RESEARCH PROPOSAL

Davide Berdin

January 29, 2016

Applicant for doctoral studentship in Computer Science, focusing on Natural Language Processing and Statistical Machine Learning

Proposed thesis title

NLP language-independent framework for Crime prediction using Convolutional Neural Networks

Background and problem description

Crime is a classical "unpredicted" problem, yet not totally random. The opportunity of finding a pattern to predict a felony is possible as demonstrated by the literature [1][2]. In fact, mining social media have been extensively used for crime prediction [3]. Also, trying to predict if someone is going to commit a felony based on the geographical area has resulted as an effective prediction method [4]. The state-of-art proposes different data mining algorithms as well as Natural Language Processing techniques in order to extract information from different sources. A particular attention has been given to the *Deep Web* where the opportunity of collecting data is practically endless.

The thesis aims to create a language-independent framework for detecting suspicious text in order to perform crime prediction. A particular attention will be given to the **Dark Web** where the likelihood of finding sex-offenders, drug dealers, etc. is very high. Part of the investigation methodology is to develop a complete profile of the offender is such a way that the detective's team in charge of a case, is able to predict the next move(s) of the criminal. In this way, the police can act accordingly in order to catch the delinquent. This thesis leverages on the creation and usage of Criminal Profiles, built in automatic fashion. Combining a profile with a set of specific words-correlations, it is possible to apply a statistical machine learning model to extract the likelihood of a possible felony (details are described in the next sections).

Despite the fact that we are mining for text in the Dark web, the framework can be used also in combo with audio files. The procedure can be easily applied using a speech recognition system that could be useful for the law enforcement agency since they can analyze several phone-calls to prevent any sort of felony in an automatic fashion.

Figure 1 represents a macro-overview of the whole system.

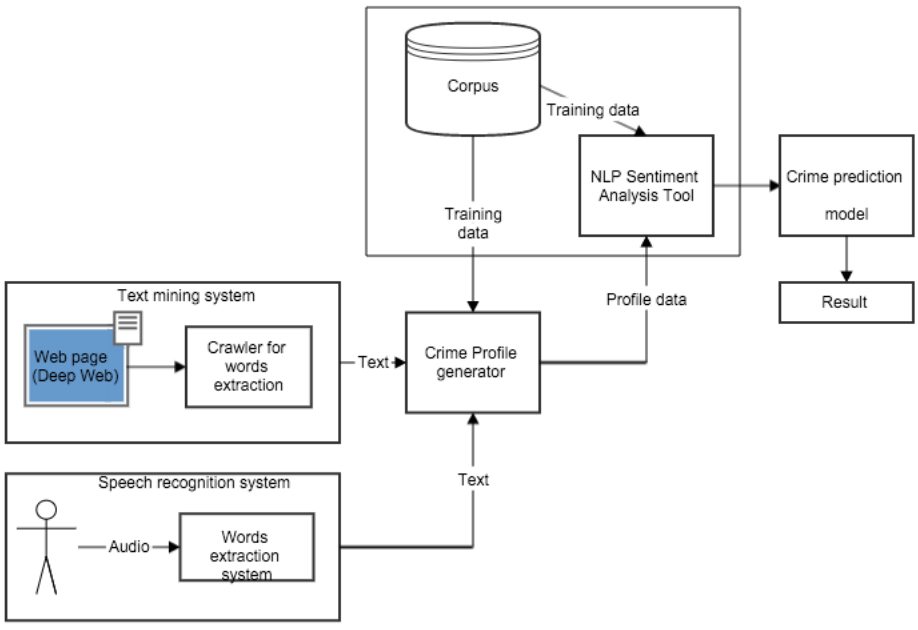

Figure 1: General overview of the whole system

Methodology

This project will utilize a methodology based on [5] for information retrieval. We can narrow down the searching process to those websites that are more suitable for our purpose. This means that, the crawler is most likely to look for those pages in which people exchange comments and opinions, such as Forums and social medias. This drastically simplify the process for text-mining because the amount of information is less dispersive.
In [6], it is used the concept of *sub-communities* in order to define if a particular member belongs to those categories that can be considered as homeland security threat. The project aims to generalize this concept by defining profiles as a container of all possible categories and not only those considered as threats.

1 - Criminal Profiles

The first phase of this project will focus on creating the corpus necessary to create Criminal Profiles. The criminal profile is the knowledge representation of the data that will be used for the prediction. There are two methods that is possible to use: the first and most accurate, would be asking the police department to grant an access to their records. The second method (less accurate) is to use a web-crawler to build criminal profiles. Generally speaking, the crawler would scan web-pages in which is described the kind of felony that the person has committed. Truth be told, it does not matter whether the person is still alive or not. The important matter is that there is the description of the crime action he/she committed during his/her life.
The reason is that, we will use the extracted information to build a behavioral model compose by two parameters:

- ID, name or unique code of the criminal
- (Action, Topic), set of actions that the person committed in relation to a topic

Here an example:

- Pablo Escobar
- (deal, drugs), (bribery, politics), (play, politics), etc.

The idea behind this type of Criminal Profile is that we want to build a simple and concise representation of criminals. This structure will be then used to model a map in which will be used as input of the second part of the project.

2 - Definition of Sentiment

The second part will focus on building the NLP tool that will act as a Sentiment Analysis system. This differs from current systems in that instead of returning a simple polarity, it will return a word cloud showing the most likely "synonomial" cluster of words relating the Action and Topic. This cluster would show our undetermined sentiment in relation to understood words, and therefore be more readily comprehended by humans. Words are understood in relation to each other as represented by the picture 2, for which data which can be mined from famous websites such as *dictionary.com* or *wordreference.com*, and their thesauri.

