo	TF-IDF	Naive Bayes	57%
(ROSSO; CRAIG; MOSCATO, 2009)	Litratura Online	Palavras Funcionais e Léxicas	Entr./Diver./Comp. Estat.	N/A	N/A
(PARAPAR; LOSADA; BARREIRO, 2012)	PAN 2012	União de chats por autor	LIWC e TF-IDF	SVM	83%

Tabela 1 – Resumo dos trabalhos relacionados.

Nenhum dos autores realizou uma análise de complexidade de seus métodos, reforçando a nossa impressão de que não há uma preocupação com detecção de conversas de aliciamento em ambientes de *hardware* limitado, por este motivo relacionamos na tabela 2 o custo computacional dos trabalhos relacionados listados neste capítulo para visualizarmos a eficiência de cada abordagem, separando os custos de treino e teste, tomamos essa decisão pois o teste é a etapa mais relevante a ser analisada em nosso problema visto que o treino pode ser realizado *offline*, mas é importante que o modelo treinado possua boa eficiência para executar em *hardware* com pouca capacidade de processamento como os *smartphones*, que é um requisito fundamental para este trabalho.

Autor	Pré-Process	Extração Carac.	Treino do Class.	Teste do Class.	F1
(PENDAR, 2007)	$O(t)$	$O(tv)$	Nenhum	$O(tk + td)$	94%
(BOGDANOVA; ROSSO; SOLORIO, 2012)	Nenhum	O(vt)	Nenum	$O(tv^2)$	N/A
(VILLATORO-TELLO et al., 2012)	Nenhum	$O(tv)$	$O(t^2v)$	$O(tv)$	95%
(MORRIS; HIRST, 2012)	Nenhum	$O(tv)$	$O(t^2v)$	$O(tv)$	83%
(PEERSMAN et al., 2012)	Nenhum	$O(tv)$	$O(t^2v)$	$O(tv)$	90%
(CHEONG et al., 2015)	$O(v)$	$O(tv)$	O(tvc)	O(tvc)	57%
(ROSSO; CRAIG; MOSCATO, 2009)	$O(tv)$	Nenhum	O(tv+t+v)	N/A	N/A
(PARAPAR; LOSADA; BARREIRO, 2012)	$O(t)$	$O(tv)$	$O(t^2v)$	$O(tv)$	83%

Tabela 2 – Análise de complexidade dos trabalhos relacionados.

Analisando as estratégias adotadas pelos autores, decidimos optar pela base de dados PAN 2012 que é a mais utilizada segundo a tabela 1 e possui *logs* de conversas online de pedofilia e não-pedofilia, que chamaremos de regular, necesários para realizar os experimentos, faremos o balanceamento da base com atécnica de *undersampling* para melhorar

os resultados de classificação como é afirmado por Machado (2009), realizaremos o pré-processamento das conversas com o objetivo de filtrar as palavras funcionais para impedir que elas causem interferência na entropia e morfologia dos histogramas de palavras que representam as conversas, sendo assim manteremos apenas as léxicas, da mesma forma como é feita em Rosso, Craig e Moscato (2009), mas utilizaremos um *parser* gramatical para automatizar esta tarefa, realizaremos a extração de características utilizando Entropia e JSD por ainda não terem sido utilizadas em problemas de detecção de conversas de pedofilia e validar a hipótese de que o uso de suas características pelo classificador aumenta a eficiência da tarefa de classificação, e por fim utilizaremos nos experimentos os classificadores SVM, Árvore de Decisão, Naive Bayes e KNN para analisar os resultados e decidir qual o classificador que possui o melhor custo-benefício de qualidade de classificação e tempo de execução. Utilizaremos como *baseline* o autor Villatoro-Tello et al. (2012), por ter elaborado o modelo de detecção de conversas online de pedofilia que venceu a competição PAN 2012 obtendo o melhor valor de medida $F_{0.5}$ e o melhor valor de $F1$ da nossa seleção de trabalhos relacionados, sendo assim, utilizaremos estas duas métricas para avaliar os resultados nos experimentos. Não relacionamos os valores de $F_{0.5}$ dos trabalhos pois nem todos os autores apresentam seus resultados de precisão e revocação.

3.4 Considerações Finais

Neste capítulo fizemos um resumo das estratégias adotadas por alguns autores que abordam o problema da identificação de conversas online de pedofilia, percebemos que não há uma preocupação de que estes modelos executem em *smartphones* então direcionamos nossas escolhas para que seja possível cumprir o nosso objetivo descrito no capítulo 1 observando as dificuldades e direcionamentos descritos nos trabalhos dos autores que consideramos em nosso estudo. Após relacionar as técnicas e resultados obtidos pelos autores, realizamos uma análise de complexidade dos métodos para obtermos uma estimativa de desempenho em relação eficiência da execução desses métodos e por fim decidimos eliminar as palavras funcionais como estratégia de pré-processamento, utilizar os quantificadores de teoria da informação para extrair características de texto e diminuir a dimensionalidade delas e testar todos os algorítimos de classificação citados para analisar seu comportamento assintótico e qualidade de classificação quando fornecemos as características de informação para que eles façam as predições das instâncias, também escolhemos como *baseline* o método do autor Villatoro-Tello et al. (2012) por possuir as melhores métricas de classificação e por utilizar o BoW que é o estado da arte para características de texto, iremos considerar os valores de medida $F1$ e $F_{0.5}$ como métricas de qualidade de classificação para analisar melhor a influência da oscilação da precisão e revocação nos métodos. Após realizar todas estas considerações e escolhas de métodos, iremos apresentar a metodologia

proposta por este trabalho no capítulo 4 a seguir.

4 Concepção da Arquitetura e Metodologia

Neste capítulo é descrita 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 5 e possui as seguintes etapas: Pré-Processamento, Extração de Características e Treino do Modelo, detalhadas a seguir.

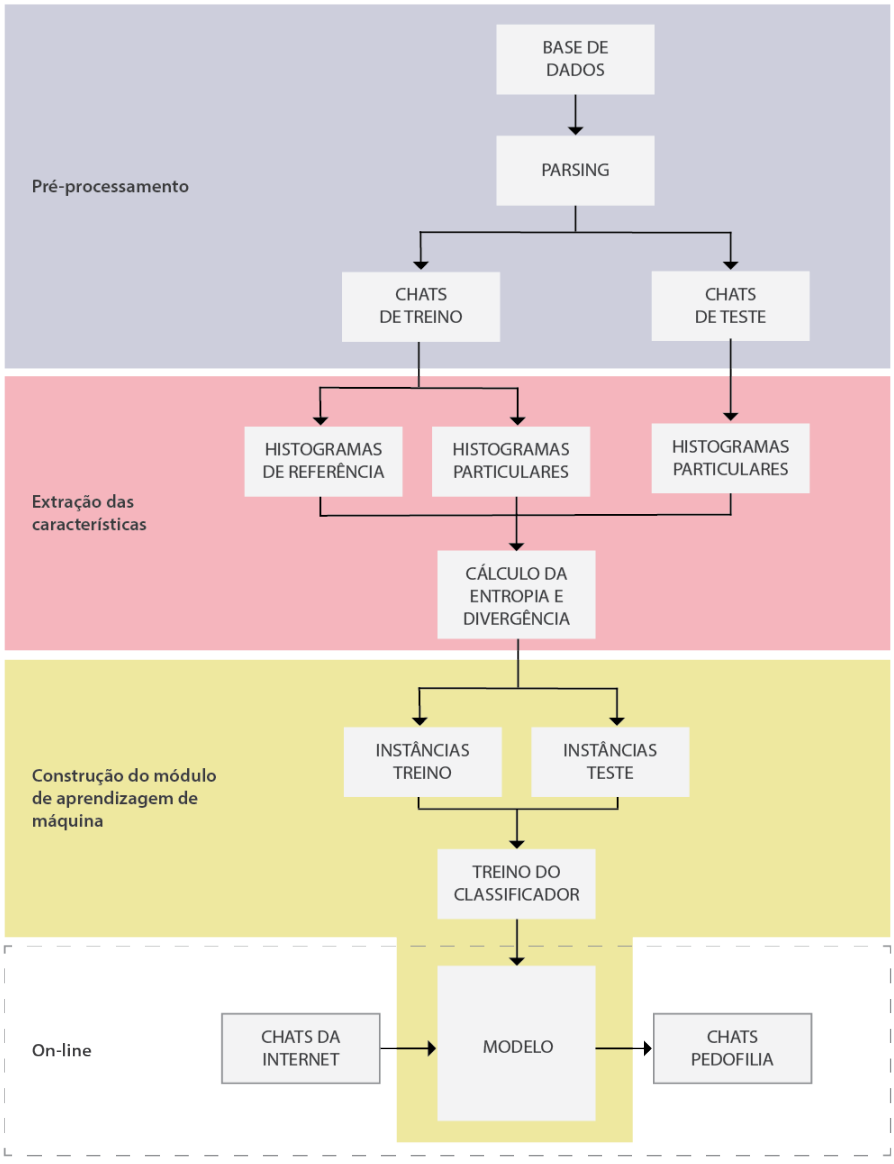


Figura 5 – Arquitetura da solução proposta.

4.1 Pré-Processamento

Utilizamos a base de dados do evento científico *International Competition on Plagiarism Detection* (PAN), do ano de 2012[1], que abordou o tema de aliciamento em *chats online*. É utilizada pelo *baseline* Villatoro-Tello et al. (2012), por Morris e Hirst (2012), Peersman et al. (2012) e Parapar, Losada e Barreiro (2012). Também utilizaremos esta base em nosso trabalho pois contém *chats* de pedofilia da única fonte disponível e para estabelecer uma experimentação justa no capítulo 5.

De forma semelhante a Rosso, Craig e Moscato (2009), o pré-processamento utilizado neste trabalho consiste no mapeamento das palavras em dois grandes grupos gramaticais: palavras funcionais e léxicas. O grupo de palavras funcionais é composto de palavras que possuem significado semântico fraco, útil apenas para estruturação das frases; as classes gramaticais que compõem esse grupo são verbos auxiliares, pronomes, conjunções, preposições, determinantes e modais. O grupo de palavras 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 *parser* para descobrir a classe gramatical de cada palavra e em seguida verificamos se ela se encaixa no grupo de funcionais ou léxicas. O passo seguinte é a filtragem das palavras funcionais, consideradas *stopwords*. Por fim, utilizaremos apenas as palavras léxicas dos *chats* na etapa de Extração de Características descrita na seção 4.2 a seguir.

4.2 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 particular. Todos os histogramas são normalizados de maneira que a soma de todas as probabilidades é sempre igual a um (distribuição de probabilidades discreta).

O histograma de referência é uma distribuição de probabilidade média de palavras calculado a partir dos *chats* da partição de treino, considerando o seu vocabulário (palavras únicas de um *corpus*) composto apenas das palavras léxicas devido ao pré-processamento realizado. Na base de dados PAN 2012 existem três tipos de entidades que são: pedófilo, vítima e regular (nem vítima e nem pedófilo), sendo assim teremos um histograma de

[1] Base de dados PAN 2012. Disponível em: <http://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-12/pan12-data/pan12-sexual-predator-identification-training-corpus-2012-05-01.zip>.

referência para cada um. Para obter estes histogramas, separamos as linhas de conversa escritas por cada entidade e aplicamos as equações 6, 7 e 8 de Rosso, Craig e Moscato (2009) contabilizando a frequência das palavras onde os valores associados a elas nos histogramas representam a sua probabilidade de aparecer em um texto daquele tipo de entidade, ou seja, os histogramas indicam o padrão de uso de palavras pelos tipos de entidades. Um exemplo de histograma de referência em pequenas dimensões é mostrado na figura 6.

