LUISS

Department of
Business and Management

Course of Cybercrime and fraud detection

Drug dealing activities on TOR: an extensive analysis of the Alphabay market

Prof. Gianluigi Me

Supervisor

Vincenzo Rosario Musto

Candidate No. 243921

Academic year 2021/2022

Table of contents

Table of Figures

Table of Listings

Introduction

The evolution of the information technologies brought new kind of threats. The cybercrime world does not only regard the frauds and the leak of personal information, but it also regards crimes that were already perpetrated physically. Drug dealing activities are one of the main activities carried out on the Dark Web, where the identities of the sellers and the customers are protected by cryptography. The aim of this paper is to give an insight into the world of online drug dealing activities. This work is possible thanks to the data collected by search engines, researchers and police forces. In particular, the realization of this paper would not have been possible without the data and the guidance provided by Professor Gianluigi Me.

Along this paper, an introduction to the Dark Web functioning will be discussed, together with the description of the Darknet Markets, their evolution, and the role of the cryptocurrencies as a medium to exchange goods and services. Then, an extensive analysis of the Alphabay market is provided, comparing data collected in 2015 and 2016, remarking the differences and the analogies, with an extensive exploratory data analysis. Data Mining and Machine Learning activities will be performed using RapidMiner and Python, powerful tools which are useful to detect patterns, extract data and show graphically the results, giving the opportunity to the readers to easily understand the work and its results. Another goal of this paper is to develop an Artificial Intelligence which is capable to understand and classify the kind of products sold on the markets, given data which can be easily extracted with crawling software from any Darknet Market, training the model over the data extracted from Alphabay.

1. The Deep Web ecosystem

The Surface Web, which serves millions of users every day, comprises almost all the contents that can be easily reached with a common browser. The Surface Web, also called Clear Web, is just the tip of a way bigger iceberg. The hidden side of this iceberg, which is not indexed from the common search engines, is called Deep Web. At the beginning of the year 2000, it has been estimated that the contents on the Deep Web were from 400 to 550 times the contents on the Surface Web. Even if the term Deep Web may be seen from the readers as a negative term, its contents include, for instance: dynamically generated pages, e-mails, contents protected with passwords, cloud stored files or websites which need a registration. In fact, search engines can index static pages through the use of automated systems which record hypertext links in each visited page. In this way, they propagate in these other links, discovering all the pages linked to each other. This kind of undiscriminated crawl can be criticized for the huge number of pages returned for a simple request on a search engine. On the other hand, there are contents that are intentionally hidden from the Surface Web. It is the case of the Dark Web, a small part of the Deep Web which can be surfed with specific software. Tor (The Onion Router) and I2P (Invisible Internet Protocol) are examples of software used to reach Dark Web contents. The development of a Dark Web platform as Tor goes back to the U.S. government, which funded its creation to guarantee the anonymity of the transmissions between its military parts. Thanks to its higher level of privacy, Tor can be used by journalists, political opponents or just from normal citizens, especially the ones living in authoritarian countries where, talking about politics, reporting human rights violations or reporting corruption can be life-threatening. It is the case of Middle East countries, where during the Arab Spring, the number of users on the Tor network increased from 7,000 to 40,000 in Iran and from 600 to 15,000 in Syria.

TOR (The Onion Router) is called in this way because of its layered structure, remembering the shape of an onion. In fact, the anonymity of the IP address of the user is reached through the use of three encrypted layers, connected to random nodes around the world. While connecting to the TOR network with a particular browser, based on Mozilla Firefox and maintained by the non-profit organization “Tor Project”, the user downloads a list of available entry nodes from the official server. The user is then randomly connected to one of the entry nodes, where the information is encrypted, then it passes through another middle node and, in the end, the packet passes through

an exit node, where the encryption is removed and the final destination is reached. Operating in this way, the sender and the receiver of the communication cannot know the IP of the other and the communication is almost impossible to be tracked back. The main issue of that system is that sending an information in different and distant parts of the world, slows the connection. This is one of the reasons why Tor websites mostly use a simple and poor graphic. This encryption is not only useful to reach Dark web sites, which have a ".onion" extension, but also to connect to Surface Web sites, bypassing the restrictions imposed by the governments, it is the case of Twitter, which was banned in Turkey in 2014, where the number of users connecting from that country to the Tor network doubled after the ban.

Even if the Tor network can be used for noble intents, like defending the democracy or provide freedom of speech and press, the other side of the coin is that the encryption and the difficulty to be tracked, can favour the spread of illegal online activities. Forums hosting child pornography, markets selling fake IDs, weapons, narcotics, counterfeit products, credit cards information and other illegal products spread on the Tor network. One of the ways to access Tor websites is to find them on special search engines or wikis like "TORCH", "the Hidden Wiki" and "notEvil". As the wikis could be modified by many users (even though they are not freely modifiable from everyone), it may be possible to find links to clone Tor websites instead of the original ones, making it more possible for the users to get scammed.

1.1 Silk Road - The first Darknet Market

In 2011, one of the first online black markets started its business on the Dark web. It was the so known "Silk Road". This market, as many others, gave the opportunity to sellers from around the world to publish their listings on the platform, just like legit marketplaces as Ebay or Amazon marketplaces do. The main difference is that any kind of illegal product was sold. When the users were connecting to the Silk Road website, they found out a form to login or register with a username, password, withdrawal PIN and a CAPTCHA to resolve. This login form and the CAPTCHA, together with anti-crawling systems could have made the crawling activities for researchers or law enforcement agencies more difficult, not only for Silk Road, but for many other markets. Once the registration or login process has finished, the users accessed the Silk Road main page. The graphic was simple as other Tor websites, products could be searched with a search bar or by checking the categories on a menu

on the left. This market, as many others, had an "ethic". In fact, listing assassinations services, weapons or paedophilia was not allowed. On the other hand, weapons were firstly allowed, then they got moved on a market related to Silk Road, The Armory. Even though the price of the products on the Dark Web markets can be higher than the traditional face-to-face sellers, the reasons behind the choice of the consumers to prefer those markets instead of the street sellers can be the anonymity of the trade, the home shipping, the avoidance to get in touch with violent street sellers and the avoidance to get caught by the police. The very positive reviews on the Dark Web markets could convince the buyers of the high quality of the products. On the other hand, analysis have been performed on drugs bought on these markets to check the actual purity of these substances. In a 2014-2015 study, "Of the total of 219 samples that were analysed by gas chromatography and mass spectrometry, the majority (91.3%) contained the drugs as advertised, but with significantly varying levels of purity." (Pergolizzi JV, LeQuang JA, Taylor R, Raffa RB, 2017).

The medical risk is not the only risk that a buyer is going to take while buying from a Dark Web market. The reviews of the products and the sellers' reputation are the only way for the buyers to trust a certain seller. The sellers could have created fake accounts to leave positive reviews and then scam the real customers, by shipping a brick or sugar but, the forums associated with most of the markets are used as a place of discussion where this kind of sellers are easily exposed. The most active and influent users on these forums have the power to make a simple seller evolve to a big and popular seller. For this reason, this kind of "super reviewers" can receive free samples from the sellers to assess the quality of the product and then leave feedback. Even sellers with a high reputation can fail, for some reasons, to deliver the products. To avoid or at least to lower the scams, escrow services are implemented in these markets.

1.2 The escrow and the role of the criptocurrencies

An escrow consists in paying the price to a third-party wallet, which is managed by the marketplace, and only when the buyer receives the product, the payment will be sent to the seller. This practice is surely discouraging sellers to scam but, on the other hand, it creates a new kind of cybercrime, the so-called "exit scam". This kind of cybercrime consists in the founders or the administrators of a platform, in this case a Dark Web market, to disappear with all the funds collected, this is the case of the "Evolution" market which disappeared with a $12 million equivalent in Bitcoin in

March 2015. The centralized escrow service is therefore another risk to take in consideration. In the Deep Web markets, scams can be seen as a business model rather than a market deviance; even an exit scam would not stop the users to trust a particular market as it is seen as an intrinsic risk. To avoid the exit scams, new kind of escrow systems have been developed. “Hansa” market implemented “multi-sig”, a way to control the escrow process directly on the blockchain, in particular on the Bitcoin blockchain, instead of relying on a centralized escrow controlled by the marketplace.
The main issue behind the use of cryptocurrencies is that the high volatility of their value cannot assure a fixed profit for the sellers. For this reason, the price of the products must be updated very often, or the economic value of the product can be pegged to the dollar. As the use of an escrow delays the reception of the payment until the products are received from the customers, the value of the cryptocurrencies can fluctuate enough to make the deal unprofitable. Some markets can allow the “Finalize Early” function for the sellers with high volumes and a strong reputation, this means that the buyers can unlock the payment for the sellers even before the products have arrived. In this way, the buyers are taking a risk and the only certainty is the seller’s word, which may incentivize this function with free samples or future discounts. The anonymity needed in this kind of business makes the use of cryptocurrencies an obvious choice. In 2008, Bitcoin, the first cryptocurrency, appeared online and its mass distribution began. The reason behind the creation of Bitcoin can be the necessity of an electronic payment system more resistant to frauds, compared to credit cards. Another feature of the cryptocurrencies is that there is no central authority which controls the transactions, providing more freedom to the cryptocurrencies’ owners. This level of freedom is possible thanks to the technology behind the blockchain, which is a distributed ledger where everybody can see the transactions, ensure the network with a particular procedure called “mining” or contribute to the development of it as, most of the times, the blockchains are based on open-source projects. On the other hand, cryptocurrencies can be the main way to exchange money and goods for cybercriminals. As stated before, Bitcoin is the main payment method on Dark Web markets, but it can be used also in other kinds of cybercrimes. For example, it can be used for money laundering or for the payment needed to unlock the digital contents crypted by a ransomware, a particular virus which asks for a ransom in order to provide a key to decrypt the files on the infected machine. During the years, many cryptocurrencies have been developed, each with their own characteristics. Monero

and Dash are two examples of cryptocurrencies oriented to privacy. Together with Zcash and Ethereum, they are gaining more popularity for cybercrime activities. The Bitcoin's ledger is, in fact, easy to analyse for the law enforcement agencies, transactions between wallets are clearly readable and they are then analysed to find out the owners of the wallets. On the other hand, the Monero blockchain, for example, encrypts the addresses and the amount of the transaction, making the money flow impossible to follow. The need to know more about the Dark Web and to stop the cybercrimes going through it has taken researchers and law enforcement agencies to explore it with data extraction and data analysis techniques. The first analysis have been performed on Silk Road but they have been eventually performed on many other markets. As mentioned before, crawling software can be used to extract pages with all the information, as descriptions, reviews and pictures. Software as HTTrack or Tor_Crawler, with the latter dedicated to the Tor crawling, are the starting point of the Dark Web analysis.

1.3 The rise of Alphabay

With the seizure of Silk Road in 2013 by the U.S. Federal Bureau of Investigation, many other markets tried to get the monopoly of the Dark Net ecosystem. One of these, Alphabay, started its journey in the end of 2014 and from then, it got more popular every day. With the closure of other concurrent markets as "Evolution" and "Agora", Alphabay got an extensive increase of insertions, becoming one of the leading marketplaces on Tor. The drug dealing was the main activity on this market, until its seizure in 2017. Given the importance of Alphabay, analysing its characteristics as sellers, products and prices, can provide a view of the international drug market over the Tor network, during its activity period. This data can be useful for police forces to better understand the world of illegal drugs traffic and consumption. Moreover, future research can compare the data of these past years with more actual data from other markets.

2. Darknet Markets analysis: the Alphabay case

The aim of this paper is, in fact, the exploration of the Alphabay market through the use of a dataset. This dataset is provided from the Professor Gianluigi Me, and contains data crawled between the 23rd and 24th March 2016. Moreover, some comparisons with a second dataset will be provided along this paper, the second dataset can be obtained from the "Gwen.net" website, which has a collection of Dark Web markets datasets. In particular, the chosen dataset has been provided from the famous Deep Web search engine "Grams", it is a dataset containing data crawled from Alphabay, between the 22nd April and 13rd July 2015. Even though the two datasets have been collected in two different timespans, the result of the pre-processing will show, in the next paragraphs, that the quantity of insertions on the market is almost the same, this can probably mean that a 24h crawling activity, with the right optimizations and adequate hardware, can give a good picture of the market's situation. Therefore, this analysis will give an insight into the Alphabay marketplace in two different periods and with different kind of data. Data mining activities and artificial intelligence techniques will be applied to give useful insights about this market ecosystem. The main tools utilized for this research are RapidMiner 9.10 and Python 3.10, the former tool provides an easy to use, complete and fast developing interface, while the latter provides more flexibility for some custom analysis.

2.1 Description of the datasets

Beginning with the dataset provided by the professor Gianluigi Me, this dataset contains 68,652 rows, each stating an insertion, crawled on the marketplace. The rows will also be called "examples" in this paper, based on the fact that RapidMiner calls the rows in this way. Going on with this reasoning, the columns of the dataset can also be called "attributes" and the dataset itself can be called "ExampleSet". The dataset has been provided in ".sql" format, which means that some more steps are needed to use it in RapidMiner. First of all, the file has been imported in a sql DBMS. In this case, MariaDB from XAMPP has been utilized. Then, a connection between RapidMiner and the DBMS has been configurated. At this point, the import of the table with the Alphabay listings was possible.

The dataset obtained has 25 columns, or "attributes". Between them, there are many attributes that do not contain any information, they are in fact filled with missing

values, this may be due to the fact that the crawling system could not detect this kind of information, or they were not present in this particular marketplace. These columns are “Quantity_g”, “Price_bc”, “Index”, “Date_visit”, “Raw_material”, “Product_review”, “Quantity_t”. Moreover, other columns are not considered useful for the analysis of this dataset even if they contain some data, these columns are:

- **Marketplace**, as the only value is Alphabay, this column may be useful for other analysis, together with other marketplaces data instead.
- **Payment**, as the only value is USD.
- **Url**, this column contains the offline location of the crawled pages.
- **Bids**, containing the bids for the auctioned products, it does not contain any value.
- **DateOCrawling**, states the date of the crawling, it spaces between 23-03-2016 at 15:25 to 24-03-2016 at 14:39.
- **Guid**, containing the guid information.
- **File**, indicating the compressed offline file of reference for the insertion.

Moreover, the other 11 columns are going to be useful for the aim of this paper, a short description of these columns is then provided:

- **Name**, this attribute contains the name of the insertions, which are the products to be sold on the market.
- **Vendor_name**, as can be imagined, it states the name of the vendor’s account.
- **Escrow**, a Boolean value indicating if the seller accepts payment through the escrow system (1) or if the seller accepts direct, unprotected payment (0).
- **Product_Type**, the most interesting column of the dataset. It contains the type of product which is sold. Mostly composed by types of drugs, this column will be the target of some prediction models over this paper.
- **Description**, contains the description of the product which is sold.
- **Refund_Policy**, when present, it states the way the seller is going to refund the buyer in case of problems during the trade.
- **Shipping_options**, containing the different kind of shipping options and their relative prices.
- **Price_t**, the price of the product, presumably in dollars.
- **Ships_from**, the country from where the product is shipped.
- **Ships_to**, the countries where buyers can receive the product.
- **Feedback**, the column containing the eventual feedback left from the buyers.

The second dataset, provided by gwern.net, is composed of 13921 rows, and 10 columns. The file is provided in ".csv" so it can easily be uploaded on RapidMiner. Looking at the columns, there are some that are not going to be useful, in particular:

- **Hash**, the result of a hash function for each row.
- **Market_name**, same as the other dataset, is only filled with "Alphabay" value.
- **Item_link**, the onion link where this page has been crawled from.
- **image_link**, the onion link where the image of the product has been crawled from.
- **add_time**, the time of crawling, expressed in Unix epochs.

Moreover, a description of the imported columns is provided:

- **vendor_name**, the nickname of the seller on the market.
- **price**, it contains the price of the product and, as it is represented as a small number with an average of 1.155, it can be assumed that it is the price in Bitcoin.
- **Name**, as before, is the column containing the name of the insertion.
- **Description**, containing the description of the product
- **Ship_from**, the shipping country for each insertion.

2.2 Anonymization of the datasets

The next step, after the import of the two datasets in RapidMiner, consists in the anonymization of them. In fact, the "Vendor_name" column contains the username of the sellers on the market for both the datasets, while the column "Description" may contain some email addresses. In order to provide an adequate level of anonymity, a Python script has been developed to mask these values. In the first place, the Faker package has been installed, as it provides an easy way to anonymize data. From RapidMiner, the datasets are passing through an "Execute Python" operator, which runs the Python script that can be seen in the Listing 1 below. RapidMiner provides the dataset to the Python script as a DataFrame object, from the Pandas library.

```python
from faker import Faker

def rm_main(data, data2):
    faker = Faker()
    Faker.seed(4321)
    dict_names1 = {name: faker.name() for name in data['Vendor_name'].unique()}
    dict_names2 = {name: faker.name() for name in data2['vendor_name'].unique()}
    dict_names = dict_names1 | dict_names2
    data['Vendor_name'] = data['Vendor_name'].map(dict_names)
    data2['vendor_name'] = data2['vendor_name'].map(dict_names)

    return data, data2
```

Listing 1: Anonymization of the "Vendor_name" column

In order to get the same result every time that the script is executed, a seed is set. Then, a Python dictionary is populated with the name of the vendors from both the datasets, together with their masked name. in this way, if the same vendor is present in both the datasets, the same will happen after the anonymization. This will be useful to understand if the same vendor was selling on Alphabay on these two different periods. Then, the anonymized vendor name is mapped to both the datasets, and they are finally returned to RapidMiner. Moreover, another Python script is needed to anonymize the emails contained in the “Description” column of the main dataset. In fact, after some tests, the dataset from Gwern.net has been found out to not containing any email address in the Description column, meaning that probably an anonymization step for this dataset has been done before the publishing of it, so it is directly stored as a new dataset, which is going to be used later for the next analysis. For this anonymization step, we can see from the Listing 2 below, that the main Alphabay dataset is delivered to the Python script in the same way as before. There, the “re” package is imported, it is necessary to use the RegEx expressions in Python. As stated from the IEEE and The Open Group, “Regular Expressions (REs) provide a mechanism to select specific strings from a set of character strings.” (The IEEE and The Open Group, 2004).

```
from faker import Faker
import re



def rm_main(data):
    faker = Faker()
    Faker.seed(4321)
    email_names_dict = {}
    descriptions = []
    for description in data["Description"]:
        email_finder = re.findall(r'([\w]*[._.-]*[\w]+[@]{1}[\w]+[.][a-zA-Z]+)',
description)
        for email in email_finder:
            email_names_dict.update({email: email})



    dict_names = {name: (faker.unique.first_name() + "@anon.com") for name in
email_names_dict}
    i = 0
    for description in data["Description"]:
        string_encode = description.encode("ascii", "ignore")
        string_decode = string_encode.decode()
        new_description = string_decode
        splitted = new_description.split()
        for word in splitted:
            if word in dict_names:
                new_description=re.sub(r'([\w]*[._.-]?[\w]+[@]{1}[\w]+[.][a-zA-
Z]+)', "", new_description)
                new_description = " ".join(new_description.split())
                new_description = new_description + " " + dict_names[word]
        data.at[i, 'Description2'] = new_description
        i += 1


    return data
```

Listing 2: Anonymization of the "Description" column

In this way, it is possible to firstly detect all the email addresses in the “Description” column and then, to replace them with their anonymized version. In particular, a first loop is used to detect all the emails present in the descriptions and to add them to a Python dictionary. Then, the detected emails are looped inside another dictionary, together with the anonymized emails formed with, a first name generated with Faker, merged with “@anon.com” domain. Maintaining the email format will be useful to understand the use of the emails among the sellers. In the last loop, the emails are removed from the descriptions and then, the anonymized emails are inserted. In the end, the dataset is returned to RapidMiner, where the dataset is going to be stored, and then used to conduct the next analysis.

2.3 Data pre-processing

The analysis of the Alphabay market datasets begins with a data pre-processing step. In fact, the anonymized datasets, coming from the Python scripts, still need some adjustments before the various analysis. As the steps are almost the same for both the datasets, the process is going to be described in a general way, with some specifications about the different steps in the end. First of all, the datasets are retrieved from the repository, then the columns that are not useful for the analysis, described in the paragraphs before, are deselected. As all the columns are imported in RapidMiner as "polynomial" value type, a change of type is needed for the columns that do not contain data useful for a distinction between categories. In fact, polynomial values can be seen as categories in the analysis, it makes more sense for the distinct countries to be seen as categories, rather than the feedbacks or the descriptions. Between the columns that contain text rather than numbers, the only columns that are not going to be converted into "text" data type, are: Product_Type, Ships_from, Ships_to and Vendor_name. With the same reasoning, the "Price_t" column is parsed to become a "numeric" column. These changes are needed in order to provide the right type of data for the next operators. Moreover, the duplicates are removed, considering every single column to determine if two or more rows should be considered duplicates, and some filters are applied. The rows with a "Shipping_options" or a "Name" consisting of a void value, which is different from a missing value, are filtered, they consist in less than 40 rows, but this step will still be useful for the next operators and Python scripts to run smoothly. The last filter regards the price of the products, a range for the price is chosen in order to avoid bulk products or unavailable products, where a high price is set to avoid orders, and products with a way too small price to be thought as a possible price. This range has been set from 1 to 1000. These pre-processing steps give, as result, that the main dataset is composed of 15,014 rows, instead of 68,636. On the other hand, the second dataset needs one more step to get the value of the insertions in dollars. In fact, the "price" column has way too small values to think that

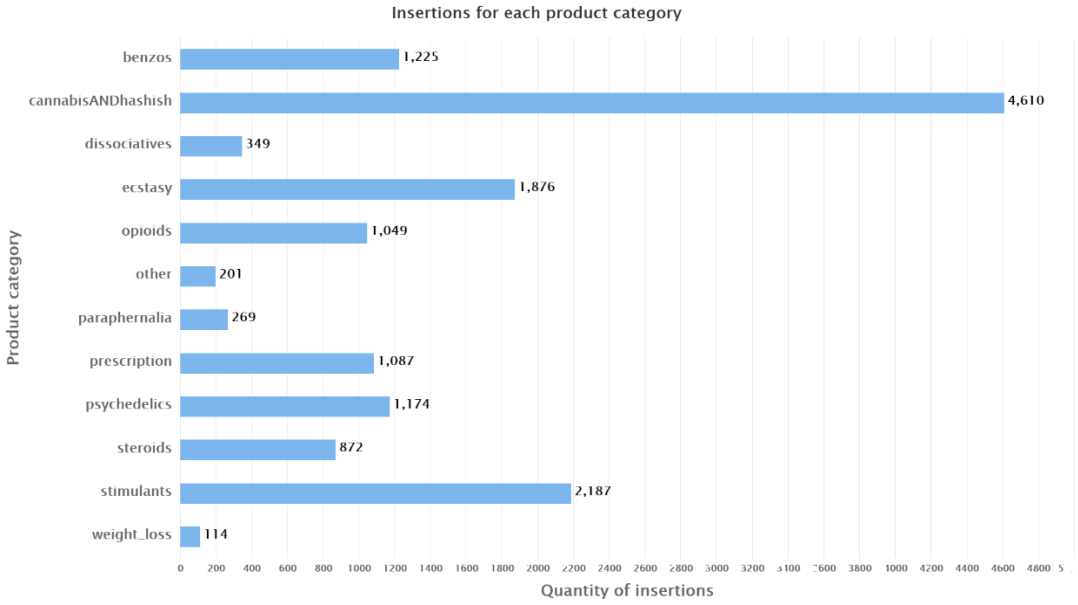

Figure 1: Listings for each product category

it represents the price in dollars. As the main cryptocurrency used on Alphabay in 2015 was the Bitcoin, a conversion in dollars is made with the average price of Bitcoin during the weeks of crawling, creating the column "price_t", this dataset is now composed of 10.930 rows instead of 13,921. Before going on with further steps, the datasets have been checked to find other pre-processing opportunities that were not easy to detect with a dataset which had many duplicates. Data visualization techniques have been used to detect interesting insights. At this point, only the analysis useful for the pre-processing will be shown, the rest of the graphs will be deeply discussed in the next paragraph. As it can be seen from the Figure 1 above, the "Product_type" column from the main dataset is mainly composed of cannabis and hashish products, with a fair share of many other kind of drugs, like stimulants, ecastasy, benzos and so on. On the other hand, tobacco products are the smallest category of products represented in this dataset. As this category does not provide enough useful information and, as tobacco products are generally not considered drugs, even if they can still be smuggled, these listings will be relabelled as "other", making this category slightly bigger and diversified. In this way, the categories are dropping from 13 to 12.

2.4 Exploratory Data Analysis

Now that the pre-processing step is complete, a visualization of the clean data contained in the main dataset can be provided, providing insights on the drug dealing activities on the Alphabay market, while the second dataset, as it is way less detailed,

will only be used for some comparisons. Giving a general look to the dataset, a bar plot can be a good choice to understand the provenience and the destination of the

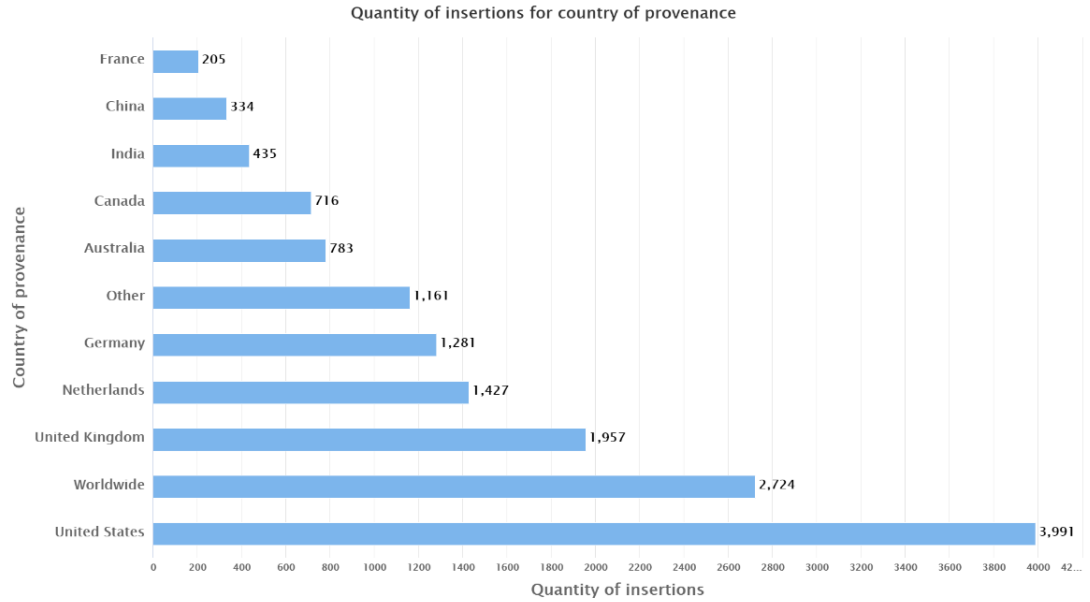

Figure 2: Quantity of insertions for location of provenance

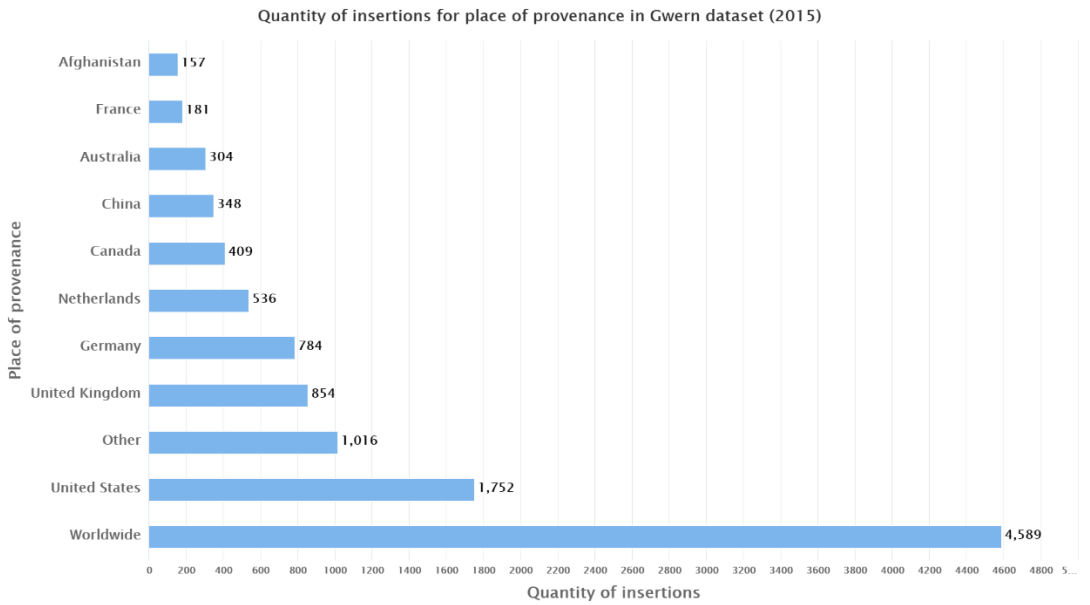

Figure 3: Quantity of insertions for place of provenance (GWERN 2015)

goods sold on the marketplace. The country of provenance of the goods are 51, in the Figure 2 above, the top 10 countries for quantity of insertions are shown, together with a "other" country, containing the aggregation of the insertions of the remaining 41 countries. The remaining countries, in fact, would lower the quality of the graph if they were going to be included. It is easy to notice the dominance of the United States, the United Kingdom, together with the "Worldwide" attribute, being the top 3 "locations" present on Alphabay. A comparison can be made with the Gwern's dataset, where the

locations where the goods are shipped from are 81 instead of 51 but, as before, only the top 10 labels will be shown, while the others will be grouped under the label "Other". From the Figure 3 above, it can be noticed the dominance of the Worldwide label in this dataset, while "Other" is at the third place, meaning that the countries of origin of the goods are more varied rather than the main dataset. It can be possible that the dataset contains many other listings more than the drugs related ones, these other listings can include digital goods, which can be shipped from everywhere and to everywhere, as they do not need a physical shipping. It can be challenging to detect only the insertions related to drugs.

A geographical heat map representing the world is then provided, in order to understand the presence of all the countries, and not only the top 10 countries, which are the origin of the products. From the Figure 4 below it can be seen that some geographical areas of the world are more represented in Alphabay while others, like Central America, Russia and, with some exceptions, Africa, do not have listings on Alphabay. It should be noticed that the second shipping location after the United States of America, is "Worldwide". This means that many drug dealers could ship from more than one country, it can be the case of cartels or criminal groups working together, or it can be just a way to preserve the privacy of the sellers, in order to make more difficult the tracking of their businesses. This means that, even if not present on this map, many sellers can still ship from the remaining countries. In any case, it is easier to understand that most of the goods are shipped from countries coloured in red, and then in yellow, rather than countries coloured in blue. The USA have the largest share of insertions on Alphabay, way more than Canada and Mexico, for the North America region. Moreover, UK, The Netherlands and Germany have the largest share in Europe. Therefore, these are the most interesting areas to analyse.

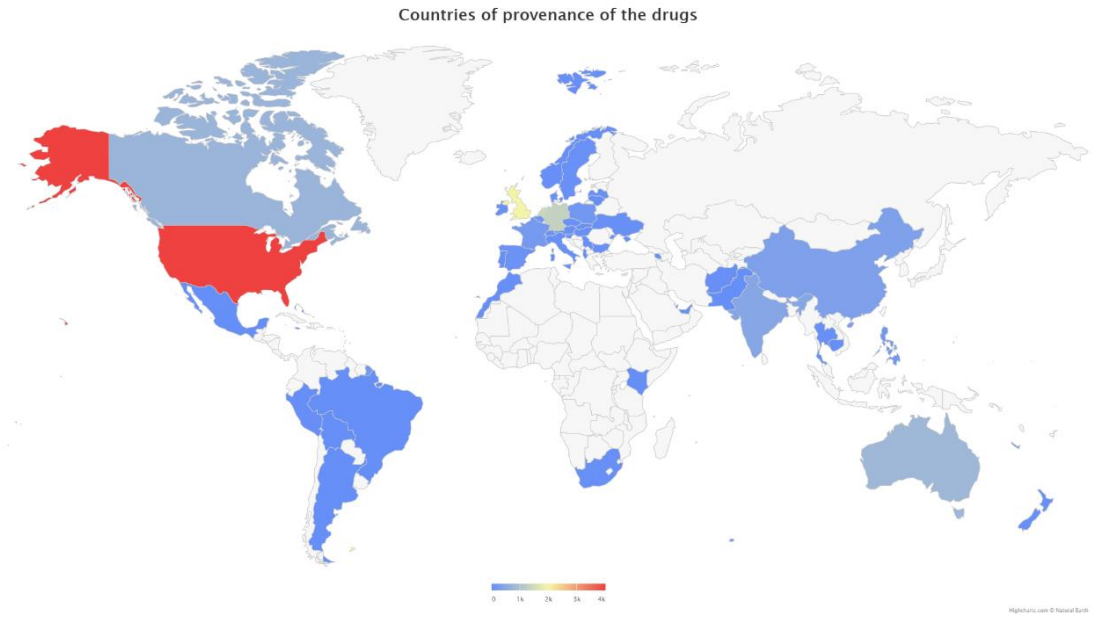

Figure 4: Geographic heat map of goods provenance

A heat map can also be used to show the different kind of products that are shipped from each country. Using the "Pivot" operator from RapidMiner, in fact, a column for each country, with the respective count of its insertion can be obtained, grouped by the product type. The result of this operation can be seen in the Figure 5 below, where only the top 10 shipping location are represented. As the heat map seen before, the yellow and then the red colour represent a bigger value, rather than the blue colour. It can be noticed that the sellers from the United States are listing many cannabis and hashish related products, the reason could be the different legislation for the dealing and consumption of cannabis related products in the different states. Moreover, the United Kingdom and the "Worldwide" label are characterized by a huge amount of cannabis and related products listings. It is interesting to see that, the only other product with more than 500 listings for a single country is the ecstasy product, in The Netherlands. Moreover, another category going near the 500 listings for a single country is the stimulants category, shipped from the United States. Lastly, the paraphernalia products are not widely shipped from many countries, but they are concentrated in the United States.

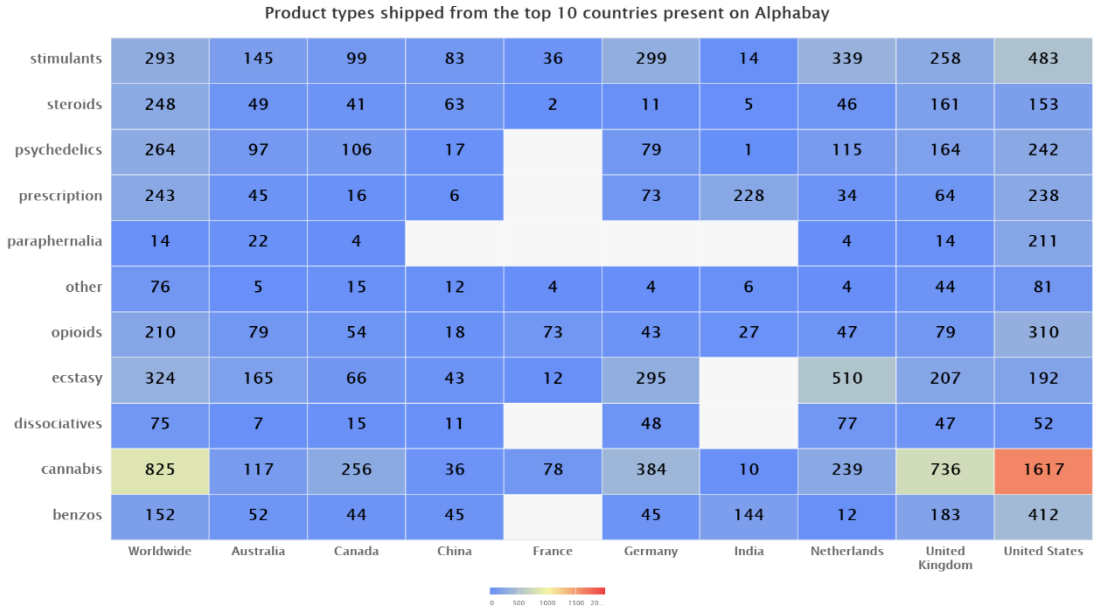

Figure 5: Heat map of the top 10 countries for shipping goods and the relative categories

Going on with the data visualization activities, the focus is moved from the countries where the products are shipped from, to the countries where the products are shipped to. From the Figure 6 below, it can be seen that the situation is quite different from the analysis of the countries where the products are shipped from. In fact, the destinations indicated in this field, can contain more than one destination. As it has been seen before in Figure 2, Worldwide location and the United States are dominating the scene, but their raking is now swapped, with a huge preference for the Worldwide shipping. It can mean that many goods shipped from the United States, can be delivered worldwide, or in other regions. moreover, new actors like "Europe" or "North America" are showing off. The reason behind this difference with the Figure 2, could be the fact that drug dealers do not encounter many barriers with the international shipping. If it is the case, it means that the customs controls between the different countries are not contrasting the international drug dealing efficiently.

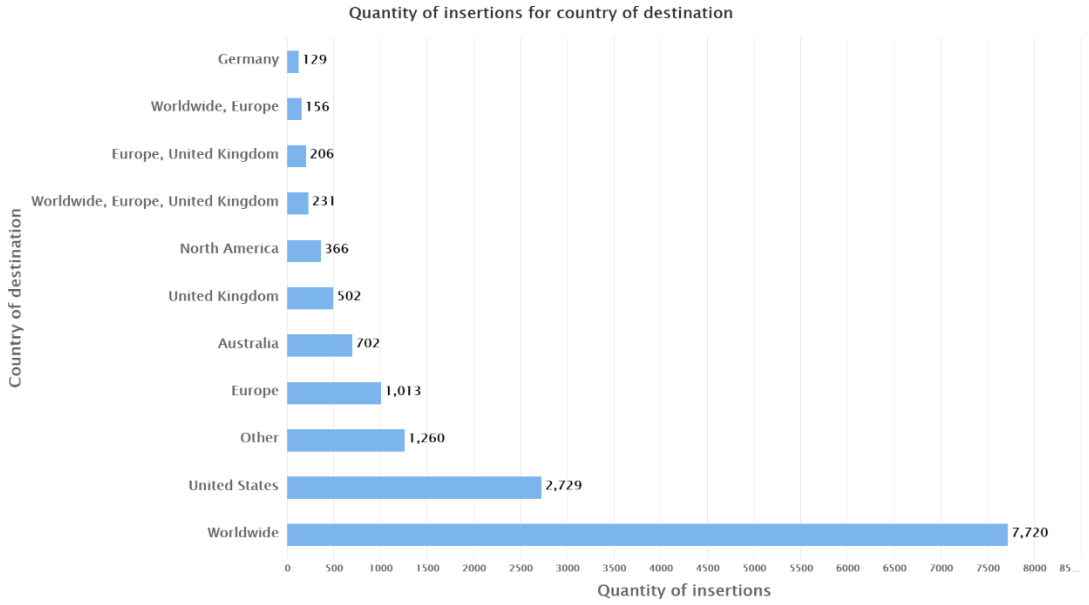

Figure 6: Quantity of insertions for country of provenance

As stated before, the shipping destinations are more varied because more than one destination is allowed for each listing, giving the result of 171 destinations. This means that the results are more diluted as hoped. By the way, a heat map for Worldwide, United States and Europe is analysed in the next page, trying to still get some useful information.

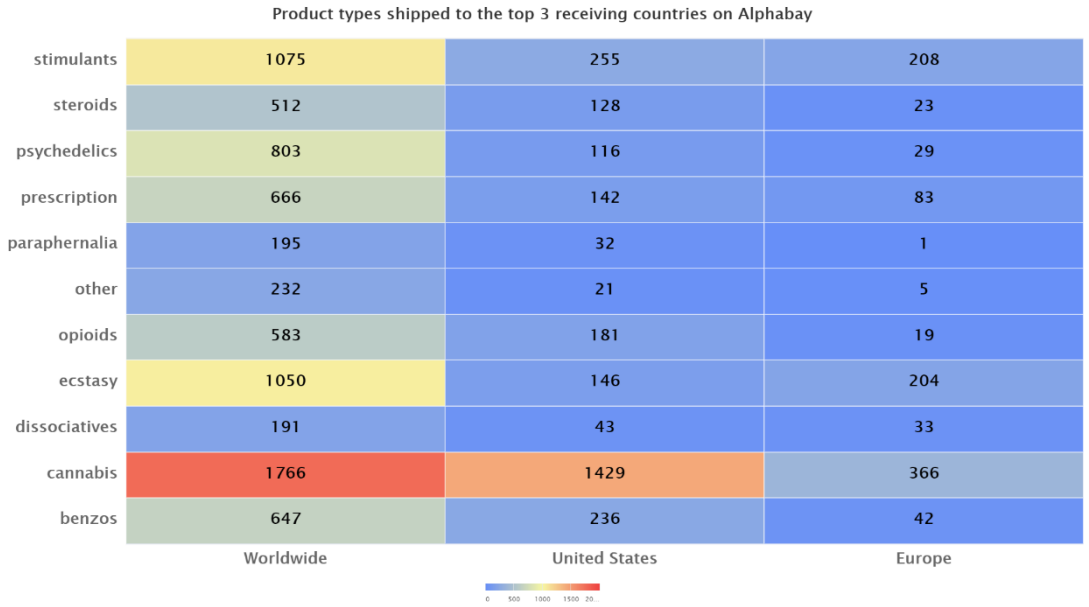

Figure 7: Heat map of the top 3 locations for receiving goods and their relative categories

As it can be noticed from the Figure 7 above, there is a large increase of products shipped worldwide. Moreover, there is a huge reduction of stimulants, psychedelics, prescription, paraphernalia, other and benzos products shipped to the United States, meaning that these kinds of products are more easily shipped to other countries rather

than other products, like cannabis and hashish products. Another reason can be that these products are easier to obtain in the United States and it is highly profitable to ship them to other countries. In the end, the Europe region does not give any useful insight. It can be caused by the huge fragmentation of the available destination (171 destinations). Moreover, further data pre-processing activities could have provided a better view of the European countries and their drug dealing activities.

Data Mining and Machine Learning applications

The data visualization step gave a good insight on many useful columns present in the dataset. On the other hand, there are other information that can be extracted from the text columns. These columns, in fact, contain unstructured data which can be analysed through Regular Expression and, furthermore, with Machine Learning techniques.

The columns that contain text data in the main dataset, which are going to be used for the next analysis are, alphabetically:

- Description: which contains the description of the goods provided by the sellers.
- Feedback: which contains the feedbacks left from the buyers of the goods.
- Name: which contains the name of the insertions.
- Vendor_name: which contains the anonymized nickname of the sellers.

Beginning with the Name column, it is easy to notice that one of the information provided in this column is the weight of the good. In fact, this is the main text a buyer is going to read, when they are looking for a product on Alphabay. The pattern to provide this information is almost always the same. The numbers, sometimes a space and then the unit of measure. This pattern can be easily detected with a Regular Expression defined in the Generate Extract operator, which creates a new column with the result of the Regular Expression. As there are 398 different combinations of quantities and unit of measure, an aggregation is made to rank the most common weights for each product type. In the end, a Python script is used to select only the 3 most popular weights sold for each product type. As it can be seen from the Figure 8 below, the columns have been stacked. This means that it can be seen if a certain quantity is used only for a certain type of goods or even for more. A stacking in percentile has been avoided as it would not make it clear which quantity is preferred in general.

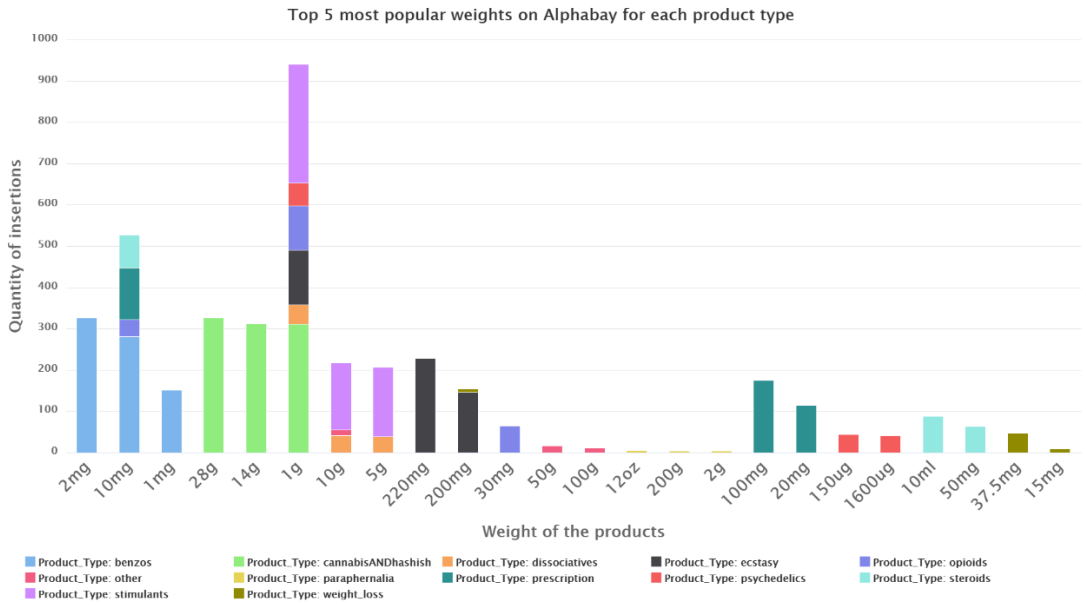

Figure 8: Top 3 most commercialized weights for insertions on Alphabay

The weight preferred to commercialize is, as it can be easily presumed, 1 gram. Six out of twelve product types are highly commercialized in this weight. It can happen that even the other product types are commercialized in this weight but, as only the most used ones for each product type are shown, these insertions may have not been included in the graph. Psychedelics products are mostly sold in a very small weight, in the unit of micrograms, making them easier to ship stealthily. As it is known, psychedelics drugs can even be shipped within a letter, thanks to their weight and as they can be produced in the form of a postage stamp. Moreover. this information is not new to the experts of illicit substances and to the police forces but, extracting this information can be useful for Machine Learning algorithms, for example to classify the product types.

Going on with the analysis of the Vendor_name column, it can be interesting to understand if the sellers on Alphabay are continuously selling over 2015 and 2016 by checking the nicknames that are present in both the datasets. As the anonymization step was made in order to preserve the correspondence between the nicknames in the two datasets, it is easy to perform this comparison. The column Vendor_name is selected from both the datasets, duplicates have been removed, then a Set Minus operator is used to get the vendors that are only present in one dataset, while the Join operator is used to perform an inner join, in order to get the vendors that are present in both the datasets. A different label has been given to the two results and, in the end, an Append operator is used to get the two results together.

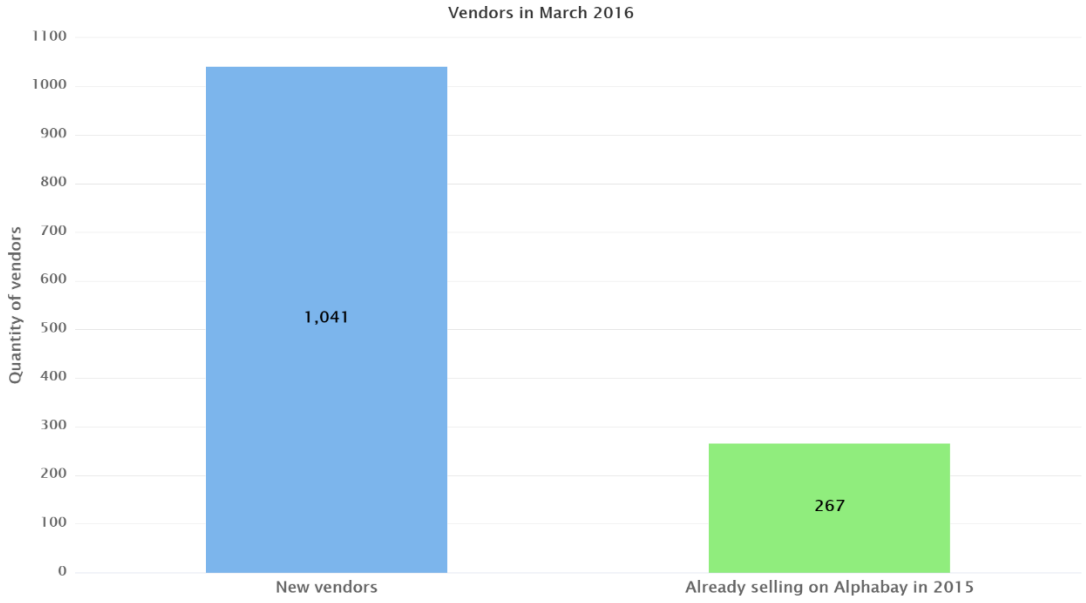

Figure 9: Vendors on Alphabay in March 2016

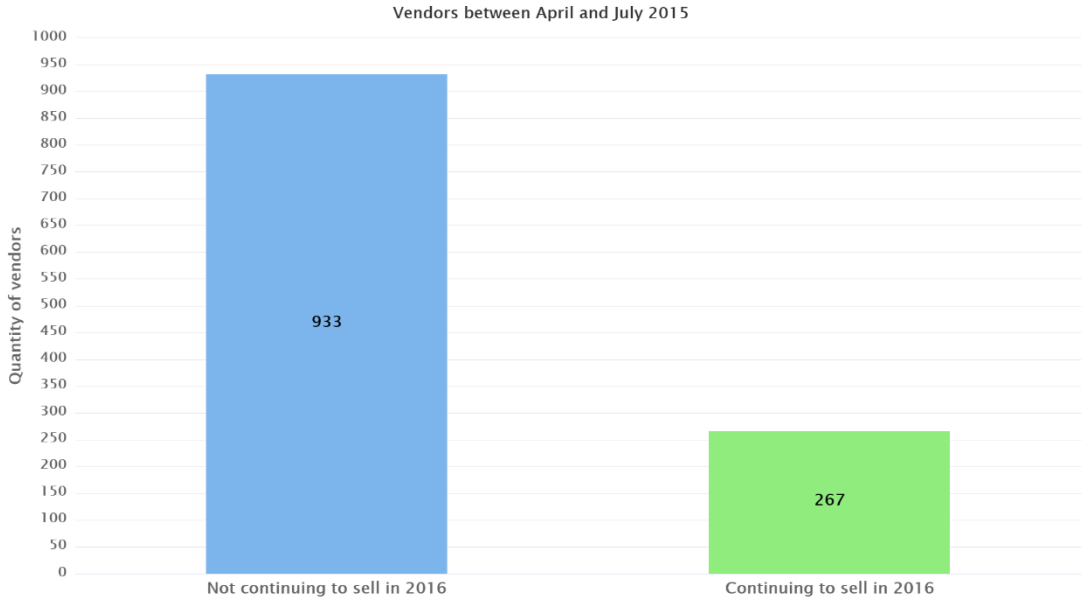

Figure 10: Vendors on Alphabay between April and July 2015

The results of these operations, carried out on both the datasets, are shown in the Figure 9 and Figure 10 above. As it can be seen, the number of vendors on Alphabay is almost the same in the two different periods. It has to be said that the data from Figure 10 have been crawled for almost 3 entire months, while the data from Figure 9 come from a 24h crawling. Therefore, there is the possibility that the number of vendors present on Alphabay in March 2016 is underestimated. In the end, it can be noticed that 267 vendors were selling on Alphabay in both the time periods. These sellers can probably be the most organized and the ones with the highest dealing activities. On the other hand, it can be assumed that the other sellers may have closed their accounts and

started to sell with new accounts, in the circumstances of bad reviews or in case they were present on Alphabay to scam the customers.

Proceeding with the analysis of the Description column, it can be interesting to search for email addresses, which have been, together with the vendor names, anonymized in the anonymization step. The aim of this analysis is to find email addresses associated with more sellers on Alphabay, in order to understand possible cartels. As before, a Regular Expression can help to perform this operation easily. Then, only the Email and the Vendor_name columns have been selected and duplicates have been removed. In the end, the missing values in the Email column have been removed. These operations led to the fact that there are 40 unique emails associated to 38 vendors. In particular, 2 vendors are each using 2 different emails, but there are no emails associated with more distinct vendors. This analysis is therefore not providing any useful information more than that the vendors are not so usual to put an email address in the description of their products, preferring to conversate with the customers on Alphabay, probably afraid to reduce their privacy.

Furthermore, a sentiment analysis has been performed over the Feedback column. This column, in fact, contains the feedbacks left from the customers after purchasing on Alphabay. This column contains 6159 missing values over 15013 rows, it is not the best situation, but it is still worth to try. In order to perform a sentiment analysis and even further analysis, the Operator Toolbox plug-in has been installed. There, the Extract Sentiment operator can be found. This operator can easily perform a sentiment analysis without having to deal with coding. Sentiment analysis consists in looking for an overall sentiment, analysing the body of a text. This operation is mostly useful for companies to understand the satisfaction of the customers and, in this case, it can be a signal to understand which insertions are fraudulent. This operator assigns a negative or positive weight to the words present in the text and gives, as a result, the sum of their weights as a score. As the context can be quite different from the average use of this operator, few words and their score have been rebalanced. For example, the term “stealth” is a positive characteristic while receiving illegal products home delivered. For these words, the value has not been changed, only the sign has been changed to the opposite. So, the word “stealth” is just changed from a minus sign to a plus sign. Following, a distinction has been made between the insertions with a negative score, the ones with a score equal to zero, reported as with a neutral sentiment, and the ones with a positive score. Creating these categories is useful to show graphically the results

of the sentiment analysis. As it can be noticed from the figure 11 below, the most populated category is the one for the insertions with a neutral sentiment. In fact, the huge number of missing values in the Feedback column and, probably, many insertions without proper feedbacks, took to this result. Another interesting fact is that only 87 insertions have been reported to have a negative sentiment in the feedbacks. A check of the negative sentiment distribution over the product types has been performed, but it did not bring to light any useful insight. On the other hand, the positive labelled insertions are 3557, indicating an overall satisfaction for the customers on a big share of insertions. It would have been interesting to see if these results were comparable with the Gwern dataset (2015) but, unfortunately, that dataset does not contain a feedback column.

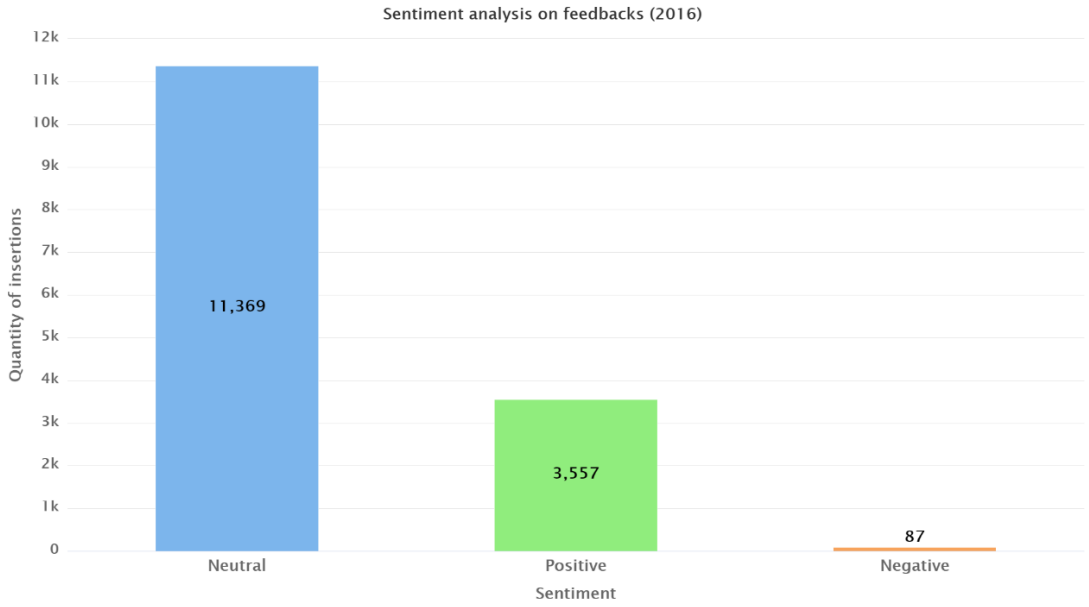

Figure 11: Sentiment analysis on the Feedback column

Going on with the analysis on the text columns, it is interesting to see how a topic extraction algorithm performs on the Description column. This column is composed by descriptions long enough to detect patterns, which can be useful to classify the different product types. The use of certain words, in fact, could be a characteristic of a specific product type. The operator LDA from the Operator Toolbox plug-in is used. This operator applies the LDA (Latent Dirichlet Allocation) method to the chosen column, in this case, the description. LDA works through a generative approach, it means that it does not require a label column to work. Instead, it generates a probabilistic model, avoiding strong assumptions between the labels and the text. The presence of the topics in each insertion is calculated with the use of a set of particular

words, chosen by the model, as their presence means a pattern for a certain topic. An optimization of the hyperparameters has been tried, in order to improve the model, considering the insertions classified with their product types, for each topic. The number of topics can be optimized against their perplexity value but, optimizing this measure may not always lead to an easier human interpretation. Measuring the perplexity of the model with the number of topics spacing from 3 to 15 showed that the perplexity is inversely proportional to the number of topics. By the way, the diminishing of the perplexity was not leading to better performance in the classification models that will be later applied. The inexperience of the author did not lead to a numerical measurement to understand the best number of topics in order to get topics that are only present in certain product types. Looking to the results graphically, 6 topics have been chosen as, an increase over this number, was not leading to the creation of topics with interesting characteristics. On the other hand, the "top words per topics" hyperparameter was easily optimized through the Optimize Parameter operator, the resulting table for the 5 top words per topic, is shown in the Figure 12 below. This means that only the 5 most influencing words found out by LDA are going to be used to calculate the confidence for the descriptions to contain a certain topic.

topicId	word	weight
0	order	5140
0	shipping	5064
0	please	4311
0	days	3384
0	orders	3320
1	product	3763
1	description	2938
1	xanax	1920
1	generic	1848
1	order	1741
2	product	8823
2	description	5805
2	mdma	4211
2	quality	4119
2	days	3279

topicId	word	weight
3	product	1568
3	lsd	1403
3	dmt	1316
3	description	1199
3	get	1027
4	product	4040
4	description	3736
4	weed	3716
4	cannabis	2874
4	high	2870
5	testosterone	2179
5	steroid	1228
5	steroids	1194
5	use	1184
5	also	1000

Figure 12: top 5 words per topics chosen by the LDA algorithm

It is therefore interesting to understand if the results of this topic extraction process are useful to classify the different product types. In the Figure 13 below, it is shown all the topics and their relative insertions, grouped by their product type.

Beginning with the topic 0, there is not much interesting to see. Recalling also the Figure 12, it can be seen that the words used for this pattern are not specific to any particular product. Looking to the topic 1, it can be noticed the huge number of benzos and prescription products, the words "Xanax" and "generic" therefore, have a huge weight to detect these kinds of products. Going on with topic 2, the stimulants and ecstasy products have a big share of this topic, but also the opioids and mostly all the dissociative products are there. The "mdma" word is recalled for this topic, which is an ecstasy product. The topic 3 contains mostly all the psychedelics and the paraphernalia products, "lsd" and "dmt", two psychedelic substances, have a heavy weight on this topic prediction. The topic 4 is dedicated to the cannabis and hashish products, with words like "weed", "cannabis" and "high" in the top 5 words parameter. In the end, the topic 5 is composed by steroids products and, in fact, some of the words used to detect this topic are "testosterone", "steroid" and "steroids". In the end, it can be said that the LDA seems to have achieved a good result in detecting patterns for the different product types.

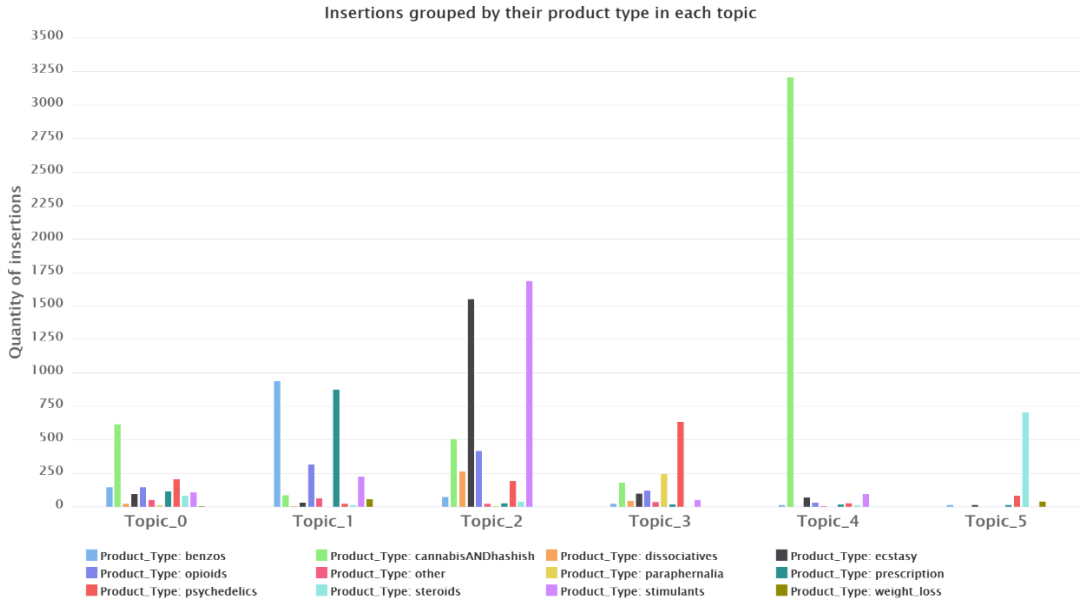

Figure 13: LDA resulting topics and their composition

Going on with the text mining activities, it is another time the turn of the Name column. This time, extracting specific words is the main focus. On the first place, the extraction of the term frequency is going to be used to detect the term with lowest frequency and the term with highest frequency for each insertion. To perform the first part of the job,

which is the calculation for the term frequency within all the rows, an easy procedure with the RapidMiner operators would have been sufficient. On the other hand, the typing errors, the slang and the difference between the singular and the plural of the terms would produce an inaccurate term frequency vector. For this reason, a mix of RapidMiner operators and a custom Python script has been used. The Levenshtein library has been used. This library makes possible to calculate the "distance" between two words, confronting the letters present in both. So, after the creation of a term frequency vector, it has been ordinated from the highest frequency to the lowest frequency. Then, each word from the vector is iterated over the same vector. If the Levenshtein ratio between the iterated word and any other one present in the vector is greater or equal to the 80%, the word with the highest frequency is summed with the frequency of the lowest, and the word with the lowest frequency is therefore dropped. An example of this process can be: the word "seed" has a frequency of 500, while the word "seeds" has a frequency of 200. In the end, only the word "seed" will stay in the vector, and its frequency will be of 700. The next step regards the detection of the term with the highest frequency and the term with the lowest frequency, for each insertion. To perform this job, another Python script has been developed. There, the term frequency vector is ordinated in ascending way, then it is looped from the term with the highest frequency to the lowest one for each word present in the insertion "Name" column, it will break the loop when the term looped is also present inside the insertion name. The same reasoning is applied to detect the term with the lowest frequency.

