A Framework for Digital Forensics Analysis based on Semantic Role Labeling

Ravi Barreira[1]
Graduate Program in Applied Informatics
University of Fortaleza (UNIFOR)
Fortaleza, Ceará, Brasil
raviff@gmail.com[1]

Vládia Pinheiro[2], Vasco Furtado[3]
Graduate Program in Applied Informatics
University of Fortaleza (UNIFOR)
Fortaleza, Ceará, Brasil
vladiacelia@unifor.br [2], vasco@unifor.br [3]

Abstract— **This article describes a framework for semantic annotation of texts that are submitted for forensic analysis, based on Frame Semantics, and a knowledge base of Forensic 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.**

Semantic Role Labeling, Forensic Analysis, Crime Analysis

I. INTRODUCTION

Forensic work is of extreme relevance in police investigation since it can produce important material evidence to support the criminal process. With the popularization of smartphones and the intensive use of instant messaging applications, it is common to find such phones at crime scenes and with content that is considered criminal evidence. According to [1], criminal investigations today increasingly have one thing in common – mobile data evidence. In this survey, 95% of respondents affirmed that mobile devices are their most significant source of information. A collateral problem is that, without the proper tools, digital forensics labs fail to keep up with the growth in demand. In [1], 80% of respondents said they had overdue examinations, and 44% said there was accumulation of more than one month's work. In a particular state of Brazil, there is an accumulation of approximately eight thousand mobile devices submitted for examination. It is estimated that this will require 5 years to be concluded. A single mobile device can have over 100 thousand lines of instant text messages from WhatsApp, Telegram, Facebook Messenger, Kik etc., that must be analyzed by the expert in search of crime evidences, and that can result in several weeks of reading.

On top of that, in instant text messages there is a predominance of informal and abbreviated language, without the proper use of capitalization, as well as many spelling and punctuation errors. For example, in the message in Portuguese "*q eu tou pidino uma presesa de fumo jk e sal.*" [in English, "I want to order weed and salt"] There are misspellings like "pidino" (correction: "*pedindo*" – in English, "asking"). It can also be observed the use of deception, as a way of hiding the real intention from someone who intercepts the message. In the previous text, the word "salt" is used with the meaning of cocaine. Another example found during this research was the message in Portuguese "*beisso era pro nego ficar era com uma caneta aki mah*." [in English, "what the heck, dude was supposed to be here with a pen, man"], in which "*caneta*" ("pen") is a substitute word for firearm, since this expression is usually followed by calibers, such as 380 and 357.

Currently, the two main tools for mobile data analysis - Cellebrite UFED [2] and Microsystemation XRY [3]- only allow the forensic expert to add keywords to be searched in text messages, not having Natural Language Processing (NLP) functions for searching according to stems, lemmas or related words, grammar correction of the original text, or the discovery of new words to compose the lexicon. In addition, in an internal survey conducted in forensic departments in Brazil, it was found that only 20% of them have a keyword vocabulary for different types of crimes, and that 40% of them use specific keywords, according to each case, without creating a uniform and standardized lexicon. It has thus been noticed that there is little use of keywords by the computer forensic experts in Brazil, and there is no standard lexicon of keywords.

An innovative alternative is the application of Semantic Role Labeling techniques (SRL) [4] in the semantic analysis of textual messages extracted from mobile devices. SRL is a NLP task, which allows for the identification of semantic roles (participants and their relationships) involved in an event, object or situation. In this paper, we propose a framework for semantic annotation of texts that are submitted for forensic analysis, based on Frame Semantics [5] and on a base of Forensic Frames - FrameFOR, developed in this work from FrameNet [6]. The Forensic Analysis framework makes it possible to identify the expressions or words, and contextualize them in the crime investigation. For example, the semantic analysis performed through the FrameFOR knowledge base allows for the identification, in addition to the act of buying and selling, of an object which was bought or sold and the subjects involved in the action – the buyer and the seller The FrameFOR base applied to digital forensics also allows for the identification of new terms or expressions used by criminals to confuse the understanding of whoever intercepts the message.

