Modelagem de Aliciamento de Menores em Mensagens Instantâneas de Texto

Priscila L. L. Santin, Cinthia O. A. Freitas, Emerson C. Paraíso, Altair O. Santin

Pontifícia Universidade Católica do Paraná (PUCPR) – Escola Politécnica – Programa de Pós-Graduação em Informática (PPGIa) – Curitiba – PR – Brasil

{priscila.santin, cinthia, paraiso, santin}@ppgia.pucpr.br

Abstract. *The approaches presented in the literature are not suitable for detection of children sexual grooming, because the proposals are usually limited to the identification of stages of communication between the predator and the victim. These approaches are inefficient due to the use of a unified profile to identify a sexual grooming stage. Our proposal takes into account the identification of the communication's stage applying detached profiles, one for the victim and another for the predator. This approach, based on a stochastic technique (i.e. HMM), aims at the individual modeling of each profile to enhance the detection hit rate. The experiments achieved promising identification rates with results closer to 91%.*

Resumo. *As abordagens existentes na literatura não são modeladas para detecção do aliciamento sexual de menores, mas apenas fazem a descoberta do estágio de aliciamento numa comunicação entre o agressor e sua vítima. Além disto, mostram baixa eficiência em função do emprego de um perfil único para a descoberta dos estágios. Este artigo considera a descoberta dos estágios de aliciamento com o perfil do agressor e da vítima separadamente. Esta abordagem, baseada em avaliação estocástica (i.e. HMM), visa a modelagem individual de cada perfil para aumentar a eficiência da detecção. Os experimentos mostram taxas de acerto promissoras com resultados próximos a 91%.*

1. Introdução

A *Internet* oferece inúmeros recursos para interação entre indivíduos das mais variadas faixas etárias. Atualmente, crianças (pessoas de até 12 anos de idade incompletos) e adolescentes (pessoas entre 12 e 18 anos) tomam contato com a rede mundial de computadores ainda nos seus primeiros anos de vida. A partir do uso rotineiro da *Internet,* estes menores de idade vão assimilando os diferentes tipos de recursos à medida que passam pelas diferentes fases de seu crescimento. A interatividade fornecida por recursos como salas de bate-papo (*chats*) traz para os menores uma espécie de local onde encontram todas as respostas para as suas dúvidas e curiosidades, porém, em suas visões, sem se expor a "censura" tradicional dos familiares.

Indivíduos mal intencionados conhecem esta tendência comportamental das crianças e adolescentes e usam as salas de bate-papo para o aliciamento (*grooming/enticement*) de menores com o intuito de praticar abuso sexual [Rashid 2008].

Este trabalho é parcialmente apoiado pela Rede Marista de Solidariedade – Projetos Sociais Nº 01/2010.

Nos Estados Unidos (EUA), o Centro Nacional para Crianças Desaparecidas e Exploradas (*National Center for Missing and Exploited Children – NCMEC*) recebe informações de casos de exploração sexual de menores, incluindo o aliciamento *online*. Nesta categoria de aliciamento *online*, desde 9 de março de 1998 até o final de 2011 já haviam sidos informados 54.492 casos ao NCMEC, sendo que em 2011 a média foi de 77 casos por semana [NCMEC 1984].

No *site* brasileiro sobre a Campanha Nacional de Combate à Pedofilia na *Internet*[1] encontra-se que: "Também constam dados de uma pesquisa realizada nos EUA, dizendo que de cada 5 crianças que navegam na *Internet*, uma recebeu proposta de um pedófilo, e uma a cada 33 já se comunicou, através de telefone e recebeu dinheiro ou passagem para se encontrar com um criminoso". O *site* ainda alerta: "Pais e filhos, inconscientes dos perigos da rede são presas fáceis de pedófilos. Uma criança ingenuamente não identifica um adulto se passando por um amiguinho da mesma idade".

Em 2006, o Brasil já ocupava o terceiro lugar entre os países do mundo em número de denúncias de crimes *online* que ferem os direitos humanos, tais como: racismo, intolerância religiosa e pornografia infantil. Nesta época, tinham ocorrido em 20 dias um total de 2,25 mil denúncias, como declarado por Thiago Tavares, presidente da SaferNet, uma organização não-governamental (ONG) que luta contra crimes virtuais, durante a cerimônia de assinatura do acordo entre a entidade e o Ministério Público Federal (MPF) no dia 29 de março de 2006 em São Paulo[2]. Em janeiro de 2012, a violência de abuso sexual de menores (crianças e adolescentes), segundo a SaferNet[3], representava 60% dos casos reportados.

Diante desta problemática, pesquisadores têm estudado e desenvolvido técnicas para analisar conversas em salas de bate-papo [Kontostathis et al. 2009a] [Ho e Watters 2004] [Kontostathis et al. 2009b] [McGhee et al. 2011], de modo a auxiliar na descoberta dos perfis de agressores e no estabelecimento e identificação dos estágios de comunicação que estes agressores utilizam no contato com menores. A definição de tais perfis tem o intuito de contribuir para detectar a agressividade do aliciamento sexual e consequente a susceptibilidade ao abuso.

O trabalho de Kontostathis [Kontostathis et al. 2009a] relata que a identificação dos estágios de aliciamento é prejudicada quando considerado o uso do modelo do agressor e da vítima em conjunto. Diante disso, tem-se como objetivo estudar a correlação entre os dois modelos com intuito de superar esta limitação. Assim, este artigo propõe um método para análise de mensagens instantâneas de texto, realizadas em salas de bate-papo na *Internet,* visando a modelagem do perfil do agressor e da vítima separadamente. Esta modelagem é baseada na identificação dos estágios do diálogo entre um agressor e sua possível vítima. Esta abordagem permite detectar a correlação (*match*) entre os estágios de cada perfil com intuito de melhorar a precisão da detecção de aliciamento e ainda, de maneira análoga, inferir a susceptibilidade (probabilidades) a exposição das vítimas ao risco do abuso sexual.

