FrameFOR ï Uma Base de Conhecimento de Frames Semânticos para Perícias de Informática

Ravi Barreira, Vládia Pinheiro, Vasco Furtado

Programa de Pós-Graduação em Informática Aplicada
Universidade de Fortaleza
Av. Washington Soares, 1321, Fortaleza, Ceará, Brasil

raviff@gmail.com, vladiacelia@unifor.br, vasco@unifor.br

Resumo. *Este artigo descreve a base de conhecimento de Frames Semânticos Forenses - FrameFOR. Demonstramos através de avaliações experimentais que a aplicação das técnicas de Anotação de Papéis Semânticos (APS) e Processamento de Linguagem Natural (PLN), em perícias de informática, aumenta o desempenho em relação as ferramentas utilizadas para análise por peritos forenses em termos de agilidade, precisão e recall.*

Abstract. *This article describes a knowledge base of Forensic Semantic Frames - FrameFOR. We demonstrate through experimental evaluations that the application of the Semantic Role Labeling (SRL) techniques and Natural Language Processing (NLP) in digital forensic increases the performance of the forensic experts in terms of agility, precision and recall.*

1. Introdução

O trabalho pericial é de extrema relevância para investigação policial, pois pode produzir importantes provas materiais para suportar o processo penal. Com a popularização dos smartphones e do uso intensivo de aplicativos de mensagens instantâneas, é comum encontrá-los em locais de crimes e contendo indícios do crime. Segundo [Cellebrite 2015], as investigações criminais hoje têm cada vez mais uma coisa em comum - evidência em dispositivos móveis. Nessa pesquisa, 95% dos pesquisados afirmam que dispositivos móveis são sua fonte de informação mais significante. Um problema colateral é que, sem as ferramentas adequadas, os laboratórios forenses de informática não conseguem acompanhar o crescimento da demanda. Em [Cellebrite 2015], 80% dos entrevistados disseram ter perícias pendentes, e 44% disseram que havia acumulo de serviço superior a 1 mês. Em um estado específico do Brasil, há um acúmulo de aproximadamente 8 mil dispositivos móveis que demandarão um prazo 5 anos para conclusão de seus exames. Um único dispositivo móvel pode chegar a ter 100 mil linhas de mensagens instantâneas de aplicativos como WhatsApp, Telegram, Facebook Messenger, Kik etc, que deverão ser revisadas pelo perito, podendo demandar várias semanas de leitura.

Além disso, em mensagens de textos instantâneas há predominância de linguagem informal e abreviada, sem o uso de letras maiúsculas e minúsculas adequadamente, com muitos erros de grafia e pontuação. Por exemplo, no trecho de mensagem ñ*q eu tou pidino uma presesa de fumo jk e sal.*ò tem-se erros ortográficos como em ñpidinoò

[correção: ñpedindoò]. Observa-se ainda o uso de *deception*, ou termos substitutos, como forma de ocultar, de alguém que interceptar a mensagem, a real intenção do falante. No texto anterior, temos o uso da palavra ñsalò com o significado de cocaína. Outro exemplo encontrado nessa pesquisa foi na frase ñ*beisso era pro nego ficar era com uma caneta aki mah.*ò, onde òcanetaò é palavra substituta para arma de fogo, pois suas menções normalmente vêm seguidas de calibres, 380 e 357.

Atualmente, as duas principais ferramentas utilizadas pelas perícias forenses para análise de dispositivos móveis ï UFED [Cellebrite] e XRY [Microsystemation] - possibilitam apenas que o perito forense adicione palavras-chave a serem pesquisadas nas mensagens de texto, não dispondo de funções de PLN para pesquisa por radicais ou palavras relacionadas, correção gramatical do texto original, ou descoberta de novas palavras para compor o léxico. Além disso, em pesquisa interna realizada em departamentos forenses do Brasil, apenas 20% possuem um léxico com palavras-chave para diferentes tipos de crimes e 40% criam léxicos de acordo com o caso, sem usar uma base uniforme e padronizada. Percebe-se então que há pouco uso de palavras-chave pelas perícias no Brasil, e não há uma padronização nas palavras utilizadas entre os estados.