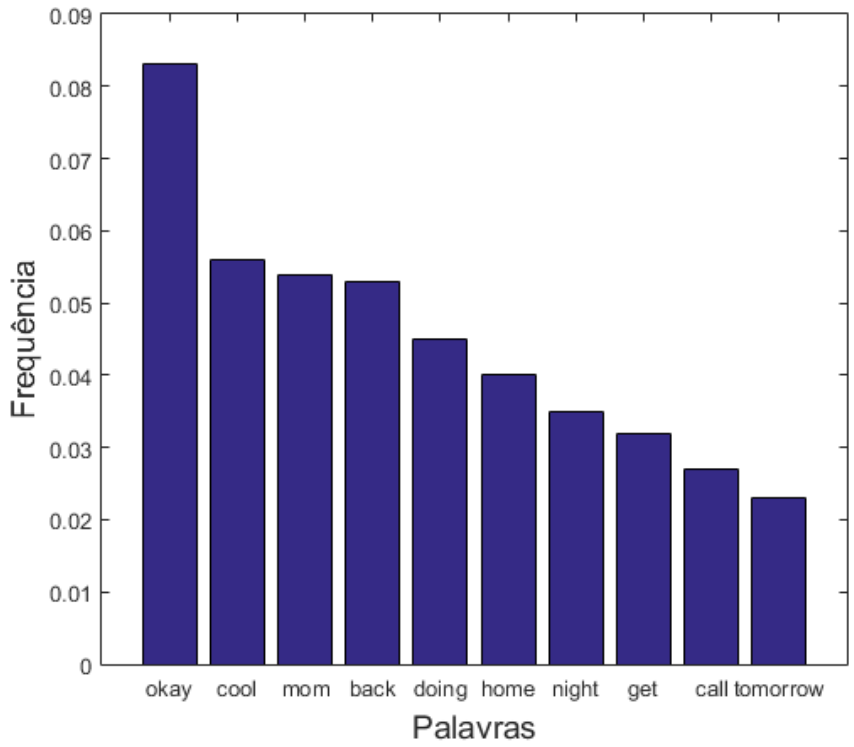


Figura 6 – Exemplo de Histograma de Referência com as 10 Palavras mais Frequentes do Corpus.

O histograma particular é uma distribuição de probabilidade de palavras que representa um *chat*. Utiliza as mesmas palavras do histograma de referência de uma determinada entidade, porém seus valores expressam a frequência relativa das palavras presentes na conversa que o originou, sendo assim, também haverão histogramas particulares para cada entidade. Estes histogramas são obtidos dos *chats* das partições de treino e teste, pois fazem parte da obtenção das características que representam as instâncias a serem utilizadas na elaboração do modelo do classificador. Nem sempre um histograma particular terá todas as palavras de um histograma de referência, ocasionando em algumas palavras com frequência igual a zero, a figura 7 mostra um exemplo de histograma particular compatível com o histograma de referência da figura 6.

No entanto, é necessário manter a compatibilidade entre os histogramas particulares e os de referência para possibiliar os cálculos dos quantificadores de Teoria da Informação

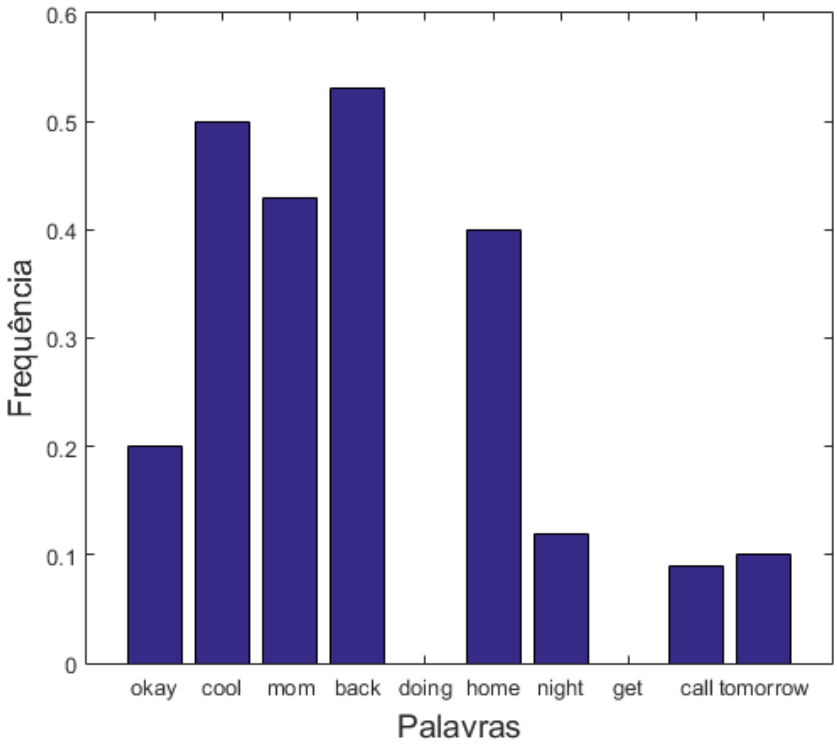


Figura 7 – Exemplo de Histograma particular compatível com o seu histograma de referência.

entropia e JSD descritos a seguir.

4.2.1 Entropia de Shannon Normalizada (H)

A Entropia em processamento de texto revela um tipo de informação que é a riqueza de vocabulário, feito por meio da análise de dispersão das palavras em um *corpus*, seu cálculo é descrito na equação 2.3 e aplicamos nas probabilidades das palavras dos histogramas particulares. Como a equação é normalizada (assume apenas valores entre 0 e 1), valores próximos a zero indicam muitas repetições de palavras, e valores próximos a um indicam extrema variação de palavras. No trabalho de Rosso, Craig e Moscato (2009) por exemplo, a análise da entropia de comédias e tragédias tem resultados notavelmente diferentes, isto se deve ao fato de que em comédias as histórias são simples, os personagens são, no geral, humildes e com vocabulário restrito e regional, resultando em entropias baixas (valores iguais ou inferiores a 0.5), ao passo que nas obras de tragédia, as histórias possuem um enredo mais elaborado, com personagens mais complexos e diálogos mais variados, resultando em entropias altas (valores superiores a 0.6).

Bogdanova, Rosso e Solorio (2012) afirmam que pedófilos tendem a manter o assunto de cunho sexual com jovens em conversas, sendo assim utilizamos a entropia para detectar esta característica de vocabulário repetitivo em relação às conversas regulares. Os valores

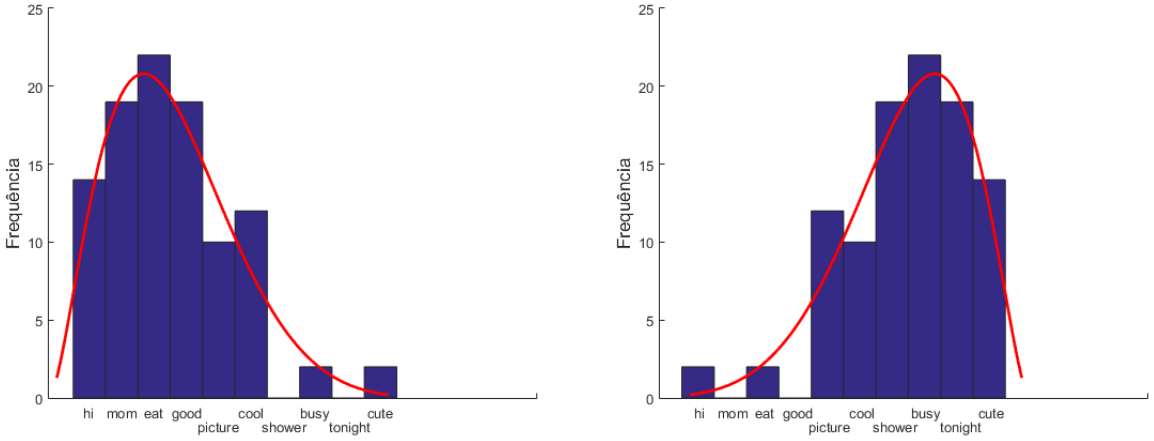
de entropia referentes à um *chat* são obtidos dos seus três histogramas particulares, por exemplo, se desejamos encontrar a entropia considerando as palavras utilizadas pela entidade pedófilo, então aplicamos a equação no histograma particular do pedófilo, e o mesmo processo é feito para o histograma de vítima e regular, logo, computamos **três valores de entropia** como característica para cada *chat*, um para cada histograma particular considerado: pedófilo, vítima e regular. Assim, a entropia indica qual tipo de entidade foi a que mais participou ou se expressou na conversa.

No entanto, é possível que em conversas haja repetição de assunto e não significa uma tentativa de aliciamento, que é o caso das conversas regulares, pois 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. Também existem casos em *chats* de pedofilia onde o maior valor de entropia é a regular, então a conversa pode ter sido inofensiva apesar de uma das entidades ser comprovadamente um pedófilo, assim como existem casos em *chats* regulares onde o maior valor de entropia é o de pedofilia, sugerindo que a conversa tenha cunho sensual entre dois adultos. Por esse motivo, a utilizamos a entropia em conjunto com a divergência de Jensen-Shannon.

4.2.2 Divergência de Jensen-Shannon (JSD)

A JSD é um quantificador em função da entropia que representa a similaridade entre distribuições de probabilidade, é capaz de perceber semelhança (ou diferença) morfológica nestes histogramas, por exemplo, o histograma na figura 8(a) possui valor de entropia idêntico ao histograma 8(b), mas seus valores de divergência são muito diferentes. O cálculo é feito por meio da equação 2.4. Em nossa proposta de extração de características utilizamos a JSD para calcular o nível de semelhança entre os histogramas particulares de um *chat* e os histogramas de referência, o valor obtido representa a semelhança do discurso presente em *chat* em relação às entidades. O valor produzido pela equação da JSD também é normalizado devido ao uso das entropias normalizadas descritas na seção 4.2.1, portanto, quanto mais similares forem os histogramas, mais próximos de zero estarão os valores e quanto mais próximos de 1, mais divergentes. Rosso, Craig e Moscato (2009) aplicaram a JSD para medir o quanto as obras literárias consideradas em seus estudos obedeciam as normas de escrita da literatura renascentista inglesa, neste contexto um histograma de referência representa o padrão de escrita da única entidade existente na base de dados, que é o autor renascentista inglês e cada obra literária é comparada ao histograma simulando um sistema de crítica literária.

Para o cálculo da JSD em nosso trabalho é necessário utilizar a entropia normalizada de histogramas de referência e regulares compatíveis entre si, por exemplo, a JSD em relação à entidade pedófilo é obtida utilizando os valores de entropia do histograma de referência e particular do pedófilo, o mesmo processo é feito para as entidades vítima e



(a) Histograma **A**. **(b)** Histograma **B**.

Figura 8 – Diferença morfológica entre os histogramas **A** e **B** que é identificada pela JSD.

regular, sendo assim, cada *chat* terá como resultado **três valores de JSD**. Portanto, cada conversa possui **seis características**, que serão as entropias e as divergências em relação ao padrão de discurso de cada tipo de entidade, indicando a dispersão de palavras se cada entidade, de cada conversa, assemelha-se mais com um pedófilo, uma vítima ou uma pessoa regular.

4.2.3 Construção do Modelo de Aprendizagem de Máquina

Para a elaboração do modelo, treinaremos um classificador SVM para realizar as predições. O motivo da escolha foi o fato de que este classificador é o mais utilizado e produz os melhores resultados de classificação em problemas de detecção de *chats* de pedofilia, segundo a revisão bibliográfica realizada no capítulo 2.

O ajuste de parâmetros e escolha dos *kerneis* foi feita de forma empírica. Os resultados produzidos pelo método proposto serão analisados no capítulo 5 para uma comparação das técnicas de extração de características BoW e Teoria da Informação verificando qual delas tem maior influência positiva nas predições de chats de pedofilia.

4.3 Considerações Finais

Neste capítulo apresentamos a arquitetura e metodologia da solução proposta, construída a partir das escolhas de técnicas resultantes do estudo bibliográfico feito no capítulo 2 e análise de estratégias dos autores no capítulo 3. Descrevemos com detalhes a composição e proveniência da base de dados que utilizamos, as etapas de pré-processamento, extração de características e classificação, também fazemos algumas observações sobre a adoção ou não de procedimentos feitos pelos autores, como a filtragem de *chats* feita pelo *baseline*

que não concordamos por eliminar informação relevante para o estudo. Conforme dito no capítulo de Introdução, a nossa contribuição é o uso de Entropia com JSD para extrair características de *chats* e verificar os impactos nos resultados de classificação, por isso enfatizamos a seção onde estão descritas. No próximo capítulo apresentaremos os resultados dos experimentos em comparação ao *baseline* para analisar seu custo benefício de uso, apontar suas vantagens e desvantagens e justificar o seu potencial como extrator de características de *chats* para resolver problemas de detecção de *chats* de pedofilia.

