EMM-JRC congress notes

Corrado Lanera

2018-01-19

Contents

Introduction

These notes are about the workshop on the Europe Media Monitoring tool as presented in Gazzada in occasion of the EMM congress provided by the Text Mining Team of the JRC (Joint Research Center).

The congress span on three days (29-30 November and 1 December 2017) and it is divided in two main part:

1. Mornings: conference in which many application and characteristics of EMM has been communicated.
2. Afternoons: Workshops, which are three parallel ones on the usability on:
    - overview of the system and the on-line platform,
    - use of the desktop app and relative API,
    - internal procedures used by EMM.

Organization of the notes: one chapter per part of the congress, each one is subdivided in sub-chapter corresponding to the talks (in the first chapter) or corresponding to the the topics of the workshop.

Chapter 1

Conference

1.1 International co-operation in Financial Criminal Investigations[1]

Abstract: Technological developments make things possible that we could not have imagined twenty-five years ago. Forms of communication, producing things, transporting goods, the way in which transactions take place are changing constantly, not only in legal, but also in illegal businesses. To effectively fight serious financial cross border criminality, international cooperation is a necessity. The past year two international initiatives were launched which had financial criminal investigations as the starting point. ENFIN, a European knowledge Network were members and partners can exchange experiences, methods and new developments on Financial Investigations, and FCInet a virtual solution that makes data matching and exchange easy and enlarge the international cooperation possibilities. FCInet is based on the proven technology and is supported by the Forum of Tax Crime Investigations (FHTCI) of the OECD. Both networks consist of Law Enforcement organisations specialised in financial investigations and Tax and Customs administrations. They collaborate with respect to the differences in working methods, (privacy) legislation, data protection and independence of the participating organisations. The presentation will contain backgrounds on the ENFIN and FCInet, the aims, partners, steps taken and the outlook for 2018.

1.2 SIRIUS as a law frameworkfor internet investigations[2]

Abstract: Nowadays, investigators working on a case cannot avoid investigating the digital footprint of those who plan terrorist attacks or are suspected of recruitment, training and financing of terrorism, as well as incitement to commit a terrorist offense, including relationships, communication means, financial aspects, logistics, centres of interest and behavioural activity. Whilst these information sets were historically under the remit of national entities, they have now acquired a global perspective and are owned by Online Service Providers, oftentimes based outside of the EU territory. Further complexity stems both from the volatility of data held across different legislations, as well as its volume and the urgency with which this information is needed in the context of a CT case In the attempt to cope with these challenges and to maximize the level and the quality of operational support provided, we have recently launched SIRIUS, a project which aims to cater for the investigators' needs in an online environment. Available only to law enforcement authorities and deployed in a closed and secured environment, SIRIUS is the place where

[1] *Eugenie de Lange* — Dutch Tax Authorities
[2] *Juan De Toledo Maetinez* — EUROPOL

all the information related to Online Service Providers, and how to increase the investigation efficiency can be found, with manuals, tips, forums, Q&A, etc. Additionally, the platform will also include a repository for the collaborative development of tools to support investigations of crimes facilitated by the Internet, developed by and intended for the Law Enforcement community.

- https://www.europol.europa.eu/newsroom/news/europol-launches-sirius-platform-to-facilitate-online-investigations

1. 26 attacks since January 2016
2. Communications:

- many to many — propaganda
- one to many — recruitment
- one to one — private communications

3. Steps:

- standardization of data collections
- identification of nodes about the *big plyers* do the (net)
- high profiles investigation

4. Big Data:

- Open
- Closed/Private

Europol Experts –> SIRIUS project: 263 members, 13 ONSIT tools

1.3 Social Media Monitoring for Awareness of Security threats against VIPs: opportunities and challenges[3]

Abstract: Identifying possible adversaries is a key element of Security Threat Assessments. When assessing threats against persons with high public visibility, monitoring Social Media may seem a promising idea in order to identify potential groups, persons, ideologies developing a specific hate narrative. Indeed, social media features all sort of opinion trends and are often used as an echo chamber for propaganda purposes. They can thus be seen as an easily accessible and abundant source for Personal Threat Assessment purposes. However, exploitation of Social Media material into actionable intelligence (e.g. to support decisions on the set up of VIP security measures) poses several methodological and technical challenges. The purpose of this presentation is to stimulate discussion on such challenges, rather than describing pre-cooked solutions.