Uma alternativa inovadora é a aplicação de técnicas de Anotação de Papéis Semânticos (APS) [Chisman 2008] na análise de mensagens extraídas de dispositivos móveis. APS é uma tarefa de PLN que permite identificar os papéis semânticos (entidades e objetos participantes e seus relacionamentos) envolvidos num evento ou situação. Nessa pesquisa, nós propomos uma base de Frames Semânticos Forenses ï FrameFOR, desenvolvida a partir da FrameNet [Baker et al, 1998]. As informações na base de conhecimento FrameFOR permite identificar, por exemplo, além de que há o relato de um evento de compra e venda, qual objeto foi comprado ou vendido e os sujeitos envolvidos na ação - quem compra e quem vende.

A base FrameFOR, aplicada a perícias de informática, permite ainda a identificação de novos termos ou expressões usadas pelos criminosos para dificultar o entendimento de quem interceptar a mensagem. Por exemplo, a seguinte mensagem extraída de um celular apreendido no estado do Ceará - ñ*ei man um brother meu chegou aqui com umas gramas do kunk.*ò, foi identificada como relevante por conter a unidade léxica ñgramasò do frame ñ**Quantidade**ò, o qual possui como elemento de frame ñentidade da medidaò, que, por sua vez, pode ser identificada como o substantivo comum que segue a unidade léxica ñgramasò. Neste exemplo, o termo ñ*kunk*ò, que se refere a um tipo de maconha, não estava presente nas palavras-chave utilizadas pelas perícias do Brasil e nem nas unidades lexicais do frame **Intoxicantes**.

Nossa proposta se baseia no argumento de que a possibilidade de investigar um maior número de fontes de informação, juntamente com uma melhoria na análise semântica de tais fontes, alavancará a qualidade das ações empreendidas para a segurança pública. Neste trabalho, realizamos uma avaliação extrínseca da base de conhecimento FrameFOR comparando com dois outros cenários - um em que se usou uma ferramenta tradicional baseada em palavras-chave e outro em que foram aplicados algoritmos de aprendizagem por máquina para inferir se uma mensagem é relevante ou não. Com base nos resultados das avaliações experimentais, podemos verificar nossa hipótese - o uso de uma ferramenta que implementa a tarefa APS e uma base de frames semânticos forenses aumentam a agilidade e cobertura do processo de análise de dados de dispositivos móveis por peritos forenses.

2. A base de conhecimento FrameFOR

Neste trabalho, propomos a base de conhecimento FrameFOR com uma centena de frames semânticos em português, adaptados da FrameNet [Baker, 1998], para o domínio de perícias de informática. A FrameFOR contém um conjunto de generalizações semânticas, denominadas frames semânticos [Fillmore 1976], expressando várias situações práticas, cada uma contendo um vocabulário específico e suas realizações em um *corpus* anotado com exemplos. Este conhecimento é utilizado principalmente na tarefa de Anotação de Papéis Semânticos (APS). O objetivo da tarefa APS é analisar um texto para identificar as entidades que participam de uma determinada situação prática ou evento e quais papéis ou funções essas entidades desempenham [Wang 2012]. Conforme relatado por [Giuglea 2006], os papéis semânticos podem ser usados para identificar palavras que não eram conhecidas anteriormente, simplesmente porque a identificação de palavras é realizada considerando o contexto do evento.

O uso da tarefa APS na área forense possibilita os seguintes avanços na análise semântica destes textos: (1) a identificação dos elementos (entidades, objetos, pessoas) envolvidos na ação ou evento e seus relacionamentos com a ação ou evento; (2) a identificação de novos termos que são constantemente atualizados pelos criminosos, principalmente para tentar disfarçar situações delituosas (uso de *deception*).