5 Comparação entre Técnicas de Extração de Características

Neste capítulo serão apresentados os resultados dos experimentos utilizando o método proposto por este trabalho que iremos referenciar como **H+JSD** por ser baseado nos quantificadores Entropia e Divergência, em comparação ao *baseline*, que iremos referenciar como **BoW** por ser baseado no método *Bag of Words*, escolhido por ter o melhor valor de medida F_1 e $F_{0,5}$, dentre os demais autores considerados em nosso capítulo 3, sendo que para o **H+JSD** faremos testes com os classificadores **SVM**, Árvore de Decisão (**DT**), Naive Bayes (**NB**) e **KNN**, para o **BoW** utilizaremos apenas o **SVM** de acordo com a arquitetura definida em Villatoro-Tello et al. (2012). Estabelecemos uma metodologia de experimentação onde detalhamos informações como quantidade de *chats* utilizados, a descoberta empírica de configurações dos algorítimos de classificação e seleção de atributos, fazemos também adaptações na metodologia do *baseline* para tornar os experimentos mais justos possíveis, por fim realizamos os experimentos que de fato analisam o desempenho de predição e de custo computacional de ambos os métodos, com as variações de classificadores para o método **H+JSD**, para que possamos fazer considerações sobre os resultados.

5.1 Metodologia (Materiais e Métodos)

Conforme dito no capítulo 4, utilizaremos a base de dados PAN 2012, que também é utilizada pelo *baseline* que referenciaremos no restante deste trabalho como **BoW** (*Bag of Words*). A base original é desbalanceada, apresentando 208.248 conversas regulares e 3.677 de pedofilia. Para evitar viés da classe com mais amostras, utilizamos um subconjunto balanceado composto pelas 3.677 conversas de pedofilia e 3.677 conversas do tipo regular, sorteadas aleatoriamente.

Para uma comparação justa entre os métodos, a estratégia de seleção de palavras do BoW também foi usada para o H+JSD, isolando apenas a forma de transformá-las em características. Esse processo é ilustrado na figura 9.

Os métodos foram implementados usando Matlab R2015b, Weka 3.8.0, e Stanford Core NLP 3.7.0 (para identificar palavras léxicas e funcionais). 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.

Os experimentos adotam 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

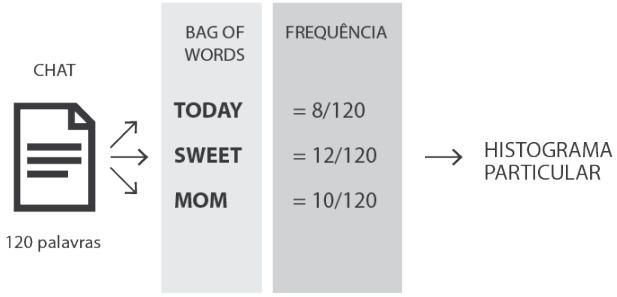


Figura 9 – Processo de seleção de palavras.

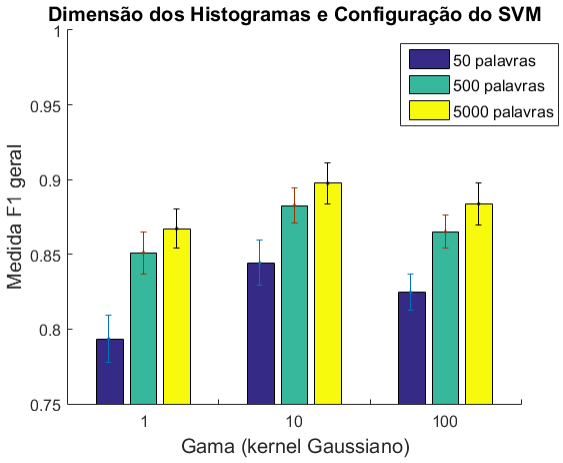
comparativa para os trabalhos relacionados e pondera igualmente a Precisão e Revocação dos métodos. A medida $F_{0,5}$ prioriza os valores de Precisão dos resultados e também foi utilizada por Villatoro-Tello et al. (2012). Os intervalos de confiança referem-se a $\alpha = 0.05$ (95% de confiança).

5.1.1 Ajuste Empírico de Parâmetros

Nesta subseção faremos o ajuste empírico de parâmetros dos algorítimos SVM, KNN e a seleção de características que consiste em encontrar o tamanho ideal do vetor de características do BoW e dos histogramas de referência dos métodos baseados em H+JSD.

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 com o método de extração de características proposto onde testamos os tamanhos de vetor 50, 500 e 5.000 e o classificador SVM com 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 10, mostram que um vocabulário de tamanho 5.000 e $\gamma = 10$ é o suficiente para obtermos um valor próximo a $90,0\%$ de F_1. Portanto, esta será a configuração utilizada no método proposto (H+JSD).

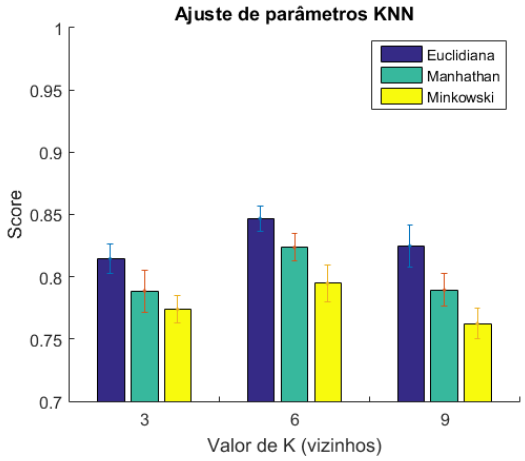
O ajuste de parâmetros do KNN será feito utilizando o mesmo método de extração de características do experimento anterior e o histograma de referência com 5.000 palavras. Iremos descobrir um valor satisfatório para o parâmetro K que indica a quantidade de vizinhos utilizadas para classificar uma amostra e para a função de distância utilizada para determinar a similaridade de uma amostra a ser classificada em relação aos seus K vizinhos. Testamos os valores K = 3, 6 e 9 e as funções de distância Euclidiana, Manhathan e Minkowski e escolhemos a configuração com base nos valores de F_1. Os resultados, dispostos na figura 11, mostram que a configuração $K = 6$ e distância Euclidiana são



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

Figura 10 – F_1 para diferentes tamanhos de histograma e diferentes valores de γ.

suficientes para produz valor de F_1 superior a $86,0\%$ e por este motivo utilizaremos esta configuração para os futuros usos do KNN neste trabalho.



Função Dist.	$K = 3$	$K = 6$	$K = 9$
Euclidiana	0,8245 ± 0,012	**0,8624 ± 0,010**	0,8347 ± 0,017
Manhathan	0,7867 ± 0,018	0,8237 ± 0,011	0,7951 ± 0,013
Minkowski	0,7786 ± 0,011	0,7978 ± 0,013	0,7691 ± 0,012

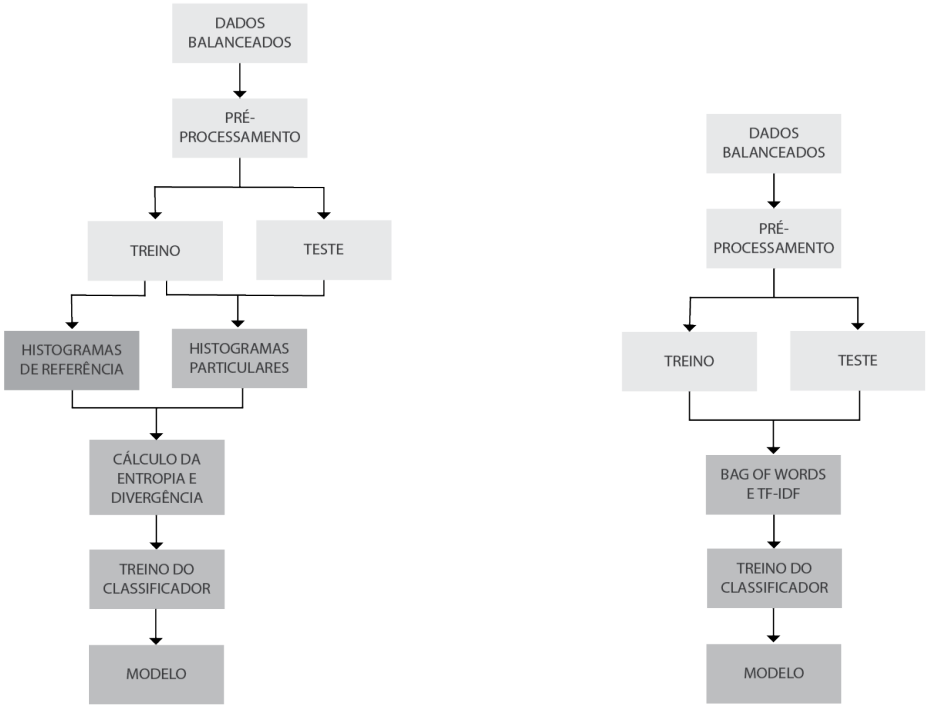
Figura 11 – F_1 para diferentes valores de K e diferentes funções de distância.

5.1.2 Adaptações no Baseline

Os seguintes pontos da arquitetura original do método destes autores precisarão ser adaptados devido a algumas concessões que fizemos na elaboração do nosso método no capítulo 4:

- Quantidade de *chats*: Balancear os dados pela quantidade de *chats* de pedofilia.
- Filtros de *chats*: Não utilizar filtros para a seleção de *chats*.
- Pré-processamento dos dados: Realizar o pré-processamento dos *chats* no experimento que tem como objetivo verificar os efeitos da remoção de pontuações e palavras funcionais nos resultados de classificação.
- Seleção de características: Realizar seleção de características no *Bag of Words*, reduzindo sua dimensionalidade ao valor ideal encontrado empiricamente no experimento da seção 5.1.1.

Caso contrário a comparação não será justa. Sendo assim, para os experimentos descritos neste capítulo, utilizaremos as arquiteturas ilustradas nas figuras 12(b) e 12(a) abaixo.



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

Figura 12 – Fluxos do método proposto (H+JSD) comparado ao baseline (BoW).

5.2 Resultados Quantitativos

Esta seção apresenta uma avaliação quantitativa entre o H+JSD, que utiliza os classificadores SVM, Àrvore de Decisão, Naive Bayes e KNN, e o BoW que utiliza apenas

o classificador SVM. As avaliações serão divididas em experimentais e analíticas (complexidade computacional). Na análise de eficácia são feitos experimentos com e sem pré-processamento dos dados, os resultados das métricas Precisão, Revocação, F_1 e $F_{0,5}$ estarão separadas por classe (Pedofilia e Regular) e também mostramos os resultados das matrizes de confusão, que são informações importantes para uma análise mais profunda da capacidade de predição dos modelos. A análise de complexidade e as medições de tempo das execuções são instrumentos complementares para demonstrar a maior eficiência do H+JSD em relação ao BoW e também identificar qual classificador possui melhor desempenho quando usado em conjunto com o H+JSD.

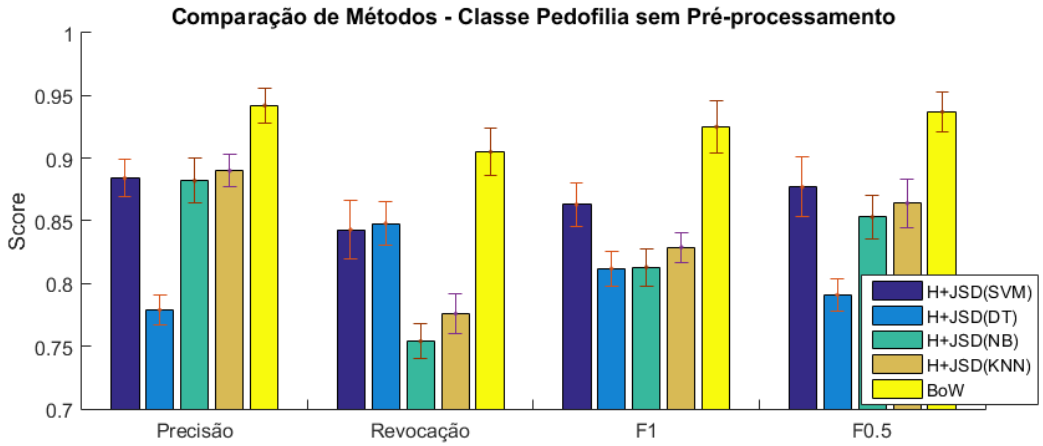
5.2.1 Avaliação Experimental de Eficácia (Qualidade da Classificação)

Neste experimento utilizaremos o resultado da extração de características obtido pelo método proposto H+JSD como *input* para o treino dos classificadores SVM, DT, NB e KNN e para o BoW, com as adaptações descritas na subseção 5.1.2, utilizaremos apenas o SVM. Em seguida iremos relacionar os resultados das métricas de classificação Precisão, Revocação, F_1 e $F_{0,5}$ e matrizes de confusão separadas por classe e pela realização ou não de pré-processamento dos dados, os objetivos são verificar se a redução de dimensionalidade de características proporcionada pelo H+JSD resultará em resultados consistentes de classificação, analisar o desempenho por classe e matriz de confusão dos métodos e suas variações de classificadores para verificar qual algorítimo possui melhores resultados quando utilizados em conjunto com o H+JSD e por fim verificar se a etapa de pré-processamento melhora os resultados de classificação.

