Analysing the digital transformation of the market for fake documents using a computational linguistic approach

Clara Degeneve [a], Julien Longhi [b,c,d,e], Quentin Rossy [a]

[a] Ecole des Sciences Criminelles, University of Lausanne, Switzerland

[b] CY Cergy Paris Université, France

[c] Institute of Digital Humanities, 33 Bd du Port, F-95000 Cergy-Pontoise, France

[d] Agora Lab, 33 Bd du Port, F-95000 Cergy-Pontoise, France

[e] Institut universitaire de France, 1 rue Descartes, F-75231 Paris, France

Analysing the digital transformation of the market for fake documents using a computational linguistic approach

Highlights

- Analyse the market of fake documents on an online anonymous market
- Explore the informative potential of computational linguistics to analyse language traces
- Highlight the digital transformations of the market of fake documents to specific types of digital document

Abstract

The market for fake documents on the Internet is a topic that has not been yet explored in depth, despite its importance in facilitating many crimes. This research explores the market of fake documents on the White House Market anonymous market with a computational linguistic methodology; more specifically using textometry. The textual corpus is composed of the data of the ads titles as well as the profiles of the sellers, which are analysed as traces of their online activities. We investigate how these remnants can help to answer general questions: what kinds of fake documents are sold, can we distinguish types of sellers based on their selling activities or profiles and can we link distinct vendors based on language traces similarities? The free software IRaMuTeQ was used to carry out the analysis. The results show that the textometric methods have a real potential in terms of classification, highlighting the different products on the market and grouping the sellers according to their offers.

Keywords

Fake documents, cryptomarket, computational linguistic, textometry, language trace

Introduction

Identity documents are needed for many everyday activities such as subscribing to a telephone service, taking out a loan from a bank, crossing borders, or buy alcohol to name a few. In addition to granting rights to their rightful holder, they can confer trust, authority, benefits, and responsibilities. This makes them highly attractive assets for individuals deprived of such benefits. Document fraud is thus a convenient solution, sometimes the only one, to get pass identity checks and controls and access to the places or services sought (Baechler, 2020). But identity documents are not the only documents used for obtaining benefits and are thus not the only ones affected by forgeries. They are a particular type of 'secure document' such as travel documents, banknotes, or diplomas, which can be defined as document giving legal or commercial function and value to the holder and have the property of allowing the confirmation of its veracity, validity and authenticity as a genuine document (Ombelli & Knopjes, 2008). This makes fake secure documents a hot product for the illicit market. In this paper, we will refer to 'fake documents' as forgeries of identity document and secure documents.

The market for fake documents has found its way to extend to online marketplaces. The marketplace chosen for this research is the 'White House Market' (WHM) cryptomarket, still active at the moment of the study from January to June 2021. It was then one of the most active cryptomarkets on the Dark Web[1], until its closure in October 2021, with nearly five hundred thousand users and almost three thousand sellers.

We focus on the textual data present in HTML traces collected on the WHM cryptomarket. These traces can be apprehended from several forensic perspectives. The one we focus on is called by Renaut and colleagues (2017) the "language trace". It is the remnant of an action (Margot, 2018; Roux et al., 2022) which is the writing of an illegal or litigious text by an author with an informative potential on its source, but also on the illicit activity itself. Language trace may result from illicit acts that can be committed through language, such as threats, defamation or even an apology for terrorism (Renaut et al., 2017). In this study, we investigate language traces resulting from the publication of illegal ads posted by vendors to reconstruct their activities and get insight upon the online market for fake documents. We investigate how these remnants can help to answer general questions: (1) "what kinds of fake documents are sold?", (2) "can we distinguish types of sellers based on their selling activities or profiles?", (3) "can we link distinct vendors based on language traces similarities?".

The analysis of 'words as traces' in the forensic context raises many questions about the objectivity, reliability and reproducibility of the methods used to analyse? language traces. Since traces are more often than not considered as silent witness, considering words as traces is not obvious. From a methodological point of view, this research is thus based on computational linguistics, which, integrated with forensic science, is commonly called "forensic linguistics". Seen as a particular field of applied linguistics it is defined as "*a branch of linguistics which applies in the field of justice technics from*

[1] https://darknetone.com/market/white-house-market-whm/

linguistics and phonetics for the analysis of evidence in court" (Renaut et al., 2017, p. 426 free translation). However, such a definition reduces the scope of the methods to the trial (Svartvik, 1968), whereas forensic science covers the exploitation of traces more broadly in policing (Roux et al., 2022). Indeed, computational linguistics approaches can be exploited for global forensic purposes, such as authorship attribution (Lam et al., 2021; Overdorf et al., n.d.; Peng et al., 2016), or the recognition and classification of illegal activities (He et al., 2019; Nabki et al., 2017).

In this case study, textometric methods have been selected to carry out the recognition and classification of illegal activities tasks. Textometry is based on "*the lexicon, that is the counting and distribution of words within the texts of a corpus, but also other levels of linguistic and textual description (morphosyntax, textual structures, etc.)*" (Pincemin, 2020). The main interest of choosing this method is that it includes both a quantitative and a qualitative dimension. Indeed, textometry is based on statistical analysis of textual data, but it integrates what Pincemin (2020) calls a "*back to the text*" step, where the scientist evaluates the results of the computational analysis by considering the surrounding context of the detected textual forms.

This paper is structured as follows: first, a review of the existing literature on the online market for fake documents is presented. Then, the research methodology and the different technical aspects are developed. Finally, the results are presented and discussed.

The Online Market for Fake Documents

The market for fake documents has found its way to extend to online marketplaces (Baravalle et al., 2016; Bellido et al., 2017; Mireault, 2016). These online markets form a specific type of 'virtual convergence settings' where offenders (i.e. sellers and buyers) interact and leave traces (Rossy & Décary-Hétu, 2017; Soudijn & Zegers, 2012). They can take multiple forms such as publicly accessible websites (e.g. online shops or platforms) or more private channel of communication such as private groups on social media or instant messaging app. Because private settings are more difficult to study due to accessibility and ethical issues, this research focuses on a specific type of public setting: online anonymous market present on the ".onion" darkweb relying on the TOR network which is also known as 'cryptomarket'. A cryptomarket is an online marketplace on the darkweb, which is quite similar to regular e-commerce platforms. Sellers post their ads and payments are carried out by cryptocurrencies. These anonymous markets allow users to engage in illegal activities while limiting the risk of being checked by the authorities (Kruithof et al., 2016; Martin, 2014).

In addition to the reasons of accessibility, this choice to analyse a cryptomarket is based on three main reasons. First, online platforms bring together a variety of sellers and buyers, allowing for analysing the activity of multiple stockholders as a whole, whereas dedicated online shops selling fake documents appears to be quite rare (Laferrière & Décary-Hétu, 2022). Second, the tracking of dedicated online shops involves gathering heterogenous data, whereas platforms have a unified internal structure. Finally, the choice of monitoring the ".onion" darkweb is adequate since illicit markets on the web are known to contain scams, while darkwebs give a higher level of anonymity.

Indeed, darkwebs such as ".onion", which is recognized as the main one, concentrate illicit activities and in particular illicit markets. They offer a high degree of anonymity for both the manager of the websites and their users. They are not regulated by the DNS system of the ICANN, but are "Special-Use Domain Names" that are auto-regulated and self-authenticating since they are solely derived from cryptographic keys (*RFC 6761 - Special-Use Domain Names*, n.d.; *RFC 7686 - The ".Onion" Special-Use Domain Name*, n.d.). Moreover, the ".onion" darkweb is settled upon the TOR network which secures the content of communication through encryption and protect anonymity with the use of multiple intermediary nodes and a dedicated communication process known as the "onion routing" to exchange information between computers without directly exchanging identifying information such as IP addresses (Loesing et al., 2010).