For example, the following message extracted from a smartphone seized in Ceará, state of Brazil - "*ei man um brother meu chegou aqui com umas gramas do kunk*." [in English, "hey man, a brother of mine just got here with some

978-1-5090-6727-5/17/$31.00 ©2017 IEEE

grams of skunk"], was identified as relevant because it contains the lexical unit "grams" from the "**Quantity**" frame, which has as the frame element the "measurement entity", which in turn can be identified as a noun that follows the lexical unit "grams". In this example, the term "kunk", which refers to a type of marijuana, was not present in the lexicon used by the Brazilian experts or in the lexical units of the **Intoxicants** frame.

Our proposal is sustained on the argument that the power to investigate a higher number of sources of information, coupled with an improvement in the semantic analysis of such sources, will leverage the quality of the actions undertaken for public safety. In this work, we compared the framework and the knowledge base FrameFOR with two other scenarios - one that used a traditional tool based on keywords and a second scenario that applied machine learning algorithms in order to infer if a message is relevant or not. Based on the results of the experimental evaluations, we can verify our hypothesis – the use of a tool that implements the SRL task and a base of forensic semantic frames increasing the agility and coverage of the actual process of mobile data analysis by forensics experts.

II. SEMANTIC ROLE LABELING APPLIED IN DIGITAL FORENSICS

The aim of Semantic Role Labeling (SRL) tasks is to analyze a text in order to identify the entities that participate in a given practical situation, and what roles or functions these entities play in the event [7]. As reported by [8], semantic roles can be used to identify words that were not previously known, simply because word identification is performed considering the context of the event. The main knowledge bases for SRL are VerbNet [9], composed of a lexicon of verbs, and FrameNet [6], from UC Berkeley. FrameNet contains a set of semantic generalizations, called semantic frames [5], expressing various practical situations, each containing a specific vocabulary, and their realizations in a corpus annotated with examples. Each frame is identified by a name, a definition of the represented situation, its core and non-core frame elements (FE) and its lexical units (LU). The latter are words used to evoke frames, consisting mostly of nouns, adjectives, and verbs.

Figure 1 presents the definition of the frame "**Intoxicants**" (name of the frame) - "*An Intoxicant is ingested in order to achieve an altered state of consciousness…*". The core frame element *Intoxicant* that represents the substance ingested, and the non-core frame elements: *Country_of_origin*, *Descriptor*, and *Type*. For example, in the sentence "We smoked some high-grade MARIJUANA.", the expression "high-grade" (highlighted in green) is annotated with the semantic role *Descriptor*, describing a quality of intoxicant, usually an adjective that precedes the noun, and the word "MARIJUANA" (highlighted in blue), which is one of the lexical units of the frame, annotated with the semantic role *Intoxicant*.

Each frame element has a set of syntactic realizations, which are the grammatical structures in which they normally appear in sentences. For example, the syntactic realization of the frame element *Descriptor* is defined as "ADJ.Dep", that is, it occurs as an adjective before the noun. The verb "smoked" is highlighted, but is not present in this frame, being a lexical unit of another frame also used in this research, "**Ingest_substance**". In this case, the verb "smoke" is a lexical unit, which normally is followed by a noun that is the ingested substance

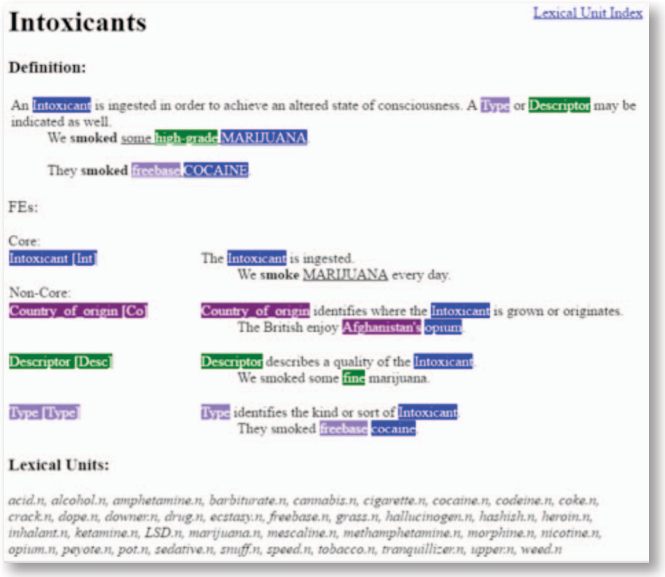


Figure 1. Definition, Frame Elements and Lexical Units of the frame **Intoxicants**.