2.1 Construção da base de conhecimento FrameFOR

A base FrameNet possui mais de 1.000 frames que, em sua grande maioria, não têm relação com atividades ilícitas. Como exemplo, tem-se o frame IDADE, que identifica quando há no texto menção de idade, ou o frame COR, que identifica a cor mencionada em uma conversa. Por esta razão, a base de conhecimento não poderia ser aplicada diretamente nos textos extraídos dos dispositivos móveis, uma vez que seriam recuperados uma grande quantidade de mensagens sem qualquer relação com os objetivos das investigações forenses (ou seja, falso-positivos).

Neste trabalho foi construída a base de frames forenses FrameFOR por um processo de análise e seleção manual realizado por um perito criminal, especialista em computação forense. O perito analisou pontualmente cada frame da FrameNet, especificamente o objetivo do frame e suas unidades léxicas, com o intuito de selecionar aqueles que tivessem relevância com os crimes que são normalmente objetos de solicitações de investigação pericial. Observou-se que em diversas situações o mesmo frame poderia ser utilizado para identificar mais de um tipo de crime, como também o mesmo tipo de crime poderia ser identificado por mais de um frame. Ao final, a base FrameFOR foi composta por 113 frames, relacionados a diversos tipos de crimes, como formação de quadrilha, tráfico de drogas, sequestro, corrupção, receptação, contrabando, pedofilia, estupro, agressão, tortura, falsificação, ameaça, porte ilegal de arma, estelionato, extorsão, entre outros. A Tabela 1 apresenta uma parte da base FrameFOR com os frames que possuem relação aos crimes mais investigados pela Perícia Forense do Estado do Ceará, Brazil.

A base FrameFOR está representada em linguagem XML e foi traduzida para a língua portuguesa, pois o conteúdo das mensagens que tínhamos disponíveis eram mensagens nesta língua natural. Portanto, a base FrameFOR constitui uma base bilíngue de frames semânticos para o domínio de perícias de informática. A Figura 1 mostra um exemplo do frame **Cenário Comércio**. Este frame contém elementos de frame principais ou fundamentais, que contém os atores ou objetos das ações: **Vendedor**, **Comprador**,

Mercadoria [o objeto que foi comprado ou vendido] e **Dinheiro**. Os elementos de frame secundários ou não essenciais consistem em elementos que podem ser encontrados na situação, mas não são obrigatórios: **Taxa de Troca** e **Unidade**. As unidades lexicais são as palavras que evocam os frames, usadas para identificar a situação ou evento em questão. Os elementos principais e unidades lexicais foram anotados com a classe gramatical de seus complementos para possibilitar a identificação das realizações sintáticas nas frases. Por exemplo, a unidade lexical "preço" pode ser complementada com um número, marcado como "Z" pelo analisador. O elemento principal cujo complemento é um número, é "dinheiro", então, se uma frase for identificada com a palavra "preço" seguida de um número, a frase será rotulada como expressando uma situação de ñCenário de Comércioò, contendo um preço, e o número é anotado com o papel semântico "**dinheiro**".

TABELA 1. FRAMES FORENSES MAIS IMPORTANTES DA BASE FRAMEFOR.

Frame	Elemento de Frame	Crime
Comércio Compra	Comprador/Mercadoria/Dinheiro	tráfico/contrabando
Comércio Cenário	Comprador/Mercadoria/Dinheiro/Vendedor/Taxa/ Unidade	tráfico/contrabando
Comércio Venda	Comprador/Mercadoria/Dinheiro/Vendedor	tráfico/contrabando
Ingestão de Substância	Dispositivo de ingestão/Ingestor/Substância	suicídio/tráfico/envenenamento
Intoxicantes	Intoxicante	suicídio/tráfico/envenenamento
Assassinato	Causa/Intrumento/Assassino/Vítima	homicídio/ameaça
Quantidade	Quantidade/Valor	tráfico/contrabando/receptação
Tiro projéteis	Agente/Arma de Fogo	homicídio/ameaça/lesão
Armas	Arma	posse de arma/ameaça/homicídio