Holt & Lee (2022) have formalize the mechanism for the online selling of fake ID documents with a crime script. By analyzing 19 sellers found both on the Clear and Dark Web, they identified four main steps:

- "*precondition of potential customers"*: these are the arguments put forward by sellers to attract buyers, such as the possibility of travelling,
- "*initiation and entry into the market*": buyers can access markets via their browsers, sometimes after viewing advertisements that allow them to choose the seller. An initial contact then takes place between the buyer and the seller,
- "v*endor actualization and doing of document creation*": the buyer pays for the order after having outlined his or her requirements to the seller, who then proceeds to create the document. The seller then proceeds to create the document,
- "*exit scripts of the customer and vendor*": once the transaction is done, contact is often broken between the seller and the buyer, except for those who are trying to build customer loyalty or who offer order tracking.

This description of the process outlines two dimensions of investigation about the Market. The first one is related to the nature of the target of the transaction (i.e. the fake document). The questions are "what types of documents are buyers looking for" and "for what purposes"? The second dimension is related to the means of contact used to enter into the market. The questions are "what are the means" and "how to detect and monitor online settings used"? Globally, there are still very few specific studies addressing these questions about the market for fake documents. This might probably be explained by the small proportion that fake documents represent among all other illicit products available on cryptomarkets.

According to the study of Baravalle and colleagues (2016), that analyses the sale of fake documents on cryptomarkets, these products are much less prevalent than others, such as drugs, which account for 80% of the products for sale on the "*Agora*" cryptomarket (N = 30,680 products and sellers pages collected). By comparing ads for drugs and fake IDs on this platform, they determined that the market for fake IDs was more concentrated, with fewer sellers and ads than for drugs.

In his book, Akhgar and colleagues (2021) consider the fake identity document market within the "*fraud and counterfeit*" category of product that can be found in the Dark Web, among 5 other major product types. The description given is limited to "*Fraud and counterfeits – the document fraud, with the online trading of fraudulent, fake, stolen and counterfeited documents and cards, such as fake passports or identification cards and cloned and stolen credit cards or accounts, is emerging and one of the fastest-growing markets, in all types of criminal activities including terrorism. 'Card shops', for example, are one of the specialty markets in the Dark Web*." (Akhgar et al., 2021, p. 101).

In Mireault's Msc thesis (2016), fifty websites selling counterfeit documents on the web were analysed to describe their visibility, products sold and the sales process. The online stores appear to exploit online forms and emails as their preferred means of communication. They also favour payments by digital currency (e.g. Bitcoin), but also international money transfers (Western Union and MoneyGram), which are well known to be used by scammers. The main types of fake documents detected were a driver's license on 68% of the websites (n=34), identity cards (28%, n=14) and student card (24%, n=12). Passports, visas, residence and civil status documents were detected on 16 percent of websites (n=8). Professional cards, diplomas and fancy documents are sold on a smaller number of sites (10%, n=5).

On the darkweb, dedicated online shops selling fake documents appears to be quite rare. (Laferrière & Décary-Hétu, 2022) identified 108 illicit online shops, but only 6 (5.5%) are dealing fake documents. Much more websites appear to sell drugs (37%, n=40) or carding credentials (31%, n=34). No information upon the products sold is detailed in this global study.

Bellido and colleagues (2017) investigated the acquisition mechanisms of fake documents in order to establish a state of the market. Using a keyword search on Bing, Yahoo and Google browsers, as well as a more extensive search for new links contained in previously crawled pages, they obtained a total of 375 URLs, 357 distinct hostnames and 223 identifiers. They determined the most common ways in which sellers make themselves visible to their buyers, via different web spaces. Dedicated videos represent "*37% of the means of selling*", publications on forums and blogs represent 27% of these methods, hidden TOR sites 19%, dedicated sites 12% and finally evaluation and advice sites represent 5% of the means of selling. The authors also detailed the sales process by first determining the main motivations invoked by sellers to induce customers to buy a fake ID, as well as the main means of contact and ordering. Their results seem to show that, regardless of the distribution medium used, email is consistently found as a means of contact, even if it is not the most frequent. They then conducted a market analysis to see which products are the most sold and at what price. These parameters seem to vary depending on the platforms used, but the driver's licence seems to be the most commonly sold and cheapest document, compared to the passport and ID card. Those results are consistent with the results found by (Mireault, 2016).

Methodology

Dataset

The data used for this research have been collected from the cryptomarket ‘White House Market’ (WHM). This cryptomarket, online from February 2019 to October 2021, was one of the major cryptomarkets in the Dark Web at the end of the study. Twenty crawls were performed from 11.08.2020 to 11.03.2021. The webpages of the advertisements as well as the sellers’ profiles have been extracted for a total of 83’516 distinct ads and 2’519 distinct vendor profiles (see Table 1). All parts of the collection process were based on open-source APIs and own developments done by the ESC.

Sections	Distinct Vendor Url	Distinct Product Url	Distinct Product Title
Drugs	2’296 (91.1%)	68’699 (82.3%)	82’618 (84%)
Online Business *(excluding SSN/DOB/PII)*	183 (7.3%)	6’681 (8%)	7’294 (7.4%)
Services *(excluding “Fake Documents”)*	163 (6.5%)	2’275 (2.7%)	2’343 (2.4%)
Software	85 (3.4%)	2’522 (3%)	2’606 (2.6%)
Forgeries/Counterfeits	81 (3.2%)	1’785 (2.1%)	1’841 (1.9%)
Online Business > SSN / DOB / Other PII	72 (2.9%)	384 (.5%)	445 (.5%)
Services > Fake Documents (Digital)	62 (2.5%)	772 (.9%)	801 (.8%)
Services > Fake Documents (Physical)	35 (1.4%)	331 (.4%)	343 (.3%)
Defense/Counter Intel	27 (1.1%)	76 (.1%)	84 (.1%)
Total	**2’519 (100%)**	**83’516 (100%)**	**98’375 (100%)**

Table 1 : Number of distinct vendors and ads for each section of the cryptomarket. The number of ads is counted based on distinct URLs of the ads, but also with the number of distinct product titles for each product since the product title might have changed over time.

The sections presented here have subsections. The subsections “*Fake Document (Digital)*” and “*Fake Document (Physical)”* are included in the section “*Services*”. As the focus of this study is on fake documents, those two subsections are treated separately from the rest, for a total of 1103 advertisements (1.3% of all ads) and 86 vendors (3.4% of all vendors).

Pretreatment

To carry out the textometric analysis, we chose to use the software IRaMuTeQ[2], which is a free software based on Python and R. It allows multiple statistical analysis and produce visualizations. It has been chosen for its ease of use and the available textometric methods.

To integrate the data into the software as corpus (i.e. a set of text units to be analysed), they have to fit with a particular format, called "*Alceste*" (Marpsat, 2010). First for the ads, each category is separated from the others and converted into a .txt document (UTF-8 encoding) containing the ad title, category and vendor's name. Every new text is introduced with four asterisks "****". These are followed by the first information, here the name of the vendor, like "**_name1*" and then the name of the corresponding category in the same format. These variables are called "*illustrative variables*", which means that they are not part of the text analysed but used to filter the dataset. The text submitted to textometric analysis is the title of the ad. The descriptions of the products in the ads have been tested in several analyses but didn't give sufficient results to be considered relevant and thus are excluded. The same process is used to prepare the corpus composed of the 86 vendors of fake documents, with their names and date of admission to White House Market as illustrative variable and their profile for the textometric analysis.

In Table 2, it is possible to see that only 69 vendors are taken into account for the "*vendor_fakedoc*". This can be explained by the fact that 17 vendors don't have any written profile. The corpora containing two categories are called "*mixed corpora*". Section specific corpora are used to obtain monothematic sets to avoid replication of the initial structure of the sections (Camargo et al., 2016).