Este artigo está organizado de tal forma que a Seção 2 traz uma introdução dos

[1] *http://www.censura.com.br/*

[2] *Disponível em: http://www.safernet.org.br/twiki/bin/view/SaferNet/Noticia20060329194811. Acesso em julho/2009.*

[3] *http://www.safernet.org.br/site/indicadores*

conceitos fundamentais estudados para a realização do trabalho proposto. A Seção 3 apresenta os trabalhos relacionados. Na Seção 4 encontra-se descrito o método proposto para o desenvolvimento deste trabalho. A Seção 5 apresenta os testes experimentais e os resultados obtidos. Por fim, a Seção 6 aborda conclusões e trabalhos futuros.

2. Fundamentação

Esta seção apresenta a fundamentação para aliciamento e modelos estocásticos Markovianos.

2.1. Aliciamento

O termo de consenso para se referir ao aliciamento na área computacional é *grooming* ou *enticement* [Michalopoulos e Mavridis 2010]. Neste caso, o aliciamento sexual de menores se refere ao ato de preparar a vítima, por meio de salas de bate-papo, via troca de mensagens instantâneas para cometer um abuso sexual [Michalopoulos e Mavridis 2010].

Pelas características de interatividade e recursos que os meios computacionais atuais oferecem o *grooming/enticement* ocorre com muita frequência em salas de bate-papo. Neste caso, a detecção do aliciamento precisa ser feita *online*, avaliando as mensagens de texto que trafegam entre o possível agressor e a possível vítima [Michalopoulos e Mavridis 2010]. A detecção do aliciamento por meio da avaliação das mensagens de textos de conversas de salas de bate-papo pode ser executada identificando os estágios da conversa que o agressor utiliza para conquistar sua vítima com o objetivo de se relacionar sexualmente com a mesma. A seguir apresenta-se a definição dos estágios de comunicação para aliciamento encontrados na literatura e depois sua adaptação para a área de computação.

2.1.1 Estágios do Aliciamento

Para entender o aliciamento deve-se entender como este ocorre, quais seus estágios e como estão caracterizados. Assim, com base nos estudos da teoria de comunicação sedutiva, Olson [Olson et al. 2007] propôs um modelo de comunicação que descreve o perfil do agressor identificando seus estágios.

Este modelo apresenta três fases mais relevantes: a persuasão (ganho de acesso) à vítima, o envolvimento da vítima numa relação enganosa e a iniciação e manutenção de um relacionamento sexualmente abusivo (Figura 1).

Olson mostra que a primeira fase, persuasão (ganho de acesso à vítima), tem três divisões. Duas referem-se à obtenção de características individuais do agressor e da vítima, que correspondem basicamente a informações como local de residência, idade, sexo etc. A terceira é o posicionamento estratégico – definido como um encontro de curto prazo, em *shopping* ou parques, ou um encontro de longo prazo, onde o agressor constrói um relacionamento com a vítima.

A segunda fase do modelo de comunicação de Olson trata de um ciclo para envolver a vítima numa relação enganosa. A preparação (*grooming/enticement*) consiste nas estratégias de comunicação que o agressor utiliza para que a vítima aceite o contato sexual. O desenvolvimento de uma verdade enganosa é entendido pelos autores deste modelo como a capacidade do agressor em cultivar um relacionamento de amizade com

sua vítima, podendo também incluir seus familiares. Normalmente o agressor é visto como alguém importante ou uma autoridade, sendo que seu comportamento é pouco questionado pela vítima e seus parentes. Então vem o isolamento, que pode ser apresentado tanto como um isolamento físico, em que o agressor se propõe a tomar conta da vítima, como um isolamento mental, onde o agressor se mostra para a vítima como a única pessoa em quem essa pode confiar. Finalizando este ciclo tem-se a aproximação, que é definida por Olson [Olson et al. 2007], como o contato físico inicial ou introdução verbal que ocorre antes do ato sexual.

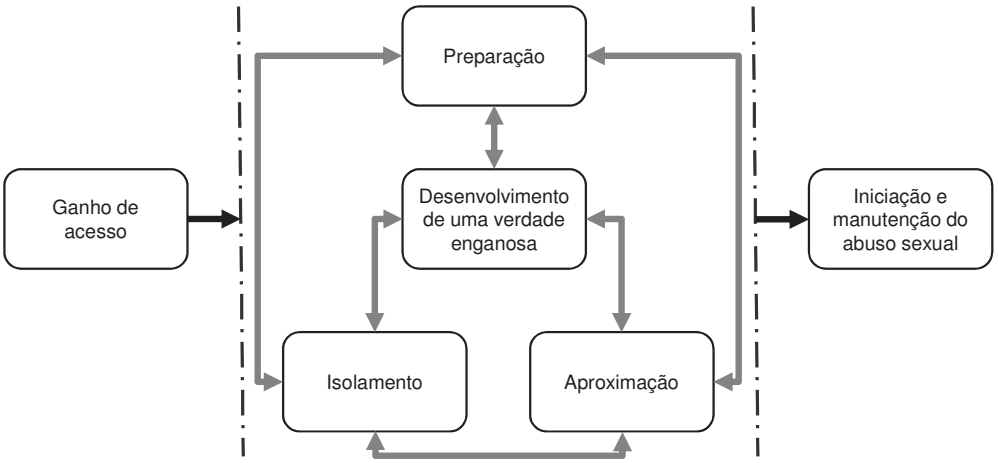

Figura 1 - Adaptação do Modelo de Comunicação [OLS07]

Na terceira fase de Olson tem-se a iniciação e manutenção do abuso sexual. A iniciação consiste em realmente praticar o ato sexual. Então, torna-se importante para o agressor o sigilo da vítima. Normalmente os agressores obrigam as vítimas ao silêncio coagindo-as ou simplesmente convencendo-as de que ao relatar o abuso não serão acreditadas ou até mesmo poderão causar um desconforto familiar.