The application of the SRL task in forensic texts analysis makes possible the following advances in the semantic analysis of these texts: (1) the identification of the elements (entities, objects, people) involved in the action or event, and their relationships with the action or event; (2) the identification of new terms that are constantly updated by criminals, mainly to try to disguise criminal situations (use of deception).

In the analysis of the Whatsapp messages extracted from a smartphone seized by the Forensic Expertise of the State of Ceará (PEFOCE), Brazil, the following message was identified as relevant: "*ei man um brother meu chegou aqui com umas gramas do kunk*." [in English, "hey man, a brother of mine just got here with some grams of skunk"]. The phrase was identified as containing the lexical unit "grams" of the "Quantity" frame, which has as frame element "measurement entity", which in turn can be identified as the common noun that follows the lexical unit "grams". In this example, the term "kunk", which refers to a specific type of marijuana, was not present in the keywords used by forensic exports or in the lexical units of the "**Intoxicants**" frame. Therefore, it would not be identified through the use of tools that use keyword search, nor by the **Intoxicants** frame. However, since it was identified by the "**Quantity**" frame, it would be a good candidate word for an update in the "Intoxicants" frame by the expert.

III. FORENSIC ANALYSIS FRAMEWORK BASED ON SEMANTIC ROLE LABELING

Figure 2 presents the pipeline of the forensic analysis framework using the SRL task, proposed in this work. In order to perform the semantic annotation in texts extracted from mobile devices, it was necessary to construct a base of Forensic Semantic Frames, called FrameFOR, and to develop a NLP framework (set of software components for NLP) that performs the semantic analysis of texts with the identification of messages of interest and of the elements involved in the event reported in the messages.

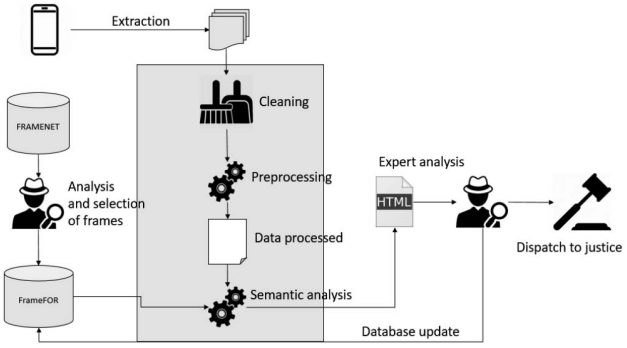


Figure 2. Pipeline of the Forensic Analysis Framework based on Semantic Role Labeling and on FrameFOR knowledge base.

Initially, all textual content, videos and images are extracted from mobile devices such as smartphones, through commercially available hardware and software (eg UFED [2]). Text messages extracted from applications such as Whatsapp, Telegram, Facebook Messenger, etc, are separated and sent for the first step of the process – Cleaning. In this step, the purpose is to prepare of the text for the next step - Preprocessing, where a morphosyntactic analyzer or POS Tagger [10] identifies the tokens and annotates them with a grammatical classification. The main step of the process is the Semantic Analysis, in which the task of SRL using the FrameFOR base is actually performed. Basically, a semantic search is made for messages that contain the lexical units of the forensic frames selected by the expert, and the recognition of the tokens that represent the frame elements (semantic roles involved in the event) takes place. The result of this step is a set of relevant messages that will be analyzed by the forensic expert (Expert analysis). After the analysis, the expert can update the database FrameFOR with new terms present in the relevant messages, in order to continuously improve the base and, consequently, the accuracy of the process.