In the end, the Name column is also used to detect the drugs or the active substance through the use of the "List of Drug Master Files" (DMFs), provided by the FDA (U.S. Food & drug administration). This file contains a column with mostly 35,250 drugs or active substances. This column is therefore imported in the RapidMiner project and used to detect drugs in the Name column. The Python script used before to extract the terms with the lowest and the highest frequency for each row can be easily adapted to this new job.

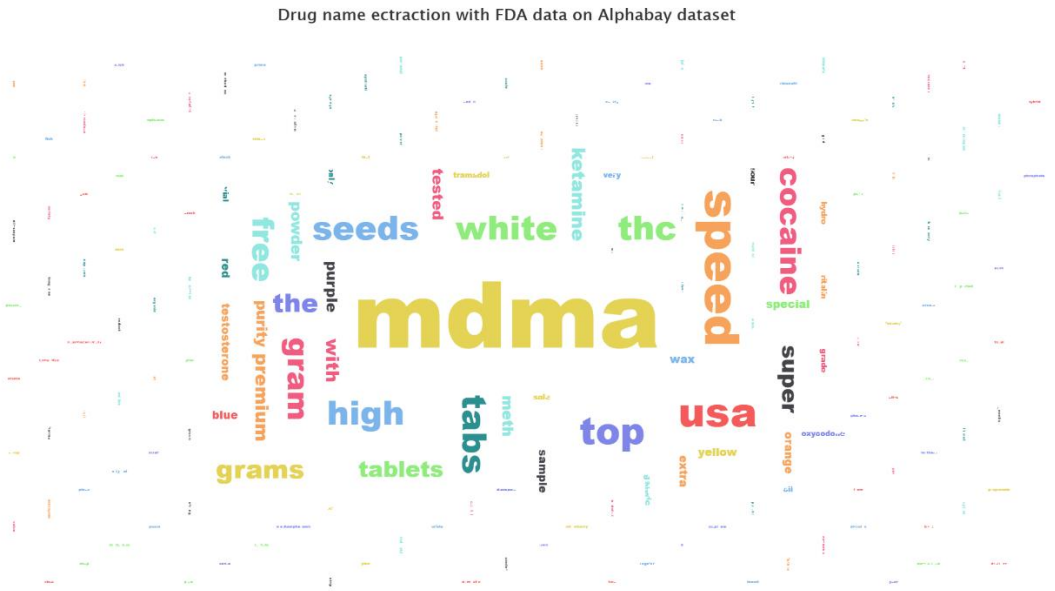

Figure 14: Wordcloud graph for drugs detected in the Name column

As it can be noticed from the Figure 14 above, the most recognized drug with the use of the FDA dataset is the mdma, followed by speed, thc, cocaine and so on. As the data from FDA are not including only the drug names or the active substances, but even small descriptions, it was necessary to exclude some not useful words that were present in many rows in the dataset, but still some detected words are not related to drugs.

2.5 Building the classification models

The last task for the analysis of the Alphabay market is the building of a classification model, useful to classify the product type label for each insertion present in the dataset. Many of the steps above produced insights which are going to be useful for the classification model. As one of the main functions of RapidMiner is the Auto Model, the dataset with all the computations made before is saved and then imported in this area of the tool. The columns selected for the Auto Model step are:

- FDA Drug
- Highest frequency term
- Lowest frequency term
- Price_t
- Ships_from
- Topic
- Weight

Escrow, Sentiment and Ships_to columns have not been included as after some tests, their contribution to the classification models is near to zero. Moreover, they are not even present in the Gwern dataset, so they have been dropped in order to apply the same model to both the datasets. The result of the Auto Model is a Gradient Boosted Trees model. This model is based on many trees, placed sequentially, which optimize their parameters at each iteration. Each tree has one split, producing two different results. At both the ends of this split, there is another tree, continuing in this way until the maximum depth, which is set in advance, is reached, or until no more improvement is done. The tuning of the parameters gives, as a result, some combinations of parameters to choose, with mostly the same error rate. So, in order to prevent the overfitting of the model, the model with 30 trees is preferred over the one with 150 trees. Moreover, the maximum depth is set to 4 and the learning rate is set to 0.1. This model has been reconstructed then in the main project, with a Bootstrapping Validation operator, and then by setting the Gradient Boosted Trees model inside of it. As it can be noticed from the Figure 15 below, the results are impressive. The accuracy of the model is near the 88.3%, with a high class recall value, for almost all the categories. Other and paraphernalia products are an exception, the reason could be the low number of insertions present in the database, therefore the training did not have enough data for these products. As the FDA Drug, Highest frequency term and Lowest frequency term columns can probably share some kind of correlation, another Gradient Boosted Trees model has been tested, with only the FDA Drug column between them, scoring near the 81.6% of accuracy, still a good result.

accuracy: 88.30% +/- 0.36% (micro average: 88.30%)

	true benzos	true cannabi...	true dissoci...	true ecstasy	true opioids	true other	true paraph...	true prescri...	true psyche...	true steroids	true stimula...	true weight_...	class precis...
pred. benzos	3910	20	5	12	65	35	0	305	11	48	40	0	87.85%
pred. canna...	59	16293	21	181	113	67	94	75	93	41	309	0	93.93%
pred. dissoc...	7	12	1020	12	7	6	8	7	27	6	26	0	89.63%
pred. ecstasy	50	79	63	6299	47	20	40	38	82	8	170	2	91.32%
pred. opioids	37	32	1	20	3159	22	38	227	44	17	88	4	85.63%
pred. other	32	21	3	12	6	349	7	20	14	4	12	0	72.71%
pred. parap...	4	41	1	1	29	16	568	9	26	5	28	0	78.02%
pred. prescr...	311	33	3	20	239	47	22	3012	17	67	204	21	75.38%
pred. psych...	59	48	40	69	62	48	138	33	3949	10	65	0	87.35%
pred. steroids	52	11	0	6	12	2	8	85	26	2908	3	54	91.82%
pred. stimul...	51	317	85	328	212	102	70	174	124	19	7076	1	82.67%
pred. weight...	0	1	0	0	0	1	2	18	0	21	3	324	87.57%
class recall	85.52%	96.36%	82.13%	90.50%	79.95%	48.81%	57.09%	75.24%	89.49%	92.20%	88.19%	79.80%	

Figure 15: Gradient Boosted Trees classification performance on Product type label

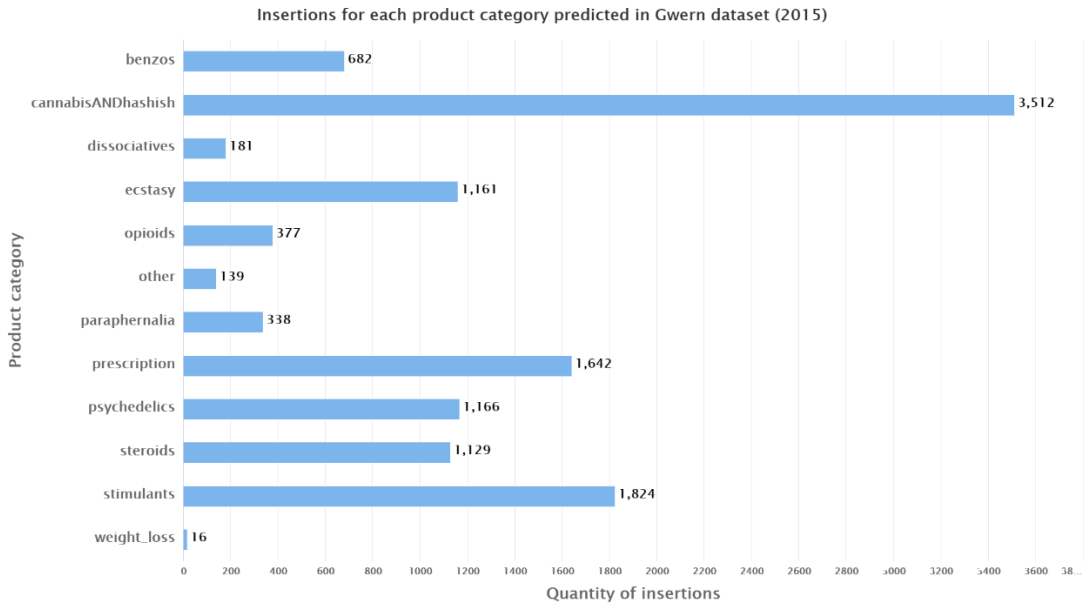

Figure 16: Classification results on Gwern dataset

The same model is then applied to the Gwern dataset, giving as a result the Figure 16 above. It can be noticed that, compared with Figure 1, the ecstasy and psychedelics products are more present in this dataset, always remembering that these are predictions and that there could be many misclassification errors.

It would be interesting to apply this classification model to other datasets which contain the Product_type label, to calculate the accuracy on a dataset different from the training one.