A seguir é mostrado como as fases do modelo de Olson foram adaptados para compor os estágios do aliciamento sexual de menores em ambientes virtuais.

2.1.2 Estágios do Aliciamento para o Contexto Virtual

Leatherman [Leatherman 2009] fez uma adaptação para o contexto virtual do modelo de comunicação proposto por Olson [Olson et al. 2007]. O resultado de seu trabalho são os estágios descritos sumariamente a seguir:

- Acesso: envolve a troca de informações pessoais tanto do agressor quanto da vítima. No aliciamento *online* o agressor ganha acesso (*persuade*) à vítima usando recursos baseados em mensagens instantâneas ou salas de bate-papo *online* e *sites* de redes sociais;
- Desenvolvimento de um confiança enganosa: refere-se a capacidade do agressor cultivar relacionamentos com potenciais vítimas e possivelmente com suas famílias para se benficiar em seus interesses sexuais. No ambiente *online* essa confiança é praticada com base na utilização de quatro recursos:
  - ✓ *Informações Pessoais*: consiste na obtenção de detalhes sobre a vítima – idades, nomes, data de aniversários, fotos etc;
  - ✓ *Informações de Relacionamento*: considera a discussão de sentimentos e atitudes em relação à manutenção, construção e dissociação de relações com amigos, outras pessoas importantes e membros da família;

- ✓ *Atividades*: é o recurso mais utilizado pelo agressor para se aproximar da vítima compartilhando com esta preferências por músicas, filmes, livros, esportes, *hobbies*, comidas favoritas etc;
- ✓ *Elogios*: consiste em envolver a vítima elogiando sua aparência, as atividades que desenvolve e caracteríticas pessoais.

- Preparação: compreende estratégias sutis de comunicação que os agressores utilizam para preparar suas potenciais vítimas a aceitar aspectos de conotação sexual. Conquistando a confiança da vítima, o agressor começa a prepará-la para aceitar ofertas de contato sexual, e também tornar a vítima menos sensível a comentários sexuais e linguagens obcenas. Há duas etapas de preparação para o aliciamento:
  - ✓ *Dessensibilização Comunicativa*: refere-se ao uso frequente de linguagem sexual vulgar, utilizada pelo agressor para dessensibilizar a vítima sobre seu uso;
  - ✓ *Reenquadramento*: ocorre quando o agressor se esforça para tornar a vítima confortável com experimentos sexuais por meio da *Internet*. Para este fim, a conversa sexual é apresentada de uma forma positiva e muitas vezes referida como uma experiência de aprendizagem, um jogo ou uma habilidade importante para aprender a fim de participar de relacionamentos amorosos futuros.
- Isolamento: pode envolver isolamento físico ou mental. O isolamento físico é definido como a dedicação de um tempo para ficar a sós com a vítima, já o isolamento mental é definido como o aumento da dependência que a vítima tem do agressor com relação a amizade e orientação. O agressor consegue o isolamento quando a vítima está em salas de bate-papo sem supervisão;
- Aproximação: o agressor tenta a abordagem da vítima, sugerindo encontros para fins sexuais. Esta é a etapa final no modelo *online*.

A caracterização destes estágios é de suma importância para o entendimento do modo de agir dos agressores visando o aliciamento sexual, bem como, para o desenvolvimento de sistemas computacionais para detecção deste tipo de problema. Considera-se que qualquer abordagem preventiva é relevante, pois na maioria dos casos no mundo real a abordagem é somente punitiva. Neste contexto, as ferramentas computacionais podem atuar preventivamente e auxiliar na detecção antecipada da intenção do abusador (agressor).

2.2. Modelos Escondidos de Markov

Os Modelos Escondidos de Markov (HMM – *Hidden Markov Models*) possuem habilidade para modelar eficientemente diferentes fontes de conhecimento, ou seja, permitem a integração entre dois diferentes níveis de modelagem: estrutural e probabilística. Para a modelagem de aliciamento tal característica é essencial, pois permite a entrada de sequências de observações O de tamanhos diversos, o que não ocorre em outros classificadores, tais como: SMV (*Support Vector Machine*) ou RN (Redes Neurais).

Esta é a razão da designação "escondido" ao modelo de Markov, pois não se sabe quantas observações O_i correspondem a cada estado no modelo Markoviano treinado. A formulação teórica de HMM está além do escopo deste artigo, porém uma excelente introdução ao assunto pode ser encontrada em [Rabiner e Juang 1993].

Assim, o HMM integra corretamente os diferentes níveis de modelagem e também fornece algoritmos eficientes para determinar valores ideais para os parâmetros do modelo. A modelagem Markoviana assume que um modelo é representado por uma sequência de observações. Estas observações devem ser estatisticamente independentes uma vez que a sequência oculta (*hidden*) de estado subjacente é conhecida. Primeiramente deve ser estabelecida uma estrutura para o modelo e então se utiliza de estimativa de parâmetros para melhorar a probabilidade de geração de dados de treinamento por parte desses modelos. O HMM absorve a sequência de observação durante a fase de treinamento, logo é capaz de associar cada estado de acordo com alguma função de densidade de probabilidade.

Os modelos de HMM para detecção de aliciamento sexual de menores são baseados em uma abordagem global e uma topologia ergódica [Rabiner 1989]. Ou seja, cada nó do modelo pode ser alcançado a partir de um outro nó qualquer por um número finito de passos.

