In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy.

Multilanguage Semantic Behavioural Algorithms to discover terrorist related online contents

Maurizio Mencarini[1] and Gianluca Sensidoni[1]
[1] Expert System S.p.A.
mmencarini@expertsystem.com, gsensidoni@expertsystem.com

Abstract

Multilanguage Semantic behavioural algorithms mapped to machine learning techniques in order to collect and analyze huge amount of heterogeneous and complex Multimedia and Multilanguage terrorist-related contents from both the Surface and the Deep Web, in order to discover (by "connecting the dots"), detect, analyze, and monitor potential terrorist-related activities and people. At the conference a live demo of to-date obtained results and some experiment coming from EU DANTE project (an H2020 EU funded research project) will be shown.

1 Introduction

Contemporary terrorists and criminal organizations increasingly exploit the Internet to spread their message and gain support throughout the world, using the Web as communication tool, in particular for **recruitment, training and propaganda activities**, but also for **disinformation, raising funds, organization, planning, and financial transactions**.

While the use of the Internet for propaganda and recruiting purposes (though not solved) has received wide publicity, terrorist groups utilize the Internet for a variety of other purposes, including fund raising. Not only Al-Qaeda, but also other terrorist groups (including ISIS) use the Internet to raise funds to support their activities.

In order to promptly face this threat, it is become urgent to put in place countermeasures to enable the Law Enforcement Agencies (LEAs) and intelligence officials to **continuously monitoring** in near real time on-line **relevant (for the purposes of counter terrorism, under lawful warrant)** communications and contents, both in the Surface Web, and **in the Deep Web**, where there is a vast amount of data (there are estimates that this is more than 95% of the available data on the Internet) not always indexed by automated search engines, but potentially providing useful contents and information for detecting and fighting terrorist activities.

This activity must absolutely be part of an effective cybersecurity strategy.

Due to the escalating number of Internet users and the increasing speed of creation/deletion of Internet contents, it is clear **that searching *terrorist-related contents*** (i.e. generated by terrorists or individuals linked to terrorists, or linked to terrorist activities, including relevant contents generated by non terrorists) **and information by keywords and manually is highly error-prone in precision and a lot time-consuming, dramatically slow and obsolete, making impractical the examination of huge amount of resources**.

2 Concept and Approach

The Multilanguage Semantic Behavioural Algorithms (also part of the EU DANTE project) are aimed to support LEAs in the most advanced intelligence processes, through big data collection and analysis. Knowing facts and events in advance is one of the key points of these proposed solutions, used to prevent potential threats by detecting relevant online **contents** over the Internet. Thus, the proposed solutions are mainly aimed at supporting the automatic discovery and analysis of relevant online sources and contents in the Surface Web, but also in the Deep Web and Dark Nets. Indeed one of the key elements of this kind of solutions aims at innovating and improving the intelligence processes in such a dark part of Internet that is only accessible with specific clients, like hidden services in TOR and I2P. These Dark Nets are used from organizations to hide their identity and publish without censorship.

Most of the detected relevant multimedia contents contain information about **activities and events**: one of the challenges of the solution is to automatically identify and cluster/classify such activities and reconstruct the chain. However the solution is also about the automatic detection and analysis of **people and groups (and relationships)**, through the **identification of identities and the understanding of capability and intentions of individuals or organizations that may be engaged in actions**, with special focus on **propaganda, training, and disinformation**. Terroristic groups, leaders and suspicious people will be detected, analyzed and monitored at different levels, including sociological, criminological, and psychological, in order to identify behavioral patterns on which to focus during the analysis processes. In this context it is crucial to identify and **recognize the real identity of people hidden behind the virtual accounts.**

3 Expected Demo at ITA-SEC 2017

This paper is focused on the following solutions, approaches and methodologies that are also used into H2020 EU funded project DANTE with a strongly activities of personalization, improving, enhancing and consolidation of:

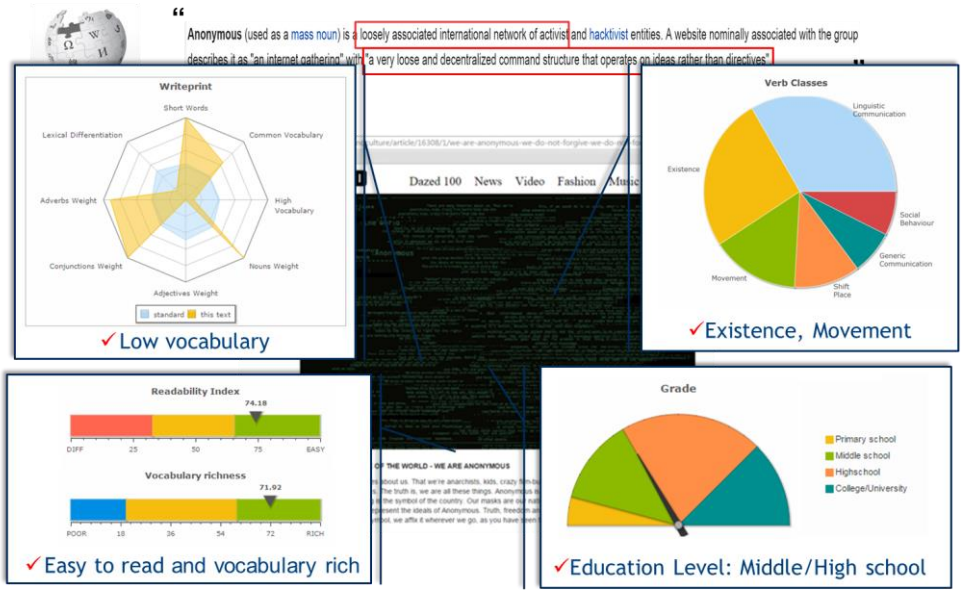
- Behavioural algorithms
- Extraction of relationships/facts
- Multilanguage approach
- Multimedia approach

3.1 Behavioural issues and Relationships

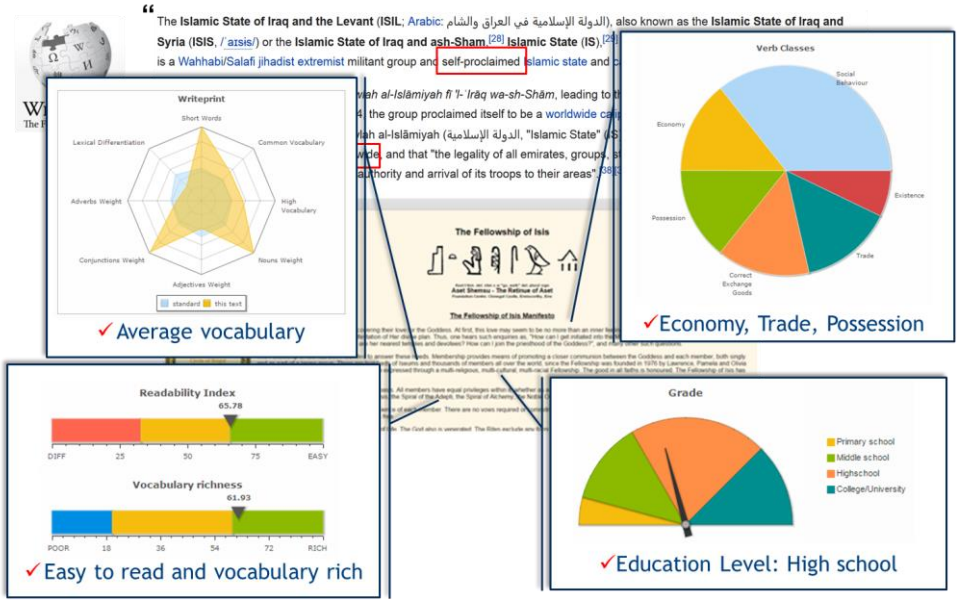
Live demo: analyzing a text coming from Internet (English language)

Some experiment coming from emotional and stylometric analysis of posts of relevant criminal organizations:

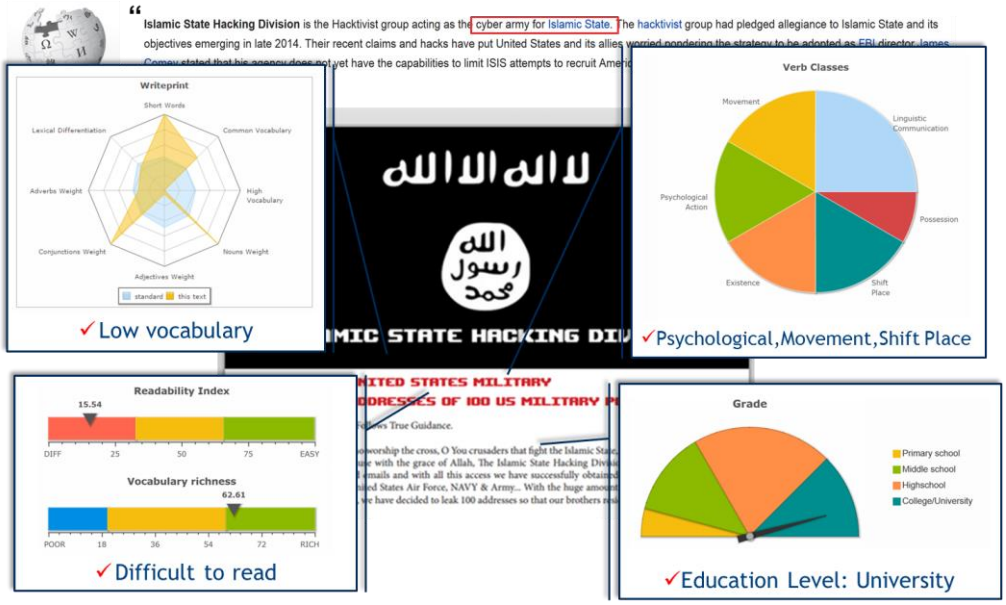
Example #1: Anonymous Manifesto



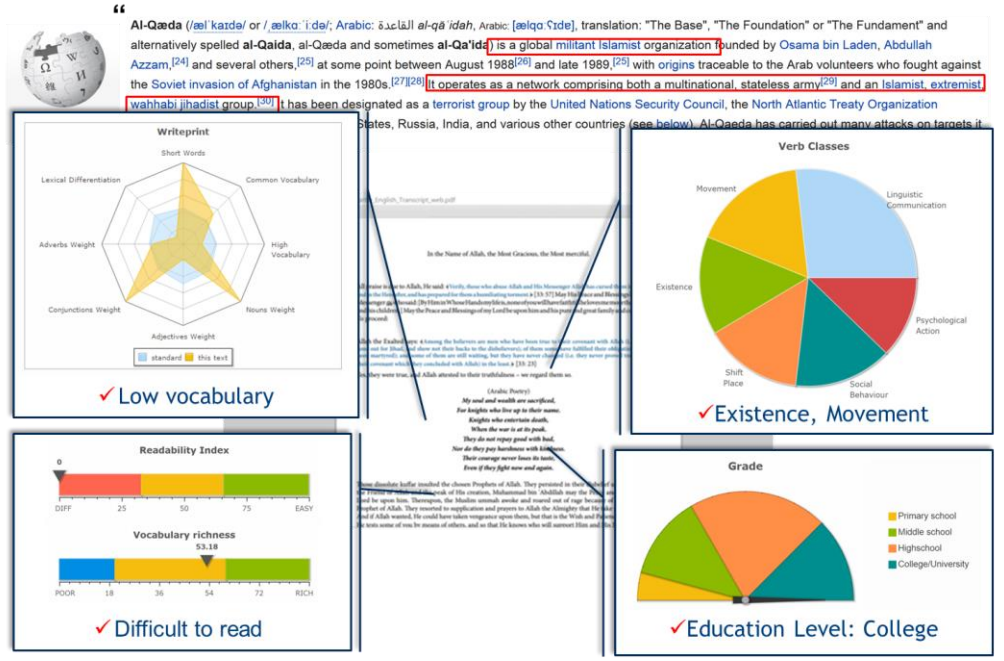
Example #2: ISIS Manifesto