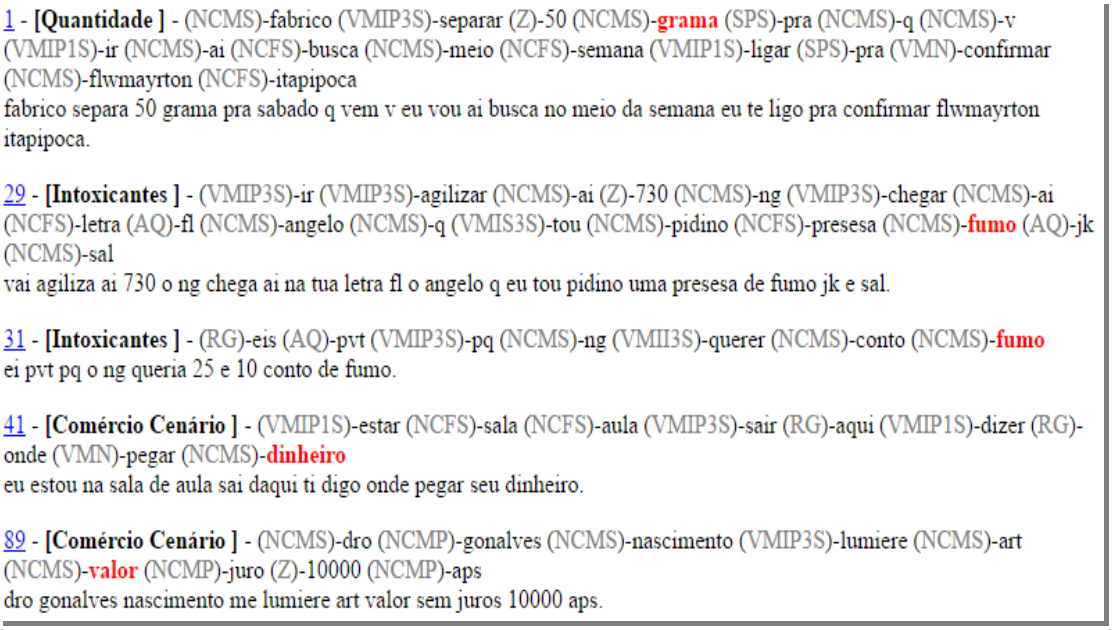
```
1  <?xml version="1.0" encoding="iso-8859-1"?>
2  <frame>
3  <nome>Comércio Cenário</nome>
4  <elementos>
5  <principais>
6  <principal complemento='NC,NP'>Comprador</principal>
7  <principal complemento='NC,NP'>Mercadoria</principal>
8  <principal complemento='Z'>Dinheiro</principal>
9  <principal complemento='NC,NP'>Vendedor</principal>
10 </principais>
11 <secundarios>
12 <secundario>Taxa de troca</secundario>
13 <secundario>Unidade</secundario>
14 </secundarios>
15 </elementos>
16 <unidadeslexicas>
17 <unidade>comércio</unidade>
18 <unidade complemento='NC,NP'>mercadoria</unidade>
19 <unidade complemento='Z'>preço</unidade>
20 <unidade complemento='Z'>valor</unidade>
21 <unidade complemento='Z'>quanto</unidade>
22 <unidade complemento='Z'>dinheiro</unidade>
23 </unidadeslexicas>
24 </frame>
```

Figura 1. Representação em XML do frame Comércio_Cenário.

Em cada frame, o perito incluiu as palavras-chave do léxico usado pelos departamentos forenses no Brasil. Por exemplo, o frame **Quantidade** não tinha as palavras "grama" ou "quilo", de modo que foram adicionados devido à sua importância para identificar a quantidade em frases relacionadas à droga. No quadro **Intoxicantes** também foram adicionadas algumas palavras, para representar melhor as expressões usadas em português, como "pó" e "pedra", comumente usadas para cocaína e crack.