A formação do modelo baseia-se no algoritmo de *Baum-Welch* [Rabiner e Juang 1993]. Neste trabalho, o processo de classificação consiste em determinar a máxima probabilidade *a posteriori* para as linhas da conversação do agressor ou da vítima, sendo que cada linha da conversa w gera uma sequência de observações O não conhecidas pelo processo de treinamento (ou seja, os exemplos de treinamento são distintos dos exemplos de teste). Sendo válida a seguinte equação:

$$\Pr(\hat{w} \mid O) = \max_{w} \Pr(w \mid O) \tag{1}$$

Aplicando-se o teorena de Bayes, tem-se a equação fundamental de reconhecimento de padrões:

$$\Pr(w \mid O) = \frac{\Pr(O \mid w).\Pr(w)}{\Pr(O)} \tag{2}$$

Então, sabendo-se que $\Pr(O)$ não depende de w, a classificação torna-se equivalente a maximizar a probabilidade conjunta dada por:

$$\Pr(w, O) = \Pr(O \mid w).\Pr(w) \tag{3}$$

Finalmente, tem-se que $\Pr(w)$ é a probabilidade *a priori* da conversa w e está relacionada com o problema modelado. A estimativa de $\Pr(O|w)$ requer um modelo probabilístico que seja capaz de considerer as variações das conversas w representadas pelas sequências de observações O. Durante os experimentos, o confronto dos *scores* entre cada modelo e as sequências de observações dos conjuntos de teste foi gerado pelo algoritmo de *Viterbi* [Rabiner e Juang 1993].

3. Trabalhos Relacionados

A seguir serão apresentados os trabalhos relacionados com o tema de modo a evidenciar as abordagens utilizadas nestes trabalhos. As abordagens podem estar baseadas no uso de técnicas com filtro de palavras, regras e classificadores.

3.1. Detecção dos estágios baseada em filtro de palavras

No estudo de Kontostathis [Kontostathis et al. 2009a] é desenvolvido um *software* denominado *ChatCoder1* que faz a busca dos estágios de comunicação adaptados por [Leatherman 2009] em transcrições de conversas de salas de bate-papo. Naquele

trabalho, os autores usaram como base de dados 288 transcrições de conversas existentes em [Perverted 2003]. Este *site* está descrito na Seção 4, pois a base de dados utilizada nos experimentos realizados também provém deste *site*. Kontostathis desenvolveu um léxico de palavras e um manual de regras para definir os estágios do modelo de comunicação. O léxico foi desenvolvido utilizando-se de termos de aliciamento, palavras, ícones, frases e linguagem virtual para cada estágio do modelo. O *software ChatCoder1* captura os diálogos e já classifica, por meio do uso do léxico, as frases em seus estágios. Porém, se as palavras ou frases não fossem apresentadas exatamente como no léxico (técnica "caça-palavras"), não eram marcadas como um dos estágios do modelo de comunicação, o que gerava baixa taxa de acerto desta abordagem.

É importante observar que as técnicas de detecção de aliciamento sexual baseadas em análise dos elementos gramaticais de uma frase apresentam-se ineficientes porque os diálogos em salas de bate-papo são muito próximos da linguagem coloquial, que não segue o rigor gramatical da linguagem escrita.

3.2. Detecção dos estágios baseada em regras

Num segundo trabalho, Kontostathis [Kontostathis et al. 2009b] faz o uso de uma técnica baseada em regras para desenvolver o *software ChatCoder2*, uma evolução do *ChatCoder1*. As regras para o desenvolvimento do *ChatCoder2* foram criadas com base na transcrição de 15 conversas disponíveis em [Perverted 2003]. O algoritmo do *ChatCoder2* foi desenvolvido utilizando-se das regras e análise da composição gramatical das frases do diálogo para cada um dos estágios do modelo de comunicação adaptados por Leatherman [Leatherman 2009]. Também foram acrescentados valores nominais para os estágios do modelo para que se uma palavra ou frase estivesse em mais de em estágio, seria considerada no estágio com maior valor nominal. Para concluir que o método baseado em regras proposto por Kontostathis apresentava resultados melhores que o método "caça-palavras" (*ChatCoder1*), os autores calcularam a confiabilidade entre os codificadores (*Intercoder Reliability*). Usando o método de [Holsti 1969] foram feitas comparações entre dois codificadores humanos, o *ChatCoder1* e o *ChatCoder2*, e os autores avaliaram ser possível obter uma melhora na confiabilidade máxima de 13,21% e na média de 5,81%.

Porém cabe ressaltar que as técnicas de detecção baseadas em regras são dependentes da linguagem cotidiana analisada e dos vícios de linguagem de seus interlocutores.

3.3. Detecção dos estágios baseada em classificadores

McGhee e seus colegas [McGhee et al. 2011] fazem uso de algoritmos de aprendizagem de máquina para classificar as linhas de conversas de salas de bate-papo. Neste estudo, o modelo de comunicação adaptado por Leatherman [Leatherman 2009] foi consolidado em apenas quatro estágios: (1) troca de informações pessoais – inclui informações pessoais, gostos/preferências e sentimentos, e aspecto de construção ou manutenção do relacionamento entre o agressor e a vítima; (2) preparação – coincide com o proposto por Olson; (3) aproximação – envolve a obtenção de telefone, endereço, marcação de encontro, compartilhamento de segredos e o isolamento da vítima e (4) outros – enunciados que não estão contemplados nos itens anteriores são classificados como do quarto estágio. Para este estudo foram utilizadas 33 transcrições de conversas

disponíveis em [Perverted 2003] como base. Três conversas eram comuns a ambos os codificadores e foram utilizadas para verificar a concordância de classificação dos codificadores. Como o número de estágios do modelo de comunicação foi reduzido neste trabalho, as regras para marcar cada linha do diálogo foram também adaptadas. O léxico utilizado em Kontostathis [Kontostathis et al. 2009b] foi também usado neste trabalho para desenvolver os atributos de entrada do algoritmo de aprendizagem de máquina. Neste trabalho os autores utilizaram a precisão (*Accuracy*) como métrica de avaliação para a comparação entre os vários experimentos realizados com diversos algoritmos de aprendizagem de máquina. Quando as 33 conversas foram submetidas aos algoritmos individualmente, pode-se dizer que a aprendizagem de máquina apresentou um melhor desempenho se comparado às regras. Quando as conversas formaram um único bloco de entrada o desempenho dos classificadores foi muito semelhante ao método baseado em regras, porém os atributos apresentados na forma numérica tiveram um desempenho um pouco melhor.

