Semantic Uplift Criminal Data from Social Networks

Eduardo F. Santos[1 2], **Fernanda Lima**[2]

[1]Lightbase Consultoria em Software Público
SCLN 309 Bl. B sala 213
Brasília/DF – Brasil, CEP 70.775-520

[2]CIC/UnB
Campus Universitário Darcy Ribeiro
Brasília/DF – Brasil, CEP 70910-900

eduardo.santos@lightbase.com.br,ferlima@cic.unb.br

ABSTRACT

The evolution of web technologies brought a new view over Web published data, making the information available to both humans and machines. However, as societies are built over citizens' needs, it is difficult to determine how governments are performing on addressing these needs. Social networks are an interesting tool for people to express their aspirations, but monitoring peoples' activities to understand their behavior is a difficult task. This paper presents an application of the semantic uplift technique on social networks to store violence and criminality data. The procedure uses the output of natural language processing techniques to classify criminal activity in a general taxonomy, so data can be compared to worldwide statistics.

KEYWORDS

Semantic Uplift; Social Networks; Criminal Data; Semantic Web

1. INTRODUCTION

According to recent data, violence is the second biggest problem for 18% of the population in Brazil. It comes right after health, which leads for 45% of the population [Leite 2014]. The insecurity perception is also part of citizens' lives in Brazil and other Latin America and Caribbean (LAC) countries: it is the region with the highest murder rate in the world. The UN study about cities in LAC explains this issue through the social inequality [ONU-Habitat 2012][p.XII]. As there are more people living in good and wealthy conditions, there even more living in total poverty. The population lives in a social tension atmosphere, which leads to violence in the end.

Even though it is an important issue for the population, only in the last 10 years Brazilian government created an unified system to centralize criminal data. Before 2003, with the creation of Unified Public Security System (*Sistema Único de Segurança Pública* – SUSP in portuguese), "the management of policy and security actions where characterized by the absence of cooperation between organizations" [Durante 2008]. The statistics relied on manual delivery from the police stations to federal government. Even in more developed countries as UK, the criminal data represent a real problem: "from deciding a crime has occurred, to reporting and recording, there are areas in which the data can mislead" [Evans et al. 2013]. The authors hypothesize a solution based on crowdsourcing Open Crime Data.

When a violence incident occurs, it is likely to find a reaction about it in social networks. However, the data should be used with caution. Even though social networks are an important tool to express society demands, they belong to private companies with private databases. Information access is usually done through API [Ko et al. 2010], but results are restricted to a limited number of events. Considering the importance of social networks data, the following hypothesis will be validated:

H_0. Social networks data can express social demands.

As violence and criminality is an emergent issue in LAC, and provide public access to social networks data is a relevant matter, the hypothesis validation will be done through a feasibility study. The example application, focusing on violence and criminality, intends to answer the research question:

Q_1. Is it possible to publish social networks data with semantic information to identify criminal activity?

The proposal goal is to build a feasibility study on using social networks data to identify information related to criminal activity. Feasibility study validation will be done through the example application in violence and criminality domain.

2. THEORETICAL BACKGROUND

Social network activity is mostly based on natural language communication data exchanged between users. Semantic Role Labeling technique (SRL) proposes that "for natural language understanding tasks to proceed beyond these specific domains, we need semantic frames and semantic understanding systems" [Gildea and Jurafsky 2002]. The technique organizes information in roles that can be domain-specific or verb-specif, depending on data usage. In [Punyakanok et al. 2008], authors analyze syntatic parsing techniques using trees in a full labeling process. Implementing the described SRL process can help to organize social data as shown by [Wang et al. 2012]. The described technique uses "events mentioned in tweets, the entities involved in the events, and the roles of the entities with respect to the events".

The semantic view over the data can only be useful if some observation can be taken from it [Hendler and Berners-Lee 2010]. Wang techniques [Wang et al. 2012] bring the idea of identifying the occurrence probability for the crime related events identified in SRL. It defines multiple events e_i associated with each day d. Every day is seen inside an abstract document containing all the words, doc_d. If n_d is the length of doc_d, the set $\{e_1, e_2, ..., e_{n_d}\}$ describe all the events on day d.

Definition of criminal activity is the first step in analyzing and understanding violence data. As stated in [Chen et al. 2004], "a criminal act can encompass a wide range of activities, from civil infractions such as illegal parking to internationally organized mass murder such as the 9/11 attacks". Police reports in Brazil use *incident* concept, the occurrence of a criminal act of any kind [LEMGRUBER et al. 2004][p.148].