2.2. Uma ferramenta de análise semântica usando base FrameFOR

Para avaliar a base FrameFOR em casos reais de análises forenses, desenvolvemos um protótipo de uma ferramenta de análise semântica em linguagem C# que possui uma interface gráfica simples, pela qual o especialista forense seleciona um ou mais frames, de acordo com o(s) crime(s) investigado(s) e solicitado(s) pela autoridade, e carrega o texto processado, após a limpeza e os processos de análise superficial. A análise semântica consiste na busca no texto pelas unidades lexicais (elementos evocativos dos frames selecionados). As mensagens de texto que contêm as unidades lexicais são anotadas - as expressões ou palavras que satisfazem a estrutura sintática dos elementos do frame (ou papéis semânticos) são marcadas com o papel semântico correspondente. Nesta fase, é possível identificar, por exemplo, o agente que ingeriu algo, a substância que é ingerida, a pessoa que compra algo e o que é comprado, etc. As linhas do texto identificadas como relevantes são apresentadas ao perito com a identificação do frame forense que justifica a anotação da mensagem. Como resultado, um relatório é gerado em um formato de arquivo HTML, com as mensagens e palavras de interesse destacadas, informando os possíveis crimes identificados e com um link que indica a linha do texto onde a mensagem foi encontrada, para que o perito possa identificar rapidamente o contexto no qual a mensagem foi escrita. A Figura 2 apresenta parte de um relatório após o processo de anotação semântica forense.


1 - **[Quantidade]** - (NCMS)-fabrico (VMIP3S)-separar (Z)-50 (NCMS)-**grama** (SPS)-pra (NCMS)-q (NCMS)-v (VMIP1S)-ir (NCMS)-ai (NCFS)-busca (NCMS)-meio (NCFS)-semana (VMIP1S)-ligar (SPS)-pra (VMN)-confirmar (NCMS)-flwmayrton (NCFS)-itapipoca
fabrico separa 50 grama pra sabado q vem v eu vou ai busca no meio da semana eu te ligo pra confirmar flwmayrton itapipoca.

29 - **[Intoxicantes]** - (VMIP3S)-ir (VMIP3S)-agilizar (NCMS)-ai (Z)-730 (NCMS)-ng (VMIP3S)-chegar (NCMS)-ai (NCFS)-letra (AQ)-fl (NCMS)-angelo (NCMS)-q (VMIS3S)-tou (NCMS)-pidino (NCFS)-presesa (NCMS)-**fumo** (AQ)-jk (NCMS)-sal
vai agiliza ai 730 o ng chega ai na tua letra fl o angelo q eu tou pidino uma presesa de fumo jk e sal.

31 - **[Intoxicantes]** - (RG)-eis (AQ)-pvt (VMIP3S)-pq (NCMS)-ng (VMII3S)-querer (NCMS)-conto (NCMS)-**fumo**
ei pvt pq o ng queria 25 e 10 conto de fumo.

41 - **[Comércio Cenário]** - (VMIP1S)-estar (NCFS)-sala (NCFS)-aula (VMIP3S)-sair (RG)-aqui (VMIP1S)-dizer (RG)-onde (VMN)-pegar (NCMS)-**dinheiro**
eu estou na sala de aula sai daqui ti digo onde pegar seu dinheiro.

89 - **[Comércio Cenário]** - (NCMS)-dro (NCMP)-gonalves (NCMS)-nascimento (VMIP3S)-lumiere (NCMS)-art (NCMS)-**valor** (NCMP)-juro (Z)-10000 (NCMP)-aps
dro gonalves nascimento me lumiere art valor sem juros 10000 aps.

Figura 2. Relatório final da análise semântica baseada na base de conhecimento FrameFOR.