As técnicas baseadas em classificadores sofrem com problemas de seleção das características, palavras ou frases, de modo a formar um léxico válido e representativo do problema em questão. Além disto, uma mesma palavra ou frase pode ser encontrada em diferentes estágios do processo de aliciamento, dificultando assim a classificação ou identificação do problema.

4. Proposta de Método para Modelagem de Aliciamento

A proposta está baseada na identificação dos estágios de comunicação de uma conversa entre um possível agressor e sua vítima, bem como na definição dos modelos que representam o perfil do agressor e da vítima. A Figura 2 ilustra uma visão geral da proposta.

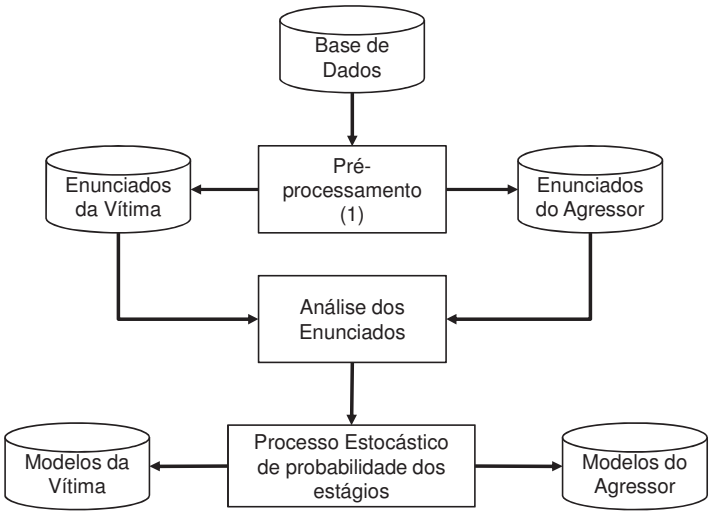

Figura 2 – Visão geral da proposta

Os esforços inicialmente se concentraram na formação da base de dados de conversações que está baseada no *site Perverted-justice.com* [Perverted 2003]. O *site* disponibiliza em seu conteúdo público transcrições de conversas de salas de bate-papo que levaram os agressores à prisão. Voluntários do *Perverted Justice* se passam por menores em salas de bate-papo com intuito de encontrar possíveis agressores que tinham como objetivo iniciar um abuso sexual de menores. Quando esse agressor é

preso e tem sua sentença definida por um juiz de direito, a transcrição da conversa é tornada publica no *site*. Um exemplo de conversação é mostrado na Figura 3, onde se pode observar os enunciados do agressor (cobbler1976) e da vítima (that_polish_chick94).

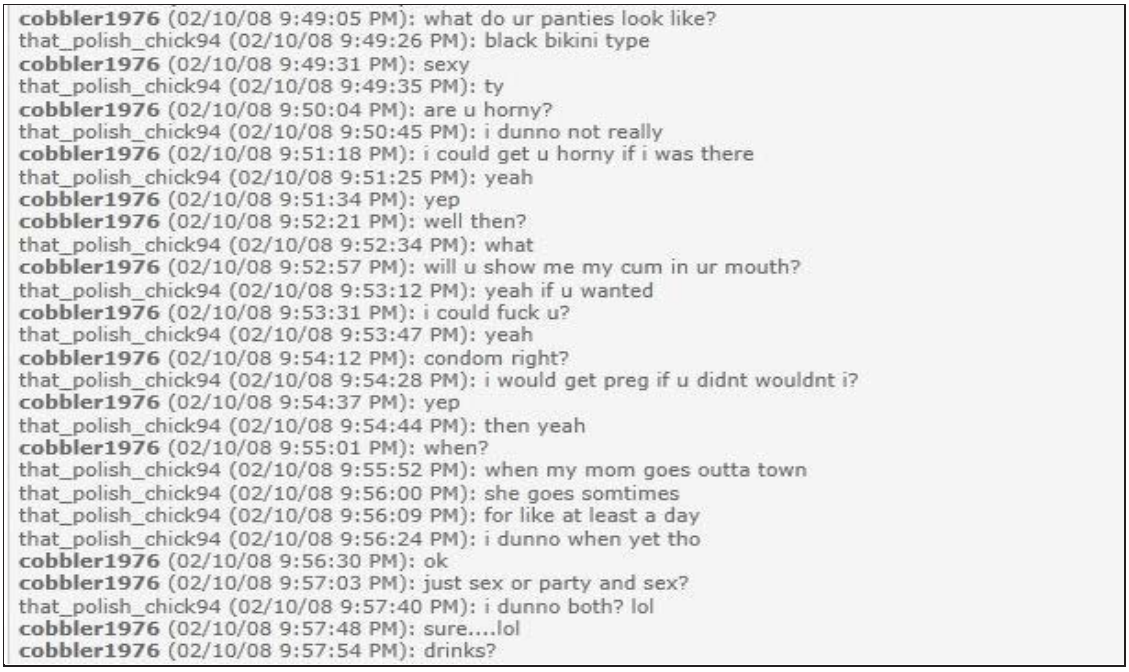

Figura 3 - Visão geral das transcrições das conversas [Perverted 2003]

A segunda fase do trabalho consistiu no pré-processamento do texto das transcrições das conversações, separando as linhas da conversa em enunciados do agressor e enunciados da vítima. O objetivo da separação dos enunciados é trabalhar com cada um independentemente no intuito de obter os modelos de agressor e vítima, sem que um venha a interferir na obtenção do outro.

Após esta separação de conteúdos as transcrições foram analisadas manualmente para que fosse possível identificar nos enunciados os estágios descritos por [Leatherman 2009].