As the proceedings shown before gave, as a result, a high importance to the topic extraction in the building of a classification model, used to classify the product type of a good sold on Alphabay, deeper research has been done in this field. Among the topic extraction algorithms developed, with their implementation mostly done in Python, the GuidedLDA package is probably the one that mostly fits with the topics discussed in this paper. As mentioned before, during the topic extraction already performed, the dropping of certain words contained in all the rows, or most of them, has been tested. Words as “product”, “description”, “order”, “days” have been dropped during some tests. By the way, the classification results were getting even worse rather than improving. In a sort of a way, it was tried to “drug” the algorithm, without having any positive result. The GuidedLDA algorithm, on the other hand, is “semi-supervising” the LDA algorithm shown before. In fact, the LDA (Latent Dirichlet Allocation) is an unsupervised machine learning technique. This means that it is not influenced by labels in any way. The GuidedLDA implementation, instead, can accept

a vector of words, called “seeds”, to influence the LDA results. It is not mandatory to provide seeds for each topic, and the number of seeds can also differ. It should also be noticed that the seeds are a starting point for the algorithm but it does not mean that all of them are going to be used. The algorithm will therefore balance its considerations over the seeds, with a given confidence. In order to work with this library, a virtual environment with Python 3.6 has been created, trying to avoid incompatibilities between the GuidedLDA library and the new versions of Python and other libraries. In order to get the seeds for each category, the LDA operator has been performed on RapidMiner, for each category took alone. The resulting words have been hand chosen in order to avoid generic words as “product”, “description” or words in common between two or more categories. The result can be seen in the Listing 3 below. Moreover, in order to optimize the algorithm, the “nltk” library has been used to drop stopwords and the punctuation. An unforeseen incompatibility between the “nltk” library and the RapidMiner project forced the author to get the “Description” column in csv format to work on Python, while the results of the script have been reimported in RapidMiner, still in csv format, rather than using the Execute Python operator. The model has been trained with 4 seeds, a seed confidence of 85%, 200 iterations and a topic for each product type.

```
benzos = ["xanax", "alprazolam", "valium", "diazepam", "anxiety"]
cannabis = ["weed", "high", "strain", "cannabis", "hash", "kush"]
dissociatives = ["ketamine", "ghb", "gbl", "mxe"]
ecstasy = ["mdma", "pills", "xtc", "molly", "ecstasy"]
opioids = ["heroin", "fentanyl", "pain", "oxycodone"]
other = ["cigarettes", "tobacco", "carton", "marlboro"]
paraphernalia = ["optiman", "pipe", "grinder"]
prescription = ["generic", "viagra", "cialis", "modafinil", "sildenafil"]
psychedelics = ["lsd", "dmt", "trip", "blotter", "pgp"]
steroids = ["testosterone", "steroid", "steroids", "anabolic", "trenbolone"]
stimulants = ["cocaine", "speed", "coke", "amphetamine", "meth"]
weight_loss = ["loss", "weight", "phentermine", "fat", "ephedrine"]
```

Listing 3: Chosen seeds for GuidedLDA

Given these premises, the GuidedLDA algorithm could have offered already labelled results, avoiding the passage between an unlabelled LDA, with a number of topics different from the number of categories, and the subsequent use of a machine learning algorithm as the Gradient Boosted Trees model. To get a confusion matrix, the results have been imported in RapidMiner and joined with the original label “Product_Type”, then the “Performance (Classification)” operator has been used. The result is visible in the Figure 17 below. The first thing to notice is the accuracy, which is near the 63%. It is easy to notice that this model is badly performing on categories as dissociatives,

prescription and weight_loss. In particular, it is misclassifying more than 500 cannabis products as “other”, while the most of prescription products is misclassified as benzos. The use of seeds made the author think about overfitting, while the results showed how the classifying of similar products, as drug categories, does not seem the main aim of the GuidedLDA model, even if the premises were promising.

accuracy: 63.16%

	true benzos	true canna...	true dissoc...	true ecstasy	true opioids	true other	true paraph...	true prescri...	true psyche...	true steroids	true stimul...	true weight...	class preci...
pred. benzos	974	29	2	2	57	21	1	508	6	3	204	58	52.23%
pred. cann...	0	3184	0	11	3	7	18	3	4	2	13	0	98.12%
pred. disso...	73	234	73	141	28	5	0	101	94	37	135	4	7.89%
pred. ecsta...	25	35	167	1368	11	15	5	9	198	0	104	0	70.62%
pred. opioids	43	54	22	4	727	25	29	177	59	8	42	0	61.09%
pred. other	43	577	15	65	31	78	7	30	168	64	33	6	6.98%
pred. parap...	12	198	7	70	31	11	183	33	21	11	81	0	27.81%
pred. presc...	18	2	0	16	0	10	0	161	52	6	14	0	57.71%
pred. psych...	0	60	1	4	0	4	2	2	481	0	0	0	86.82%
pred. steroi...	14	3	5	10	4	12	3	51	11	735	7	44	81.76%
pred. stimu...	21	141	57	168	122	12	21	11	68	6	1516	0	70.74%
pred. weig...	2	93	0	17	35	1	0	1	12	0	38	2	1.00%
class recall	79.51%	69.07%	20.92%	72.92%	69.30%	38.81%	68.03%	14.81%	40.97%	84.29%	69.32%	1.75%	

Figure 17: Confusion matrix for the GuidedLDA classification over the product types

Conclusions

The project shown it this paper empathises how unstructured data can hide information which, combined with structured data, gives insights that are not easily visible by eye. Moreover, the use of RapidMiner has surprised the author about the easy-to-use interface, which can help researchers to handle data mining activities and machine learning models without coding. On the other hand, the use of Python has been fundamental to provide deeper insights, with personalized scripts. The way to think about data and the techniques applied in this paper can therefore be useful to inspire data scientists and police forces which investigate in this field, to think outside of the box. In particular, the exploratory data analysis has brought to light how international drug dealing activities regard differently each country, emphasizing the places where each kind of drug is produced or consumed the most. The sentiment analysis over the feedbacks showed that, most of the times, the goods bought on Alphabay are successfully delivered to the customers, meaning that the controls over the national and international shipping are not effective in countering the delivery of drugs. The topic extraction over the description of the products has been revealed to be very effective to build classification models but, on the other side, the application of newly developed and not well documented models like the semi-supervised LDA did not show an improvement over the use of the conventional LDA. In the end, the aim of the paper to provide easily interpretable insights and to build a classification model to classify the products present on Alphabay and, in general, on the Darknet markets, has been accomplished.

Despite the achievements obtained in this paper, there are some limitations of the study. In particular, the inexperience of the author may have led to naive errors, providing conclusions and applying methods that are not appropriate.

On the other hand, this study provides a basis for further research. In particular, to apply social network analysis techniques on the emails discovered in the paper, using open-source intelligence software. Moreover, comparing the results of this study with data collected on other markets, can provide a more general view of the drug dealing activities, not only on Alphabay, but on a bigger portion of the Darknet markets.

Bibliography

- Nastuła, A. (2019). New threats in the cyberspace based on the analysis of the TOR (the onion router) network. ASEJ Scientific Journal of Bielsko-Biala School of Finance and Law, 22(4), 28-31. https://doi.org/10.5604/01.3001.0012.9839
- Bergman, M. K. (2001). White paper: The deep web: Surfacing hidden value. The Journal of Electronic Publishing, 7(1), 147. https://doi.org/10.3998/3336451.0007.104
- Christin, N. (2013). Traveling the silk road: A measurement analysis of a large anonymous online marketplace. Paper presented at the 213-224. https://doi.org/10.1145/2488388.2488408
- Pergolizzi, J. V., LeQuang, J. A., Taylor, R., Raffa, R. B., & NEMA Research Group. (2017). The “Darknet”: The new street for street drugs. Journal of Clinical Pharmacy and Therapeutics, 42(6), 790-792. https://doi.org/10.1111/jcpt.12628
- Minnaar, A. (2017). ONLINE 'UNDERGROUND' MARKETPLACES FOR ILLICIT DRUGS: THE PROTOTYPE CASE OF THE DARK WEB WEBSITE 'SILK ROAD'. https://www.researchgate.net/publication/333646270_ONLINE_'UNDERGROUND'_MARKETPLACES_FOR_ILLICIT_DRUGS_THE_PROTOTYPE_CASE_OF_THE_DARK_WEB_WEBSITE_'SILK_ROAD'
- Reddy, E. & Minnaar, A. (2018). CRYPTOCURRENCY: A TOOL AND TARGET FOR CYBERCRIME. https://www.researchgate.net/publication/338572871_CRYPTOCURRENCY_A_TOOL_AND_TARGET_FOR_CYBERCRIME
- Bancroft, A. (2019). The darknet and smarter crime: Methods for investigating criminal entrepreneurs and the illicit drug economy. Springer International Publishing. https://doi.org/10.1007/978-3-030-26512-0
- Chakraborty, R. (2018). the deep web: For the nefarious or the democratic? Harvard International Review, 39(4), 18-21. https://www.jstor.org/stable/10.2307/26617373

- Branwen, G. et al. (2015). “Dark Net Market archives, 2011–2015”. https://www.gwern.net/DNM-archives
- Christin, N. (2017). An EU-focused analysis of drug supply on the AlphaBay marketplace. EMCDDA commissioned paper. https://www.drugsandalcohol.ie/28855/1/An%20EU-focused%20analysis%20of%20drug%20supply%20on%20the%20AlphaBay%20marketplace.pdf
- Baravalle, A., & Lee, S. W. (2018). Dark Web Markets: Turning the Lights on AlphaBay. Lecture Notes in Computer Science, 502–514. https://doi.org/10.1007/978-3-030-02925-8_35
- The Ultimate Guide to the Invisible Web. https://oedb.org/ilibrarian/invisible-web/ Last visited: 07/05/2022
- The Open Group Base Specifications Issue 6 IEEE Std 1003.1, 2004 Edition. https://pubs.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html Last visited: 25/07/2022
- Latent Dirichlet Allocation component. https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/latent-dirichlet-allocation Last visited: 22/08/2022
- Evaluate Topic Models: Latent Dirichlet Allocation (LDA) https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 Last visited: 22/08/2022
- A Visual Guide to Gradient Boosted Trees (XGBoost) https://towardsdatascience.com/a-visual-guide-to-gradient-boosted-trees-8d9ed578b33 Last visited: 22/08/2022
- List of Drug Master Files (DMFs) https://www.fda.gov/drugs/drug-master-files-dmfs/list-drug-master-files-dmfs Last visited: 22/08/2022
- GuidedLDA: Guided Topic modeling with latent Dirichlet allocation https://github.com/vi3k6i5/GuidedLDA Last visited: 15/09/2022