Towards an Automated Analysis of the Online Supply Chain of Novel Psychoactive Substances

Clara Bacciu[1], Fabio Del Vigna[1], Andrea Marchetti[1], Maurizio Tesconi[1] and Paolo Deluca[2]

[1]Istituto di Informatica e Telematica, CNR, Pisa, Italy
[2]Addictions Department, King's College, London, U.K.
{name.surname}@iit.cnr.it, paolo.deluca@kcl.ac.uk

Keywords: Web Intelligence, Web Mining, Web Applications, Data Visualization.

Novel Psychoactive Substances (NPSs), also known as *legal highs* or *smart drugs*, are legal alternatives to illegal drugs. Many drugs consumers are appealed by the opportunity of buying substances without any legal consequences. Online shops, virtual marketplaces and other trade channels thrive in this legal grey area. The health risks connected to this phenomenon are high: every year hundreds of people present symptoms deriving to the use or abuse of those unknown chemicals, and health professionals may struggle to provide the appropriate treatments.

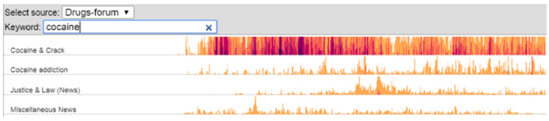
EU is taking some countermeasures, forbidding the sale of the NPS as soon as it is established their risk for the health, but natural or synthetic new substances are continuously discovered or manufactured, and it is difficult for legislators to keep up. To cope with the lack of regulation in this market sector, EU is funding the CASSANDRA project, to study and comprehend NPSs lifecycle and supply chain, through the automatic analysis of user generated content (forums and social media) and online markets.

During the first year of activity, we combined data gathering, analysis, and visualization techniques to i) provide an insight over two large forums, Bluelight and Drugs-forum, that host discussions about drugs since more than a decade; ii) investigate how substances sold by online shops of the NPSs supply chain map inside the forums and iii) investigate how social networks like Facebook and Twitter are used to avertise and discuss drug consumption.

In order to gather as much data as possible from forums, we developed an ad-hoc web scraper. The system keeps track of the forum hierarchy and structure, keeping all tags and other metadata associated to posts, threads and forums. All the content is anonymized and stored in a relational database with an associated text indicization. We got a snapshot of Drugs-forum and Bluelight, whose content spans respectively from 2003 and 1999 to today, with more than 1 million and more than 3 millions of posts, and about 200 thousands and 350 thousands of users respectively. A selected set of 10 online NPSs shops underwent a similar scraping and storage phase, while we crawled the social media pages connected to those shops through the provided APIs. We extracted a list of the advertised products in those shops and pages, finding more than 250.

The forums have been the starting point and the core of the analyses so far. We developed some interfaces to investigate their structure, the number of posts per thread and the number of posts per user (which, as expected, follow a power low distribution).

Moreover, we analysed the textual content of the posts, showing the number of occurrences of terms over time (Figure 1), in which sections a series of known NPSs are first mentioned, the terms co-occurring with other terms. In particular, this last analysis is leading to an automated system to show the most frequent symptoms mentioned together with the name of a substance. We also analyzed the hyperlinks appearing in the forums, and compared them with a comprehensive list of online NPSs shops and related social media accounts, finding that they don't quite overlap, and which NPSs sold in those shops are mentioned in the forums, finding that almost every substance is mentioned.



Select source: Drugs-forum
Keyword: cocaine

Figure 1: A web interface that shows the number of occurrences of a selected term in each section of the forums over time. The colour gets darker as the number of occurrences increases.

In the future we plan to extend the analysis to dark web marketplaces. Future work will also involve the development of an automatic system to detect the mention of unknown substances, in order to monitor the discussion about them from the start, to understand where substances are first mentioned and sold, and how the supply chain evolves.