Os modelos do agressor e da vítima foram definidos usando técnicas estocásticas com a sequência de estágios que as conversas analisadas apresentavam. Estas técnicas constituem os Modelos Escondidos de Markov, que foram utilizados neste trabalho para a modelagem de aliciamento sexual de menores. Com o HMM (*Hidden Markov Models*) se pode obter a probabilidade da ocorrência (susceptibilidade) do aliciamento sem ter todos os estágios da sequência de observação, pois não se sabe *a priori* o tamanho que uma conversação pode atingir, nem quantas destas linhas de enunciados permitirão a identificação de um estágio do aliciamento.

Outra característica importante dos HMMs para o problema do aliciamento é que as observações são geradas independentemente para o agressor e vítima, não havendo interdependência entre os elementos a serem modelados. Por exemplo, em uma mesma conversação, o agressor pode utilizar-se de 1.000 linhas de enunciados e a vítima somente de 900 linhas. Nos experimentos realizados isto não interfere nem no processo de treinamento nem no de classificação, simplesmente tem-se que a cada linha de enunciado é associado um dos nove possíveis estágios definidos por [Leatherman

2009] ou nenhum estágios é observado. Assim, tem-se que o modelo utiliza da estrutura conceitual de [Leatherman 2009], mas não está vinculada a essa. Ou seja, os estágios do aliciamento definidos por [Leatherman 2009] não obrigam que os modelos treinados tenham nove estados, mas sim um número ideal de estados que permita a cada um desses absorver parte da sequência de observações *O* (Seção 2.2.).

5. Testes Experimentais e Análise dos Resultados

Para a realização dos experimentos, 21.989 linhas de enunciados de transcrições de conversas foram analisadas por quatro avaliadores, separando-os em linhas de enunciados do agressor e da vítima. A análise de cada enunciado se fez necessária para identificar os estágios de aliciamento definidos por [Leatherman 2009] para o contexto virtual de aliciamento sexual de menores. Estes 21.989 enunciados foram observados em 20 conversas disponíveis em [Perverted 2003]. Nas conversar analisadas as idades dos agressores variaram entre 19 e 49 anos e das vítimas entre 13 e 15 anos.

Tabela 1 - Representação numérica para os estágios de aliciamento

ESTÁGIO	REPRESENTAÇÃO
Acesso	*1*
Informações Pessoais	*2*
Informações de Relacionamento	*3*
Atividades	*4*
Elogios	*5*
Dessensibilização Comunicativa	*6*
Reenquadramento	*7*
Isolamento	*8*
Aproximação	*9*

Para cada um dos estágios de aliciamento observado foi atribuído um valor numérico conforme ilustra a Tabela 1, para que fosse possível o treinamento usando HMM. Nos enunciados em que não foi possível identificar nenhum estágio foi marcada a sigla NEO – Nenhum Estágio Observado.

Tabela 2 – Estágios observados nos enunciados dos agressores

ENUNCIADOS DOS AGRESSORES												
Conversa Analisada	*Estágio Observado*											*Total de Enunciados*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Mais de um estágio*	*NEO*	
Conversa 01	10	26	2	3	7	98	8	9	16	3	350	**532**
Conversa 02	41	60	6	48	27	92	14	7	24	5	255	**579**
Conversa 03	17	12	6	97	41	39	3	5	17	19	330	**586**
Conversa 04	7	21	2			6	61	21	27	3	323	**471**
Conversa 05	9	14	19	47	31	36		7	11	16	194	**384**
Conversa 06	10	35	16	75	23	106	13	1	15	14	316	**624**
Conversa 07	4	19	7	52	19	38		2	33	0	258	**432**
Conversa 08	11	23	6	4	27	25	5	3	26	4	551	**685**
Conversa 09	12	23	6	11	12	93	12	17	34	4	605	**829**
Conversa 10	16	28	14	57	10	15	4	6	79	8	355	**592**
Conversa 11	8	25	4	7	33	23		4	10	5	630	**749**
Conversa 12	9	9		9	3	123	5	13	29	13	220	**433**
Conversa 13	3	16	6	8	26	40	17	26	26	13	384	**565**
Conversa 14	5	7			6	17		8	16	0	613	**672**
Conversa 15	3	15	8	52	8	19		2	21	6	289	**423**
Conversa 16	22	24	9	49	5	50		12	16	0	391	**578**
Conversa 17	7	10	1		3	70	2	8	5	1	333	**440**
Conversa 18	19	14	15	45	2	26		2	14	4	294	**435**
Conversa 19	11	26	6			134		11	10	1	383	**582**
Conversa 20	4	27	8	82	10	25	2	1	20	2	409	**590**
Total de Estágios	**228**	**434**	**141**	**646**	**293**	**1.075**	**146**	**165**	**449**	**121**	**7.483**	**11.181**

A Tabela 2 mostra que todos os nove estágios foram observados nas conversas analisadas, assim como o respectivo número de ocorrência para cada um. No total foram analisadas 11.181 linhas com enunciados de agressores. A Tabela 3 é similar a Tabela 2

e mostra o número de ocorrências para cada estágio considerando os enunciados da vítima. Neste caso, foram analisadas um total de 10.808 linhas com os enunciados da vítima.