5.2.1.1 Experimento sem Pré-Processamento dos Dados

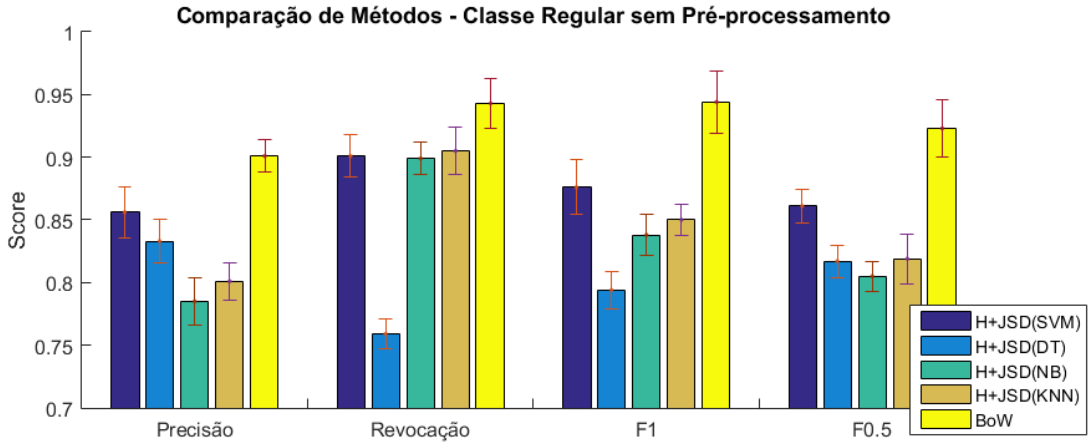
As figuras 13 e 14 apresentam os resultados de Precisão, Revocação, F_1 e $F_{0,5}$ dos métodos separados por classe e com suas variações de algorítimos de classificação do experimento onde os dados não são pré-processados e na tabela 3 estão as matrizes de confusão dos métodos.

Nos resultados referentes à classe pedofilia, o H+JSD(SVM) possui o segundo maior valor de F_1 que é igual a 86,3%, apenas 6,70% abaixo do BoW (0,863 contra 0,925), enquanto os métodos H+JSD(DT), H+JSD(NB) e H+JSD(KNN) tiveram desempenho semelhante entre si, com valores em torno de 81%, devido aos seus intervalos de confiança que se cruzam. Apesar de ter a segunda melhor média de $F_{0,5}$, o intervalo de confiança revela que seu desempenho pode ser parecido com o H+JSD(NB) e H+JSD(KNN), onde seus valores estão entre 85,0% e 87,0%. O H+JSD(DT) teve o menor valor de $F_{0,5}$, igual a 79,1%, por causa do seu valor de precisão que também foi o menor de todos. Teve o desempenho de F_1 igual a 81,2%, equiparado ao H+JSD(NB) e H+JSD(KNN) devido ao seu ao seu valor de revocação superior a 80,0%. Como a revocação é uma métrica que



	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
Precisão	0,884 ± 0,015	0,779 ± 0,012	0,882 ± 0,018	0,890 ± 0,013	0,942 ± 0,014
Revocação	0,843 ± 0,023	0,848 ± 0,017	0,754 ± 0,014	0,776 ± 0,016	0,905 ± 0,019
F1	**0,863 ± 0,017**	0,812 ± 0,014	0,813 ± 0,015	0,829 ± 0,012	**0,925 ± 0,021**
F0.5	**0,877 ± 0,024**	0,791 ± 0,013	0,853 ± 0,017	0,864 ± 0,019	**0,937 ± 0,016**

Figura 13 – Comparação dos métodos Sem pré-processamento classe pedofilia.



	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
Precisão	0,856 ± 0,020	0,833 ± 0,017	0,785 ± 0,019	0,801 ± 0,015	0,901 ± 0,013
Revocação	0,901 ± 0,017	0,759 ± 0,012	0,899 ± 0,013	0,905 ± 0,019	0,943 ± 0,020
F1	**0,876 ± 0,022**	0,794 ± 0,015	0,838 ± 0,016	0,850 ± 0,012	**0,944 ± 0,025**
F0.5	**0,861 ± 0,013**	0,817 ± 0,013	0,805 ± 0,012	0,819 ± 0,020	**0,923 ± 0,023**

Figura 14 – Comparação dos métodos Sem pré-processamento classe regular.

indica boa capacidade de predição, então os métodos H+JSD(SVM) e H+JSD(DT) são os que mais acertam instâncias da classe pedofilia.

Nos resultados da classe regular, o H+JSD(SVM) possui valor de $F_{0,5}$ igual a 86,1% o segundo maior inclusive em intervalo de confiança, novamente apenas 6,71% abaixo do BoW (0,861 contra 0,923). Também possui a segunda maior média de F_1 que é 87,6%, porém seu intervalo de confiança se cruza com os métodos H+JSD(NB) e H+JSD(KNN).

	Predição	
Original	Pedofilia	Regular
Pedofilia	84,3%	15,7%
Regular	9,9%	90,1%

(a) H+JSD(SVM) sem pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	84,8%	15,2%
Regular	24,1%	75,9%

(b) H+JSD(DT) sem pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	75,4%	24,6%
Regular	10,1%	89,9%

(c) H+JSD(NB) sem pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	77,6%	22,4%
Regular	9,5%	90,5%

(d) H+JSD(KNN) sem pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	90,5%	9,5%
Regular	5,7%	94,3%

(e) BoW sem pré-processamento.

Tabela 3 – Comparação das matrizes de confusão sem pré-processamento.

O H+JSD(DT) tem o menor valor de revocação na classe pedofilia e isso se refletiu no seu valor de F_1 que foi também o menor de todos, igual a 79,4%, este comportamento indica que o método cometeu muitos erros de falso positivo para esta classe. Seu valor de $F_{0,5}$ só não foi tão prejudicado por causa da sua precisão que está acima de 80,0%.

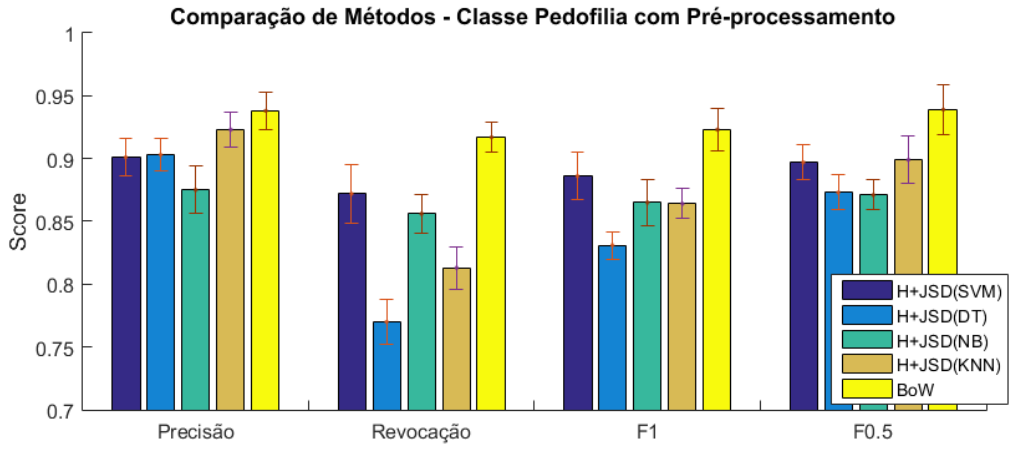
Por fim, no experimento sem pré-processamento dos dados o H+JSD(SVM) é o método que mais acerta instâncias de um modo geral por estar entre as maiores médias de revocação em ambas as classes (84,3% na classe pedofilia e 90,1% na classe regular), seus valores de F_1 em relação ao BoW é de 6,70% a menos na classe pedofilia (0,863 contra 0,925) e 7,20% a menos na classe regular (0,876 contra 0,944), os valores de $F_{0,5}$ são 6,40% menores na classe pedofilia (0,877 contra 0,937) e 6,71% menores na classe regular (0,861 contra 0,923). Os métodos H+JSD(NB) e H+JSD(KNN) tiveram desempenho muito semelhante em ambas as classes para todas as métricas, as diferenças não chegaram a 3,00%. O H+JSD(DT) tem boa capacidade de predição da classe pedofilia que é a classe de interesse, o fato de cometer muitos erros de falsos positivos poderia ser relevado se caso nenhum outro método tivesse desempenho de classificação melhor, pois em nosso problema é mais crítico cometer erros do tipo falso negativo, que significa classificar erroneamente uma conversa online de pedofilia como regular.

Comparando os valores de F_1 e $F_{0,5}$ dos métodos baseados em H+JSD com o BoW, temos que **H+JSD(SVM)** para a classe pedofilia possui F_1 **6,70%** (0,863 contra 0,925) menor que o BoW e $F_{0,5}$ **6,40%** (0,877 contra 0,937) menor que o BoW, na classe regular o F_1 é **7,20%** (0,876 contra 0,944) menor que o BoW e o $F_{0,5}$ **6,41%** (0,861 contra 0,923) menor que o BoW. O **H+JSD(DT)** na classe pedofilia possui F_1 **12,2%** (0,812 contra 0,925) menor que o BoW e $F_{0,5}$ **15,5%** (0,791 contra 0,937) menor que o BoW, na classe

regular o F_1 é **15,8%** (0,794 contra 0,944) menor que o BoW e o $F_{0,5}$ **11,4%** (0,817 contra 0,923) menor que o BoW. **O H+JSD(NB)** na classe pedofilia tem F_1 **12,1%** (0,813 contra 0,925) menor que o BoW e $F_{0,5}$ **8,96%** (0,853 contra 0,937) menor que o BoW, na classe regular o F_1 é **11,2%** (0,838 contra 0,944) menor que o BoW e o $F_{0,5}$ **12,7%** (0,805 contra 0,923) menor que o BoW. O **H+JSD(KNN)** na classe pedofilia possui F_1 **10,3%** (0,829 contra 0,925) menor que o BoW e o $F_{0,5}$ **7,79%** (0,864 contra 0,937) menor que o BoW, na classe regular o F_1 é **9,95%** (0,850 contra 0,944) menor que o BoW e o $F_{0,5}$ **11,2%** (0,819 contra 0,923) menor que o BoW. No experimento sem pré-processamento dos dados, o **H+JSD(SVM)** é o método baseado em H+JSD com o melhor desempenho de classificação por ter a menor diferença no valor das métricas F_1 e $F_{0,5}$ em relação ao BoW.