	Corpus	**Description**
Section specific	listing_defense	« Defense » section of the cryptomarket
	listing_drugs	« Drugs » section of the cryptomarket
	listing_forgeries	« Forgeries » section of the cryptomarket
	listing_onlinebusiness	« Online business » section of the cryptomarket
	listing_services	« Services » section of the cryptomarket
	listing_software	« Software» section of the cryptomarket
	listing_fakedoc	« Fake Document Digital/Physical» subsections
	listing_all_without_drugs	All listings except the drugs section
	listing_all	All listings
Mixed	listing_fakedoc/drugs	Combination of the "fakedoc" and "drugs" corpora
	listing_fakedoc/forgeries	Combination of the "fakedoc" and "forgeries" corpora
	listing_fakedoc/onlinebusiness	Combination of the "fakedoc" and "online business" corpora
	listing_fakedoc/services	Combination of the "fakedoc" and "services" corpora
	listing_fakedoc/software	Combination of the "fakedoc" and "software" corpora
	listing_fakedoc/defense	Combination of the "fakedoc" and "defense" corpora
	vendor_fakedoc	Fake documents vendors with a written profile on the cryptomarket

Table 2 – Description of all the corpora created from the data and integrated in IRaMuTeQ

[2] http://iramuteq.org/

Since most of the texts analysed are written in English, the English dictionary is used. For the other parameters of the software, the default values are used.

All the texts are then lemmatized, i.e. all the forms are reduced, "*so that a conjugated verb can be reduced to its infinitive, plural and singular forms, masculine and feminine forms can be grouped together, and, more generally, forms corresponding to the same root with different inflections can be grouped together*" (Guérin-Pace, 1997, p. 867). The interest of this step is to be able to group the main 'forms' and their derivatives under a single label to have a more robust statistical analysis.

The next paragraph describes the textometric methods used on the corpora.

Descending Hierarchical analysis (DHA)

Marpsat describes DHA as a method that allows to "*give an account of the internal organization of a discourse*" (Marpsat, 2010, p. 1). After separating the forms thus obtained into two categories, the "*analysable forms*" (i.e. terms of the text taken into account during the analysis) and the "*illustrative forms*" (Marpsat, 2010, p. 2) that having a purely descriptive value for the classes obtained from the analysable forms, the text is cut into segments. These are parts of the text of fixed size often delimited by punctuation or special characters. These text segments are then grouped together so that they contain enough analysable forms for analysis. They constitute the context of the words. They are created automatically by the software (three lines) (Camargo et al., 2016). A "*lexical table*" (Marpsat, 2010, p. 2) is then formed with the groups of segments in rows and the analysable forms in columns. Finally, the DHA is carried out, gathering the groups of segments into classes. The values of the table contain '1' if the analysable shape is present in the segment group and '0' if it is absent. The algorithm then produces a successive division of the groups into classes, first two, then two more from the largest and so on. The aim is to obtain clusters based on form frequencies representing "*lexical worlds*" (Reinert, 1993) of the texts classified. They are traces of the own 'world' (i.e. discourse universe) of the reconstructed class (Reinert, 1993). They are reconstructed solely upon the forms (and segments) independently of any semantic interpretation.

DHA is performed automatically with IRaMuTeQ, and has been applied to the product's corpus "*listing_fakedoc*". The program took into account 1187 texts over 1321. One hypothesis that could explain this exclusion of certain texts is that the software performs a pre-arbitration in the texts, if some are too heterogeneous compared to the rest of the corpus and are therefore excluded before the analysis. It has been applied to the mixed corpora too (see Table 2), in order to see which categories of products can be found with this method.

Once the DHA is done, several statistics are automatically performed. The number of occurrences of every studied form (i.e. a bag-of-word model) is used to examine each form in a concordance table. It allows observing the form in its original context (i.e. text segments) in regards to the illustrative variables (i.e. the section to which the product belongs or the vendor for instance). Ads published in the wrong sections can thus be identified.

The analysis is finally performed with the "*listing_all_without_drugs*" corpus. The choice to remove all drug ads is made because there are too many drug ads compared to the rest of the products. Then the first five most frequent words of each class created with the DHA are compared with the classification made with the "*listing_fakedoc*" corpus.

Specificities and correspondence factor analysis (CFA)

CFA is a complementary analysis of DHA, which allows associating texts with variable. The DHA table is projected on the axis defined by chosen variables (e.g. the vendor id). It gives a graphical representation of the distance between the different groups according to the analysable forms (Lefer et al., 2016). CFA process a statistical analysis (in our case a hypergeometric law) based on the selected variable.

It is automatically generated successively to the DHA analysis applied to the "*listing_fakedoc*" corpus, showing the distance between the different classes found by the DHA, then to the mixed corpora. In order to find groups of vendors based on their catalogue and then based on their profile, CFA has been applied several times in succession to the corpora "*listing_fakedoc*" and "*vendor_fakedoc*". It was produced using the name of the vendors as the variable. Before each new analysis, the vendors furthest from the core group (named "*outliers*") were removed until no more outliers are detected. The groups of vendors are finally defined based on their position on the axes. Finally, the outliers are analysed separately, in order to understand what makes them different from the main set of vendors.

Similarity analysis

This analysis aims to "*study the proximity and relationships between the elements of a set, in the form of trees*" (Moreno et al., 2015, p. 3). The links between forms are visualized with a graph model. Nodes are forms and links are based on their presence in the same text, which leads to a typical cooccurrence graph. Since the readability and interpretability of a cooccurrence graph are complex due to the multicity of links between nodes, the maximum spanning tree is used to visualize the results (Camargo et al., 2016).

Similarity analysis is applied to the "*listing_fakedoc*" corpus, conserving default settings of IRaMuTeQ. The visualization of the result has been made using the "*yEd*" software[3], IRaMuTeQ providing only a "*.png*" image of the graph. Clusters of words are detected with the "natural clusters" algorithm where each word is only in one group, maximizing the number of edges within it and minimizing the number of edges between other groups (Girvan & Newman, 2002).

Ethical consideration

The collection process relies on online open data gathered with ad hoc web-crawling and web-scraping technologies. The cryptomarket of interest can be considered as public space in regard to the massive number of users and sellers, with data available for every user. The access to the website is conditioned by an account creation but anybody can create one without any condition. To respect privacy, all the vendor's name have been anonymized and no other identifying information was used during the study. All the analyses were based on the texts and the results are presented in such a way that no link can be established with the virtual identity of the sellers. The vendor's profiles were crawled but are not presented in the results. The collected data is intended exclusively for research purposes and cannot be used in any way that could be harmful to the users since no personal data is shared.

Results

Classification of fake documents

Three distinct classes have been found based on the title of the ads with the DFA. The dendrogram in Figure 1 shows that the classes are quite balanced in terms of percentage of forms: 44,5% for class 1, 25% for class 2 and 30,5% for class 3 (N = 1187 ads).

Class 1 gather terms linked to documents sold in digital format, with terms like "*psd*" (which corresponds to the Photoshop format), "*template*" (i.e. a base that can be modified by the user) or "*gimp*" (which refers to a tool for image edition like Photoshop). The presence of the term "*passport*" is linked to the presence of the expression "*passport psd template*" in 75 ads. The term "*card*" is also present, as well as many country names, which may be linked to advertisements offering passports for each particular country. Beside that, most of the terms are not specific to fake documents.

[3] https://www.yworks.com/products/yed

It is harder to find a main topic for the terms gathered in class 2. Nonetheless, terms linked to photos and scans seem to emerge. For instance, the term "*selfie*" corresponds to an image of a person holding an identity document. This type of photo is increasingly required in online authentication processes. The expression "*custom listing*" is also present. Custom ads are specific ads created for a specific client that are often deleted after the sale is made. It is usually a personalized ad without description, as a result of a prior agreement between the seller and the customer (Soska & Christin, 2015). It is noticeable that the term "*passport*" is also present in this class. The term "*identity*" is present but is not directly linked to the term "*card*".