Tabela 3 - Estágios observados nos enunciados das vítimas

ENUNCIADOS DAS VÍTIMAS												
Conversa Analisada	*Estágio Observado*											*Total de Enunciados*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Mais de um estágio*	*NEO*	
Conversa 01	2	19	2	1	13	18	1	10	11	4	201	**282**
Conversa 02	23	42	15	46	22	9	2	6	8	2	325	**500**
Conversa 03	15	10	20	106	42	16		10	13	20	481	**733**
Conversa 04	8	20	8		5	1	7	9	15	0	610	**683**
Conversa 05	15	6	14	45	11	13		3	19	3	347	**476**
Conversa 06	11	12	11	46	11	6	1	1	8	11	490	**608**
Conversa 07	3	13	19	43	29	9		4	26	2	285	**433**
Conversa 08	8	12	4	5	18	5		4	15	1	543	**615**
Conversa 09	19	19	2	15	4	21		17	15	3	486	**601**
Conversa 10	9	15	26	37	7	5		4	18	1	357	**479**
Conversa 11	10	13	2	4	22	3		1	3	0	541	**599**
Conversa 12	17	7	4	12	3	11		18	16	0	508	**596**
Conversa 13	7	4	1	5	4	3		7	8	0	422	**461**
Conversa 14	3	9		1	3	4		10	13	2	385	**430**
Conversa 15	15	18	12	92	11	14	1		16	0	458	**637**
Conversa 16	9	20	33	68	1	6		3	13	0	360	**513**
Conversa 17	9	13	1	2	7	28		12	3	1	505	**581**
Conversa 18	30	30	18	77	3	3		1	16	0	383	**561**
Conversa 19	11	21	10		13	23	1	14	7	1	613	**714**
Conversa 20	4	13	13	42	11	4			12	7	200	**306**
Total de Estágios	**228**	**316**	**215**	**647**	**240**	**202**	**13**	**134**	**255**	**58**	**8.500**	**10.808**

Para a modelagem do aliciamento dos agressores e das vítimas foram utilizados os algoritmos dos Modelos Escondidos de Markov disponíveis no *toolkit* HTK [HTK 2012] considerando HMM discreto e topologia ergódiga [Grundy 1997] [Szoke 2004]. Para o treinamento dos modelos foram utilizadas as ferramentas do *toolkit HInit* e *HRest*, que implementam o algoritmo de *Baum-Welch*. Para o teste dos modelos resultantes foi utilizada a ferramenta *HVite* do *toolkit* HTK, que implementa o algoritmo de *Viterbi*.

Os 21.989 enunciados foram separados em quatro conjuntos (Tabela 4) e para execução de dois experimentos diferentes. No primeiro o objetivo foi estudar se a composição das sequências de observação considerando diferentes critérios influenciaria no resultado da classificação resultante, considerando cada um dos modelos (agressor e vítima) – teste (*i*).

Tabela 4 - Composição dos dados para os conjuntos de treinamento e teste

CONJUNTO	DADOS
Conjunto 1	*Sequência dos estágios observados, excluindo-se os enunciados em que mais de um estágio foi observado e também os NEO.*
Conjunto 2	*Sequência dos estágios observados, excluindo-se os NEO.*
Conjunto 3	*Sequência dos estágios observados, excluindo-se os enunciados em que mais de um estágio foi observado.*
Conjunto 4	*Sequência com todos os eságios observados.*

Para obtenção dos modelos (oito no total, dois para cada conjunto), primeiramente foi executado o treinamento com *HInit* e *HRest* para as sequências de observação do agressor e da vítima, separadamente. O treinamento foi executado utilizando-se de 80% das transcrições das conversas analisadas (e.g. 8.945 estágios nas sequências de observação do agressor e 8.647 estágios nas sequências de observação da vítima). A Tabela 5 mostra o número de estados HMM para os modelos treinados

considerando cada conjunto de dados (Tabela 4). Observa-se que o número de estados resultante do treinamento é variável, sendo maior à medida que as sequências de observações são maiores (Conjunto 2 e Conjunto 4). Isto corresponde ao esperado e mostra que o procedimento de treinamento ajusta os modelos às sequências de observações.

Tabela 5 - Número de estados resultantes do treinamento com HMM

	CONJUNTO 1		CONJUNTO 2		CONJUNTO 3		CONJUNTO 4	
	Agressor	Vítima	Agressor	Vítima	Agressor	Vítima	Agressor	Vítima
Nº de Estados	8	8	11	12	6	8	10	12

Para testar os modelos resultantes foi usado o *HVite*, utilizando os 20% das transcrição de conversações não utilizadas no treinamento (e.g. 2.236 estágios nas sequências de observação do agressor e 2.161 estágios nas sequências de observação da vítima). Os testes foram executados de maneira tradicional, *all-against-all*. Estes testes permitiram não somente verificar a eficácia do treinamento, mas também o grau de confusão entre os comportamentos do agressor e da vítima de modo a constatar se observações distintas guardavam correlações intrínsecas entre os modelos obtidos – teste (*ii*).

A Tabela 6 apresenta a matriz de confusão com o resumo dos resultados dos testes (*i*) e (*ii*). Pode-se observar pelo percentual de acerto que a separação dos modelos (agressor e vítima, leitura da Tabela 6 pela linha) não trouxe benefícios importantes, porque os resultados são muitos similares em cada um dos conjuntos. Já a separação em conjuntos apresentou uma leve variação, sendo menor a taxa de acerto para os conjuntos com maiores sequências de observações. Logo, pode-se inferir que o Conjunto 1, mais simples – constituído de menores sequências de observação (Tabela 2 e Tabela 3) e menos estados HMM (Tabela 5) é equivalente ao Conjunto 2, mais complexo (maiores sequências de observação e número de estados). Analogamente, o mesmo ocorre entre os Conjuntos 3 e 4. Assim, um resultado importante deste trabalho é a constatação que o Conjunto 1 pode representar com vantagem os demais por apresentar as sequências de observação com os estágios puros (únicos no mesmo enunciado) e com melhores taxas de acerto.

Tabela 6 - Matriz de confusão

	CONJUNTO 1		CONJUNTO 2		CONJUNTO 3		CONJUNTO 4	
	Agressor	Vítima	Agressor	Vítima	Agressor	Vítima	Agressor	Vítima
Agressor	89,18%	90,48%	88,22%	89,65%	77,74%	79,99%	77,23%	79,28%
Vítima	88,61%	91,15%	86,39%	90,07%	76,08%	80,66%	75,20%	79,91%

No segundo experimento, os conjuntos de sequências de observações do agressor e da vítima foram considerados combinados (agrupados). Os resultados obtidos podem ser observados na Tabela 7. O objetivo deste experimento foi avaliar se a separação dos modelos (agressor e vítima) trazia benefícios efetivos à modelagem ou não de aliciamento. A Tabela 7 mostra que ao se aplicar os dados de entrada do agressor e da vítima como um único conjunto se obtém uma redução nas taxas de acerto, porém as taxas de acerto nos conjuntos são analogamente proporcionais aos modelos em separado (Tabela 6). Assim, pode-se concluir que a modelagem considerando agressor e vítima em separado é mais efetiva que o modelo combinado. Observa-se também que o

modelo combinado é mais complexo que o modelo separado (composto de aproximadamente a metade dos enunciados), deve-se então preferir este último na escolha entre os dois, pois o modelo em separado é mais simples e resulta em melhor eficiência (taxa de acerto) que o combinado.