The innovative character of the proposed process lies in the semantic analysis module, providing ease and agility for interaction and extraction of information from natural-language texts that have content of interest to public safety parties. The proposal of the FrameFOR knowledge base is also innovative since there is no specific base for the domain of digital forensics, which balances the trade-off between precision and recall metrics in the relevant message retrieval.

A. *Construction of the Knowledge Base FrameFOR*

The FrameNet base has more than 1,000 frames, which, for the most part, are unrelated to illicit activities. As an example, the AGE frame, which identifies when there is in the text a mention of age, or the frame COLOR, which identifies colors mentioned in a conversation. For this reason, the FrameNet could not be directly applied to forensic texts extracted from mobile devices, since a large number of messages would be identified without any relation to the objectives of the forensic investigations (i.e. false-positive).

In this research, the FrameFOR base was built by a process of analysis and selection, manually performed by a computer crime forensic expert. The expert analyzed each frame from FrameNet, specifically analyzed the purpose of the frame and its lexical units, in order to select those that had relevance to the crimes that are normally the object of the expert's investigations requests. It was observed that, in several situations, the same frame could be used to identify more than one type of crime, as well as the same type of crime could be identified by more than one frame. Finally, the FrameFor base was composed of 113 frames, related to various types of crimes, such as gang formation, drug, kidnapping, corruption, receiving stolen property, smuggling, pedophilia, rape, aggression, torture, falsification, threat, illegal possession of a weapon, larceny, extortion, among others. Table 1 presents a part of the FrameFOR base with the frames that are related to the most investigated crimes by the Forensic Department of the State of Ceará, Brazil.

TABLE I. THE MOST IMPORTANT FORENSICS FRAMES OF THE FRAMEFOR BASE.

Frame	Frame element	Crime
Commerce_buy	Buyer Goods Money	trafficking/smuggling
Commerce_scenario	Buyer Goods Money Seller Rate Unit	trafficking/smuggling
Commerce_sell	Buyer Goods Seller Money	trafficking/smuggling
Ingest_substance	Delivery_device Ingestor Substance	suicide/trafficking/poisoning
Intoxicants	Intoxicant	suicide/trafficking/poisoning
Killing	Cause Instrument Killer Victim	homicide/threat
Quantity	Quantity Value	trafficking/smuggling/receipt
Use_firearm	Agent Firearm	homicide/threat/injury
Weapon	Weapon	illegal possession of a weapon/threat/homicide

The FrameFOR base is represented in XML language and has been translated into Portuguese since the content of the text messages we had available were in this natural language. Therefore, the FrameFOR base constitutes a bilingual database of semantic frames for the digital forensics domain.

B. *Components of the Forensic Analysis Framework using FrameFOR base*

The components of the forensic analysis framework are described below (see Figure 2):

1) Cleaning. Due to the frequent use of informal language and abbreviations, spelling and grammatical errors, this component implements PLN routines (in Java language) for cleaning and orthographic transformations in texts extracted from mobile devices. A text is defined as the set of messages

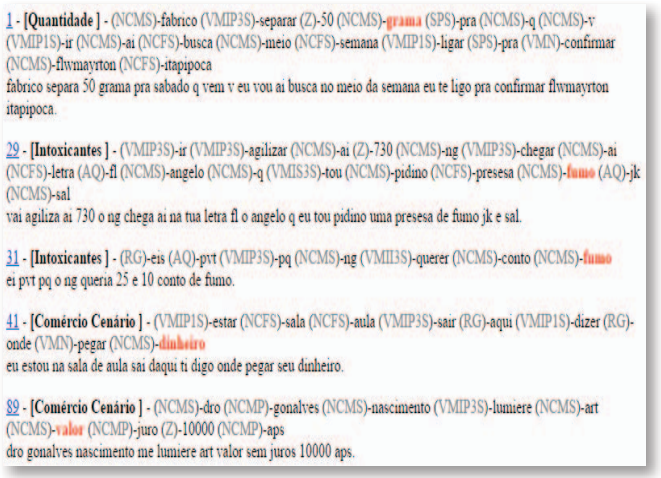
extracted from a specific mobile device. The routines perform the following operations: removal of inappropriate punctuation, symbols and emojis; removal of all characters other than letters and numbers; removal of blank lines and double spacing; removal of repeated sentences; conversion of all words to lowercase letters; removal of repeated characters in words; transforms of abbreviations - vc, q, tou, aki, etc .; addition of final punctuation in sentences. The result of this component are cleaned-up messages, corrected orthographically, ready for the next step.