For class 3, the main topic is driver licenses. Ads for driver licenses are more often than not linked to American driver licenses (some American states even emerge as the main words). Some terms linked to security features like "*secu*", "*hologram*", "*uv*" or "*holo*" are also frequently present in this class.

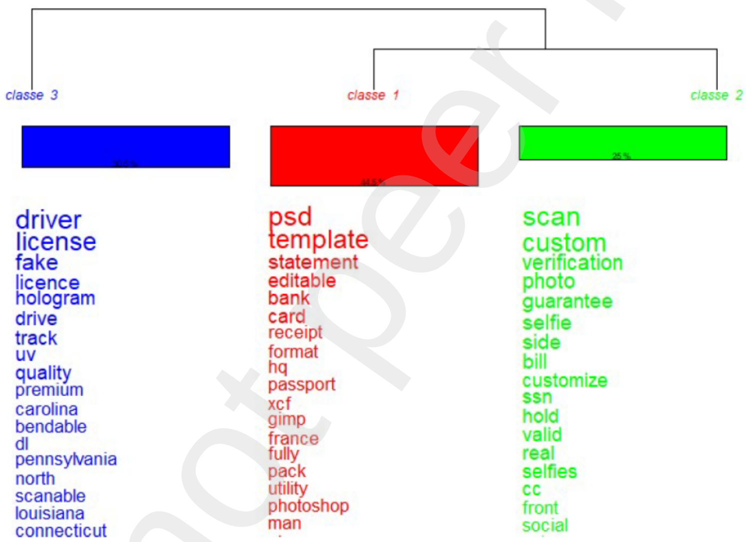

Figure 1 Dendrogram representing the distribution of the analysable forms between the three classes detected by theclassification (N=1187 ad titles analysed)

Combined with the dendrogram, the bag-of-word analysis reveals similarities between classes. In particular, the terms in common are "*passport*", "*fake*", "*license*" and "*quality*".

The CFA (see Figure 2), confirm that the three classes are well separated from each other. Class 3 is quite well isolated. This is particularly true for the American states, which seems to be very specific to this class.

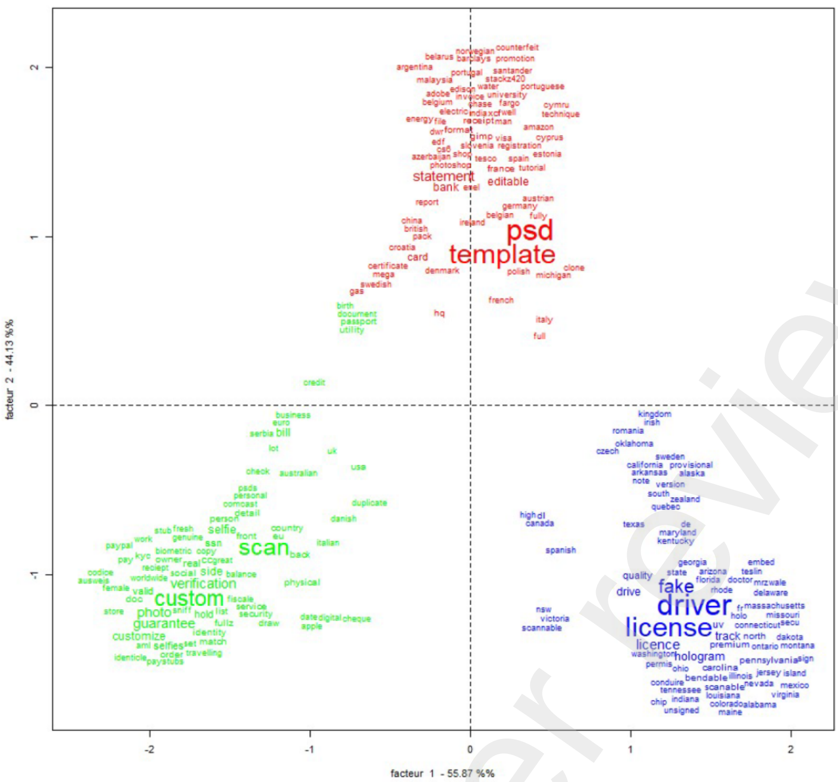

Figure 2 CFA on the classes identified from the DHA (class 1 in red, class 2 in green and class 3 in blue)

Some terms like "*passport*" stands between class 1 and 2, because it appears in the text segment from the two classes. It can be explained by the fact that some ads propose *"passport scan"* in class 2, and "*passport psd template*" is one of the major n-grams from class 1.

Similarity analysis

The analysis detects the terms that are the most frequently used in the ad titles to describe the products, as well as their relationships. The most frequent terms seem to match with the types of counterfeit documents (see Figure 3): "*id*", "*card*", "*passport*", "*driver*" and "*license*". Certain terms are very often used together. For instance, the term "*id*" seems very central and rather generalist, as it leads to different types of documents, not only "*id cards*". Moreover, the analysis leads to the detection of the different digital forms in which products can be found, like "*psd*", "*template*", and "*scan*".