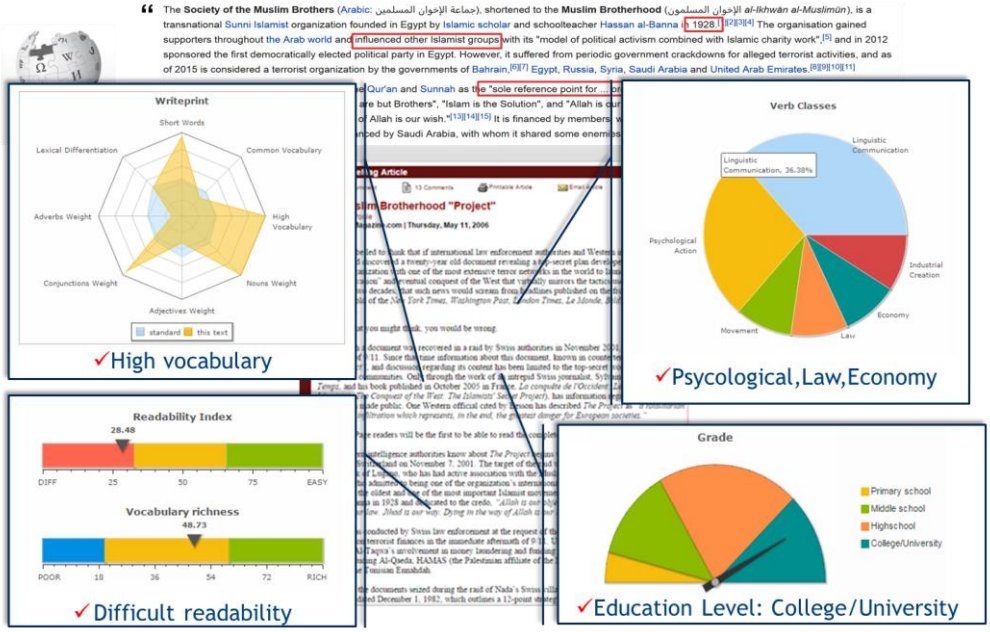
Example #3: Islamic State Hacking Division



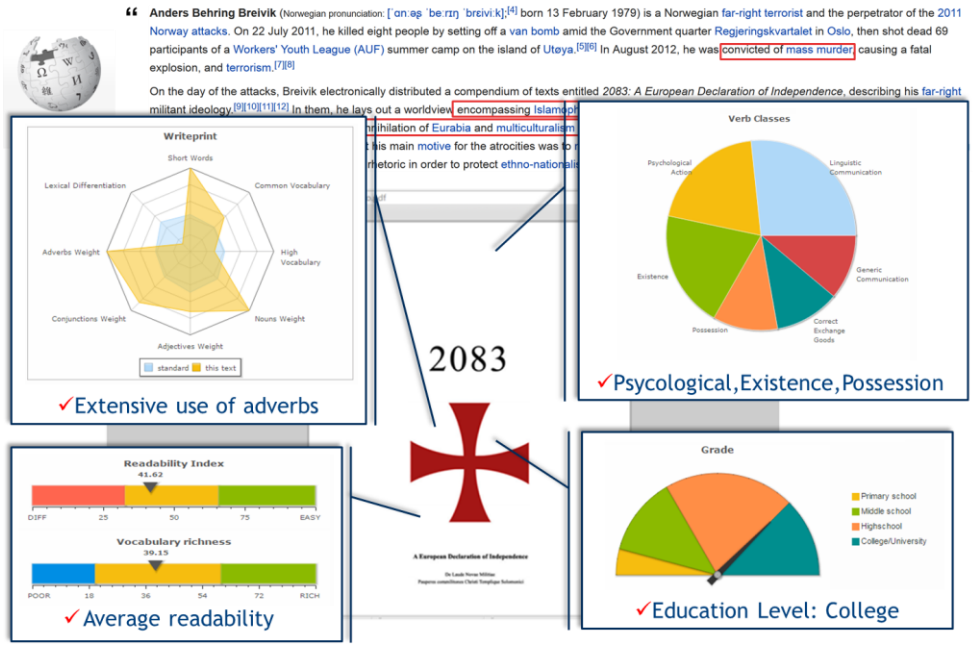
Example #4: Al-Qaida - Charlie Hebdo Attack