VIPs: - 28 European commissioners - any staff member exposed to high visibility because of his/her functions - any staff member exposed to a security threat because of his/her functions

80%/90% of false positives (for vocabulary search for the most violent languages used)

1.4 Computer support for analysing violent extremism in online environments[4]

Abstract: This talk gives an overview over the research done at the Swedish Defense Research Agency (FOI) on developing and applying tools for analysing violent extremism in online environments. The talk focuses on analysis of text data, and presents some of the core technologies we use to deal with large-scale and noisy data. We also provide examples from two recent studies

[3] *Bertrand De Longueville* — DG HR, European Commission

[4] *Magnus Sahlgren* magnus [dot] sahlgren [at] ri [dot] se — FOI – Swedish Defence Research Agency

where we have applied our tools to large collections of propaganda material from IS, and to large collections of web data from Swedish right-wing extremist groups.

Analyses: - theory driven (warnings defined by experts) - data driven

Main topics: 1. Lone wolf actors (eg, fixation, leakage) 2. Radicalization (eg, in-group/out-group, dichotomous thinking)

1.4.1 Linguistic marker:

1. word list
2. vocabulary variation (synonyms)
3. Semantic memories

1.4.2 Thematic analyses:

Theme and word list (eg, BRUTALITY: exclude, punish, behead …)

Count occurrences of word in data Monitor theme over time

1.4.3 Processing pipeline

segmentation - Language - Tokenization - Normalization

normalization improve recall but can reduce precision

- overestimation problem: polysemous (eg, “execute”): disambiguation
- underestimation problem: synonymy (“IS/ISIS/ISIL/Daesh”): semantic memory

1.4.4 End-to-end system

Character-based (deep) neural network Eliminate the need for preprocessing, more accurate then lexical analyses (if trained with sufficient data)

eg, thematic analyses of IS propaganda

- prior polarity list (sentiment analyses)

English: opinion lexicon, MPQA, sentiwordnet, LIWC, saifmohammed.com (domain specific lexicon)

how to detect ironia?!?!

Lexicon VS MLT

Annotation: - correlation as measure of reliability (>.8, for someone >.9)

- demographic: gender, age, origin classification
- author identification, alias matching
- socio-political: education, …

Analysis platform (not data collection!)

dark WEB (TOR): agora market (main drug market)

1.5 Use of Open-Source Information in the IAEA Safeguards Department[5]

Abstract: The IAEA Department of Safeguards makes extensive use of open-source information in support of its mission to verify the compliance of Member States with their safeguards obligations. Open-source information is one of several data streams that facilitate the ongoing State evaluation process. This presentation will review the department's work with open-source information, which includes both routine monitoring of news and other information sources and targeted searching, to support verification activities both in headquarters and in the field. Sources used include websites, newsfeeds, scientific and technical literature, and databases containing information on imports and exports of commodities. The ongoing, productive relationship between the IAEA and JRC Ispra has significantly strengthened the Department of Safeguards' capabilities in routine monitoring of open sources, and the IAEA is now exploring the possibility of using the EMM OSINT Suite to streamline and improve the department's capability to perform targeted searches of open sources for information of safeguards relevance.

- credible assurance to the international community that States are honoring their safeguards obligations
- correctness & *completeness*
- video: inspecting the nuclear fuel cycle

[5] *Chris Eldrige* International Atomic Energy Agency

Chapter 2

EMM workshop

> Overview: This course provides a general overview of EMM technology and related tools. It shows with practical examples how Text Mining and Analysis (TMA) can effectively support the daily work of analysts. The platform processes every day about 300,000 news articles providing: language detection, categorization, language recognition, entity extraction, quote extraction, geo-tagging, tonality, duplicate detection, categorisation, indexing and searching, clustering, statistics and event extraction. Dedicated Graphical User Interfaces allow analysts to display and browse all metadata and create reports and/or newsletters. The course structure includes hands-on sessions based on a common use case: participant will learn how to configure the platform in order to capture news related to their topics of interest, browse the results, produce newsletters and send notifications. This course is targeting people who has just started using EMM or who need to assess whether to adopt it as media monitoring platform.

Big data Pilot Project, *Text and Data Mining unit* - EMM: Europe Media Monitor - SITAF: Statistical and Information for Anti Fraud TIM: Tools for Innovation Monitoring

> for every people: do not store anything and do not need to be an expert

https://newsdesk.emm4u.eu [osint2017]

2.1 Aim: what EMM can and cannot do.

- have the platform on our server (they don't have no more resources to supply external request)

1. Analyze unstructured text:

- Natural language
- Ambiguous/incomplete
- Multi-language coverage

2. ~70 languages
3. DO NOT SCAN ANYTHING, it is not Google: it scan 70 source every 5 minutes. Every source is as any of one which decide to collaborate to the project. If our interested source is not in EMM they cannot add it! They need a (open source) contract.
4. They provide statistics or insides on the text but if we want the real text we have to click on the link! (There are the email to WHO too)

2.2 What they do with the text:

1. First thing is to asses the language! (detected with proprietary tools! 80% accuracy overall)
2. Second, category: e.g. "Taxation, Economy, EU-US trade" — 3k categories/topics, based on more 60k keywords. The quality of the results comes from the quality of the translation of the keywords (do not use Google translate!).

   we will do an exercise on this categorization process

- once you have set the keywords of the topic EMM redo every day 24/7 the categorization of sources on your topic!

4. next, there are extracted the entities: people and organization. (for each entities they search for every variant of the name). 1.7M entities.
5. It extract also the quotes! from the entities. (note: the name of the street do not become entities […he said…mmm])
6. Geo detection from the text, and add meta data with the lat and lon
7. sentiment analysis: positive or negative (based on positive and negatives keywords). Usually they do not show this results because taken by itself do not have a great meaning. on the other hand if we collect all the article of a resource over a year and we consider the average level of sentiment for that resource in that topic, then you can have a value that represent the trend of the sentiment of a specific paper into the subject in his context.it is also difficult to separate the sentiment of the people who wrote with the sentiment of the news or the topic (e.g. a very good people that work in a very bad immigrant situation)

   Extract event metadata, only available for specific topic. Anyway it only aims to reduce the ammount of work it is not magic.

first: on the website we have to detect the correct part of text of interest. 80% of quality (1/5 of text retrieved are fake/wrong/out-of-topic text)

8k news sites 300k article/day 70 languages 3k categories 60k keywords runs 24/7 25k/day visitors

Domains: Border security, Cyber-crime and [?]

2.3 Notes

the system is *live* so the new overwrite the older… if you are not on the screen you loose them

subscribe for topics: e.mail, rss, sms, …: subscribe for a one every 24 hours or even (do not did this) for a prompt alert for something

translation of 20 languages in English. Why: because someone do not want do tell to Google our keywords. there are only two security: to not disclose the translation, the click, the search… and the copyright.

If you want to have the translation service provide by the EMM the have to give you the servers (real..they are 4)

everything look at the interface is no-moderated!!!

Is it also possible to filter the article retrieved by *the same stories* but stories are language specific!

Multilingual aspect are very important. Even English do not cover all the region of the word

you can look at the stories AFTER the selection on the topic in a specific context

they can do past stories for the past but they do not have a graphical interface for this. If we have a specific request (title of the article, dates, etc…) they can retrieve the information about its story.

http://medisys.newsbrief.eu/medisys/homeedition/it/home.html

2.4 second service (not yet publicy) for expert…

…in which you can define your research and store them

Use of twitter to extend the information not to create or compose them: usually people read the news and next talk about the new. (newspaper first!)

For twitter there are two level categorization, the first from a search next a second screen based on a user search strings: top-ten user, hashtags keywords and links (note: links are very important because the content of the link are often out of the repository of the EMM resources and so this list of links can inform on something before it appear in the EMM)

2.5 Main process-flow

- Short document!!(1 page)

documenti entrano nel processo (dal web, dalle mail, …) e iniziano il processo: 1. identificazione della lingua 2. topic 3. geo location 4. entities 5. quotations …. …

alla fine esce il blocco coi metadata e il testo e qui il testo viene immediatamente cancellato! tutto il resto viene mantenuto per sempre, inclusi il link, e tutti i metadata. e quindi possono essere interrogati dal sistema come una sorta di google

2.5.1 Exercice: create a category

one file on topics based on keywords and one file on metadata

2.6 Produce newsletter

Di fatto un account consiste in due file Alert e Filter. Il primo serve a elencare e memorizzare le *categorie* il secondo serve per fare combinazioni opportune di categorie. Per esempio si possono combinare categorie diverse, considerare solo determinate lingue, determinate nazioni di origine, determinate risorse (source di informazioni da cui pescano)

Main desks: https://newsdesk.emm4u.eu/ND1/ CategoryEditor: https://newsdesk.emm4u.eu/CategoryEditor/AlertEditor.html# WorkSpace: https://newsdesk.emm4u.eu/ND1/?ws=true EMMnewws: http://emm.newsbrief.eu/NewsBrief/sourceslist/it/list.html Wiki; https://wiki.emm4u.eu/confluence/display/CE/Category+Definitions MediSys: http://medisys.newsbrief.eu/medisys/alertedition/en/Chagasdisease.html

Per vedere le proprie categorie andare in una a caso dal EMMnews (ma entrarci, fare attenzione che nell'indirizzo compaia "alertedition") e sostituire la parte finale dell'indirizzo con il nome della propria categoria. A questo punto ci si puo sottoscrivere alla categoria e quindi si possono avere aggiornamenti giornalieri. (o istantanei, me sconsigliatissimo!!), per esempio il nostro

2.7 Contacts

charles [dot] macmillan [at] ec [dot] europa [dot] eu jens [dot] linge [at] ec [dot] europa [dot] eu marco [dot] verile [at] ec [dot] europa [dot] eu eleonora [dot] mantica [at] ext [dot] ec [dot] europa [dot] eu