3. Avalição Experimental

Nesta avaliação experimental, queremos verificar nossa hipótese de que a aplicação da base FrameFOR, na análise de evidências, aumenta o desempenho dos peritos forenses em termos de agilidade, precisão e recall. Para isso, definimos três cenários de avaliação:

- CENÁRIO 1 ï Uso do software Physical Analyzer, da Cellebrite, que procura palavras-chave em textos extraídos de dispositivos móveis pela ferramenta UFED [Cellebrite]. Neste software, não há funcionalidades avançadas disponíveis como correção gramatical, extração de radical, agrupamento de palavras, entre outros. As únicas funções disponíveis são a possibilidade de diferenciar entre maiúsculas e minúsculas e a busca exata ou parcial da palavra-chave. O conjunto de palavras-chave utilizadas foi o léxico com 156 palavras-chave, usado por departamentos forenses no Brasil.
- CENÁRIO 2 ï Uso de algoritmos de aprendizado de máquina para classificação supervisionada, treinados em um conjunto de exemplos cujas características são as palavras unigramas das mensagens, e a classe indica se a mensagem é ou não de interesse para a análise por um perito forense. Foram selecionados os seguintes algoritmos - Naïve Bayes, J48, Random Tree e Sequential Minimal Optimization (SMO).
- CENÁRIO 3 ï Uso da base FrameFOR, descrita e proposta neste documento. Utilizamos apenas os nove frames semânticos forenses (ver Tabela 1), relacionados aos crimes mais analisados na Perícia Forense do Estado do Ceará.

Para desenvolver um padrão-ouro (*gold standard*) para comparação dos resultados obtidos nos três cenários de avaliação, foram selecionados doze (12) *smartphones* reais e as mensagens de texto extraídas deles foram analisadas e anotadas manualmente por um perito forense. O perito forense identificou as mensagens de interesse para investigação policial e os possíveis tipos de crimes cometidos. No total, o perito identificou 89 mensagens de interesse, de um total de 5491 linhas de mensagens (ou seja, apenas 1,6% das mensagens). A Tabela 2 apresenta, para cada smartphone, o número de mensagens existentes, o número de mensagens relevantes identificadas pelo perito, o crime inicialmente investigado e os possíveis crimes identificados pela leitura das mensagens.

A Tabela 3 apresenta, em termos de precisão (P), cobertura (R) e F1-score, os resultados do CENÁRIO 2 para cada algoritmo de classificação, com conjunto de dados de treinamento balanceado, e os resultados do CENÁRIO 1 e CENÁRIO 3. Comparando os três cenários, o resultado da cobertura (R) do CENÁRIO 3 (FrameFOR) foi de 87% (média) e foi maior do que dos outros cenários (56%). A cobertura do algoritmo baseado na FrameFOR (CENÁRIO 3) foi melhor na maioria dos smartphones, chegando a ser 100% em três smartphones. Em termos de precisão, o melhor resultado foi alcançado pelo CENÁRIO 2 com o algoritmo SMO ï balanceado, com 91% de precisão. O CENÁRIO 3 alcançou 60% em termos de precisão. Em conclusão, avaliamos que o CENÁRIO 3, com o valor de rcobertura mais alto - 87%, foi o melhor cenário. Argumentamos que, para a análise forense, é de grande importância que uma ferramenta de análise de texto forense recupere o máximo possível de mensagens relevantes, aumentando a confiabilidade do perito forense de que poucas ou nenhuma mensagem relevante tenha sido deixada de fora. Além disso, mesmo com alto valor de recuperação, apenas 187 mensagens recuperadas foram consideradas como

relevantes, ou seja, 3,40% do volume original de mensagens contidas em smartphones (5491 mensagens), reduzindo consideravelmente o trabalho de análise de mensagens pelos peritos. Comparando o CENÁRIO 3 com o CENÁRIO 1 (ferramentas tradicionais baseadas em palavras-chave), que atingiu apenas 26% em termos de F1-Score, podemos verificar nossa hipótese inicial - o uso de uma ferramenta que implementa a tarefa APS e uma base de frames semânticos forenses, aumenta a agilidade e cobertura do processo real de análise de dados móveis por especialistas em forense.

TABELA 2. DATASET EXTRAÍDO DE SMARTPHONES E ANOTADO PELO PERITO (GOLD STANDARD).

