

Deep Semantic Behavioural algorithms mapped to Machine Learning techniques in order to discover terrorist related online contents

Gianluca Sensidoni

Expert System S.p.A.

gsensidoni@expertsystem.com

Abstract

Deep 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. The final target is to discover (by “connecting the dots”), detect and monitor potential terrorist-related activities and people. At the conference a live demo of such algorithms and some experiment coming from EU DANTE project (an H2020 EU funded research project in 2015) 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**.

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.

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 Deep 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 (**Ex-Ante scenario**).

At the same time analysing and understanding the best practise and events after the threat occurs is a fundamental step of the investigation and reasoning tasks (**Ex-Post scenario**).

The target is 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 WSIS 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:

- **Deep Semantic Behavioural algorithms:** live demo (www.intelligenceapi.com) of emotional and stylometric (writeprint) analysis of content coming from social network. The analysis and extraction of semantic tags will be done by the synergy of deep semantic algorithms (based on ontologies) and machine learning ones (hybrid engine). This hybrid approach so to give smartness on creating new semantic rules and on enhancing the ontology (knowledge base) of the system in a semi automatic way
- **Multilanguage/Multimedia approach:** the added value of analyzing content (not only text based) also coming from audio/video sources and into different languages/dialects. Having a Multilanguage/Multimedia 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.
- **New challenges of behavioural algorithms:** The output of behavioural algorithms can be the input of the following scenarios:
 - Fake messages/authors and Disinformation
 - 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 H2020 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 (multimedia).