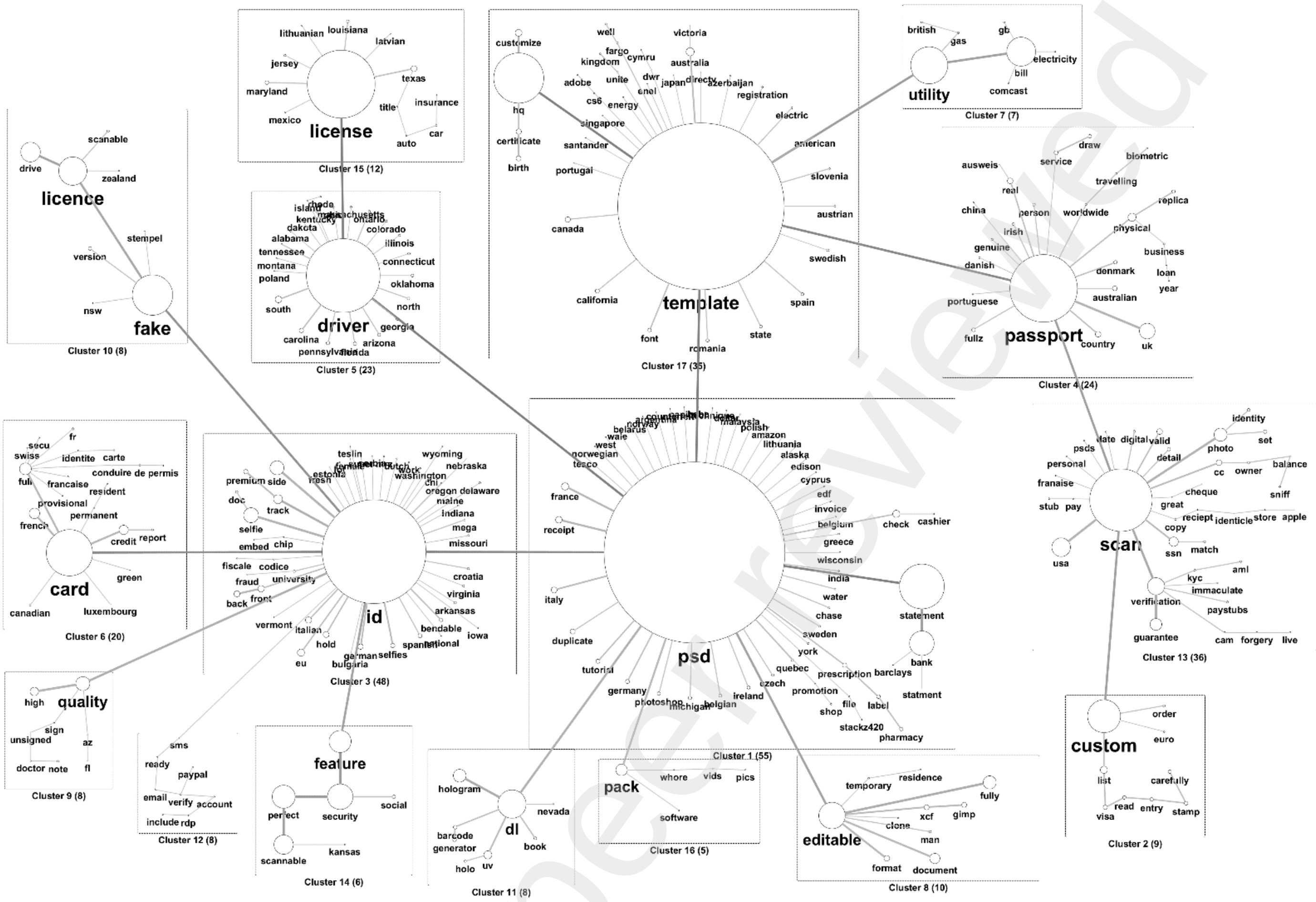

Figure 3 Similarity analysis, cooccurrence spanning tree, and clustering of the main terms (N = 322 forms analysed). The size of the nodes and width of the edges are proportional to the frequency of occurrences

Driver Licenses

A strong link is detected between both "*license*" and "*psd*", with "*driver*" as the central term. This analysis highlighted many forms of expressions to nominate driver licenses, including spelling differences ("*licence*", "*drive*") or diminutive forms ("*dl*"). The majority of the terms linked to the two principal ones are American States, which is consistent with the result found during DHA analysis. Finally, the term "*dl*" is mostly associated with security features, like "*UV*" or "*hologram*".

ID Cards

The term "*id*" is frequently linked to the words "*card*" and "*fake*". The cluster centered around the term "*id*" is mostly composed of American states and country names. Another interesting thing is that the cluster containing the term "*card*" is also composed of French terms like "*conduire*" "*identité*" or "*carte*". The "*id*" word is also frequently linked to the "*psd*" term.

Passports

The term "*passport*" is also central and linked to 24 other forms. It is frequently linked to "*template*" and "*scan*", which gives an indication of the type of counterfeiting. It is interesting to notice that the term "*physical*" is also linked to "*passport*". We can also find "*biometric*" passports, which is indirectly related to "*passport*" (with "*world*" and "*travelling*" between them).

Other types of documents

The similarity analysis also highlights other types of products categorized as fake documents, such as birth certificates (N = 16 and N = 17), utility bills (N = 119 and N = 96), bank statements (N =82 and N = 147), or apple store receipts (N = 4, N = 6 and N = 21).

Digital forms of documents

The first thing to notice is the strong link between "*psd*" and "*template*", which is coherent with the observations that those two terms are often used together in the same texts and seems to be a very current format for digital documents. Passports also seems to be frequently linked to the term "*scan*". Terms like "*gimp*" or "*editable*" or "*xcf*" give other information about the digital documents format.

If digital forms seem to be central for fake documents, one term gives more insights about the context of their potential usage: the term "*selfie*". It corresponds to photos showing a person, holding an ID document. Indeed, the digital transformation of services like neo-banking allows users to validate their accounts completely digital without any physical validation. Clients are requested to send a photo of themselves holding their ID document. Sometimes a piece of paper with the current date is also required in the picture. The identity control process is thus completely digital and might explain the appearance of new forms of illicit market for fake documents. In conclusion, the analysis shows that it is possible to detect specific kinds of fake documents, that appears to be a different kind of document compared to the one described in the literature.(Baravalle et al., 2016; Bellido et al., 2017).

Comparing fake documents with other products

A DHA has been performed on every mixed corpus to see if the method can allow discovering new categories of products. For each class, we can distinguish a main topic that links all the words of the class together. In every mixed corpus, a specific category containing the forms linked to fake documents was also detected, except for the mixed corpus "*Fakedoc_drugs*". This can be explained by the huge proportion of drug ads compared to fakedoc ads (85'262 and 1'321). This is also observed with the "*fakedoc_defense*" mixed corpus where fake documents are predominant (1321 and 88). Observing the CFA generated successively to the analysis, it is possible to notice that, in two cases ("*Services*" and "*Online business"),* the class containing terms linked to fake documents are confounded with other classes. It can be explained by the fact that some terms are common among the products proposed in those categories, like "*card*" (which can fit with "*gift card*" or "*id card*" for example). Moreover, the two subsections of fake documents were originally a part of the "*Services*" category, so it makes sense that the proposed products are close. For the online business section, it is possible to see that some terms are also semantically close. For example, this category contains a lot of "*bank drops*" (i.e. accounts that can be used for money laundering or illegal transfers) or credit cards. Termes linked to payment methods were also detected in the analysis of the fake document sections, such as "*paypal*".

Corpus	**Number of classes**	**Top 10 words in each class (by number of occurrences)**
		account ; warranty ; premium ; porn ; lifetime ; extra ; market ; cheap ; bonus ; month
		hq ; psd ; template ; card ; id ; scan ; license ; driver ; passport ; dl
		hq ; usa ; card ; bank ; cc ; fullz ; fresh ; balance ; email ; verify
Fakedoc_onlinebusiness (N = 9813)	4	database ; record ; hack ; leacked ; plaintext ; million ; dtabase ; leak ; voter ; log
		psd ; template ; id ; driver ; license ; passport ; scan ; hq ; card ; statement
		replica ; perfect ; shoe ; vuitton ; louis ; lv ; gucci ; black ; bag ; dior
		series ; gold ; black ; watch ; rolex ; box ; pro ; counterfeit ; clone ; max
Fakedoc_forgeries (N = 2812)	4	fakemoney ; series ; eur ; test ; pen ; pass ; uv ; usd ; version ; stripe
		pro ; full ; crack ; program ; macos ; adobe ; x64 ; window ; pack ; hack
		full ; software ; mac ; source ; tool ; code ; bitcoin ; rat ; android ; stealer
		premium ; porn ; video ; account ; lifetime ; movie ; book ; private ; spotify ; proxifier
Fakedoc_software (N = 3489)	4	psd ; template ; id ; driver ; license ; passport ; scan; card ; hq ; statement
		psd ; template ; id ; driver ; license ; passport ; statement ; fake ; utility ; usa
		id ; scan ; passport ; utility ; custom ; usa ; quality ; dl ; high ; bill
		complet ; credit ; full ; uk ; pack ; list ; delivery ; real ; utter ; service
		card ; hq ; egift ; pdf ; restaurant ; grill ; pizza ; italian ; group ; bar
		account ; lifetime ; premium ; warranty ; porn ; quality; vpn ; high ; instagram ; guarantee
Fakedoc_services (N = 3436)	6	book ; video ; mastery ; academy ; market ; figure ; facebook; amazon ; trade ; dan
		id ; driver ; license ; fake ; licence ; drive ; quality ; track ; high ; australia
		psd ; template ; passport ; hq ; card ; statement ; utility ; editable ; bank ; fully
Fakedoc_defense (N = 1173)	3	passport ; scan ; card ; utility ; custom ; bill ; verification ; usa ; uk ; selfie
		gram ; quality ; ship ; free ; mdma ; cocaine ; pure ; high ; 5g ; ketamine
		quality ; pill ; mdma ; high ; top ; xtc ; mg ; europe ; dutch ; import
		ship ; pill ; mg ; xanax ; sale ; usa ; duplicate ; 10mg ; bar ; price
		free ; 5g ; uk ; top ; thc ; indoor ; sale ; aaa ; grade ; haze
Fakedoc_drugs (N = 78725)		ship ; free ; thc ; 1g ; new ; premium ; fast ; cannabis ; g ; day
		hq ; card ; usa ; cc ; full ; bank ; fullz ; fresh ; scan ; email
		account ; premium ; warranty ; hq ; extra ; market ; cheap ; bonus ; month ; access
		account ; premium ; warranty ; porn ; lifetime ; extra ; bonus ; video ; movie ; include
		hq ; psd ; template ; perfect ; full ; bank ; id ; scan ; license ; driver
		hq ; card ; egift ; pdf ; gift ; money ; save ; lot ; checker ; code
		full ; pro ; pack ; crack ; complete ; vpn ; security ; program ; gold ; adobe
		database ; record ; hack ; leacked ; plaintext ; million ; dtabase ; leak ; voter ; log
All_sauf_drugs (N = 15844)	8	perfect ; replica ; shoe ; high ; quality ; vuitton ; louis ; lv ; gold ; gucci

Table 3 : Number of classes obtained by CHD per corpus and distinction of a class related to false documents. N indicates the number of analyzed ads for every corpus. The first ten words of each class found are also reported.

Detecting fake documents in other sections

The major interest of using concordance table is to determine if it is possible to detect bad categorization of fake documents in other sections. For this analysis, the first five words of each class (by number of occurrences) found from the DHA analysis of the "*listing_fakedoc*" corpus have been searched in the "*listing_all_without_drugs*" corpus. It seems important to notice that the terms studied in this analysis have been selected according to their number of occurrences in the corpus. They are thus not necessarily specific to the field of fake documents. Then, every category different from the two fake document subsections (physical / digital) have been identified. Table 4 shows all the detected categories.

			Class 1					**Class 2**					**Class 3**		
	psd	template	passport	hq	statement	passport	scan	utility	custom	bill	driver	license	fake	licence	dl
Online business, other fraud related	x	x	x	x	x	x	x	x	x	x	x	x		x	x
Online business, SSN/DOB/OtherPII	x	x	x	x		x	x	x	x	x	x	x		x	x
Online business, drops others		x							x						x
Online business, dumps		x	x			x	x		x		x	x		x	
Online business, card and CVV			x	x		x	x	x	x	x		x		x	x
Online business, various logins			x	x		x				x	x	x		x	
Online business, corporate intel				x							x				
Online business, drops bank				x			x		x	x		x			x
Online business, bank login				x			x			x	x			x	x
Services, carding	x	x					x		x		x	x			x
Services, Hosting										x					x
Services, Operational management							x								
Services, Other services			x	x		x	x	x	x	x	x	x		x	
Services, social engineering		x	x			x									
Services, VPN				x										x	
Services, SOCKS				x					x						
Services, security												x			
Forgeries/counterfeit, currency	x	x							x	x			x		
Forgeries/counterfeit, other forgeries			x			x		x	x				x		
Forgeries/counterfeit, electronics									x						
Forgeries/counterfeit, watches													x		
Software, other software		x					x	x	x		x	x	x	x	
Software, commercial software				x				x				x		x	
Software, botnet and malware				x											
Software, exploit kit							x								
Software, security software							x		x			x		x	
Defense counter intel, frequency scanner/bug detector							x								
Defense counter intel, operational security									x						
Total per word	4	8	8	13	1	8	13	7	15	9	9	12	4	11	8

Table 4 : presence/absence of the term in a category other than "Fake Document (Physical)" or "Fake Document (Digital)"

28 categories containing the first five words of our "*fake document*" classes have been identified. The term with the highest diversity is "*custom*", present in 15 categories. As previously described, this can be explained by the particular usage of this term within the cryptomarket ecosystem. "*passport*" is nevertheless present in 8 other categories. The term "*statement*" is the one with the least other categories.

The specificity of the terms can also be analyzed with the proportion of their occurrences in the "*fakedoc*" corpus compared to their total number of occurrences (Table 5).

	Word	Total number of occurrences	Number of occurrences in the "listing_fakedoc" corpus	Proportion
Class 1	psd	648	562	87%
	Template	616	516	84%
	Passport	311	205	66%
	Hq	1265	161	13%
	Statement	148	144	97%
Class 2	passport	311	205	66%
	Scan	399	199	50%
	Utility	147	119	81%
	Custom	201	105	52%
	bill	183	93	51%
Class 3	driver	321	236	74%
	License	343	217	63%
	Fake	219	130	59%
	Licence	162	91	56%
	dl	280	81	29%

Table 5 : proportion of occurrences in the "fakedoc" corpus compared to the total corpus (except drugs) (N = 15844 titles for the "listing_all_without_drugs" corpus and N = 1187 titles for the "fakedoc" corpus)

The terms with the highest rate of occurrences in the "*listing_fakedoc*" corpus are "*statement*" (97%), and "*psd*" (87%), which is consistent with the previous results, in particular concerning the most common format of selling. The terms with the lowest rate of specificity are "*hq*" (13%) and "*dl*" (29%). "*hq*" is an abbreviation of "*high quality*", which is an expression that can be used in many other contexts than fake documents. "*dl*" can be translated by "*driver license*", but also "*download*".

Grouping sellers

Based on the ad titles

Seven successive CFA have been performed during which 19 outliers were removed. Outliers ads are mostly written in other languages, like French or German. Products like Netflix accounts, Walmart receipt, Apple store subscriptions, and biometric passports and visas were also detected as very specific selling activities related to outliers.

Figure 4 shows the result of the last CFA, where no obvious outliers remain visible. Each square on the graph represents a group according to the dimensions selected by the algorithm.

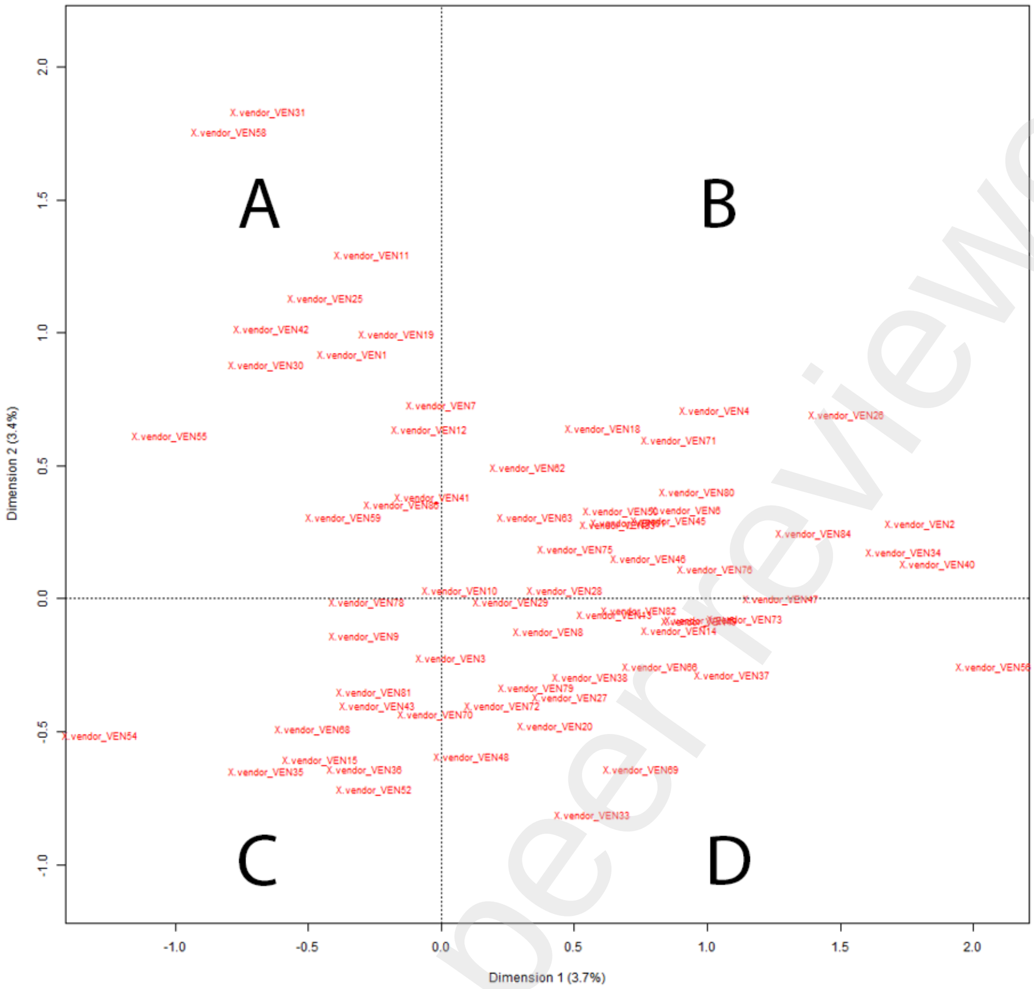

Figure 4 : Final CFA obtained as a result of successive analyses of the specificities of sellers of false documents based on ad titles (N = 67 sellers present on the CFA at the last analysis)

By consulting the catalogue of the vendors from each group, we were able to extract a main topic of products for each group:

- Group A: physical fake documents. This is the group with the fewest vendors, which is consistent with the other analysis showing that digital fake document are more common than the physical ones.
- Group B: "*scan*", "*custom*", "*selfie*" are the most common products in this group. This result is also consistent with class 2 of DHA analysis discussed earlier.
- Group C : "*psd template*". Given the omnipresence of this digital form of products on the market, it is not surprising to find a group comprising mainly the sellers of this type of document.
- Group D : "*pack*" and other products. This group is more difficult to define in terms of products sold. The term "*pack*" is present in an important part of the ads but this group contains some diversity.

We can see that some vendors are really close to the axis, which can be explained by the proximity of their offer with vendors from other groups. For example, vendor 10 (group A) is really close to group C and by looking at his catalogue, we can see that his offer is mostly composed of "*psd template*" ads, among other products.

The contribution of each group can be visualized in Table 6. The distribution of the vendors through the groups is balanced.

Group	Headcount	Percentage of total
1	15	17,4%
2	20	23,3%
3	13	15,1%
4	19	22,1%
Outliers	19	22,1%
Total	**86**	**100,0%**

Table 6 : Contribution of each group and outliers to the total number of vendors (N = 86)

Based on their profile

The same analysis was performed using the vendor's profile textual description. This analysis revealed two major issues: the first is that 17 vendors didn't have any written profiles, so they can't be taken into account. The second one is that, after 6 successive analyses, 29 vendors were excluded as they appear as outliers, resulting in a total of 46 vendors (53% of the total) that weren't analysed. The reading of the profiles did not reveal any insights upon the groups formed with the remaining sellers. The profiles of the outliers did not reveal anything conclusive either to explain their exclusion from the others.

Conclusive discussion

Is it possible to set up a classification of fake documents using textual data? DHA analysis leads to a classification of fake documents and highlighted other types of documents than fake identity documents described in literature, which distinguish between three main categories : passport, ID cards and driver licenses. The highlighting of other products like utility bills, bank statements, but also a novel category related to "selfies", shows a bigger diversity in the market than expected. The similarity analysis is informative on the most common format of selling for the products: the "*psd template*" format. Based on the observation that driver licenses are mostly linked to American state names, it can be hypothesized that the demand for this kind of document is higher. Indeed, driver licenses are much more used to check identity in the USA than the id card or passport. The discovery of the selfie brings to light new issues concerning identity control on the Internet. Indeed, today, many sites require a photo of the user holding an ID in order to access their services. The availability of these selfies therefore offers a new way of evading these controls.

However, during DHA, IRaMuTeQ showed its first limits. The term "*id*" was absent from the analysis. The assumption made about this fact is that that term was systematically contained in the texts that weren't taken into account. Another hypothesis was suggested by Loubère (2016). She suggests that the software does not take every form as "*full forms*". The major problem is thus that the operator has no control of the forms or texts analysed, which is a real issue from a forensic point of view. In order to test the hypothesis, the term "*id*" was replaced with "*identity*" in the corpus. After another DHA, the term "*identity*" appeared in the class associated with driver licenses, with a higher number of occurrences. This finding raises the hypothesis of small words being excluded just like stop words. They may not be taken into account because of their size. However, terms such as "*hq*" and "*dl*" were taken into account in the analysis. This observation led to the fundamental methodological proposition recall by Pincemin (2020): "*back to the text*". As it helps to identify these gaps induced by analyses over which the operator's control is limited, it compensates for the "black box" effect inherent in some algorithms. In our study, this problem appears to be specific to DHA analysis. The modification of the corpus made in the test may not be a viable solution, because, depending on the context, this action could be perceived as a modification of the textual trace.

Can fake documents be distinguished from other products? In the majority of cases of mixed corpora studied, it was possible to distinguish a specific theme for the classes found with DHA on mixed corpus, and to get a separate class containing forms linked to fake documents from other classes. The main issue for the comparison is the variation of the sizes of the corpus. If one of the two categories used to create the mixed corpus has many more texts than the other, the second one is hardly detected. Following this, the concordance table led to the detection of forms that can be used in different contexts and also wrong categorizations of fake documents. Freeing oneself from the sections used by the sellers to select the product to analyse is a key issue for the analysis of online marketplaces. This was not the main aim of this study, but results show the interest of the tested approaches in order for instance to evaluate the results of fully automated IA approaches like deep-learning ones.

Can sellers of fake documents be grouped based on the textual data from the advertisements? Four main groups of vendors were detected. Globally, an important proportion of digital fake documents are observed compared to physical ones. The effort required for making physical documents and the ease of transferring digital documents may explain this result. Indeed, the manufacture of fake documents requires know-how as well as equipment and materials in order to produce a document that is of satisfactory quality. There is also consistency between results found with the products and with the vendors, which might be the sign of a certain degree of specialization. The main issue of this analysis is the exclusion of the outliers during the successive CFA. Indeed, this part of the process is based on a visual analysis of the graphs. It's the operator who decides which vendor is an outlier based on its graphical distance from the main group. In that case, there was always a compact group in the center, so it was easy to determine the outliers.

Is it possible to find groups of sellers from the analysis of their profiles? This analysis suffers from the subjectivity required for the exclusion of the outliers. On the contrary to the corpus of ads titles, the distributions obtained after the successive specificity analysis for the vendor's profiles were more shattered. This led to the exclusion of 29 vendors (34%), knowing that 17 vendors have no profile, 53% of vendors were not taken into account in the analysis. Vendors profiles should thus be considered with cautiousness and further analysis are required in order to evaluate their informative content. It was indeed impossible to identify a main topic for the groups formed. This can be explained by the fact that every vendor chooses to write whatever he wants in his profile, and it doesn't necessarily have a link with what they sell. It could be interesting to try the method with a corpus of vendors of other types of products, to see if it is an inherent problem to vendors of fake documents.

Globally, several steps of the methodology used required manual work, which leads to a certain risk of error. In the concordance table analysis, for example, it would have been difficult to estimate the number of products listed outside the fake document categories for each term studied, due to the high proportion of occurrence of each word. IRaMuTeQ did not allow for an automatic numerical estimate. The size of the corpus is also a limitation for some analyses, such as the product classification. This method requires the assess the construction of the corpus itself, in order to ensure that all forms are taken into account. Finally, some limitations come from the software used. Indeed, IRaMuTeQ is an easy-to-use software that allows to obtain good results for an exploratory analysis and provides very relevant global information. However, it does not allow us to go deeper into the details of the data, at least not in an automatic way. Furthermore, the operator has little control over the forms used. It could therefore be interesting to place it in sequence with other techniques, where it would allow an initial sorting to be carried out before continuing with more elaborated methods and tools.