5.2.1.2 Experimento com Pré-Processamento dos Dados

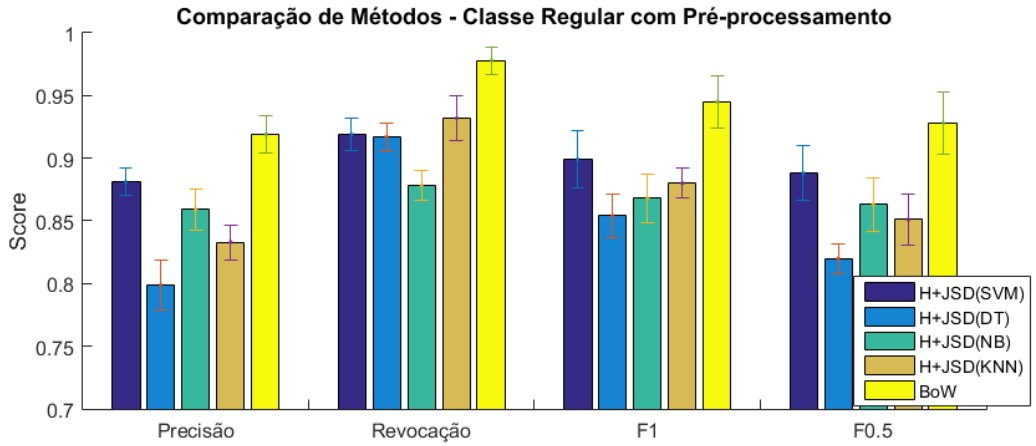
Nesta subseção veremos os resultados das métricas do experimento com pré-processamento dos dados. As figuras 15 e 16 apresentam a Precisão, Revocação, F_1 e $F_{0,5}$ dos métodos também separados por classe como no experimento anterior e a tabela 4 mostra as matrizes de confusão dos métodos.



	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
Precisão	0,901 ± 0,015	0,903 ± 0,013	0,875 ± 0,019	0,923 ± 0,014	0,938 ± 0,015
Revocação	0,872 ± 0,023	0,770 ± 0,018	0,856 ± 0,015	0,813 ± 0,017	0,917 ± 0,012
F1	**0,886 ± 0,019**	0,831 ± 0,011	0,865 ± 0,018	0,864 ± 0,012	**0,923 ± 0,017**
F0.5	0,897 ± 0,014	0,873 ± 0,014	0,871 ± 0,012	**0,899 ± 0,019**	**0,939 ± 0,020**

Figura 15 – Comparação dos métodos Com pré-processamento classe pedofilia.

Em relação à classe pedofilia, O H+JSD(SVM) continua possuindo a segunda média mais alta de F_1 que é igual a 88,6% abaixo apenas do BoW, mas a maior média de $F_{0,5}$ foi do H+JSD(KNN) que é igual a 89,9%. Neste experimento nenhum método conseguiu se destacar totalmente em F_1 e $F_{0,5}$ sem que seu intervalo de confiança se cruzasse com pelo menos um dos outros métodos. O H+JSD(DT) passou a ter o menor valor de F_1



	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
Precisão	0,881 ± 0,011	0,799 ± 0,020	0,859 ± 0,016	0,833 ± 0,014	0,919 ± 0,015
Revocação	0,919 ± 0,013	0,917 ± 0,011	0,878 ± 0,012	0,932 ± 0,018	0,978 ± 0,011
F1	**0,899 ± 0,023**	0,854 ± 0,017	0,868 ± 0,019	0,880 ± 0,012	**0,945 ± 0,021**
F0.5	**0,888 ± 0,022**	0,820 ± 0,012	0,863 ± 0,021	0,851 ± 0,020	**0,928 ± 0,025**

Figura 16 – Comparação dos métodos Com pré-processamento classe regular.

	Predição	
Original	Pedofilia	Regular
Pedofilia	87,2%	12,8%
Regular	8,1%	91,9%

(a) H+JSD(SVM) com pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	77,0%	23,0%
Regular	8,3%	91,7%

(b) H+JSD(DT) com pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	85,6%	14,4%
Regular	12,2%	87,8%

(c) H+JSD(NB) com pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	81,3%	18,7%
Regular	6,8%	93,2%

(d) H+JSD(KNN) com pré-processamento.

	Predição	
Original	Pedofilia	Regular
Pedofilia	91,7%	8,3%
Regular	2,2%	97,8%

(e) BoW com pré-processamento.

Tabela 4 – Comparação das matrizes de confusão com pré-processamento.

devido ao seu menor valor de revocação, mas seu valor de precisão de 90% influenciou positivamente no seu valor de $F_{0,5}$. O H+JSD(NB) possui valor de revocação superior ao H+JSD(KNN), 85,6% contra 81,3%, um aumento de 5,00%, demonstrando uma mudança de comportamento em relação ao experimento sem pré-processamento onde todas as métricas de ambos os métodos possuiam diferença de desempenho entre si inferior a 3,00%. Os métodos H+JSD(SVM), H+JSD(NB) e o BoW são os que mais acertam a classe pedofilia por estarem entre as melhores médias de revocação.

Na classe regular, novamente o H+JSD(SVM) tem a segunda maior média de F_1 e $F_{0,5}$, mas o intervalo de confiança mostra que seu desempenho pode ser semelhante ao H+JSD(NB) e H+JSD(KNN), os valores das métricas desses métodos ficaram entre 85,0% e 89,0%. O H+JSD(DT) teve a menor média de precisão (79,9%) e isso influenciou negativamente no seu valor de $F_{0,5}$, porém teve uma das maiores médias de revocação contribuindo para um bom valor de F_1 indicando que este método possui boa capacidade de predição para instancias da classe regular. O H+JSD(NB) teve o menor valor de revocação, mas ainda assim é 2,57% superior ao valor obtido na classe pedofilia, além disso teve diferença de 6,15% em relação ao H+JSD(KNN), novamente mostrando uma mudança de comportamento entre os métodos. Os métodos H+JSD(SVM), H+JSD(DT) e H+JSD(KNN) são os que mais acertam a classe regular.

Por fim o pré-processamento dos dados melhorou os valores de F_1 e $F_{0,5}$ de todos os métodos baseados em H+JSD, o **H+JSD(SVM)** teve suas métricas da classe pedofilia aumentadas em aproximadamente **2,00%** e da classe regular em **2,55%** para o F_1 e **3,00%** para o $F_{0,5}$. O **H+JSD(DT)** na classe pedofilia teve melhoria no F_1 de **2,28%** e $F_{0,5}$ de **9,39%** e na classe regular as melhorias foram de **7,00%** no F_1 e **0,36%** $F_{0,5}$. O **H+JSD(NB)** teve melhorias na classe pedofilia de **6%** para o F_1 e **2,00%** para o $F_{0,5}$ e na classe regular as melhorias foram de **3,45%** no F_1 e **6,72%** no $F_{0,5}$. O **H+JSD(KNN)** teve melhorias na classe pedofilia de **4,00%** para o F_1 e **3,89%** para o $F_{0,5}$ e na classe regular a melhoria foi em torno **3,00%** para ambas as métricas. O BoW teve baixa no F_1 de 0,21% e aumento de 0,21% no $F_{0,5}$, ambos da classe pedofilia, na classe regular houve aumento de 0,10% no F_1 e aumento de 0,53% no $F_{0,5}$, visto que não houve nem 1,00% de melhoria, então podemos concluir que o pré-processamento melhorou as métricas dos métodos baseados em H+JSD mas não teve influência significativa no BoW.

O H+JSD(DT) passou a acertar mais a classe regular após o pré-processamento dos dados, sugerindo que a maioria das palavras funcionais presentes nas conversas de pedofilia estavam sendo selecionadas pelo ganho de informação para compor os nós do modelo da Árvore de Decisão no experimento sem pré-processamento e ao utilizar dados pré-processados, as palavras selecionadas para os nós da árvore eram mais discriminantes para a classe regular. O H+JSD(NB) apresentou menor diferença de valores entre sua precisão e revocação em cada classe neste experimento equilibrando os valores de F_1 e $F_{0,5}$ e indicando que este método tem capacidade de predição semelhante para as duas classes. Com o pré-processamento o H+JSD(NB) tem diferença entre a precisão e a revocação de 2,21% (0,875 contra 0,856) na classe pedofilia e 2,16% (0,859 contra 0,878) na classe regular, ao passo que no experimento sem pré-processamento dos dados esta diferença é de 16,9% (0,882 contra 0,754) na classe pedofilia e 12,6% (0,785 contra 0,889), isto significa que assim como ocorre em H+JSD(SVM), no H+JSD(NB) as palavras funcionais também não contribuem para uma melhor predição e se comportam como ruído.

Comparando os valores **com pré-processamento** de F_1 e $F_{0,5}$ dos métodos baseados em H+JSD com o BoW, temos que **H+JSD(SVM)** para a classe pedofilia possui F_1 **4,00%** (0,886 contra 0,923) menor que o BoW e $F_{0,5}$ **4,47%** (0,897 contra 0,939) menor que o BoW, na classe regular o F_1 é **4,86%** (0,899 contra 0,945) menor que o BoW e o $F_{0,5}$ **4,31%** (0,888 contra 0,928) menor que o BoW. O **H+JSD(DT)** na classe pedofilia possui F_1 **9,96%** (0,831 contra 0,923) menor que o BoW e $F_{0,5}$ **7,02%** (0,873 contra 0,939) menor que o BoW, na classe regular o F_1 é **9,62%** (0,854 contra 0,945) menor que o BoW e o $F_{0,5}$ **11,6%** (0,820 contra 0,928) menor que o BoW. O **H+JSD(NB)** na classe pedofilia tem F_1 **6,28%** (0,865 contra 0,923) menor que o BoW e $F_{0,5}$ **7,24%** (0,871 contra 0,939) menor que o BoW, na classe regular o F_1 é **8,14%** (0,868 contra 0,945) menor que o BoW e o $F_{0,5}$ **7,00%** (0,863 contra 0,928) menor que o BoW. O **H+JSD(KNN)** na classe pedofilia possui F_1 **6,39%** (0,864 contra 0,923) menor que o BoW e o $F_{0,5}$ **4,25%** (0,899 contra 0,939) menor que o BoW, na classe regular o F_1 é **6,87%** (0,880 contra 0,945) menor que o BoW e o $F_{0,5}$ **8,29%** (0,851 contra 0,928) menor que o BoW. Novamente o método **H+JSD(SVM)** teve a menor diferença nos valores das métricas em relação ao BoW, quanto menor a diferença entre as métricas mais similar é o desempenho, portanto é o método baseado em H+JSD com melhor desempenho.

Após analisar o desempenho dos métodos, o próximo passo é avaliar o custo de execução destes métodos que é um ponto chave em nosso trabalho, pois para um bom funcionamento em ambientes de hardware limitado é necessário um baixo custo computacional sem muita interferência na capacidade de predição das amostras.

5.2.2 Análise de Eficiência (Análise de Complexidade)

Na etapa de extração de características todos 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 dos métodos que utilizam o classificador SVM, que é o caso do H+JSD(SVM) e BoW, 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; esse custo é referente às operações de produto escalar que o classificador realiza para encontrar o hiperplano de separação das instâncias. O teste tem custo $O(kv)$ onde v é o tamanho da coleção de teste, como o H+JSD(SVM) utiliza apenas seis características, o custo do treino se torna $O(t^2)$ e do teste $O(v)$. O custo de treino da Árvore de Decisão é $O(tk^2)$, onde t representa a coleção de treino e k as características. Este custo se deve às operações de entropia e ganho de informação calculadas para cada característica, onde o objetivo é escolher a característica com maior ganho de informação para compor cada nó da árvore e o custo de teste é $O(vk)$, onde v é o tamanho da coleção de teste. Por fim, o custo do H+JSD(DT) será $O(t)$ no treino e $O(v)$ no teste pois o custo que envolve a

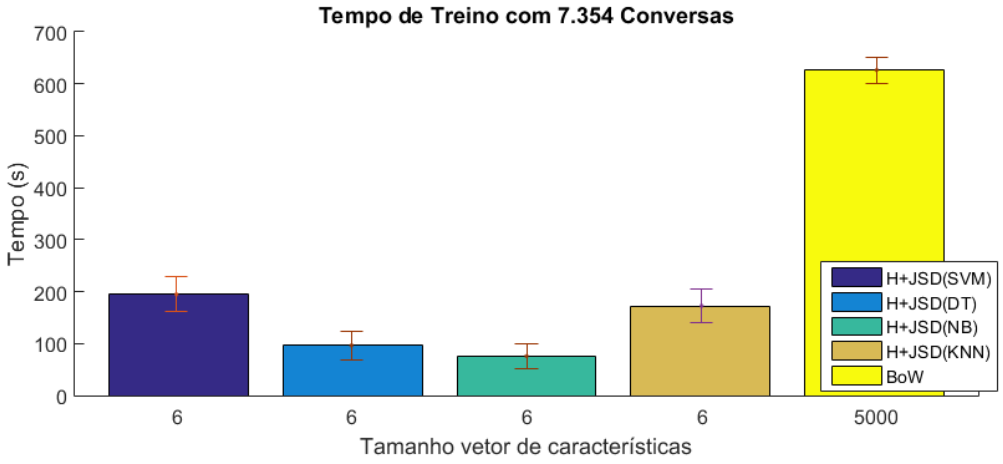
quantidade de características é constante. O custo de treino do Naive Bayes é $O(tkc)$, onde t é a coleção de treino, k são as características e c é a quantidade de classes. Este custo se deve aos cálculos de probabilidade condicional de cada característica, de cada instância e este cálculo é feito para cada classe considerada no problema, que em nosso trabalho são duas: Pedofilia e Regular. O custo de teste é $O(vkc)$, onde v é a coleção de teste. Sendo assim, o custo do treino H+JSD(NB) é $O(t)$ e o teste $O(v)$. O KNN não possui treino, todo o custo deste algorítimo está no teste, que é $O(vx+vk)$, onde v é a coleção de teste, x é a quantidade de vizinhos próximos e k é a quantidade de características. Este custo consiste no cálculo de similaridade de cada instância em relação à quantidade de vizinhos escolhida que em nosso trabalho é 3, o cálculo de similaridade é feito utilizando a distância Euclidiana e seu cálculo está em função das características das instâncias.