Example #5: Muslim Brotherhood Project



Example #6: Breivik - The Norway Lone Wolf



The previous examples uses algorithms based on **deep semantic rules**. This scenario can be expanded and enhanced with an HYBRID approach containing also **machine learning techniques**. So semantics and machine learning in order to:

- give smartness on creation of the knowledge base (ontology) of the final solution; take a look for example at the numerous way of speaking and writing existing into common channels of communication/social networks (see point 3.4)
- give more parameters/indexes to improve and reach a new innovative stylometric approaches. Following some experiments coming from EU DANTE project:

Starting from the output of Stylometric analysis we saw during the last live demo, so:

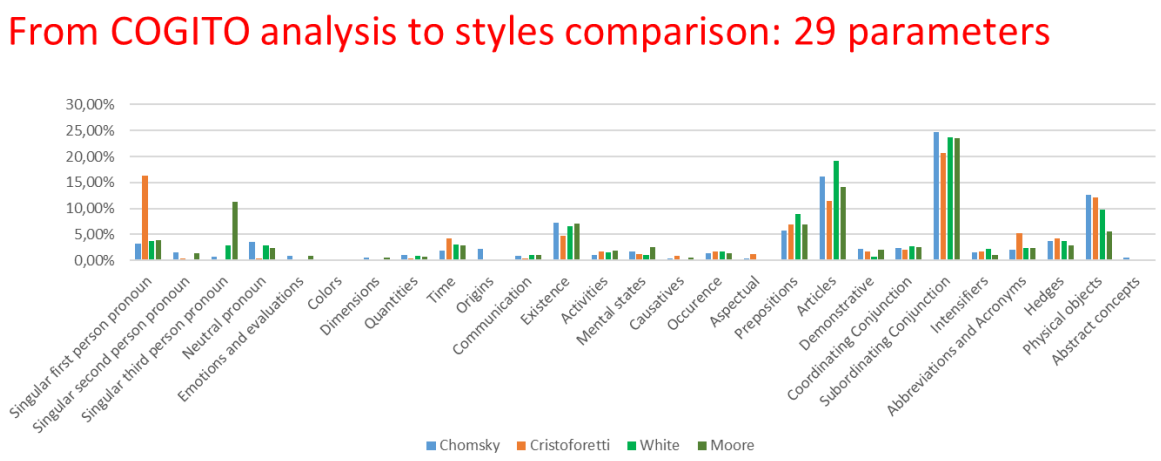
- Readability
- Vocabulary richness
- Registers and slangs
- Grammatical tenses
- and so on…..

in order to understand if it is possible to recognize the **true author of a message and/or a document**.