2) Preprocessing (or Shallow Parsing). This component performs the morphological analysis or shallow parsing of the texts. Morphological analysis consists of the separation of sentences, tokenization, lemmatization and grammatical classification (noun, verb, adjective, adverb, etc) and their modifiers of gender, number, time and verbal mode. Shallow parsing is applied to facilitate and normalize the search for frame elements (disregarding verbal, gender and number variations). In this component, we use the Freeling parser [10], version 3.8, as well as a routine for the removal of stop words, since these do not contribute to the content of the messages. The end result is that this component generates the normalized text, annotated with morphological information and without special characters.

3) Semantic Analysis. This component was developed in C# language and has a simple graphical interface, whereby the forensic expert selects one or more frames (according to the crime(s) investigated and requested by the authority) and loads the text processed from the previous steps. The semantic analysis consists of the search in the text for the lexical units (evocative elements of the selected frames). Text messages that contain the lexical units are then annotated - the expressions or words that satisfy the syntactic structure of the frame elements (or semantic roles) are marked with the semantic role. In this stage it is possible to identify, for example, the agent who ingests something, the substance that is ingested, the person who buys something, and what is bought, etc. The lines of the text identified as relevant are presented to the expert, with the identification of the forensic frame that justifies the annotation of the message. As a result, a report is generated in a HTML file format, with the messages and words of interest highlighted, informing the possible crimes identified, and with a link indicating the line of the text where the message was found, so that the expert can quickly identify the context in which the message was written. Figure 3 presents part of a report after the forensic semantic annotation process.

IV. Experimental Evaluation

In this experimental evaluation we want to verify our hypothesis that the application of the SRL task in forensic analysis increases the performance of the forensic experts in terms of agility, precision and recall. For this, we define three evaluation scenarios:


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)-pvr (VMIP3S)-pq (NCMS)-ng (VMII3S)-querer (NCMS)-conto (NCMS)-fumo
ei pvr 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.

Figure 3. Final Report of the Forensic Analysis Framework based on Semantic Role Labeling and on FrameFOR knowledge base..

- SCENARIO 1 – Use of the software Physical Analyzer, from Cellebrite, which searches for keywords in texts extracted from mobile devices by the UFED [2]. In this software, there are no advanced functionalities available like grammar correction, stem extraction, grouping of words, among others. The only available functions are the possibility of differentiating between upper and lower case and the search for the whole keyword or for a part of the keyword. The set of keywords used was the lexicon with 156 keywords, used by forensic departments in Brazil.
- SCENARIO 2 – Use of machine learning algorithms for supervised classification, trained in a set of examples whose features are the unigrams words of the messages, and the class indicates if the message is or not of interest for the analysis by a forensic expert. The following algorithms were selected: Naïve Bayes, J48, Random Tree e Sequential Minimal Optimization – SMO.
- SCENARIO 3 – Use of the Forensic Analysis Framework and the FrameFOR base, described and proposed in this paper. We used only the nine forensic semantic frames (see Table I), related to the most investigated crimes in the Forensic Department of the State of Ceará, Brazil.

In order to develop a gold standard for comparison of the results obtained in the three evaluation scenarios, twelve (12) real smartphones were selected and the text messages extracted from them were manually analyzed and annotated by a forensic expert. The forensic expert identified the messages of interest for police investigation and the possible types of crimes committed. In all, the expert identified 89 messages of interest, out of a total of 5491 lines of messages (i.e. only 1.6% of messages).

Table II presents, for each smartphone, the number of existing messages, the number of messages of interest

identified by the expert, the investigated crime, and possible crimes that were identified by reading the messages.

TABLE II. DATASET EXTRACTED FROM SMARTPHONES AND ANNOTATED BY A FORENSIC EXPERT (GOLD STANDARD).