Por fim, todos os classificadores que utilizaram as características extraídas pelo método H+JSD possuem custo computacional de treino e teste mais baixos que o BoW. Este comportamento dos métodos baseados em H+JSD proporciona vantagem significativa em relação ao BoW, pois apesar de fixarmos o tamanho do vetor de características do BoW em 5.000 palavras por motivos de eficiência, essa quantidade é, a princípio, igual ao tamanho do vocabulário e sendo assim, o custo de treino de quase todos os métodos baseados em H+JSD é linear, com exceção do H+JSD(SVM) que é quadratico, mas ainda assim é menor que o custo cúbico de treino do BoW. No teste todos os métodos baseados em H+JSD tem custo linear enquanto que o BoW tem custo quadrático.

5.2.3 Avaliação Experimental de Eficiência (Tempo de Execução)

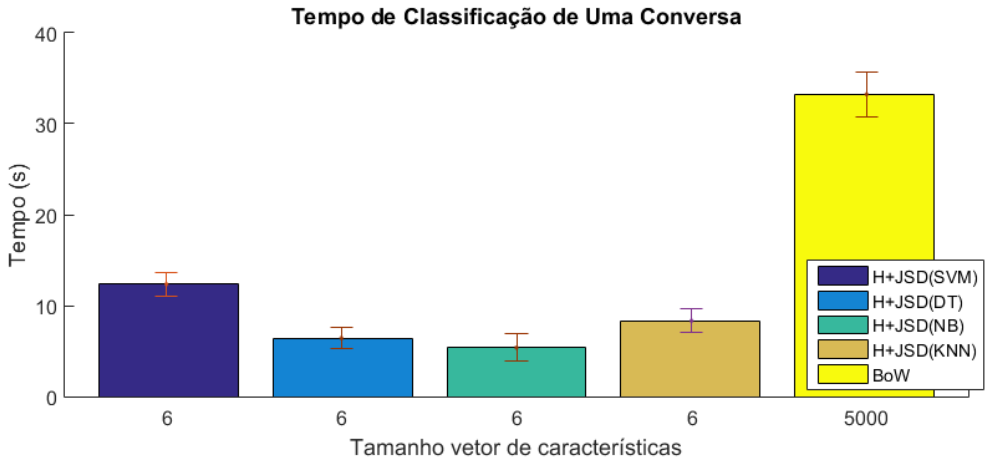
O custo computacional e o tempo de execução dos métodos são métricas fundamentais, principalmente considerando redes de mensagens instantâneas ponto-a-ponto baseadas em dispositivos móveis, tais como o Whatsapp. A figura 17 exibe tempo de execução total dos métodos (7.354 conversas, ten-fold cross-validation), enquanto a figura 18 apresenta o tempo médio de classificação de uma conversa escolhida aleatoriamente da coleção das 7.354 conversas. Os classificadores que utilizaram as características do método H+JSD tiveram eficiência (velocidade) muito maior que o BoW. Isto ocorre porque o H+JSD resume o vocabulário da coleção de treino/teste em apenas 6 características, reduzindo o custo computacional da classificação que é diretamente proporcional à quantidade de características além outros fatores. 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(SVM) é 62,8% menor que o BoW (12,35s contra 33,24s), seguindo este exemplo o H+JSD(DT) é 80,6% menor (6,42s contra 33,24s), o H+JSD(NB) é 83,8% menor (5,37s contra 33,24s) e o H+JSD(KNN) é 74,9% menor (8,33s contra 33,24s). O próximo passo é analisar o tempo de execução dos métodos variando gradativamente o tamanho

do vocabulário para perceber o comportamento assintótico dos mesmos.



H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
194,5s ± 33,1s	96,2s ± 28,1s	75,6s ± 23,5s	172,4s ± 31,5s	626,3s ± 25,4s

Figura 17 – Tempo de execução de treino dos métodos com 7.354 conversas.

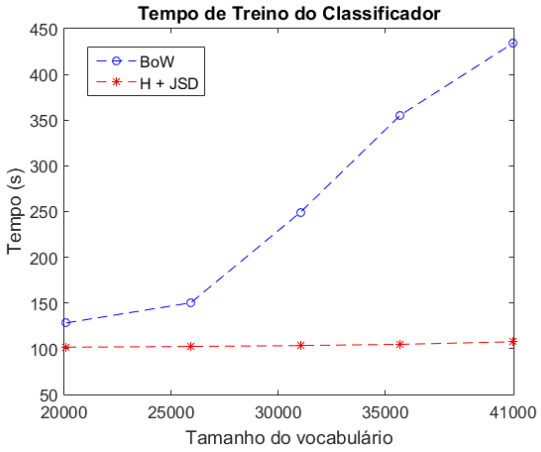


H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)	BoW
12,3s ± 1,27s	6,42s ± 1,19s	5,37s ± 1,52s	8,33s ± 1,26s	33,2s ± 2,46s

Figura 18 – Tempo de classificação de uma conversa pelos métodos.

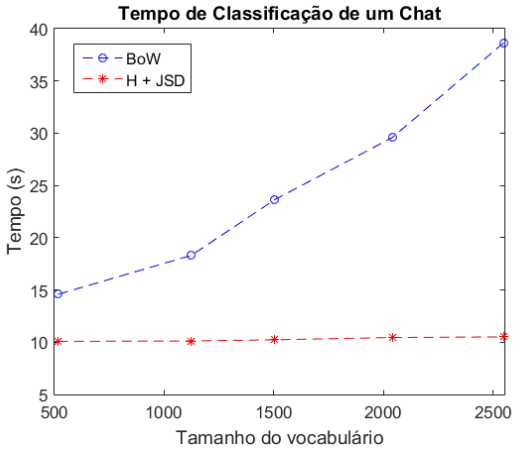
As figuras 19, 20, 21 e 22 apresentam o tempo de execução de todos os métodos para o treino do modelo e para a classificação de uma conversa. Esse experimento avalia o impacto do tamanho do vocabulário utilizado pelo BoW e nos histogramas de referência dos métodos baseados em H+JSD. Os resultados mostram que os métodos baseados em H+JSD são muito mais escaláveis que o BoW. Para conversas com mais de 2.500 palavras

o H+JSD(SVM) chega a ser 72,8% mais rápido que o BoW (10,51s contra 38,67s), O H+JSD(DT) é 87,4% menor (4,85s contra 38,67s), o H+JSD(NB) é 88,9% menor (4,27s contra 38,67s) e, para conversas maiores, esta diferença será ainda maior, o H+JSD(KNN) é 81,9% menor (6,98s contra 38,67s), mas como ele possui custo de classificação muito relacionado à quantidade de instâncias, que influencia no aumento do vocabulário, então ele terá tendência de diminuir a diferença de custo computacional em relação ao BoW ao invés de aumentar. A figura 23(a) e 23(b) relaciona os custos de treino e classificação de uma conversa respectivamente dos métodos baseados em H+JSD e verificar qual deles possui a melhor eficiência de classificação.



Vocab.	H+JSD(SVM)	BoW
20.092	101,7s	128,3s
25.914	102,4s	150,2s
31.036	103,4s	249,1s
35.711	104,8s	355,4s
40.996	107,6s	434,5s

(a) Treino.



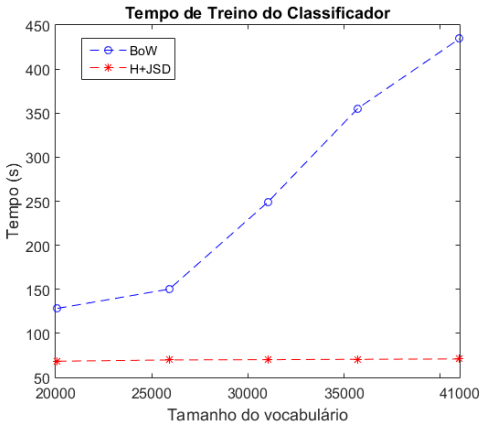
Vocab.	H+JSD(SVM)	BoW
517	10,0s	14,5s
1.124	10,1s	18,3s
1.503	10,2s	23,6s
2.039	10,4s	29,5s
2.548	10,5s	38,6s

(b) Classificação de uma conversa.

Figura 19 – Tempo de treino e classificação de uma conversa pelo vocabulário com o método H+JSD(SVM).

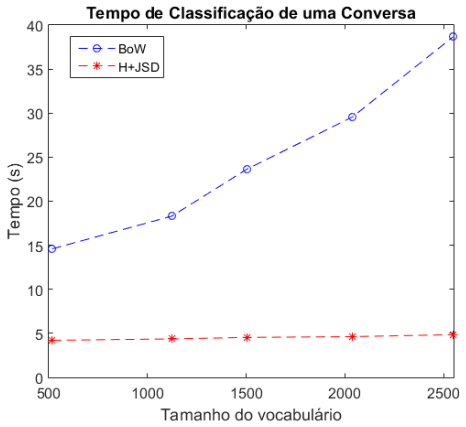
F1		
Vocab.	H+JSD(SVM)	BoW
20.092	0,556	0,598
25.914	0,583	0,634
31.036	0,637	0,672
35.711	0,652	0,699
40.996	0,675	0,722

Tabela 5 – F1 do tempo de treino do experimento 19(a).



Vocab.	H+JSD(DT)	BoW
20.092	68,3s	128,3s
25.914	69,8s	150,2s
31.036	70,1s	249,1s
35.711	70,5s	355,4s
40.996	70,9s	434,5s

(a) Treino.



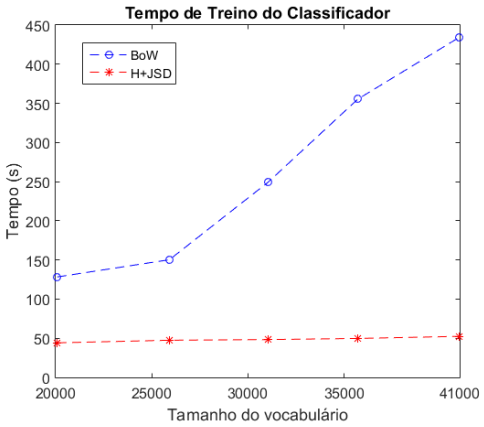
Vocabulário	H+JSD(DT)	BoW
517	4,21s	14,5s
1.124	4,37s	18,3s
1.503	4,54s	23,6s
2.039	4,63s	29,5s
2.548	4,85s	38,6s

(b) Classificação.

Figura 20 – Tempo de treino e classificação de uma conversa pelo vocabulário com o método H+JSD(DT).

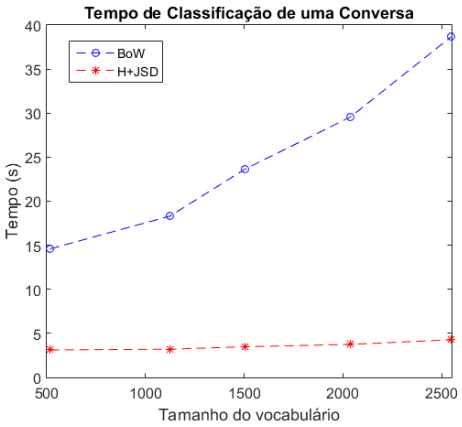
F1		
Vocab.	H+JSD(DT)	BoW
20.092	0,481	0,598
25.914	0,533	0,634
31.036	0,562	0,672
35.711	0,591	0,699
40.996	0,608	0,722

Tabela 6 – F1 do tempo de treino do experimento 20(a).