Celular	***Linhas de msgs (atd)***	***Msgs relevantes (qtd)***	***Crime investigado***	***Crime identificado pelo perito***	*Celular*	***Linhas de msgs (atd)***	***Msgs relevantes (qtd)***	***Crime investigado***	***Crime identificado pelo perito***
1	92	5	Tráfico	Tráfico	**7**	495	6	Porte de arma	Homicídio, tráfico
2	700	20	Tráfico	Tráfico, porte de arma, homicídio	**8**	84	0	Homicídio	Não identificado
3	42	4	Tráfico	Tráfico	**9**	1037	5	Tráfico	Tráfico, porte de arma, homicídio
4	1891	8	Homicídio	Homicídio	**10**	91	0	Tráfico	Não identificado
5	522	32	Homicídio	Homicídio, tráfico, porte de arma	**11**	53	0	Tráfico	Não identificado
6	97	0	Tráfico	Não identificado	**12**	387	9	Porte de arma	Porte de arma, tráfico
					Total	**5491**	**89**		

TABELA 3. RESULTADOS DA AVALIAÇÃO EXPERIMENTAL EM TERMOS DE PRECISAO (P), COBERTURA (R) E F1-SCORE (F1).

CENÁRIO	***Mensagens identificadas (qtd)***	***Mensagens identificadas corretamente (qtd)***	***Métrica***		
			P%	***R%***	***F1%***
Naïve Bayes Balanceado	77	50	65	56	60
J48 Balanceado	17	10	59	11	19
Random Tree Balanceado	83	48	58	54	56
SMO Balanceado	55	50	**91**	**56**	**69**
RESUMO DOS RESULTADOS					
CENÁRIO 1 (média)	304	34	27	56	27
CENÁRIO 2 (SMO balanceado)	55	50	**91**	56	**69**
CENÁRIO 3 (média)	187	72	60	**87**	63

4. Trabalhos Relacionados

De acordo com [Ferrara 2014], um aspecto importante da análise de dispositivos móveis é a possibilidade de encontrar grupos de indivíduos relacionados que cometem crimes. Com o uso de dados relacionados, Ferrara diz que é possível criar redes de contato que facilitam a identificação de organizações criminosas, grupos terroristas e gangues, entre outros. Em [Belbeze 2009], o autor discute como, a partir de um dispositivo celular, é possível montar essa rede de contatos e grupos estabelecidos. Para permitir exames mais rápidos, seria importante adotar metodologias confiáveis e ferramentas computacionais na análise de textos extraídos de dispositivos móveis, que é uma das tarefas mais demoradas no processo de elaboração do laudo dos peritos. Houve alguns trabalhos na área de identificação de pedofilia, nas comunicações na Internet [Pendar 2007] e na troca direta de arquivos entre usuários (P2P) [Belbeze 2009].

Em [Pendar 2007], os dados foram utilizados a partir de conversas coletadas de um site especializado em identificar e levar à justiça predadores sexuais de crianças e adolescentes. No banco de dados, houve várias conversas entre um adulto que fingia ser uma criança e predadores sexuais, que posteriormente foram condenados usando essa conversa como evidência. Para realizar a identificação, as palavras mais importantes foram extraídas e, em seguida, utilizou-se um classificador de Support Vector Machine (SVM).

Em [Belbeze 2009], os autores pretendiam identificar novas palavras-chave para a identificação da pedofilia em nomes de arquivos, já que é comum que novos termos sejam usados na identificação desses arquivos, para mascarar seus verdadeiros conteúdos. A solução adotada foi analisar a freqüência das palavras em arquivos que já tinham um termo conhecido e depois tentar identificar novos termos que estavam sendo usados para complementar o nome desse mesmo arquivo.