Tabela 7 - Taxas de acerto para o modelo combinado

CONJUNTO 1	CONJUNTO 2	CONJUNTO 3	CONJUNTO 4
82,02%	81,27%	62,66%	62,10%

Avaliando os resultados dos experimentos pode-se concluir que o conjunto puro (Conjunto 1, Tabela 5) é o mais simples, como mostrado no primeiro experimento. O mesmo ocorre com o modelo em separado, como mostrado no segundo experimento. Assim, os experimentos mostram que se pode trabalhar apenas com um dos modelos, o do agressor por exemplo – pois o objetivo deste trabalho é detecção de aliciamento, e com o conjunto puro, pois os resultados serão equivalentes a trabalhar com o modelo com tudo combinado.

6. Conclusão

Este artigo propôs um método para análise de mensagens instantâneas de texto, realizadas em salas de bate-papo na *Internet*. O objetivo da proposta foi modelar o perfil do agressor e da vítima, usando HMM, baseando-se na identificação dos estágios de aliciamento observados nas transcrições de conversas entre ambos.

Os experimentos realizados mostraram que o uso dos modelos do agressor e da vítima em separado fornecem resultados mais promissores que quando combinados. Além disso, foi mostrado que há equivalência entre os pares de conjuntos das sequências de observação. Assim, concluiu-se que se pode trabalhar com o modelo do agressor e com o conjunto de sequências de observações simples (puras), resultando em modelagem mais simples que as relatadas nos trabalhos da literatura.

É importante observar que os resultados obtidos com as sequências de teste permitem inferir que a susceptibilidade ao abuso é obtida de maneira análoga, dado que as sequências de diferentes tamanhos foram bem identificadas pelo modelo do agressor e conjunto de sequências de observações simples. Assim, esse trabalho contribui inovando com o uso do HMM na detecção de aliciamento. Esta abordagem permite inferir a susceptibilidade ao abuso sexual de menores, diferentemente do que é proposto na literatura que apenas se limita a detectar os estágios de aliciamento.

A susceptibilidade ao abuso, calculada em tempo real (durante o desenvolvimento da conversação na troca de mensagens instantâneas), pretendida como trabalhos futuros, só é possível devido ao emprego do HMM. Pois, com o HMM é possível estimar a susceptibilidade (usando probabilidades) ao abuso com diferentes tamanhos de sequências de observações.

7. Agradecimento

Os autores agradecem aos bolsistas de PIBIC Caio Carnelos, Paola Sayuri e Tiago Betiato pela participação no tratamento das conversações das salas de bate-papo.

8. Referências

Grundy, W. N.. "Modeling Biological Sequences Using HTK". Technical Report, 1997.

Holsti, O. R.. "Content Analysis for the Social Sciences and Humanities". Addison-Wesley, 1969.

Ho, W. H.; Watters, P. A.. "Statistical and Structural Approaches to Filtering Internet Pornography". IEEE International Conference on Systems, Man and Cybernetics, 2004. p. 4792-4798.

HTK toolkit. Disponível em <http://htk.eng.cam.ac.uk/>. Acesso em 04 de junho de 2012.

Kontostathis, A.; Edwards, L.; Leatherman, A.. "ChatCoder: Toward the Tracking and Categorization of Internet Predators". Proceedings of 2009 Text Mining Workshop – Society for Industrial and Applied Mathematics International Conference on Data Mining, May 2009.

Kontostathis, A.; Edwards, L.; Bayzick, J.; McGhee, I.; Leatherman, A.; Moore, K.. "Comparison of Rule-based to Human Analysis of Chat Logs". International Workshop on Mining Social Media, November 2009.

Leatherman, A.. "Luring language and virtual victims: Coding cyber-predators online communicative behavior". Technical report of Ursinus College, 2009.

McGhee, I.; Bayzick, J.; Kontostathis, A.; Edwards, L.; McBride, A.; Jakubowski, E.. "Learning to Identify Internet Sexual Predation". International Journal of Electronic Commerce, 2011. Vol. 15, p. 103-122.

Michalopoulos, D.; Mavridis, I.. "Towards Risk Based Prevention of Grooming Attacks". International Conference on Security and Cryptography, July 26-28, 2010. p. 1-4.

National Center for Missing and Exploited Children. Disponível em <http://www.missingkids.com/en_US/documents/CyberTiplineFactSheet.pdf>. Acesso em 17 de fevereiro de 2012.

Olson, L. N.; Daggs, J. L.; Ellevold, B. L.; Rogers, T. K. K.. "Entrapping the Innocent: Toward a Theory of Child Sexual Predators' Luring Communication". Communication Theory, 2007. Vol. 17, p. 231-251.

Perverted Justice Foundation Incorporated. Disponível em < www.perverted-justice.com >. Acesso em janeiro de 2012.

Rabiner, L.R.; Juang, B-H.. "Fundamentals of speech recognition". Prentice Hall Inc., Englewood Clifss, New Jersey, 1993.

Rabiner, L. R.. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, February, 1989. Vol. 77, No. 2, p.257-286.

Rashid, A.. "The Talk of Crime". Investigative Practice Journal, July 10, 2008. p. 28-29.

Szoke, I.. "Speech Units Automatically Generated by Ergodic Hidden Markov Model". Proceedings of 10th Conference and Competition Student EEICT, 2004. Vol. 1.