Vocab.	H+JSD(NB)	BoW
20.092	44,1s	128,3s
25.914	47,4s	150,2s
31.036	48,3s	249,1s
35.711	49,7s	355,4s
40.996	52,4s	434,5s

(a) Treino.



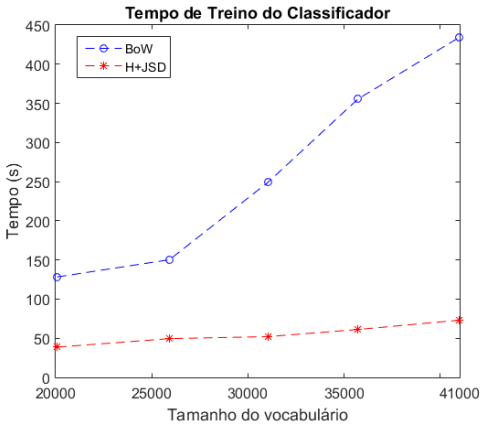
Vocabulário	H+JSD(NB)	BoW
517	3,12s	14,5s
1.124	3,20s	18,3s
1.503	3,48s	23,6s
2.039	3,75s	29,5s
2.548	4,27s	38,6s

(b) Classificação.

Figura 21 – Tempo de treino e classificação de uma conversa pelo vocabulário com o método H+JSD(NB).

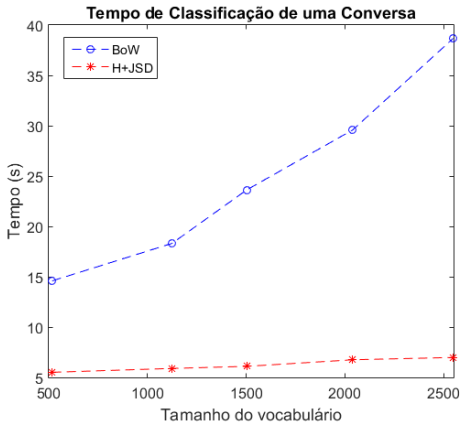
F1		
Vocab.	H+JSD(NB)	BoW
20.092	0,516	0,598
25.914	0,547	0,634
31.036	0,573	0,672
35.711	0,605	0,699
40.996	0,622	0,722

Tabela 7 – F1 do tempo de treino do experimento 21(a).



Vocab.	H+JSD(KNN)	BoW
20.092	38,7s	128,3s
25.914	49,5s	150,2s
31.036	52,1s	249,1s
35.711	61,2s	355,4s
40.996	73,1s	434,5s

(a) Treino.



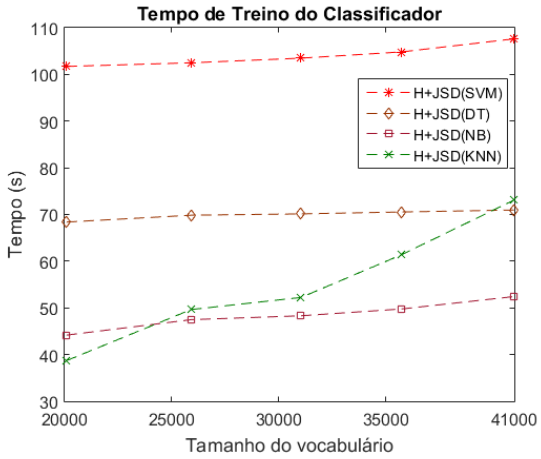
Vocabulário	H+JSD(KNN)	BoW
517	5,51s	14,5s
1.124	5,89s	18,3s
1.503	6,10s	23,6s
2.039	6,75s	29,5s
2.548	6,98s	38,6s

(b) Classificação.

Figura 22 – Tempo de treino e classificação de uma conversa pelo vocabulário com o método H+JSD(KNN).

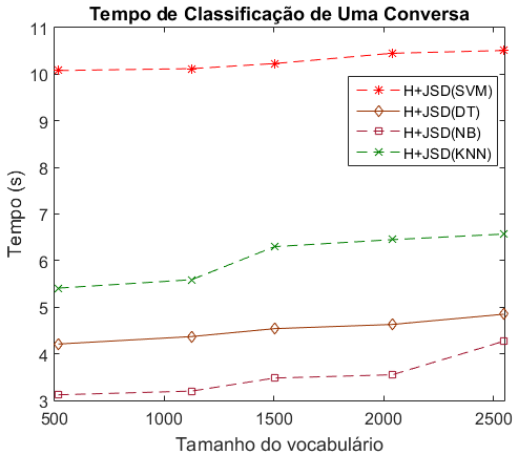
F1		
Vocab.	H+JSD(KNN)	BoW
20.092	0,505	0,598
25.914	0,553	0,634
31.036	0,594	0,672
35.711	0,612	0,699
40.996	0,634	0,722

Tabela 8 – F1 do tempo de treino do experimento 22(a).



Vocab.	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)
20.092	101,7s	68,3s	44,1s	38,7s
25.914	102,4s	69,8s	47,4s	49,5s
31.036	103,4s	70,1s	48,3s	52,1s
35.711	104,8s	70,5s	49,7s	61,2s
40.996	107,6s	70,9s	52,4s	73,1s

(a) Treino.



Vocab.	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)
517	10,0s	4,21s	3,12s	5,51s
1.124	10,1s	4,37s	3,20s	5,89s
1.503	10,2s	4,54s	3,48s	6,10s
2.039	10,4s	4,63s	3,75s	6,75s
2.548	10,5s	4,85s	4,27s	6,98s

(b) Classificação de uma conversa.

Figura 23 – Tempo de classificação de uma conversa pelos métodos baseados em H+JSD.

F1				
Vocab.	H+JSD(SVM)	H+JSD(DT)	H+JSD(NB)	H+JSD(KNN)
20.092	0,556	0,481	0,516	0,505
25.914	0,583	0,533	0,547	0,553
31.036	0,637	0,562	0,573	0,594
35.711	0,652	0,591	0,605	0,612
40.996	0,675	0,608	0,622	0,634

Tabela 9 – F1 do tempo de treino do experimento 23(a).

O H+JSD(SVM) foi o método com o maior custo de treino, isso é esperado pois sabemos que o custo de treino do SVM é o maior de todos mesmo com as 6 características do H+JSD. O H+JSD(KNN) tende a ser tão custoso quanto o H+JSD(SVM) e possivelmente o BoW se a quantidade de instâncias for grande o suficiente, pois conforme dito anteriormente, seu custo é muito relacionado à quantidade de instâncias utilizadas na classificação, que para o caso deste experimento é a validação do modelo treinado, uma vez que não há treino. O H+JSD(DT) e H+JSD(NB) possuem custo linear, a diferença é que no custo do H+JSD(DT) existe um expoente 2 associado à variavel da característica, que não é o caso do custo do H+JSD(NB), por isso o H+JSD(DT) teve valores levemente maiores que o H+JSD(NB). O treino é uma etapa importante mas não é crítica pois para um sistema de classificação de conversas online de pedofilia, esta etapa pode ser feita *offline* em um intervalo de tempo estipulado e apenas o modelo obtido é utilizado *online*. Na classificação de uma conversa, novamente o **H+JSD(SVM)** teve o maior custo, comparado aos outros métodos para um vocabulário de 2.548 palavras por exemplo, o **H+JSD(DT)** é **53,8%** mais rápido (4,85s contra 10,51s), o **H+JSD(NB)** é **59,3%** mais rápido (4,27s contra 10,51s) e o **H+JSD(KNN)** é **33,5%** mais rápido (6,98s contra 10,51s).

Por fim, existem dois métodos baseados em H+JSD viáveis para solucionar o problema de detecção de conversas *online* de pedofilia em ambientes de *hardware* limitado, o primeiro é o **H+JSD(SVM)** que possui o melhor desempenho de classificação tanto no experimento sem pré-processamento (em torno de 6,00% inferior ao BoW) quanto no com pré-processamento (em torno de 4,00% inferior ao BoW), apesar do maior custo de execução entre os outros métodos baseados em H+JSD, mas que ainda assim é **72,8%** mais rápido que BoW e o segundo é o **H+JSD(NB)** ambos na condição de pré-processamento onde suas métricas de classificação são entre **6,00%** e **8,00%** inferiores ao BoW, pois no experimento sem pré-processamento esta diferença está entre 8,00% e 12,0%, mas em questões de eficiência (tempo de execução) o H+JSD(NB) chega a ser **59,3%** mais rápido que o H+JSD(SVM) e **88,9%** mais rápido que o BoW. O **H+JSD(KNN)** apesar de ter demonstrado bom desempenho de classificação demonstrou não ser escalável em custo computacional quando aumentamos a quantidade de instâncias no estudo, por isso não recomendamos seu uso para resolver o problema e o H+JSD(DT) teve os piores resultados

de classificação nos experimentos.

5.3 Considerações Finais

Neste capítulo, inicialmente foram feitas formalizações de procedimentos para prover uma experimentação justa entre o método de extração de características proposto por este trabalho e o *baseline* que referenciamos como BoW, adaptando as arquiteturas para manter compatibilidade entre as etapas de processamento dos dados. Em seguida foram feitos experimentos que visavam otimizar alguns destes processos, como por exemplo a descoberta do tamanho ideal do vetor de características e histogramas de palavras, bem como a configuração dos classificadores SVM e KNN. Após isso, foram realizados experimentos de desempenho de classificação e custo computacional onde o método proposto variou entre os classificadores SVM, Árvore de Decisão, Naive Bayes e KNN em comparação ao BoW, os experimentos de classificação foram divididos em experimento sem pré-processamento dos dados e com pré-processamento. O objetivo dos experimentos neste capítulo era testar o método de extração de características proposto com os classificadores mencionados e comparar ao *baseline* em relação à desempenho de classificação e custo computacional para escolher o método que melhor atende a estes dois requisitos que são de extrema importância para a implementação de sistemas que executem em *hardware* com baixa capacidade de processamento. Os experimentos revelaram dois métodos baseados no método de extração de características proposto com potencial para o uso em *smartphones* que são o **H+JSD(SVM)** e **H+JSD(NB)** ambos com pré-processamento por possuírem desempenho de classificação próximos ao BoW considerando as métricas F_1 e $F_{0,5}$, com custo computacional inferior. No capítulo 6 a seguir, são feitas as conclusões finais deste trabalho, relacionando o resultado do estudo feito com os objetivos propostos no capítulo 1, apontando dificuldades enfrentadas durante o desenvolvimento da pesquisa e direcionando trabalhos futuros para a continuidade deste estudo.

6 Conclusões

Neste trabalho abordamos o problema de detectar conversas online de pedofilia em ambientes com *hardware* de baixa capacidade de processamento, nossa hipótese era que a elaboração de um método de extração de características de texto utilizando quantificadores de teoria da informação para condensar um vocabulário em apenas seis características de de informação, seria capaz de diminuir o custo computacional do processo de classificação das conversas sem causar muita perda na qualidade das predições. Por este motivo conduzimos nossos experimentos para validarmos esta hipótese e submetemos o método proposto, que referenciamos como H+JSD e o *baseline*, que referenciamos como BoW, a testes que envolviam a análise da qualidade de classificação com e sem pré-processamento dos dados e análise do tempo de execução com análise de complexidade, o H+JSD foi utilizado em conjunto com os classificadores SVM, árvore de decisão, Naive Bayes e KNN e o BoW apenas com o SVM. Para permitir uma experimentação justa, fizemos ajustes na arquitetura do BoW como a inclusão da etapa de pré-processamento, o balanceamento da base de dados e a exclusão da etapa de filtro de amostras, com o objetivo de tornar ambas as arquiteturas H+JSD e BoW compatíveis, em seguida fizemos pré-experimentos para descobrir empiricamente o tamanho igual a 5.000 para o vetor de características do BoW que também se aplica aos histogramas de referência do H+JSD, também fizemos ajustes de parâmetro nos classificadores SVM e KNN para que eles alcancem resultados próximos a 90,0%, para o SVM esta condição foi satisfeita com *kernel* Gaussiano com $\gamma = 10$ e o KNN com $K = 6$ e distância Euclidiana. Por fim fizemos os experimentos de classificação e tempo de execução utilizando vocabulário fixo e variável, obtida por meio de diferentes amostras da base de dados, comparando o H+JSD e o BoW e elegemos dois métodos baseados em H+JSD que possuem o melhor custo-benefício em eficácia e eficiência.