Há também alguns trabalhos na área de análise de conversação, mas não diretamente relacionados à área forense, como, por exemplo, em [Reynolds 2011], onde vários algoritmos de aprendizado de máquina foram usados para identificar o comportamento do cyberbullying em conversas na internet. Enquanto isso, [Hancock et al. 2009] desenvolveu um método para identificar o uso de *deception* na troca de mensagens instantâneas. No entanto, primeiro foi necessário anotar manualmente várias situações em que isso ocorreu. Este trabalho foi aplicado principalmente em mensagens que continham mentiras, não necessariamente palavras substitutivas, ou *deceptions*, que são mais importantes para a presente pesquisa. Em [Derrick et al. 2013], tem-se um método que não requeria um *corpus* anotado para detecção de *deception*. Em seu experimento, foi utilizado um robô de conversação para entrevistar voluntários que deveriam responder de forma verdadeira ou falsa de acordo com as instruções transmitidas na tela. No final, um algoritmo fez a classificação de verdadeiro ou falso com base no tempo de resposta, número de edições feitas nas respostas, quantidade de palavras usadas e diversidade lexical.

5. Conclusão

Neste artigo, apresentamos uma base de conhecimento de Frames Forenses - FrameFOR, que permite identificar expressões ou palavras, e contextualizá-las na investigação criminal. A proposta da base de conhecimento do FrameFOR é inovadora, uma vez que

não existe uma base específica para o domínio de perícias de informática, que equilibra o trade-off entre precisão e cobertura na recuperação de mensagens relevantes. Em uma avaliação experimental, comparamos a análise semântica com base no conhecimento representado na FrameFOR com outros dois cenários - um que usou uma ferramenta tradicional baseada na pesquisa de palavras-chave e um segundo cenário que aplicou algoritmos de aprendizado de máquina para inferir se uma mensagem é relevante ou não. A base FrameFOR obteve os melhores resultados em termos de cobertura - 87%. Argumentamos que, para a análise pericial de informática, é de grande importância que uma ferramenta de análise de texto forense recupere o máximo possível de mensagens relevantes, aumentando a confiabilidade do perito forense de que poucas, ou nenhuma, mensagens relevantes tenham sido deixadas de fora.

Referências

CELLEBRITE. Cellebrite Predictions Survery 2015. Publicado em 2015. Availiable at: http://www.cellebrite.com/Media/Default/Files/Forensics/Cellebrite-Predictions-Survey-2015.pdf Acessado em: 20 out. 2016

CELLEBRITE UFED. http://www.cellebrite.com

MICROSYSTEMATION XRY. https://www.msab.com/

CHISHMAN, Rove et al. Corpus e Anotação Semântica: um Experimento para a Língua Portuguesa a partir da Semântica de Frames. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. ACM, 2008. p. 321-325.

FILLMORE, Charles J. Frame semantics and the nature of language. Annals of the New York Academy of Sciences, v. 280, n. 1, p. 20-32, 1976.

BAKER, Collin F.; FILLMORE, Charles J.; LOWE, John B. The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 1998. p. 86-90.

WANG, Xiaofeng; GERBER, Matthew S.; BROWN, Donald E. Automatic crime prediction using events extracted from twitter posts. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer Berlin Heidelberg, 2012. p. 231-238.

GIUGLEA, Ana-Maria; MOSCHITTI, Alessandro. Semantic role labeling via framenet, verbnet and propbank. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006. p. 929-936.

KIPPER, K.; DANG, H.T.; PALMER, M. Class-based construction of a verb lexicon. In: Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, 2000.

FERRARA, Emilio et al. Detecting criminal organizations in mobile phone networks. Expert Systems with Applications, v. 41, n. 13, p. 5733-5750, 2014.

PENDAR, Nick. Toward Spotting the Pedophile Telling victim from predator in text chats. In: ICSC. 2007. p. 235-241.

BELBEZE, Christian et al. Automatic Identification of Paedophile Keywords. Measurements and Analysis of P2P Activity Against Paedophile Content Project, 2009.

REYNOLDS, Kelly; KONTOSTATHIS, April; EDWARDS, Lynne. Using machine learning to detect cyberbullying. In: Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on. IEEE, 2011. p. 241-244.

HANCOCK, Jeff et al. Butler lies: awareness, deception and design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2009. p. 517-526.

DERRICK, Douglas C. et al. Detecting deceptive chat-based communication using typing behavior and message cues. ACM Transactions on Management Information Systems (TMIS), v. 4, n. 2, p. 9, 2013.