Smart phone ID	*Messages lines (qty)*	*Relevant Messages (qty)*	*Investigated crime*	*Identified crime by Forensic Expert*
1	92	5	Drug traffic	Drug traffic
2	700	20	Drug traffic	Drug traffic; firearm; homicide
3	42	4	Drug traffic	Drug traffic
4	1891	8	Homicide	Homicide
5	522	32	Homicide	Homicide; Drug traffic; firearm
6	97	0	Drug traffic	Not identified
7	495	6	Firearm	Homicide; Drug traffic
8	84	0	Homicide	Not identified
9	1037	5	Drug traffic	Drug traffic; firearm; homicide
10	91	0	Drug traffic	Not identified
11	53	0	Drug traffic	Not identified
12	387	9	Firearm	Firearm; Drug traffic
Total	**5491**	**89**		

Table III presents the results of the SCENARIO 1 and SCENARIO 3 in terms of the evaluation metric – precision (P), recall (R) and F1-score. Table IV presents the results of the SCENARIO 2 for each classification algorithm, with non-balanced and balanced training dataset. At the bottom of Table IV the results of the three scenarios are summarized

Comparing the three scenarios, as can be seen in tables III e IV, the result of the Recall (R) of SCENARIO 3 (SRL and FrameFOR) was 87% (average) and was higher than the recall of the other scenarios (56%). The recall result of SCENARIO 3 was 50% or more in all smartphones and was 100% in three smartphones. In terms of the Precision, the best result was achieved by SCENARIO 2 with the SMO algorithm and balanced training dataset – 91%. SCENARIO 3 achieved 60% in terms of precision. In conclusion, we evaluated that SCENARIO 3, with the highest recall value – 87%, was the best scenario. We argue that, for the forensic analysis, it is of great importance that a forensic text analysis tool recover as much as possible from relevant messages, increasing the reliability of the forensic expert so that few or no relevant messages have been left out. Besides this, even with high recall value, only 187 retrieved messages are seen as relevant, that is, 3.40% of the original volume of messages contained in smartphones (5491 messages), considerably reducing the number of messages to be analyzed by experts. Comparing SCENARIO 3 with SCENARIO 1 (traditional tools based on keywords), which achieved only 26% in terms of F1-Score, we can verify our initial hypothesis – the use of a tool that implements the SRL task and a base of forensic semantic frames increasing the agility and coverage of the actual process of mobile data analysis by forensics experts.

TABLE III. RESULTS OF SCENARIO 1 AND 3 FOR EACH SMARTPHONE IN TERMS OF PRECISION (P), RECALL (R) AND F1-SCORE (F1).

Smart phone ID	*Relevant Messages (qty)*	*SCENARIO 1*			*SCENARIO 3*		
		P%	*R%*	*F1%*	*P%*	*R%*	*F1%*
1	5	60	60	60	60	60	60
2	20	18	50	26	47	80	59
3	4	100	50	66	100	50	67
4	8	2	25	3	10	75	18
5	32	15	40	22	73	78	75
6	0	0	100	0	100	100	100
7	6	12	33	18	30	100	46
8	0	0	100	0	100	100	100
9	5	0	0	0	23	100	38
10	0	0	100	0	0	100	0
11	0	100	100	100	100	100	100
12	9	22	22	22	75	100	85
Total	**89**	**27**	**56**	**26**	**60**	**87**	**63**

TABLE IV. RESULTS OF SCENARIO 2 FOR EACH CLASSIFICATION ALGORITHM, IN TERMS OF PRECISION (P), RECALL (R), AND F1-SCORE (F1).

Algorithm	*Identified Messages (qty)*	*Correctly Identified Messages (qty)*	*METRIC*		
			P%	*R%*	*F1%*
Naïve Bayes	94	12	13	13	13
Naïve Bayes Balanced	77	50	65	56	60
J48	0	0	0	0	0
J48 Balanced	17	10	59	11	19
Random Tree	48	20	41	22	29
Random Tree Balanced	83	48	58	54	56
SMO	47	37	78	42	54
SMO Balanced	55	50	91	56	69
SUMMARY OF RESULTS					
SCENARIO 1 (avg)	304	34	27	56	27
SCENARIO 2 (SMO Balanced)	55	50	**91**	56	**69**
SCENARIO 3 (avg)	187	72	60	**87**	63

V. RELATED WORKS

According to Ferrara [11], one important aspect of the analysis of mobile devices is the possibility of finding linked groups that commit crimes. With the use of linked data, Ferrara says it is possible to create contact networks that facilitate the identification of criminal organizations, terrorist groups, and

gangs, among others. In [13], the author discusses how, from a cellular device, it's possible to mount this network of contacts and established groups. In order to enable faster examinations, it would be important to adopt reliable methodologies and computational tools in the analysis of texts extracted from mobile devices, which is one of the most time-consuming tasks. There have been some works in the area of pedophile identification, in internet communications [12] and in direct exchange of files between users (P2P) [13]. Pendar [12] used data from conversations collected from a website specialized in identifying and bringing to justice sexual predators of children and adolescents. In the database there were several conversations between an adult pretending to be a child and sexual predators, who were subsequently convicted using this conversation as evidence. To perform the identification, the most important words were extracted and then a support vector machine (SVM) classifier was used. In [13], the authors intended to identify new keywords for identification of pedophilia in filenames, given that it is common for new terms to be used in the identification of these files, to mask their true contents. The solution adopted was to analyze the frequency of the words in files that already had a known term, and then attempt to identify new terms that were being used in complementing the name of this same file. There were also some works in the area of conversation analysis, but not directly related to the forensic area, such as, for example, in [14], where several machine learning algorithms were used to identify the behavior of cyberbullying in Internet conversations. Meanwhile, Hancock et al. (2009) [15] developed a method to identify the use of deception in the exchange of instant messages. However, it was first necessary to manually annotate various situations where this had occurred. This work was applied mainly to messages that contained lies, not necessarily substitute words, or deception, which is more important to this present research. Derrick et al. [16] used a method that did not require an annotated corpus for deception detection. In his experiment, he used a conversation robot to interview volunteers who should respond truthfully or falsely according to the instruction passed on the screen. In the end, an algorithm made the classification of what was true or false based on response time, number of edits made on responses, amount of words used and lexical diversity.

VI. CONCLUSION

In this paper, we propose a framework for semantic annotation of texts that are submitted for forensic analysis, based on Frame Semantics, and a knowledge base of Forensic Frames - FrameFOR, developed from FrameNet. The Forensic Analysis framework makes it possible to identify the expressions or words, and contextualize them in the crime investigation. The innovative character of the proposed framework lies in the semantic analysis module, providing ease and agility for interaction and extraction of information from natural language texts that have content of interest to public safety parties. The proposal of the FrameFOR knowledge base is also innovative since there is no specific base for the domain of digital forensics, which balances the trade-off between precision and recall metrics in the relevant message retrieval. In an experimental evaluation, we compared the framework and the knowledge base FrameFOR with two others scenarios - one that used a traditional tool based on search of keywords and a second scenario that applied machine learning algorithms in order to infer if a message is relevant or not. The framework for SRL and the FrameFOR base achieved the best results in terms of recall - 87%. We argue that, for digital forensic analysis, it is of great importance that a forensic text analysis tool recover as much as possible from relevant messages, increasing the reliability of the forensic expert so that few or no relevant messages have been left out.

REFERENCES

[1] CELLEBRITE. Cellebrite Predictions Survery 2015. Published in 2015. Availiable at: <http://www.cellebrite.com/Media/Default/Files/Forensics/Cellebrite-Predictions-Survey-2015.pdf> Access in: 20 out. 2016

[2] CELLEBRITE UFED. http://www.cellebrite.com

[3] MICROSYSTEMATION XRY. https://www.msab.com/

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] PADRÓ, Lluís et al. Freeling 2.1: Five years of open-source language processing tools. In: 7th International Conference on Language Resources and Evaluation. 2010.

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

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

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

[14] 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.

[15] 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.

[16] 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.