Verificamos nos experimentos de qualidade (eficácia) de classificação que a etapa de pré-processamento dos dados melhora as métricas de classificação dos métodos, isso ocorre porque a presença das palavras funcionais desestabiliza os valores de entropia pois elas possuem frequência alta e fraco valor semântico, fazendo com que as conversas no geral tenham entropias mais altas em relação ao uso de dados pré-processados, no caso da divergência estas palavras causam interferência na morfologia dos histogramas de referência e particulares prejudicando a interpretação do cálculo de similaridade. Verificamos também que todos os classificadores que utilizaram as características produzidas pelo H+JSD foram mais rápidos (eficientes) na execução que o BoW, isso ocorreu porque para qualquer base de dados de conversas o H+JSD produz apenas seis características, sendo três relacionadas a entropia e três relacionadas à divergência, como consequência, na análise de complexidade desses métodos, todo o custo computacional relacionado a quanti-

dade de características se tornou constante, sendo assim o custo do método H+JSD(SVM) ficou $O(t^2)$ no treino, onde t é o tamanho da coleção de treino e no teste $O(v)$, onde v é o tamanho da coleção de teste. o custo do H+JSD(DT) ficou $O(t)$ no treino e $O(v)$ no teste, o H+JSD(NB) ficou $O(t)$ no treino e no teste $O(v)$. como O KNN não possui treino e todo o custo deste algorítimo está no teste, então o custo do H+JSD(KNN) é $O(vx+v)$, sendo x a quantidade de vizinhos próximos que nos pré-experimentos definimos o valor 6, e por fim o custo do BoW no treino é $O(kt^2)$ onde k é a quantidade de características que representa cada instância e no teste é $O(kv)$, ou seja, todos os métodos baseados em H+JSD tem custo linear no treino e no teste, com exceção do H+JSD(SVM) que possui custo quadrático no treino mas que ainda assim é menor que o custo cúbico do BoW. Para um sistema de monitoramento, o treino do modelo para detectar conversas de pedofilia poderá ser feito *offline*, então o custo mais relevante é o de teste (classificação).

Por fim os métodos baseados em H+JSD H+JSD(SVM) e H+JSD(NB) se destacaram nos experimentos de classificação e tempo de execução. O H+JSD(SVM) teve valores de F_1 e $F_{0,5}$ na classe pedofilia iguais a 88,6% e 89,7% apenas 4,00% (0,886 contra 0,923) e 4,47% (0,897 contra 0,939) inferiores ao BoW respectivamente e na classe regular o F_1 e $F_{0,5}$ foram iguais a 89,9% e 88,8% apenas 4,86% (0,899 contra 0,945) e 4,31% (0,888 contra 0,928) inferiores ao BoW respectivamente, o tempo de classificação de uma conversa com aproximadamente 2.500 palavras, o H+JSD(SVM) foi **72,8%** mais rápido que o BoW (10,5s contra 38,6s) e continuará mais rápido caso o vocabulário aumente assim como demonstrado nos experimentos do capítulo 5. O H+JSD(NB) teve valores de F_1 e $F_{0,5}$ na classe pedofilia iguais a 86,5% e 87,1% apenas 6,28% (0,865 contra 0,923) e 7,24% (0,871 contra 0,939) inferiores ao BoW respectivamente e na classe regular o F_1 e $F_{0,5}$ foram iguais a 86,8% e 86,3% apenas 8,14% (0,868 contra 0,945) e 7,00% (0,863 contra 0,928) inferiores ao BoW respectivamente, o tempo de execução considerando os mesmos parâmetros do H+JSD(SVM) resulta em 88,9% (4,27s contra 38,6s) de maior eficiência em relação ao BoW. A diferença de qualidade na classificação entre o H+JSD(SVM) e H+JSD(NB) considerando o F_1 e $F_{0,5}$ na classe pedofilia, o H+JSD(NB) é 2,37% (0,865 contra 0,886) e 2,89% (0,871 contra 0,897) inferior ao H+JSD(SVM) respectivamente e na classe regular ele é 3,44% (0,868 contra 0,899) e 2,81% (0,863 contra 0,888) inferiror respectivamente, mas classifica uma conversa 59,3% (4,27s contra 10,5s) mais rápido que o H+JSD(SVM) nas mesmas condições especificadas anteriormente, por este motivo o **H+JSD(NB)** é a melhor escolha para implementação em *smartphones*.

6.1 Considerações Finais

Apresentamos uma experimentação variada que nos levou a eleger um método identificação de conversas de pedofilia em redes sociais de mensagens instantâneas. O método H+JSD(NB) baseado no método de extração de características proposto H+JSD alcança

valores de F_1 e de $F_{0,5}$ próximos a $90,0\%$, comparados a $94,0\%$ do estado-da-arte (BoW). Contudo, conforme demonstramos, o BoW não é escalável e seu custo computacional é proibitivo para dispositivos móveis. Em contraste, o H+JSD(NB) é escalável, com um custo computacional reduzido comparado ao BoW. Esta escalabilidade é um requisito chave para implementação de filtros de pedofilia para 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. O H+JSD é significativamente mais eficiente, seu custo de classificação é linear, comparado ao custo quadrático do BoW. Em termos de tempo, o H+JSD(NB) chegou a ser 88,9% mais eficiente que o BoW e 59,3% mais eficiente que o H+JSD(SVM) que foi o segundo método que se destacou nos experimentos. Todavia, ainda há espaço (e necessidade) para aumento da eficiência (redução do custo computacional) e da eficácia (qualidade da classificação).

6.2 Limitações do Método

Primeiramente, uma limitação crítica do método é a carência de dados do tipo pedofilia para a construção do modelo de aprendizagem de máquina devido a questões éticas e judiciais. A única fonte cientificamente válida de *chats* de pedofilia é o site da organização *Perverted Justice*, com aproximadamente 600 conversas, e a consequência da escassez de dados no método são as deficiências na capacidade de generalizar o conhecimento para a classe de pedofilia.

O segundo ponto é o fato de que o *parser* de *stanford* também comete erros no *POS Tagging*, ele nao possui uma categoria para abrigar palavras que não se encaixam em nenhuma classe gramatical, como por exemplo as gírias, os *emojis* e as palavras escritas incorretamente, sendo assim, estas palavras acabam sendo associadas arbitrariamente a classes gramaticais válidas para o estudo e acabam se tornando um ruído difícil de tratar.

6.3 Trabalhos Futuros

Uma implementação futura objetivando maior eficácia, é o uso de outras métricas de dissimilaridade com maior capacidade discriminativa que a divergência de Jensen-Shannon para o problema de identificação de conversas de pedofilia. Por esse motivo, 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. Outro trabalho futuro importante consiste na adaptação do método para operar de forma incremental (linha-a-linha). Uma versão incremental linha-a-linha proporcionará a identificação mais veloz de uma conversa de pedofilia, permitindo o seu bloqueio imediato e, consequentemente, reduzindo a exposição das vítimas

aos pedófilos, ou seja, o desenvolvimento de um método que execute em tempo real. Também seria interessante a inclusão de um *parser* gramatical português para que o método seja capaz de detectar conversas de pedofilia em nosso idioma.

Referências

BARZILAY, R.; ELHADAD, M. Using lexical chains for text summarization. *Advances in automatic text summarization*, p. 111–121, 1999.

BOGDANOVA, D.; ROSSO, P.; SOLORIO, T. 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*. Jeju, Korea: Association for Computational Linguistics, 2012. p. 110 – 118.

CHEONG, Y. et al. Detecting predatory behavior in game chats. *Transactions on Computational Intelligence and AI in Games*, IEEE, v. 7, n. 3, p. 220 – 232, 2015.

CHIN, A.; ZHANG, D. *Mobile Social Networking-An Innovation Approach*. [S.l.]: Springer, 2013.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273 – 297, 1995.

HANCOCK, J. T. et al. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, Taylor & Francis, v. 45, n. 1, p. 1 – 23, 2007.

JAGGI, M. An equivalence between the lasso and support vector machines. *Regularization, Optimization, Kernels, and Support Vector Machines*, CRC Press, v. 2, n. 1, p. 1 – 26, 2014.

KAMIŃSKI, B.; JAKUBCZYK, M.; SZUFEL, P. A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*, Springer, p. 1–25, 2017.

KONTOSTATHIS, A.; EDWARDS, L.; LEATHERMAN, A. *Text Mining and Cybercrime*. West Sussex, United Kingdom: John Wiley & Sons, Ltd, 2010. 149 – 164 p.

LANNING, K. V.; CHILDREN, N. C. for M. . E. et al. *Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children*. Virginia, USA: National Center for Missing & Exploited Children with Office of Juvenile Justice and Delinquency Prevention, 2010. 212 p.

LEWIS, D. D. Naive (bayes) at forty: The independence assumption in information retrieval. In: SPRINGER. *European conference on machine learning*. [S.l.], 1998. p. 4–15.

LIU, W.; CHAWLA, S. *Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. 345 – 356 p.

LIVINGSTONE, S. et al. Risks and safety on the internet: the perspective of european children. *Full Findings*, LSE: EU Kids Online, London, United Kingdom, 2010.

MACHADO, E. L. Um estudo de limpeza em base de dados desbalanceada e com sobreposiçao de classes. 2009.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008. 506 p.

MANNING, C. D.; SCHÜTZE, H. et al. Foundations of statistical natural language processing. MIT Press, v. 999, 1999.

MIHALCEA, R.; STRAPPARAVA, C. The lie detector: Explorations in the automatic recognition of deceptive language. In: *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP), Conference Short Papers*. Suntec City, Singapore: Association for Computational Linguistics, 2009. v. 3, n. 1, p. 309 – 312.

MORRIS, C.; HIRST, G. 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*. Rome, Italy: The CLEF Initiative, 2012. v. 12, n. 1, p. 1 – 29.

NEWMAN, M. L. et al. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, Sage Publications, v. 29, n. 5, p. 665 – 675, 2003.

NIVRE, J. et al. *Universal Dependencies 1.2*. Universal Dependencies Consortium, 2015. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. Disponível em: ⟨http://hdl.handle.net/11234/1-1548⟩.

PANG, B.; LEE, L. et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008.

PARAPAR, J.; LOSADA, D.; BARREIRO, A. 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*. Rome, Italy: The CLEF Initiative, 2012. v. 12, n. 1, p. 1 – 12.

PEERSMAN, C. et al. 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*. Rome, Italy: The CLEF Initiative, 2012. v. 12, n. 1, p. 1 – 13.

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

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, Erlbaum Publishers, v. 71, n. 3, p. 1 – 23, 2001.

PEREIRA, M. J. S. Natural language processing for safe products. *Computer Speech & Language*, Elsevier, v. 15, n. 4, p. 403–434, 2014.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011.

REIS, J. et al. Uma análise do impacto do anonimato em comentários de notícias online. In: *Anais do 13o. Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*. [S.l.]: SBC, 2016. p. 1290–1304.

ROBERTSON, S. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, City University, London, UK, v. 60, n. 5, p. 503 – 520, 2004.

ROSSO, O. A.; CRAIG, H.; MOSCATO, P. Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 388, n. 6, p. 916 – 926, 2009.

ROSSO, O. A.; OSPINA, R.; FRERY, A. C. Classification and verification of handwritten signatures with time causal information theory quantifiers. *PloS one*, Public Library of Science, v. 11, n. 12, p. e0166868, 2016.

SALTON, G.; MICHAEL, J. Mcgill. *Introduction to modern information retrieval*, McGraw Hill, New York, p. 24–51, 1983.

SANTORINI, B. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). 1990.

SILVA, C. et al. 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)*. [S.l.]: SBC, 2016. p. 1245–1259.

STUBBS, M. *Words and phrases: Corpus studies of lexical semantics*. [S.l.]: Blackwell publishers Oxford, 2001.

VILLATORO-TELLO, E. et al. 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*. Rome, Italy: The CLEF Initiative, 2012. v. 12, n. 1, p. 1 – 12.

ZHANG, M.-L.; ZHOU, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, Elsevier, v. 40, n. 7, p. 2038–2048, 2007.