As explained in [7], it is possible to use the semantic linking in order to establish the relationship between words. The proposed algorithm is a semantic-linking-context-based model where the value of action-topic pair is defined according to path-distance and to some context of analysis.

Need to expand and explain better the correlation

3 - Crime Prediction Model

The third part of this thesis will focus on the Crime Prediction model. The prediction will made based on a Convolutional Neural Network that is trained in order to perform a sentiment analysis on action-topic pair. In [8] they described a method of using CNN for predicting the sentiment on short text. Given this, we can push the limits using the synonymous model and the correlation action-topics as training set of the tool.

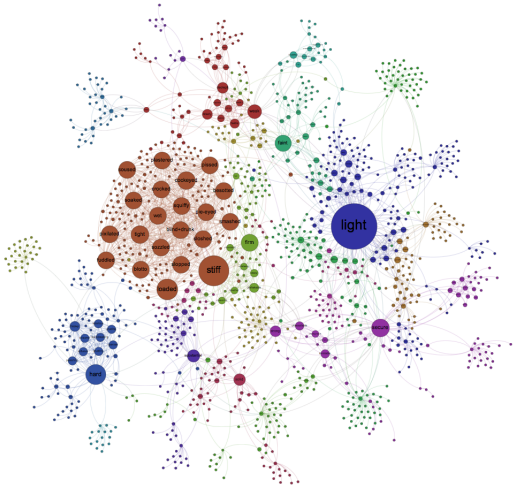

Figure 2: Example of synonymous clustering[1]

Using this model, we can correlate different action-topic pairs giving a weight to each of them. The weight is based on the distance between the two and the meaning of the words. In fact, since the meaning of a word can change based on the context, we need to establish rules that will provide the right weight for each couple. The state-of-art of word-meaning-context-based is described in [9][10] but other techniques can be used such as: *semantic similarity*, *latent semantic similarity*, *n-gram frequency analysis*, etc. From here we can use a statistical approach to assign the right weight for each pair.

This system can be used not only for the English language, but also to any other language where it is possible to build a synonymous model.

An example is given below:

- (play, politics) = 0.3
- (prescription, drugs) = 0.1
- (deal, drugs) = 0.7

where 0.3, 0.1 and 0.7 are the crime rates estimated by the using the Criminal Profile and the NLP Tool. Given these results, we can potentially claim whether the person is about to commit a crime in the next future or not. In positive case, Police can start a new investigation. In addition, to improve the weighting model, we could use those profiles that have been used for training the system. The system can be considered as **semi-supervised** machine learning approach because it utilizes the initial criminal profiles as *"reference"*.
Generally speaking, the system can potentially speed-up the process of finding criminals either across the web or from audio calls without depending on a particular language.

[1] http://allthingsgraphed.com/2015/04/09/a-matter-of-degrees/

Work plan

Work plan to complete the project and writing the dissertation.

Key Item	Date
Engage with literature and refine the scope of the project. This will involve conducting an initial literature review in order to narrow down the research questions of the project, as well as potentially altering the methodology in light of new information uncovered.	September - December 2016
Prototype the text-miner in order to extract information from Dark Web	January - March 2017
Identify the appropriate method for building criminal profiles and code the related part starting from the information extracted with the crawler	April - September 2017
Build dataset and engage with literature about Convolutional Neural Networks. Prototype a system capable of using short text as train dataset	October - December 2017
Engage with literature in speech recognition and prototype a continuous speech-recognition system in order to extract words. Build (or complete) a criminal profile based on those information	January - March 2018
Literature in Sentiment Analysis and code the NLP tool for establishing the pairs correlation	April - August 2018
Define the statistical model methodology and prototype the crime predictor	September - December 2019
Overlay gathered data with the selected frameworks and engage with final tests	January - April 2019
Write up Results chapter	May - July 2019
Write up Discussion chapter	August - October 2019
Write up the remaining sections of the thesis	November - January 2020
Draft submission of thesis to supervisor and rework based on feedback	Early February 2020
Thesis presentation/defense and submission to assessors. Potentially rework based on feedback	March 2020
Final submission of thesis	April 2020

Table 1: Estimated work plan

Acknowledgment

The project has been designed with the help of Leonard Brown from University of Rochester in which can lead to a collaboration between the two universities.

References

[1] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 231–238, Springer, 2012.

[2] X. Zheng, Y. Cao, and Z. Ma, "A mathematical modeling approach for geographical profiling and crime prediction," in *Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on*, pp. 500–503, IEEE, 2011.

[3] X. Chen, Y. Cho, and S. young Jang, "Crime prediction using twitter sentiment and weather," in *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pp. 63–68, IEEE, 2015.

[4] C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 779–786, IEEE, 2011.

[5] Y. He, D. Xin, V. Ganti, S. Rajaraman, and N. Shah, "Crawling deep web entity pages," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 355–364, ACM, 2013.

[6] S. A. Ríos and R. Munoz, "Dark web portal overlapping community detection based on topic models," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2012). Association for Computing Machinery, Beijing*, 2012.

[7] Z. Ren, D. van Dijk, D. Graus, N. van der Knaap, H. Henseler, and M. de Rijke, "Semantic linking and contextualization for social forensic text analysis," in *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pp. 96–99, IEEE, 2013.

[8] C. N. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland*, 2014.

[9] K. Erk and S. Padó, "A structured vector space model for word meaning in context," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 897–906, Association for Computational Linguistics, 2008.

[10] K. Erk and S. Padó, "Exemplar-based models for word meaning in context," in *Proceedings of the acl 2010 conference short papers*, pp. 92–97, Association for Computational Linguistics, 2010.