The analysis of words as a trace in the judicial context is an issue that still raises many questions. Indeed, words are more often than not considered as subjective and sensible to a lot of variation and interpretation, an aspect that the statistical methods tend to mitigate. But the potential of these methods during investigation and for intelligence purpose appears to be very high. This research work is intended to be a starting point and, above all, an open door to explore how the statistical analysis of textual data might help to answer crime analysis questions.

References

Akhgar, Babak., Gercke, Marco., Vrochidis, Stefanos., & Gibson, Helen. (2021). *Dark Web Investigation*.

Baechler, S. (2020). Document Fraud: Will Your Identity Be Secure in the Twenty-first Century? *European Journal on Criminal Policy and Research*, *26*(3), 379–398. https://doi.org/10.1007/s10610-020-09441-8

Baravalle, A., Lopez, M. S., & Lee, S. W. (2016). Mining the Dark Web: Drugs and Fake Ids. *IEEE International Conference on Data Mining Workshops, ICDMW*, *0*, 350–356. https://doi.org/10.1109/ICDMW.2016.0056

Bellido, L., Baechler, S., & Rossy, Q. (2017). La vente de faux documents d'identité sur Internet. *Revue Internationale de Criminologie et de Police Technique et Scientifique*, *70*(2), 233–249.

Camargo, B. v., Justo, A. M., & Forte, T. (2016). *IRAMUTEQ tutorial. R interface for multidimensional analysis of texts and questionnaires*. http://www.iramuteq.org/documentation/…chiers/IRaMuTeQ Tutorial translated to English_17.03.2016.pdf

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Guérin-Pace, F. (1997). La statistique textuelle. Un outil exploratoire en sciences sociales. *Population*, *52*(4), 865–887. https://doi.org/10.2307/1534617

He, S., He, Y., & Li, M. (2019). Classification of illegal activities on the dark web. *ACM International Conference Proceeding Series*, *Part F1483*, 73–78. https://doi.org/10.1145/3322645.3322691

Holt, T. J., & Lee, J. R. (2022). A Crime Script Analysis of Counterfeit Identity Document Procurement Online. *Deviant Behavior*, *43*(3), 285–302. https://doi.org/10.1080/01639625.2020.1825915

Kruithof, K., Aldridge, J., Décary-Hétu, D., Sim, M., Dujso, E., & Hoorens, S. (2016). *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. https://www.research.manchester.ac.uk/portal/en/publications/internetfacilitated-drugs-trade(9a2980f2-f8f3-46ba-9aa7-cd42d5573551).html

Laferrière, D., & Décary-Hétu, D. (2022). Examining the Uncharted Dark Web: Trust Signalling on Single Vendor Shops. *Https://Doi.Org/10.1080/01639625.2021.2011479*. https://doi.org/10.1080/01639625.2021.2011479

Lam, T., Demange, J., & Longhi, J. (2021). Attribution d'auteur par utilisation des méthodes d'apprentissage profond. *EGC 2021 Atelier "DL for NLP : Deep Learning Pour Le Traitement Automatique Des Langues."* https://hal.archives-ouvertes.fr/hal-03121305

Lefer, M.-A., Bestgen, Y., & Grabar, N. (2016). Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances. *Actes de La Conférence Conjointe JEP-TALN-RECITAL 2016. Volume 2 : TALN (Posters)*, 555–563. https://aclanthology.org/2016.jeptalnrecital-poster.31

Loesing, K., Murdoch, S. J., & Dingledine, R. (2010). A case study on measuring statistical data in the Tor anonymity network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6054 LNCS*, 203–215. https://doi.org/10.1007/978-3-642-14992-4_19

Loubère, L. (2016). L'analyse de similitude pour modéliser les CHD. *JADT 2016 : 13ème Journées Internationales d'Analyse Statistique Des Données Textuelles*, 9. http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83440/83440.pdf

Margot, P. (2018). Traceology, the Bedrock of Forensic Science and its Associated Semantics. In Q. Rossy, D. Décary-Hétu, O. Delémont, & M. Mulone (Eds.), *The Routledge International Handbook of Forensic Intelligence and Criminology* (pp. 29–39). Routledge.

Marpsat, M. (2010). La méthode Alceste. *Sociologie*, *N°1, vol. 1*. http://journals.openedition.org/sociologie/312

Martin, J. (2014). Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs. In *Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. Springer. https://doi.org/10.1057/9781137399052

Mireault, C. (2016). *La vente en ligne de faux documents d'identité. Une recherche exploratoire.* [Travail aux cycles supérieurs / Graduate student work, Université de Montréal]. https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/16373/Mireault_Camille_2016_travaildirige.pdf?sequence=4&isAllowed=y

Moreno, M., Marchand, P., & Ratinaud, P. (2015). Analyse d'un corpus multilingue : visualisations textométriques des convergences et divergences dans l'écriture journalistique. *SHS Web of Conferences*, *20*, 01015. https://doi.org/10.1051/shsconf/20152001015

Nabki, M. W. al, Fidalgo, E., Alegre, E., & de Paz, I. (2017). Classifying illegal activities on tor network based on web textual contents. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *1*, 35–43. https://doi.org/10.18653/v1/e17-1004

Ombelli, D., & Knopjes, F. (2008). *Documents : the developer's toolkit* [Book]. Via Occidentalis International Organisation for Migration.

Overdorf, R., Technol., R. G.-Proc. Priv. E., & 2016, undefined. (n.d.). Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. *Cyberleninka.Org*. https://doi.org/10.1515/popets-2016-0021

Peng, J., Choo, K. K. R., & Ashman, H. (2016). Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, *70*, 171–182. https://doi.org/10.1016/j.jnca.2016.04.001

Pincemin, B. (2020). La textométrie en question. *Le Français Moderne - Revue de Linguistique Française, CILF (Conseil International de La Langue Française), 2020, Linguistique et Traitements Quantitatifs*, *88*(1), 26–43.

Reinert, M. (1993). Les "mondes lexicaux" et leur 'logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société*, *66*(1), 5–39. https://doi.org/10.3406/lsoc.1993.2632

Renaut, L., Ascone, L., & Longhi, J. (2017). De la trace langagière à l'indice linguistique : enjeux et précautions d'une linguistique forensique. *Ela. Études de Linguistique Appliquée*, *188*, 423–442.

RFC 6761 - Special-Use Domain Names. (n.d.). Retrieved March 24, 2022, from https://datatracker.ietf.org/doc/html/rfc6761

RFC 7686 - The ".onion" Special-Use Domain Name. (n.d.). Retrieved March 24, 2022, from https://datatracker.ietf.org/doc/html/rfc7686

Rossy, Q., & Décary-Hétu, D. (2017). Internet traces and the analysis of online illicit markets. In Q. Rossy, D. Décary-Hétu, O. Delémont, & M. Mulone (Eds.), *The Routledge International Handbook of Forensic Intelligence and Criminology* (1st ed., pp. 249–263). Routledge. https://doi.org/10.4324/9781315541945

Roux, C., Bucht, R., Crispino, F., de Forest, P., Lennard, C., Margot, P., Miranda, M. D., NicDaeid, N., Ribaux, O., Ross, A., & Willis, S. (2022). The Sydney declaration – Revisiting the essence of forensic science through its fundamental principles. *Forensic Science International*, *332*, 111182. https://doi.org/10.1016/j.forsciint.2022.111182

Soska, K., & Christin, N. (2015). Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. *Proceedings of the 22nd USENIX Security Symposium (USENIX Security 2015)*, 33–48.

Soudijn, M., & Zegers, B. (2012). Cybercrime and virtual offender convergence settings. *Trends in Organized Crime*, *15*(2–3), 33–48.

Svartvik, J. (1968). THE EVANS STATEMENTS: A Case for Forensic Linguistics. In *Gothenburg Studies in English*. University of Gothenburg.