In USA, FBI keeps National Incident-Based Reporting System [FBI 2014] to organize the incident-based criminal data in two categories: Group A, for which extensive crime data are collected and Group B, for which only arrest data are reported. In Brazil, it is still a

challenge to build an uniform database of criminal incidents. The design of an Unified Public Security System (*SUSP* in Portuguese), presented in [LEMGRUBER et al. 2004][p.141], supposes a three level data gathering. The data flow order is presented at Figure 1. Brazilian law also defines different police types for different activities: military police is responsible for crime prevention, while civil police answers for crime investigation. Both of them use different systems and maintain different databases. The incident registration process is described in Figure 3 [BRASIL 2009].

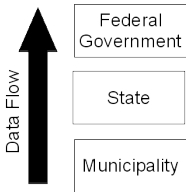

Figure 1. Data flow in criminal data

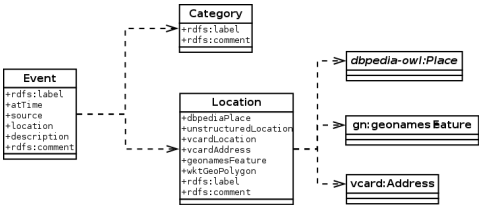

Figure 2. UML View for semantic violence dataset

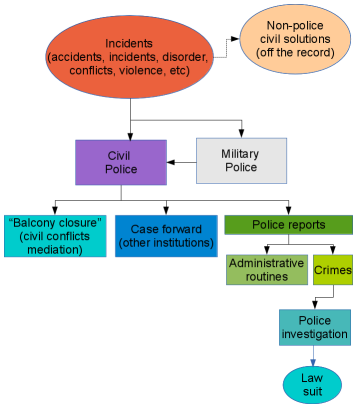

Figure 3. Crime reporting in Brazilian police

3. PROPOSAL DESCRIPTION

Considering the long way from social networks activity to a crowdsourced database, implementation architecture will be divided in three layers, as demonstrated in Figure 4. The diagram is built on top of linked open data applications architecture definition for Euclid Project [EUCLID 2014].

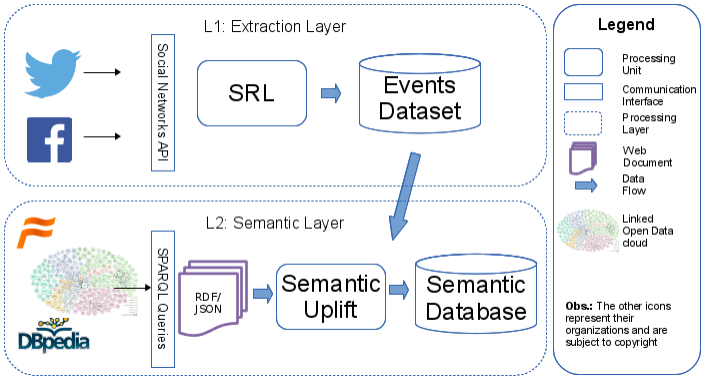

Figure 4. Semantic Uplift implementation adapted from Euclid Project [EUCLID 2014]

3.1. L1: Extraction layer

This layer describes how data will be extracted from social networks API and processed to generate an event database. An example SRL usage is presented below, based on the previous sentence from Ceilândia Muita Treta facebook page:

(1) Monday in Ceilândia.. Tense week beginning.. 3 homicides today..! Crazy day. Peace on the streets..!! [1]

The sentence will be analyzed using *nlpnet* SRL module for Brazilian Portuguese [Fonseca and Rosa 2013]. After applying SRL technique the sentence will be organized as follow:

$$[_{predicate}\text{começou}]([_{e_1:description}\text{tense}][_{e_1:number}3][_{e_1}\text{Homicides}][_{e_1:time}\text{today}])$$

3.2. L2: Semantic Layer

This layer shows how data coming from Extraction layer (L1) will be processed to add semantic information. The procedure is the *Semantic Uplift* technique described in [Brennan et al. 2013]. Only data from violence and criminality related events will be considered. After extraction models will contain events as attributes. Crime identification will be done crossing events with the statistics published in [FBSP 2013]. Considering the data in [FBI 2014] database, used in more than 3000 United States organizations, after merging data from Brazil, the following taxonomy will be used [2]:

Homicide Offenses Intentional and violent death incidents, or *crimes violentos letais e intencionais (CVLI)* in Portuguese

Larceny/Theft Offenses Pocket-Picking, Purse-Snatching, Shoplifting and other non-violent robberies

Robbery All violent theft offenses, such as kidnapping and robbering, armored car robbery, gunfire robbery, etc.