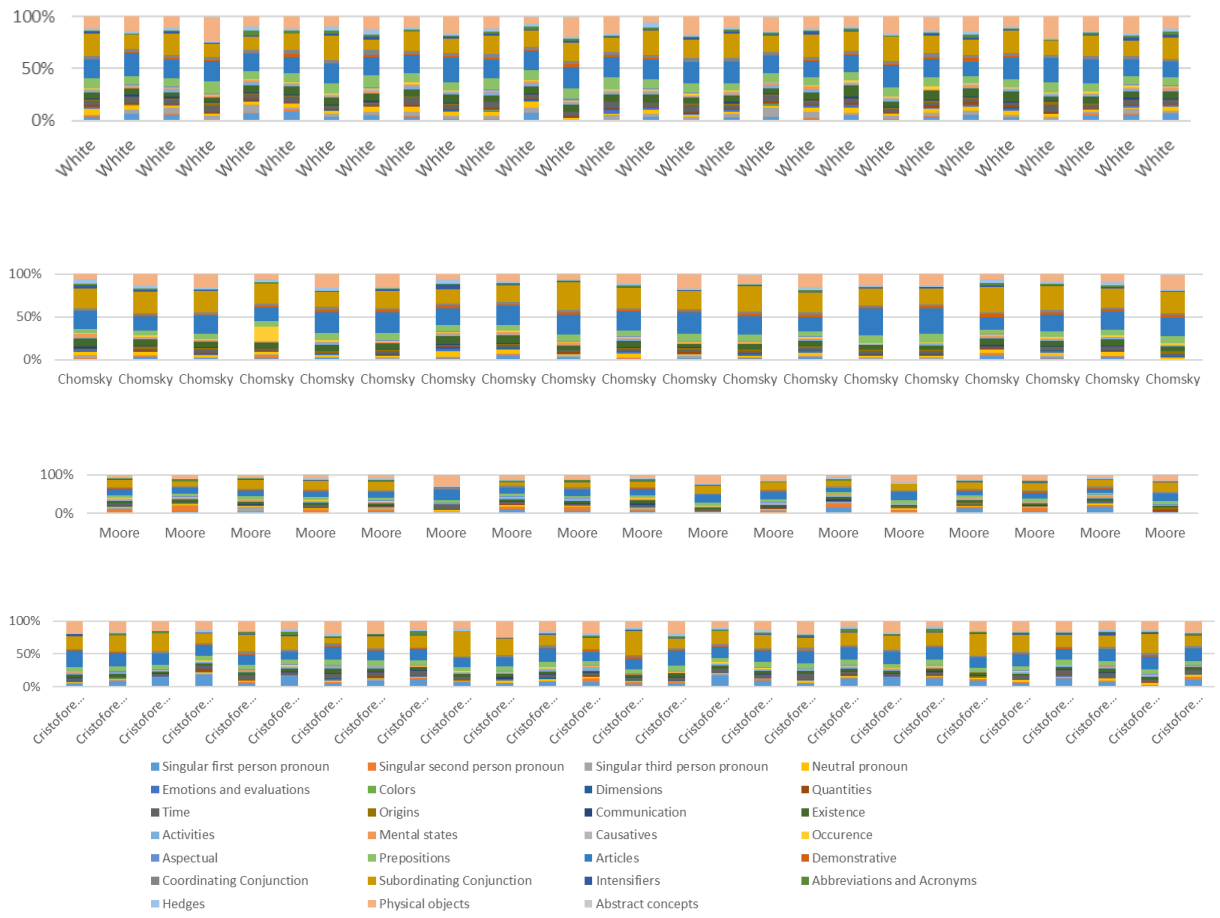
All the analysed parameters have been used for training a Machine Learning system (**Weka algorithm**) which is in charge of finding similar features among various input texts.
This is a **supervised technique**, so with the knowledge created from specific documents such as a training set.

Training set is related to: 96 texts of 4 different authors (Michael Moore, Samantha Cristoforetti, Michael White and noam Chomsky).

Below there is the Stylometric DNA linked to the four different authors (with different views):



.



Testing set is related to 11 new texts and the final result is:

Correct classes	Predicted classes (ML)
Cristoforetti	Cristoforetti
White	White
Cristoforetti	Cristoforetti
White	White
Chomsky	Chomsky
Moore	Cristoforetti
Moore	Moore
White	White
Chomsky	Chomsky
Chomsky	Chomsky
Cristoforetti	Cristoforetti

3.2 Multilanguage approach

Having a Multilanguage approach is a must in the analysis of terrorist and criminal contents.

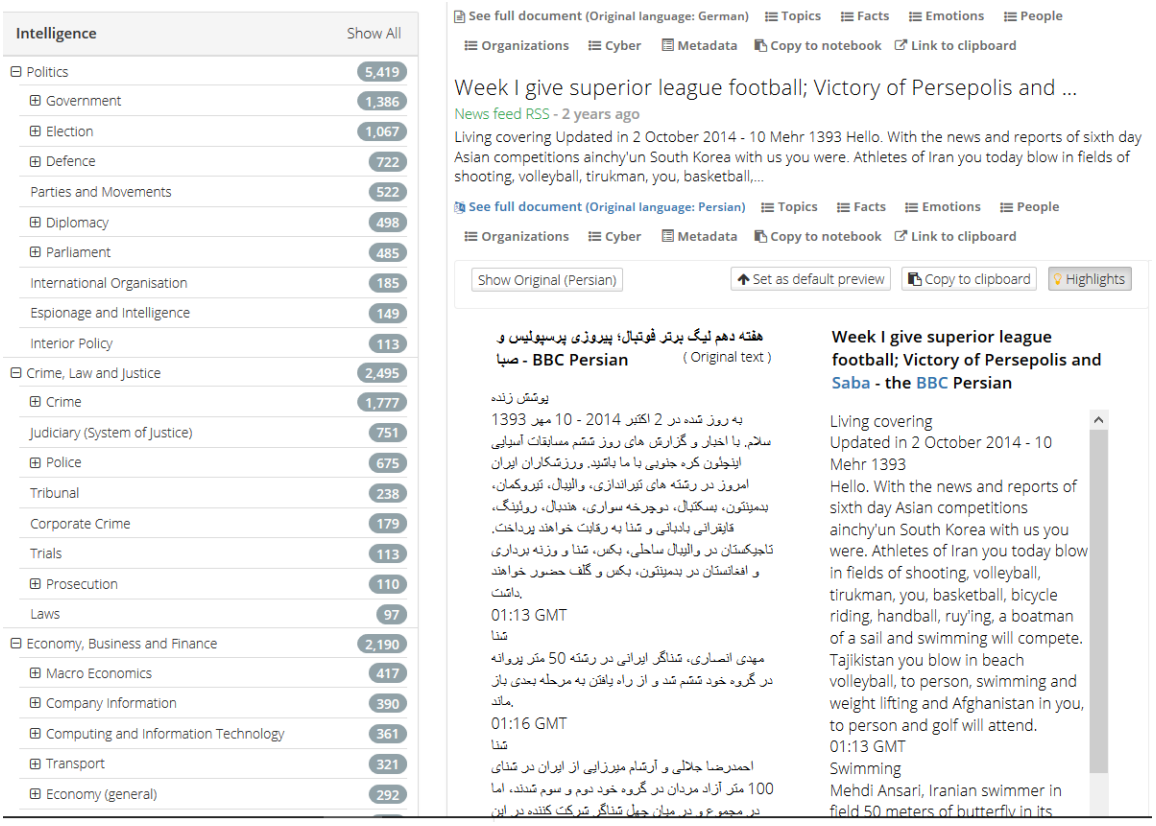
Different languages can be used in the same communication thread, cover terms done in another language considering the main language of the text can be found and multiple dialects terms. These are the Multilanguage issues an analyst has to face during his navigation into either Surface or Dark Web.

Multilanguage approach can be faced into 2 different ways:

- Deep semantic technology to analyze each specific language
- Automatic translation to normalize multiple langue into “English” for example

Both approaches can be used too in order to propose an HYBRID approach (again as previously mentioned about stylometric analysis); so an automatic translation for example, of Persian content into English and analysis by a semantic technology strongly focused on the destination language (so English.

Live demo: analyzing a text coming from Internet (Persian language)

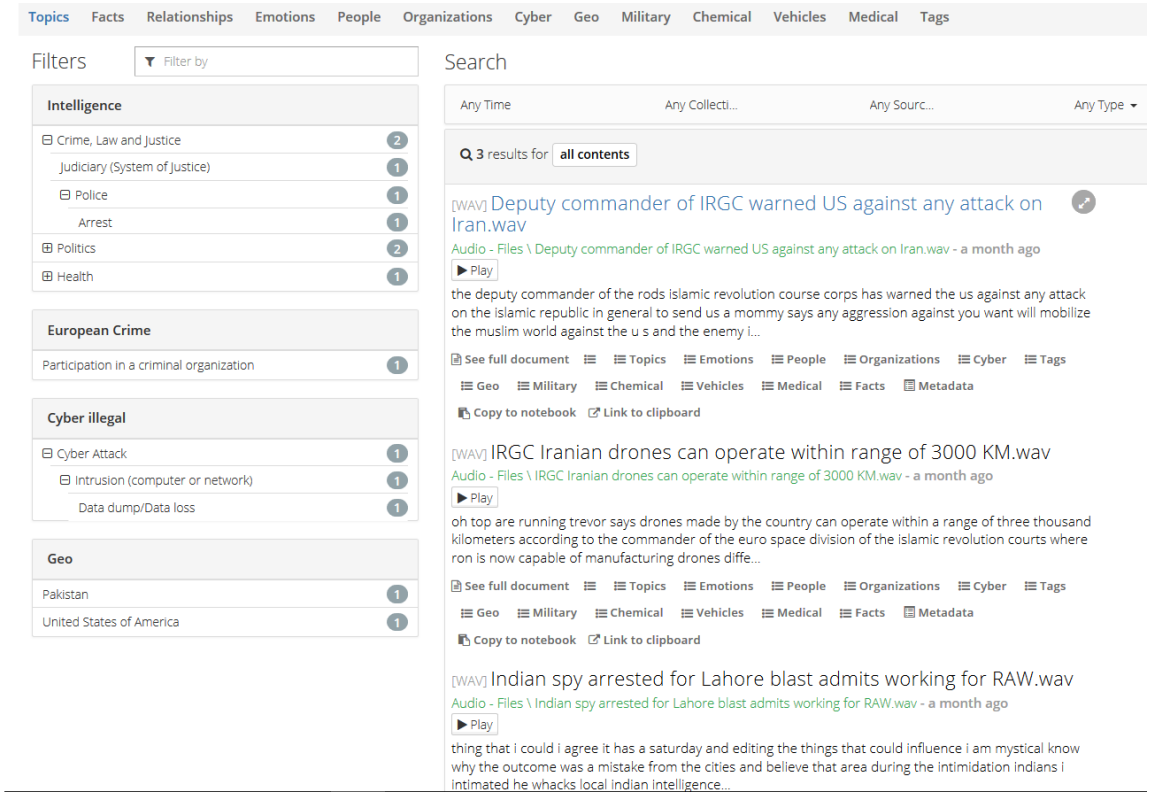


3.3 Multimedia approach

There are different countries where local radios and broadcast sources has a relevant importance compared with writings coming from social network. So Speech To Text (STT) technology, mainly tuned on the specific sensor, has a key value into the final solution

Also in this case multilanguage issue is an important aspect to be considered; in particular, referred to analyze local dialects.

Live demo: analyzing an audio/video coming from Internet (English language)



3.4 New challenges of behavioural algorithms and multilanguage approach

The output of behavioural algorithms can be the input of the following scenarios:

- Fake messages/authors and Disinformation (see previous example regarded Weka machine learning algorithm)
- Encoded messages (also related to multilangauge approach)
- Mapping virtual identity with physical identity

- Propaganda (radicalization process)

In order to reach these goals, also in relation of EU DANTE project, particular focus is given to the managing and comprehension of slangs, acronyms and abbreviations included in the content and also misspellings. This approach can be more relevant if we use multilanguage approaches and transcription ones.

As the end, last but surely not least, new stylometric cool features requested from the Law Enforcement Agencies(LEAs):

- Understanding the gender of the writer/speaker
- Understanding the age of the writer/speaker
- Discovering the leaderships
- Linking to mother tongue speaker/writer

4 References

Aa.Vv. *The use of the Internet for terrorist purposes*, 2012 report from UNODC (UNITED NATIONS OFFICE ON DRUGS AND CRIME) Retrieved from http://info.publicintelligence.net/UNODC-TerroristInternet.pdf

Robert Anderson, Jr., 2014, *Cyber Security, Terrorism, and Beyond: Addressing Evolving Threats to the Homeland*, Statement Before the Senate Committee on Homeland Security and Governmental Affairs Washington, D.C. Retrieved from https://www.fbi.gov/news/testimony/cyber-security-terrorism-and-beyond-addressing-evolving-threats-to-the-homeland

Ben Saul, 2012, *Terrorism*, Bloomsbury Publishing

H.M. Virupakshiah, 2009, *Terrorism Challenge Diplomacy*, Concept Publishing Company

Aa.Vv, 2015, *Cyber Counterterrorism, Cyber International Conflict, Virtual Cyber War Crimes*, Journal of Legal Technology Risk Management

Haim Assa, 2014, *When Marx Meets Nietzsche in Cyberspace: Revolutionary Praxis and the Will to Power in Twenty-First-Century Revolution*, Contento de Semrik