Drugs – traffic In Brazilian law, only traffic is considered a crime.

Gunfire possession Unauthorized gunfire possession

Sex offenses Forcible - Forcible Rape, Forcible Sodomy, Sexual Assault With An Object, Forcible Fondling

Assault Offenses Aggravated Assault, Simple Assault, Intimidation

Others Any other crime not in above categories

Figure 2 shows an UML representation of the semantic domain, a reduced version of the model presented by [Brennan et al. 2013]. RDF publication makes it easier for other researchers to adapt data to their needs. This data exchange will be done using RDF/JSON documents [Davis et al. 2013].

4. FINAL CONSIDERATIONS

In June 2013, when FIFA Confederations Cups started, Brazil saw a general uprising explode in an wave of protests that began in south region and spread almost everywhere in the country. The biggest impact fact is that none of country's ruling powers and institutions was able to predict it was coming, as "social media boost a protest's transmission rate through susceptible societies" [MacKenzie 2013]. Mackenzie talks about a leaderless network away from hierarchy based on self-organization.

[1] From the original text in Portuguese: *Segunda-feira na Celândia.. A semana começou tensa.. 3 Homicídios hoje..! Bagulho tá é louco. Paz na quebrada..!!.* Original link: `https://www.facebook.com/ceilandiamuitatretta/posts/753732047981961`

[2] The proposed taxonomy contains incidents from Group A in [FBI 2014]

Semantic uplift implementation should provide social networks data about criminal activity. The backend should work as a tool so other researches can build Linked Open Data applications. The extraction procedure should also work as reference model to identify and qualify other social demands, helping governments to identify demands which are not available through regular statistics.

REFERENCES

BRASIL (2009). Crime records map. URL: http://migre.me/kHPO5.

Brennan, R., Feeney, K. C., and Gavin, O. (2013). Publishing social sciences datasets as linked data: a political violence case study. *ENRICH 2013 Conference Proceedings*.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., and Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56.

Davis, I., Hors, A. L., and Steiner, T. (2013). RDF 1.1 JSON alternate serialization (RDF/JSON). W3C note, W3C. URL: http://www.w3.org/TR/2013/NOTE-rdf-json-20131107/.

Durante, M. O. (2008). Avanços e desafios na implantação do sistema nacional de estatísticas de segurança pública e justiça criminal (SINESPJC). *Anuário de Segurança Pública*.

EUCLID (2014). Chapter 1: Introduction and application scenarios. URL: http://www.euclid-project.eu/modules/chapter1.

Evans, M. B., O'Hara, K., Tiropanis, T., and Webber, C. (2013). Crime applications and social machines: crowdsourcing sensitive data. In *SOCIAM: The Theory and Practice of Social Machines*. URL: http://eprints.soton.ac.uk/351275/.

FBI (2014). National incident-based reporting system (NIBRS) – General Information. URL: http://www2.fbi.gov/ucr/faqs.htm.

FBSP (2013). Segurança pública em números. *Anuário Brasileiro de Segurança Pública 2013*.

Fonseca, E. R. and Rosa, J. L. G. (2013). A two-step convolutional neural network approach for semantic role labeling. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–7. IEEE.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Hendler, J. and Berners-Lee, T. (2010). From the semantic web to social machines: A research challenge for ai on the world wide web. *ARTINT*, 174:156–161. doi:dx.doi.org/10.1016/j.artint.2009.11.010.

Ko, M. N., Cheek, G. P., Shehab, M., and Sandhu, R. (2010). Social-networks connect services. *Computer*, 43(8):37–43.

Leite, M. (2014). Datafolha aponta saúde como principal problema dos brasileiros. URL: http://folha.com/no1432478.

LEMGRUBER, J. et al. (2004). Arquitetura institucional do sistema único de segurança pública. *Acordo de Cooperação Técnica: Ministério da Justiça, Secretaria Nacional de Segurança Pública, Federação das Indústrias do Rio de Janeiro, Serviço Social da Indústria e Programa das Nações Unidas para o Desenvolvimento. Distrito Federal*.

MacKenzie, D. (2013). Brazil's uprising points to rise of leaderless networks. *New Scientist*, 218(2923):9.

ONU-Habitat (2012). *Estado de las ciudades de América Latina y el Caribe 2012*. ONU-Habitat. URL: http://www.onuhabitat.org/index.php?option=com_docman&task=cat_view&gid=362&Itemid=538.

Punyakanok, V., Roth, D., and Yih, W.-t. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer.