Universitat Oberta de Catalunya (UOC)

Data Science master's degree (*Data Science*)

MASTER'S DEGREE FINAL WORK

Area: 5

Sale of access IoT devices in underground forums

Author: Fernández García, Sergio

Tutor: Hernández Gañan, Carlos

Professor: Solé Ribalta, Albert

Ibi, January 2, 2022

Copyright

FINAL WORK SHEET

Title:	Sale of access IoT devices in underground forums
Author name:	Sergio Fernández García
Teaching collaborator name:	Carlos Hernández Gañan
Responsible professor name:	Albert Solé Ribalta
Date of delivery (mm/yyyy):	01/2022
Degree or program:	Data Science master's degree
Área del Trabajo Final:	M2.882 - TFM - Area 5
Work language:	English
keywords:	IOT, sale, underground forums

Dedication / Quote

To my family, infinite source of happiness.

Thank you for giving meaning to my life.

There are no favorable winds for those who do not know where they are going.

Seneca

Abstract

Insecure configuration and lack of consistent updating in devices in the Internet of Things has led to widespread malware infection. Infected devices are added to botnets, commonly used to perform Distributed Denial-of-Service (DDoS) attacks. Underground forums allow criminals to interact, share knowledge and trade in botnets usage through Command and Control servers. Given the great potential for harm of these botnets, it is important to understand how this information is shared, and the economics behind it.

Current work analyzes already collected information about underground forums in order to know how IoT accesses are sold and its evolution over last years.
Largest underground forum on the *clear web* is *Hack Forums*. Given the impossibility of analyzing all the forums, it has been the chosen forum to carry out the study.

Collected data is extracted and classified between *supply* and *demand* by using a classification model. That model is selected, by its performance, from a set of well known classification models. Classified data is then analyzed as time-series in order to extract conclusions.

Evolution over last ten years shows that *supply* has surpassed *demand*. Common terms used to refer to these services have evolved too, being *botnet, DDOS and stresser* more common nowadays. Preferred payment methods evolution shows that Bitcoin has passed even USD and is currently the most used.

Keywords: IOT, botnet, ddos, access, sale, underground, forums, evolution, Hack Forums

Resumen

Las configuraciones inseguras y la falta de actualizaciones en los dispositivos IoT, han favorecido la expansión de malware. Los dispositivos infectados suelen formar parte de Botnets, las cuales se utilizan para realizar ataques de denegación de servicio distribuidos. Los foros underground, sirven a los criminales para compartir conocimiento y vender servicios de botnet en forma de servidores tipo *Command and Control*.
Dado el gran potencial para causar daño de estas botnets, es importante conocer el modo en que esta información es compartida y los aspectos económicos detrás de las mismas.

El presente trabajo analiza información existente sobre foros clandestinos, con el fin de conocer el modo en que son vendidos los accesos a dispositivos infectados y su evolución en los últimos años.
EL foro clandestino más grande de la *clear web* es *Hack Forums*. Dada la imposibilidad de analizar todos los foros, este ha sido el elegido para llevar a cabo el presente estudio.

Los datos recogidos son extraídos y clasificados en dos categorías principales, *oferta* y *demanda*, mediante un modelo de clasificación. Dicho modelo es seleccionado por su precisión de entre un conjunto de modelos comúnmente utilizados para tareas de clasificación. La información clasificada es analizada en forma de series temporales para extraer conclusiones.

La evolución en los últimos diez años, muestra que la oferta a superado a la demanda. Los términos utilizados para referirse a este tipo de servicios, ha evolucionado también, siendo *botnet, DDOS and stresser* los más comunes en la actualidad. Los medios de pago preferidos también han evolucionado, siendo el Bitcoin el preferido en la actualidad, incluso por delante del USD.

Palabras clave: IOT, botnet, ddos, acceso, venta, foros, underground, evolución, Hack Forums

Contents

Abstract vii

Resumen ix

Index xi

List of Figures 1

1 Introduction 3

1.1 Problem description . . . 3
1.2 Objectives . . . 3
1.3 State of the art . . . 4
1.3.1 Forums . . . 4

2 Design and Development 7

2.1 Data source . . . 7
2.1.1 CrimeBB . . . 7
2.1.2 Database structure . . . 8
2.2 Project Architecture . . . 9
2.2.1 Overview . . . 9
2.2.2 Architecture . . . 9
2.2.3 Testing . . . 10
2.3 Dataset generation helper software . . . 11
2.3.1 Generating pre-annotated dataset . . . 12
2.3.2 Generating *ground truth* and *full* datasets . . . 13
2.4 Datasets creation process . . . 14
2.5 Data analysis . . . 15

3 Data analysis and results 17

3.1 Data source . . . 17

3.2 Data pre-process . . . 18
3.2.1 Tokenization . . . 18
3.2.2 Term Frequency . . . 18
3.3 Model training . . . 19
3.3.1 LinearSVC . . . 19
3.3.2 Stochastic Gradient Descent (SGD) . . . 19
3.3.3 K-nearest neighbors . . . 20
3.3.4 Multinomial Naïve Bayes . . . 20
3.4 Model evaluation . . . 21
3.4.1 Metrics . . . 21
3.4.2 ROC curve . . . 21
3.4.3 Model selection . . . 22
3.5 Data Analysis . . . 22
3.5.1 Supply vs Demand . . . 22
3.5.2 Relevant tech terms . . . 23
3.5.3 Preferred payment methods . . . 23

4 Conclusions **25**

5 Future work **27**

6 Source code **29**

Bibliography **29**

List of Figures

1.1 IoT botnet for sale on Hack Forums 4

2.1 CrimeBB docker folder tree 7
2.2 CrimeBB database structure 8
2.3 Hexagonal Architecture 10
2.4 Pre-annotate process 12
2.5 Final datasets generation process 13
2.6 Final datasets generation detailed process 14
2.7 Final datasets generation detailed process 15

3.1 Ground truth data preview 18
3.2 LinearSVC confusion matrix 19
3.3 Stochastic Gradient Descent confusion matrix 20
3.4 K-nearest neighbors confusion matrix 20
3.5 Multinomial Naïve Bayes confusion matrix 20
3.6 Training metrics report 21
3.7 ROC curve 21
3.8 Labeled documents example 22
3.9 Supply vs Demand posts volume in HF market 22
3.10 Sale of access IoT devices - DDoS Tech terms 23
3.11 Preferred payment methods 24

Chapter 1

Introduction

1.1 Problem description

The use of IoT devices is growing rapidly due to their low price, connectivity capabilities, and variety of sensors and applications.
A large number of brands and manufacturers use electronics and software from OEM suppliers which are, sometimes, vulnerable to known attacks and exploits. These known vulnerabilities makes that kind of devices perfect candidates for being infected, controlled and added as part of botnets, commonly used to perform DDoS attacks [16].

Online hacker forums have been, since their inception, meeting places in which to share information. When it comes to DDOS attacks, they can be a valuable source of information on how to protect. But, as stated by Yue et al. [18], they have often also been a source of information on how to carry out attacks.

Given the great potential for harm of these botnets and the increasing number of DDoS attacks that currently occurs, it is important to understand how this information is shared, and the economics behind it. That knowledge could be valuable in order to build *prevent & protect* strategies.

1.2 Objectives

Current work main purpose is know how the sale of accecss to IoT devices for DDoS has evolved over time. To do so, the key points to be analyzed are:

- Identifying main underground forum sites on which IoT access are being sold.
- Know the importance of these services vs other type of criminal offered services.

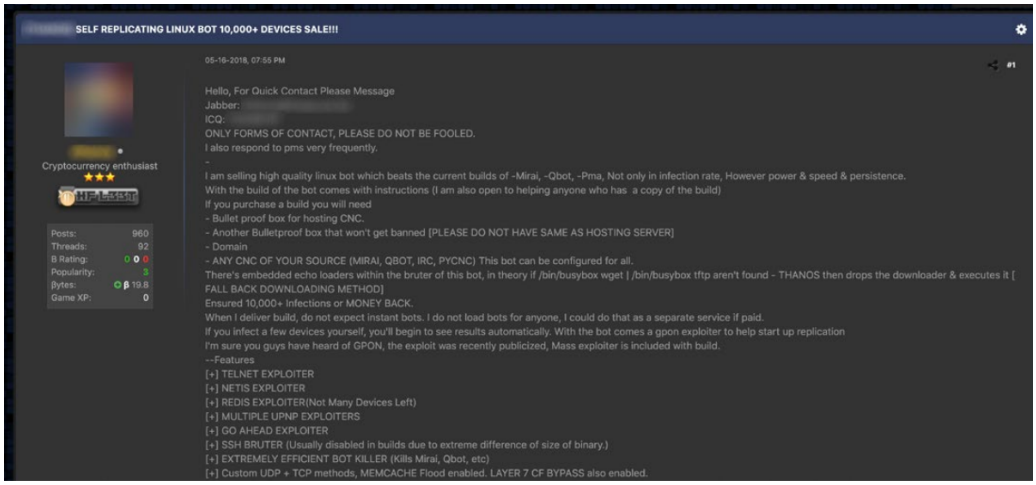

Figure 1.1: IoT botnet for sale on Hack Forums

- Measure how much underground communities are interested in these services in terms of supply vs demand.

- Getting details about how these services are monetized.

1.3 State of the art

There are works focused on measuring how *Cybercrime-as-a-Service* (CaaS) are sold in underground forums [1]. These works are a valuable source in order to understand how CaaS are sold in underground forums, but doesn't focus on sale of IoT devices access for DDoS.
Also there are some studies regarding *Cybercrime-as-a-Service* (CaaS) focused on Sale of access IoT devices for DDoS, main topic of this work. But these works are focused in other channels like Telegram and Discord [2], not just for underground forums.

Some private security companies, have analysed underground communities around the world in order to get a view on how IoT devices are being used for criminal purposes. Hilt et al. in their analysis about "IoT in the Cybercrime Underground" [8] (Trend Micro research), stated that "nation-states and more dangerous threat actors are also infecting IoT devices to use them as DoS platforms". In that article they shows some good examples as the sale of an IoT botnet in Hack Forums [9].

1.3.1 Forums

First it is needed to identify most relevant underground forums. As stated by work done by Akyazi, U., van Eeten, M. J. G., & Hernandez Ganan, C [1], based on *CrimeBB* [17] database

from Cambridge Cybercrime Centre, most relevant underground forums (in order of relevance) are:

Forum	Language	Members	Threads	Posts	Oldest
Hackforums	EN	573925	3856143	40196641	01/2007
Multiplayer Game Hacking	EN	452186	739527	8907938	12/2005
Antichat	RU	77865	242408	2449221	05/2002
RaidForums	EN	43278	33100	124776	03/2015
Offensive Community	EN	10593	18436	58779	06/2012
SafeSkyHacks	EN	7378	12892	26842	03/2013
Kernelmode	EN	1441	3144	25024	03/2010
Garage4Hackers	EN	872	2096	7697	07/2010
Stresserforums	EN	764	708	7069	04/2017
Greysee	EN	440	1239	6969	06/2015

The information collected in crimeBB, with the activity of the underground forums, is a vast data set with millions of records. Given the impossibility of covering all that information in this work, it is decided to analyze the information contained in Hack Forums, as it is the largest.

Chapter 2

Design and Development

2.1 Data source

2.1.1 CrimeBB

Data source for current project is *CrimeBB* [17] database from Cambridge Cybercrime Centre. This database is made up of information collected from the main underground forums, through web scrapping.
Content of the database has been provided in the form of PostgreSQL dumps (.sql files). In order to make it possible to handle and query DB registries, these dump files have been loaded in a dockerized PostgreSQL database engine.
Steps needed for creating and Hydrating DB are described in project's source code `README.md` file. Files needed to do these steps are under `docker/postgres_crimebb` folder.

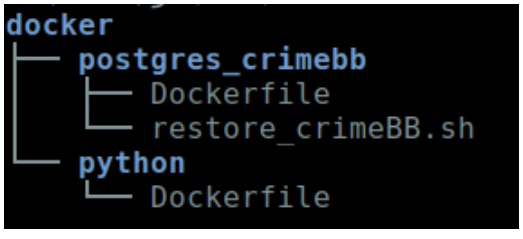

Figure 2.1: CrimeBB docker folder tree

2.1.2 Database structure

Schema shown in image 2.2 describes how data models have been defined in it's tables, data field types and relations between them.

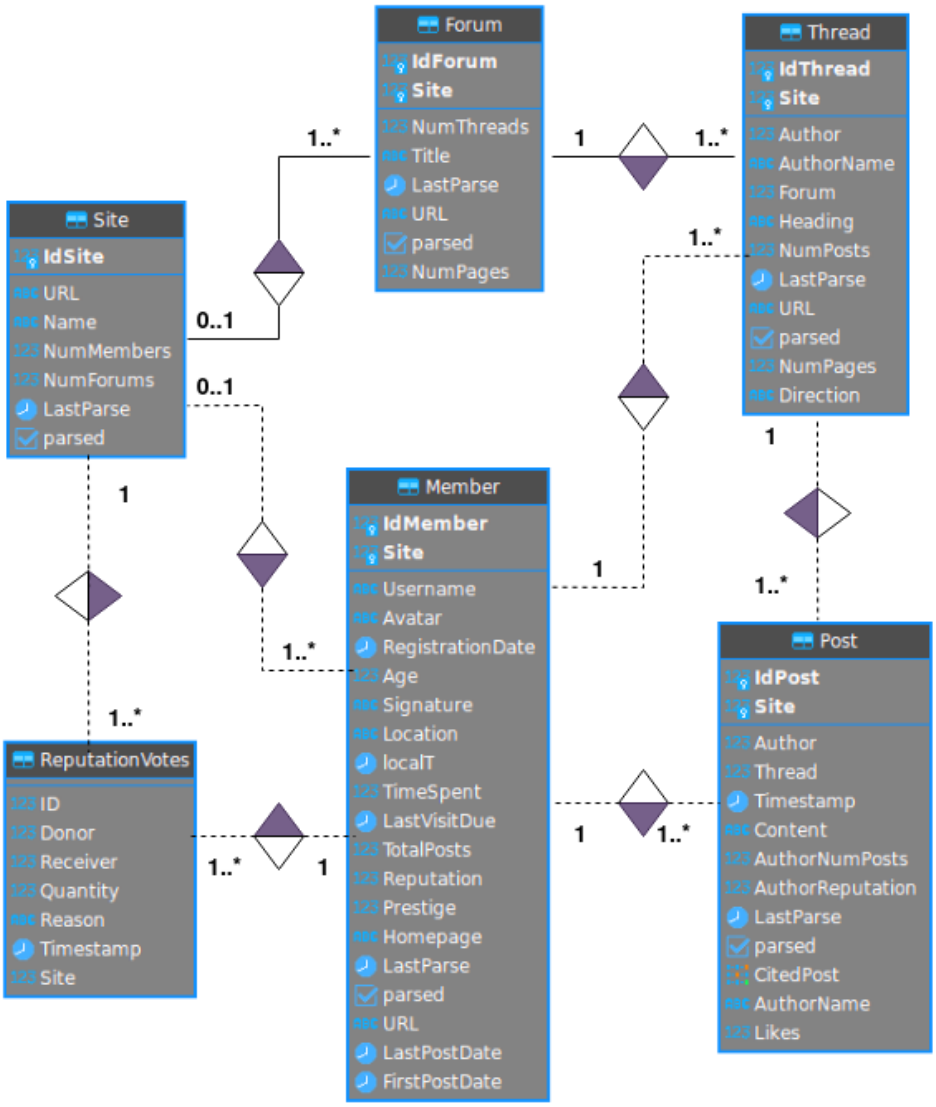

Figure 2.2: CrimeBB database structure

NOTE:
Schema shows relations between models, but provided DB data does not include foreign keys that ensure these relationships. That would be desirable in order to ensure data integrity.

2.2 Project Architecture

2.2.1 Overview

All needed project components are self contained in a folder. It includes all needed *Docker* and *Docker-compose* stuff, *Python* source code, generated datasets, *Jupyter Workbook* ad so on ... It is intended for allowing anyone to clone project's git repository and build up project in order to reproduce all results. For details on how to run it, see `README.md` in project root folder.
Just to prevent database content from being accidentally shared without permission, CrimeBB `.sql` dump files are not included in project folder. For the same reason, the container folder for the docker volume, from PostgreSQL, is also outside the project's root directory.

2.2.2 Architecture

Curent project source code architecture follows *Hexagonal Architecture principles* (see figure 2.3), also known as *Ports and Adapters*. This architecture was firstly described by Dr. Alistair Cockburn [3] and adopted by Steve Freeman, and Nat Pryce in their book *Growing Object-Oriented Software Guided by Tests* [6].
It is meant to be a flexible, changes ready way of structuring software projects. It is based in the *Clean Code* principles, described by Robert C. Martin (Uncle Bob) in his well known *The Clean Architecture* [14] article.
This architecture is divided in three main layers:

- **Infrastructure**: The outer layer. Controllers and all I/O related stuff (DB access, file readers/writers, ...). Anything that can change by an "external" cause (not by your decision), is in this layer. It includes repositories specific implementation, known as *adapters*.
- **Application**: Use cases represented by application services. In essence, these are actions launched from outside, which aim to solve use cases typical of the business logic that this project intends to support.
- **Domain**: Inner layer. Business context and rules goes here, represented by models and domain services. Repository Interfaces, known as *ports*, belongs to this layer.

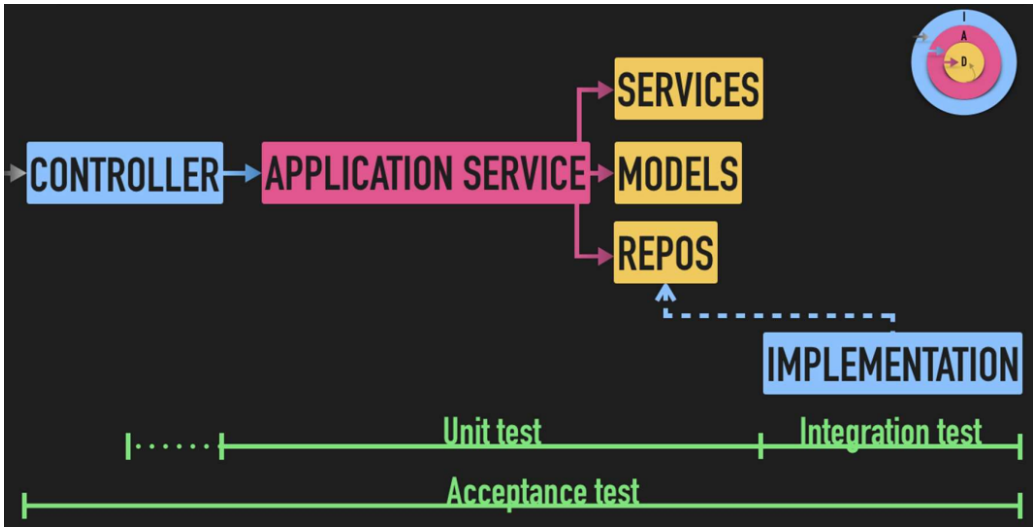

Figure 2.3: Hexagonal Architecture

This project has an important part of data exploration that makes it necessary to be able to change, in an agile way, the way to extract and process data.
Requirements could change along development process. According to that, main reasons for choosing this architecture are:

- Layers distribution is intuitive. So it eases finding software component in the place that it is figured to be.
- Testing code is mandatory in order to build good quality software, reliable and easy to maintain. This architecture makes it easy to build software using *TDD* (Test Driven Development) methodology.
- It allows to change the objectives, re-using as much code as possible, without the project structure ending up being a disaster.

2.2.3 Testing

In accordance with good practices widely accepted by the software developer community, Development has been done by following TDD (Test Driven Development) principles. TDD is now an established technique for delivering better software faster. TDD is based on a simple idea: writing tests for the code before writing the code itself.
For every layer, main component directories, contains a `tests/` folder with complete tests cases. Tests have been built with `unittest` Python library.
Running full tests battery is as easy as running the following command inside `python` docker container or within a Python virtual environment that meets `requirements.txt` packages:

```
python -m unittest -v
```

2.3 Dataset generation helper software

Ground truth and *full* datsets will contain *HackForums* data. For this project, relevant data is placed in sub-forums under *Market* section (the palce were services are supplied and demanded). So, first step for building our datasets is to extract relevant data from Threads in these sub-forums. To do so, a Python tool has been built in order to connect to PostgreSQL DB and extract Threads and Posts from these subforums.
Once relevant information has been extracted from these Threads, DDoS related information is filtered and pre-annotated.

Resulting data is saved in a CSV dataset, that will be the seed of final datasets (see subsection 2.3.1). So, this pre-annotated datset,will be the starting point for the manual work necessary to prepare the *supply* and *demand* datasets that will be used to form the *ground truth* and the *full* datasets.
Both datasets, *supply.csv* and *demand.csv*, are the result of manual work. These datsets are then mixed and used for building *ground truth* and *full* (see subsection 2.3.2). The former will be used for training and evaluating classification models and the latter is meant to be annotated by the selected model. That annotated full dataset will finally be used for data analysis in Jupyter Workbook.

NOTE: Manual work needed for creating *supply.csv* and *demand.csv* datasets is explained in section 2.4.

2.3.1 Generating pre-annotated dataset

For generating pre-annotated dataset just execute Python tool entrypoint main file:

```
python main.py annotate
```

Resulting dataset CSV file is saved in project folder `datasets/ddos_auto_annotated_dataset.csv`
The process can be summarized as shown in figure 2.4.

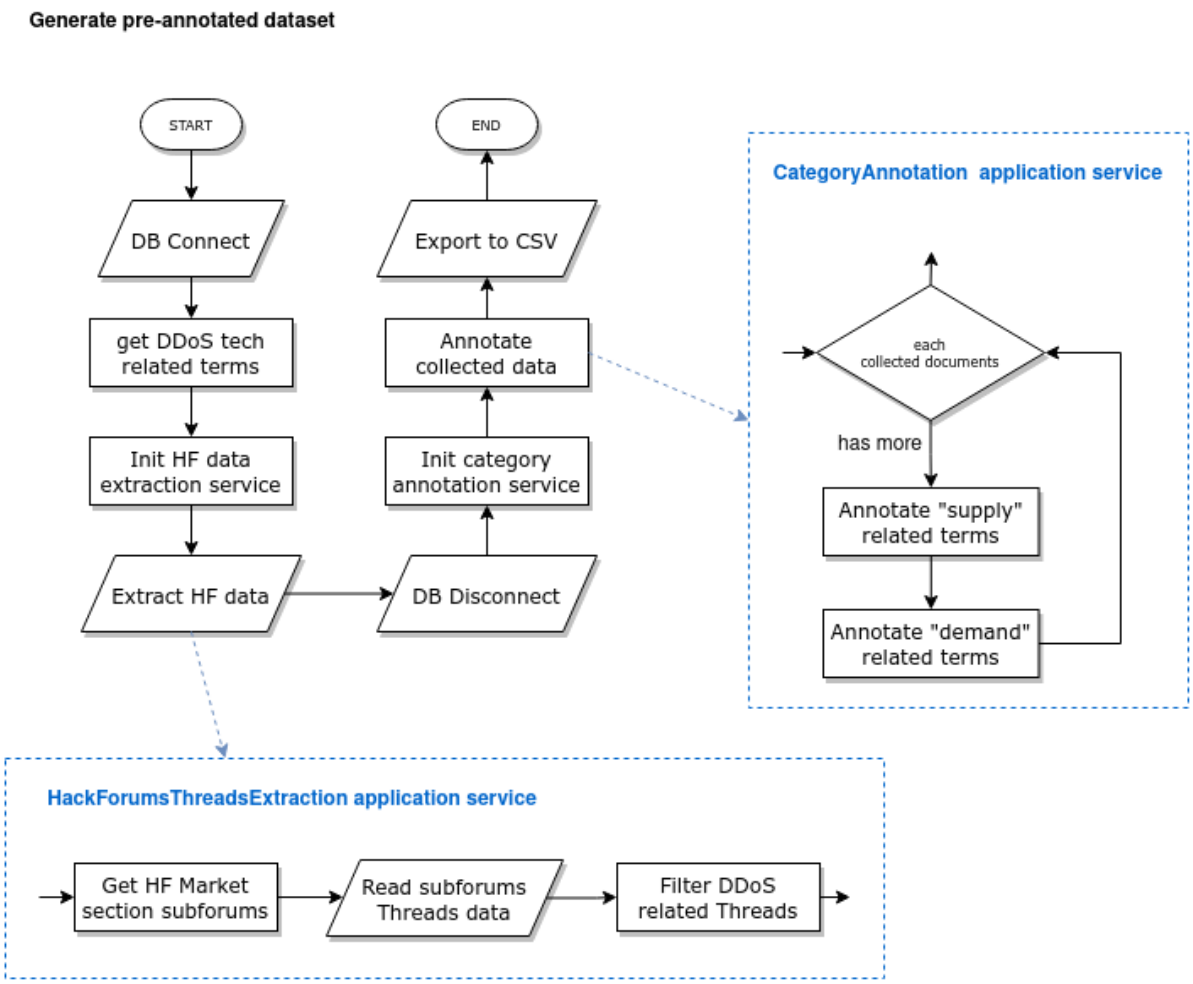

Figure 2.4: Pre-annotate process

2.3.2 Generating *ground truth* and *full* datasets

The result of manual work is *supply.csv* and *demand.csv* datasets. These datasets are then mixed to obtain *ground truth* and *full* datasets. To do that, Python tool is used as follows:

```
python main.py generate_datasets
```

In addition to the two above, two other datasets will be generated: *market_section_posts_count_dataset.cs* and *ddos_posts_count_dataset.csv*. The former is a resume of HF market section posts count, grouped by month. The latter is the same but only DDoS related posts are included. These two datasets are intended to be used, later, in Jupyer Workbook. Process overview is shown in 2.5

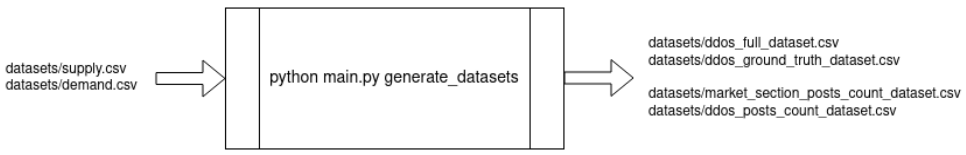

Figure 2.5: Final datasets generation process

Figure 2.6 shows diagram flow on how all those four datasets are generated.

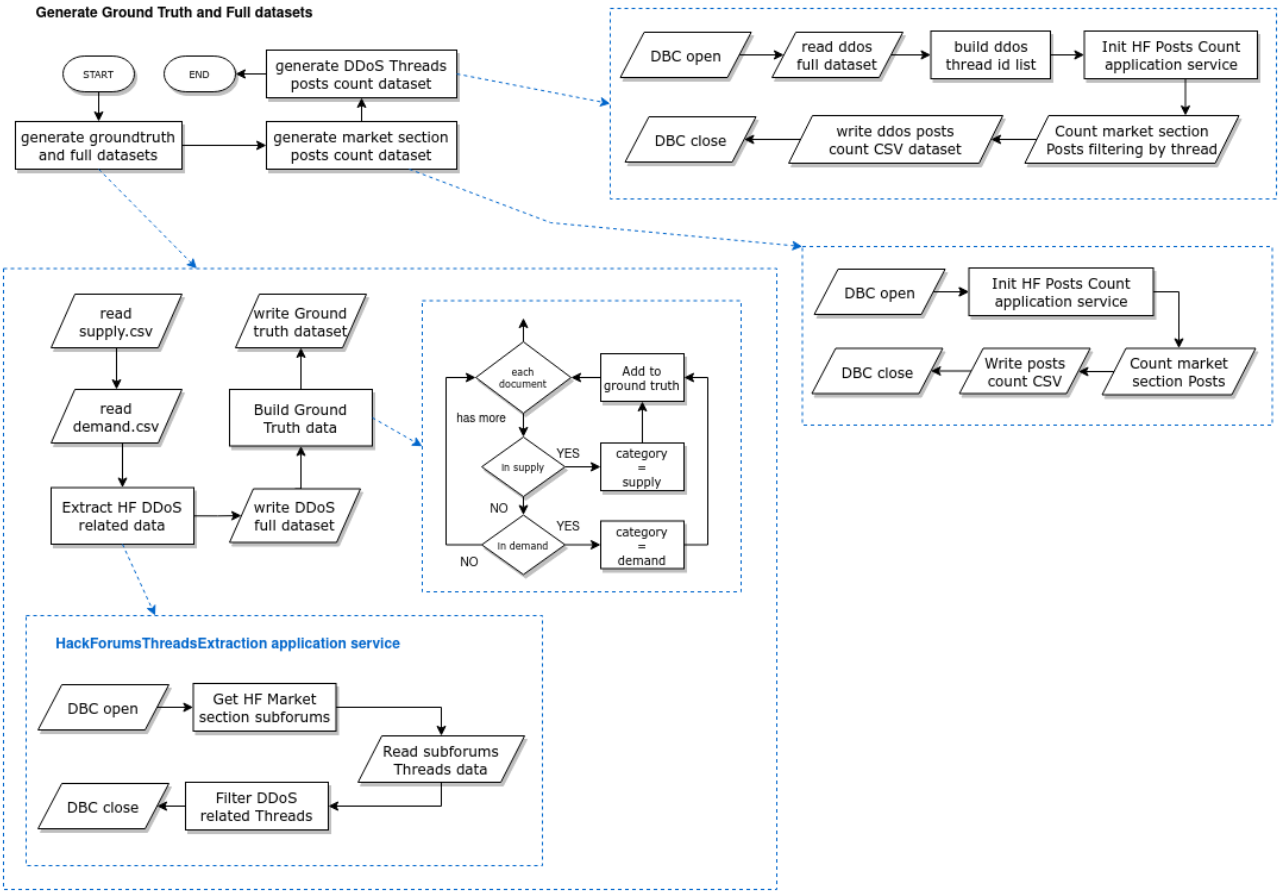

Figure 2.6: Final datasets generation detailed process

2.4 Datasets creation process

Datasets creation process consists in a mix of automated tasks and a lot of manual work. Automated processes have been covered in section 2.3. This section shows the full process. Diagram 2.7 describes the workflow from when we generated the first pre-classified dataset, through manual refining and classification, to the construction of the final datasets.

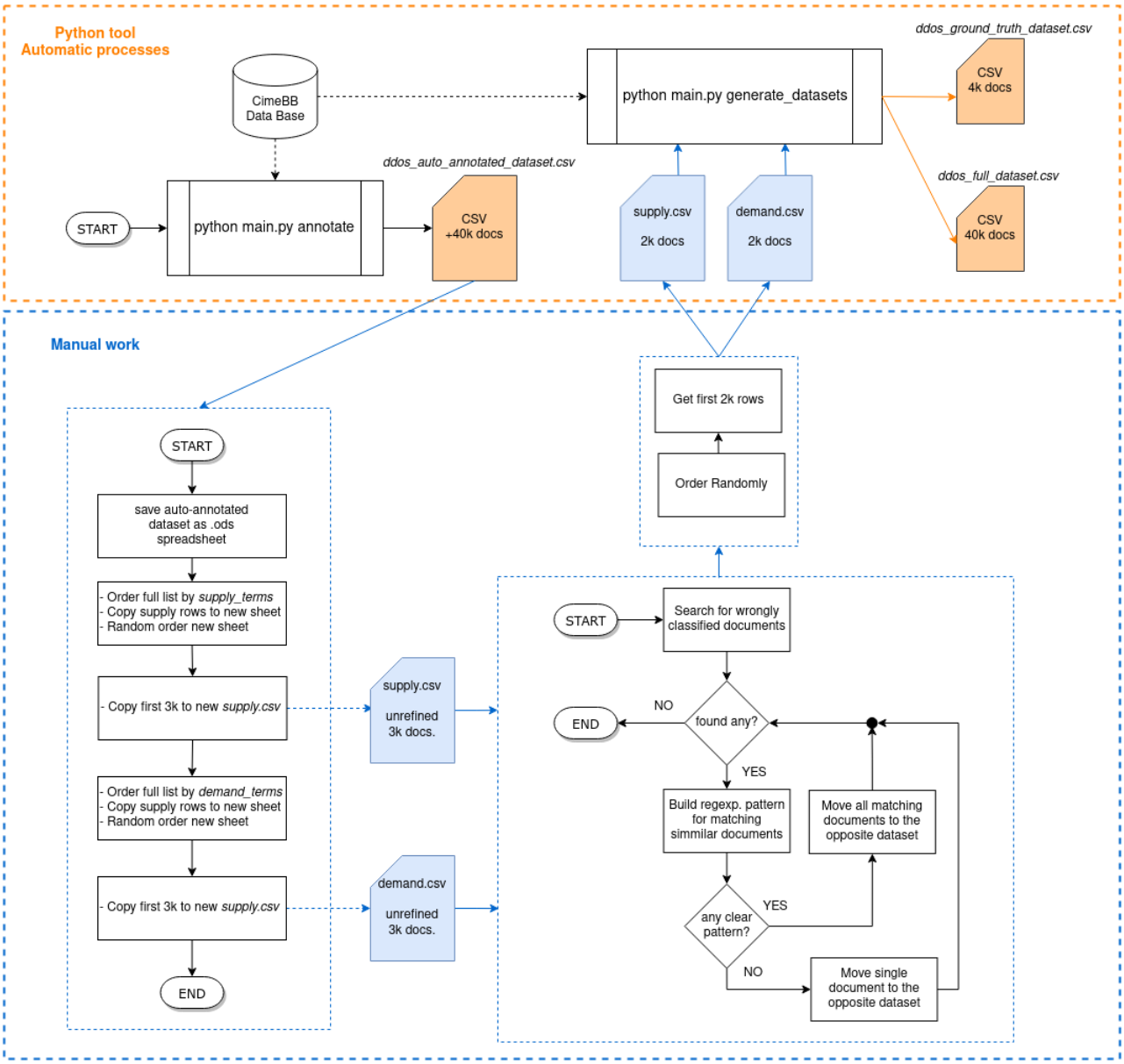

Figure 2.7: Final datasets generation detailed process

2.5 Data analysis

Final purpose of development phase is to analyse cimeBB DB data in order to know how *Sale of access IoT devices* for DDoS is traded in underground forums. Once annotated dataset and a full dataset have been built, data analysis tasks can be performed.

Jupyter Workbook has been selected as the tool for this final stage. It allows to write Python code and importing external libraries in order to load and analyse data and graph results.

Despite data analysis and results obteined in Jupyter workbook are explained in chapter 3, this section describes main aspects of it.
Workbook sections are:

- **Data source**: It loads datasets in pandas dataframes [15] and shows its structures.
- **Data pre-process**: Data clean, Tokenization and calculate *tf and tf-idf* (Term Frequency times Inverse Document Frequency).
- **Model training**: Train some well known model types, commonly used in classification tasks (Linear Support Vector Classification, Stochastic Gradient Descent, K-nearest Neighbor, Multinomial Naïve Bayes).
- **Model evaluation**: Calculate *accuracy*, *precision*, *recall* and *F1 score*, and graph ROC curve in order to select best performing model.
- **Data Analysis**: In order to know how *Sale of access IoT devices* has evolved in last years, we graph timeseries for main topics, segmenting by *supply* and *demand*. These main topics are: Supply *vs* demand ratio, most used DD0S related tech terms, payment methods.

Chapter 3

Data analysis and results

3.1 Data source

Given all threads from Hack Forums Market section, each document is a mix of data: Thread heading and First post content and timestamp.
All threads from sub-forums in Market section have been pre-processed. Two datasets have been built by filtering all DDoS related threads:

- *ddos_full_dataset.csv*: (>40K docs, full dataset with all market section documents related with IoT DDoS.

- *ddos_groundtruth_dataset.csv*: (4K docs), data sample. Randomly extracted from full dataset. Every doc category has been annotated by differenciating beteen *supply* and *demand*. This dataset is intended for training and testing text classification models.

All contents have been cleaned when building the datasets. Cleaning operations have been:

- Lower case all text.

- Removed: punctuation, accents, non-textual content like citing, images, urls, ...

Ground truth dataset preview can be shown in figure 3.1.

	site	thread	post	content	tstamp	matching_terms	category	target
0	0	16115	93673	selling botnet zombies i'm selling a botnet 15...	2008-05-03 12:51:00+00:00	['botnet']	supply	0
1	0	67711	593690	i want to buy a botnet i am paying 0 20$ for e...	2009-04-16 02:19:00+00:00	['botnet']	demand	1
2	0	68382	600494	i want to buy a botnet i want to buy a botnet ...	2009-04-18 01:48:00+00:00	['botnet']	demand	1
3	0	81356	725485	selling fud copies of bandook rat 6 months upd...	2009-05-27 09:35:00+00:00	['ddos']	supply	0
4	0	82357	735031	selling up to 300 bots $0 50 each hello i am s...	2009-05-30 05:07:00+00:00	['botnet']	supply	0

Figure 3.1: Ground truth data preview

Ground truth dataset has been split in two datasets for training and testing with 80/20 ratio. Resulting datasets have 3257 documents and 814 documents respectively.

3.2 Data pre-process

Since there are a lot of different words in corpus, result is a high dimensional dataset (too much different features). Building feature vectors with full word list is inefficient and can be unmanageable. This is why only non-zero parts of feature vectors are maintained in memory during analysis process. Main purpose of data pre-process phase is to reduce dataset dimensions. To do that, *Tokenization* and *Term Frequency calculations* techniques have been applied to dataset.

3.2.1 Tokenization

It consists in breaking down text in term vectors and filtering stop words. *CountVectorizer* [4] sklearn text feature extraction tool has been used for that purpose. It converts a collection of text documents (ground truth dataset *content* column) to a matrix of token counts. This implementation produces a sparse representation of the counts.

3.2.2 Term Frequency

Some thread contents are longer than others, it can result in a higher average count values than shorter ones, but they really talk about same topic (category). Among that, some words appears in many documents in the corpus. These common words are less informative, so downscaling weights for these words is needed. In order to avoid this problem, a calculation tf and tf-idf (*Term Frequency times Inverse Document Frequency*) has been done. Both (tf and tf-idf) are computed by TfidfTransformer [5], a tool from sklearn text feature extraction.

3.3 Model training

Accuracy has been tested for four well-known classification models. These models are widely used in classification tasks:

- LinearSVC: Linear Support Vector Classification [10].
- SGD: Stochastic Gradient Descent [11].
- K nearest Neighbors [13].
- Multinomial Naïve Bayes [12].

Performance metrics to be analysed are: precision, recall, F1-score and occurrences of each class (support). Confusion matrix will show, for each model, ratio of correct and wrong classification between categories (only two categories in our case).

3.3.1 LinearSVC

LinearSVC classification model accuracy has been 0.96. Figure 3.2 shows its confusion matrix.

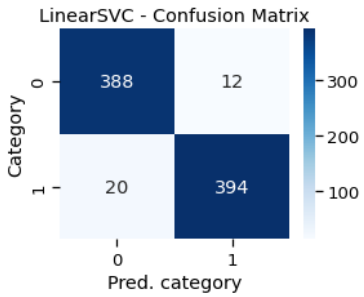

Figure 3.2: LinearSVC confusion matrix

3.3.2 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent classification model accuracy has been 0.93. Figure 3.3 shows its confusion matrix.

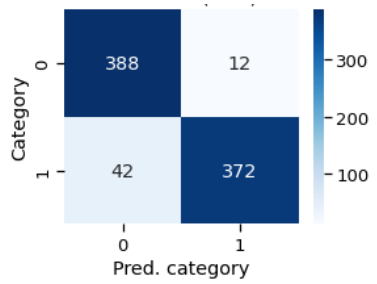

Figure 3.3: Stochastic Gradient Descent confusion matrix

3.3.3 K-nearest neighbors

K-nearest Neighbors classification model accuracy has been 0.79. Figure 3.4 shows its confusion matrix.

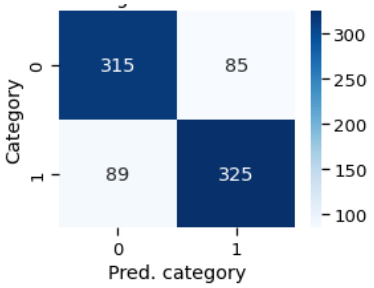

Figure 3.4: K-nearest neighbors confusion matrix

3.3.4 Multinomial Naïve Bayes

Multinomial Naïve Bayes classification model accuracy has been 0.89. Figure 3.5 shows its confusion matrix.

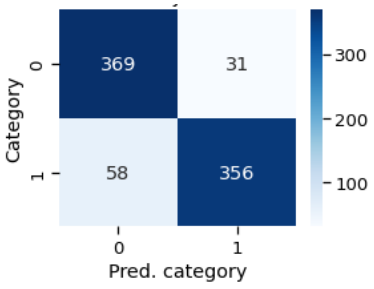

Figure 3.5: Multinomial Naïve Bayes confusion matrix

3.4 Model evaluation

3.4.1 Metrics

As stated in section *Model training* 3.3, performance metrics to be analysed are: precision, recall, F1-score and occurrences of each class (support). Results obtained in training stage are shown in figure 3.6.

	model	accuracy	precision	recall	F1 score
0	Linear Support Vector Classification.	0.960688	0.960712	0.960845	0.960686
1	Stochastic Gradient Descent	0.933661	0.935538	0.934275	0.933635
3	Multinomial Naïve Bayes	0.890663	0.892033	0.891202	0.890635
2	K-nearest Neighbor	0.786241	0.786193	0.786262	0.786209

Figure 3.6: Training metrics report

3.4.2 ROC curve

The receiver operating characteristic (ROC) curve [7] is frequently used for evaluating the performance of binary classification algorithms. It provides a graphical representation of a classifier's performance, rather than a single value like most other metrics.

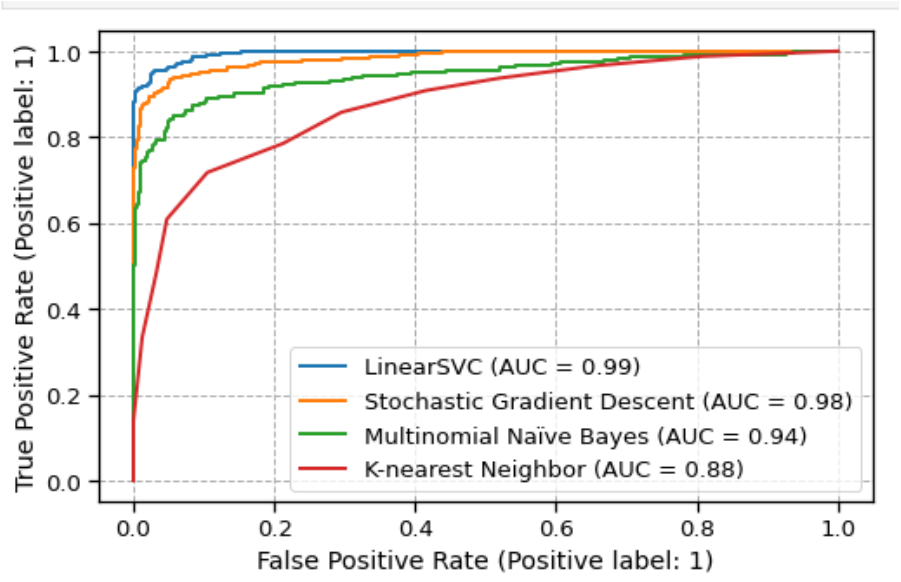

Figure 3.7: ROC curve

3.4.3 Model selection

According to evaluation results, best performing model is LinearSVC (Linear Support Vector Classifier) 3.3.1. Area under the curve in ROC curve figure 3.7 and training metric values are close but better than Stochastic Gradient Descent model. The rest of the models clearly have worse results.

3.5 Data Analysis

After selecting a prediction model, full dataset has been loaded and its documents have been labelled according to selected model prediciton results. An example on how documents have been labeled can is shown in figure 3.8

	site	thread	post	content	tstamp	matching_terms	category	category_name
0	0	11481	53361	buying botnet im buying a good botnet with alo...	2008-02-18 08:37:00+00:00	['botnet']	1	demand
1	0	12919	61516	botnet with undetected server 4sale hey people...	2008-03-21 08:47:00+00:00	['botnet']	0	supply
2	0	13901	70758	interested in purchasing a botnet im intereste...	2008-04-09 04:41:00+00:00	['botnet', 'ddos']	1	demand
3	0	14432	77296	trojans botnet source keyloggers packers crypt...	2008-04-15 07:57:00+00:00	['botnet']	0	supply
4	0	15291	85997	professional bulletproof hosting service from ...	2008-04-24 12:11:00+00:00	['botnet']	0	supply

Figure 3.8: Labeled documents example

3.5.1 Supply vs Demand

Analysing evolution of supply and demand over time allows to know how community members interest in that type of services have evolved of underground. It is the first step in analysing relevance of Sale of IoT devices for DDoS services in underground forums.

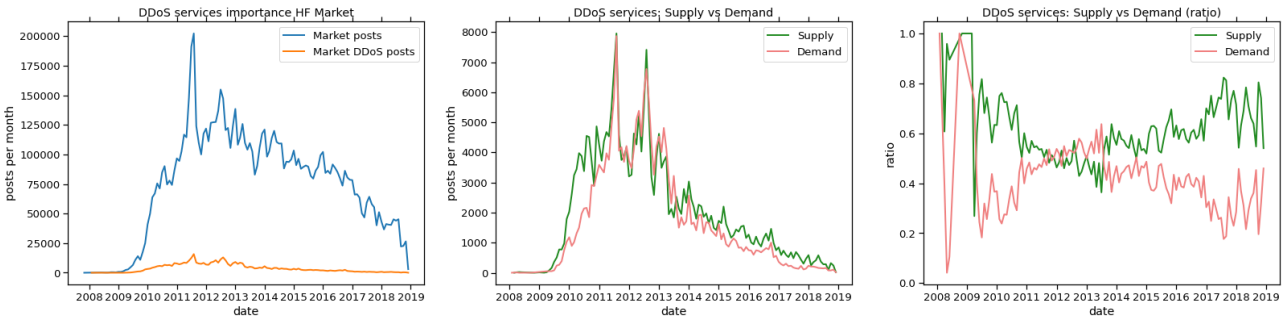

Figure 3.9: Supply vs Demand posts volume in HF market

As shown in fig. 3.9, supply did pass demand in 2012 and nowadays is clearly higher. However, according to results shown in fig. 3.10 depending on the intention of the service, the demand may be higher than the supply.
This is happening today with *stresser* related services. *Stressers* are DDoS systems controlled by a professional operator. They are used to test the resistance of network systems against DDoS attacks.

3.5.2 Relevant tech terms

It is common to refer to *Sale of access IoT devices services for DDoS* by using different tech terms like booter, stresser, ddoser, and so on. The use of some terms or others, when referring to these services, varies depending on the technology used to develop them, the intended use or simply the slang of the moment.

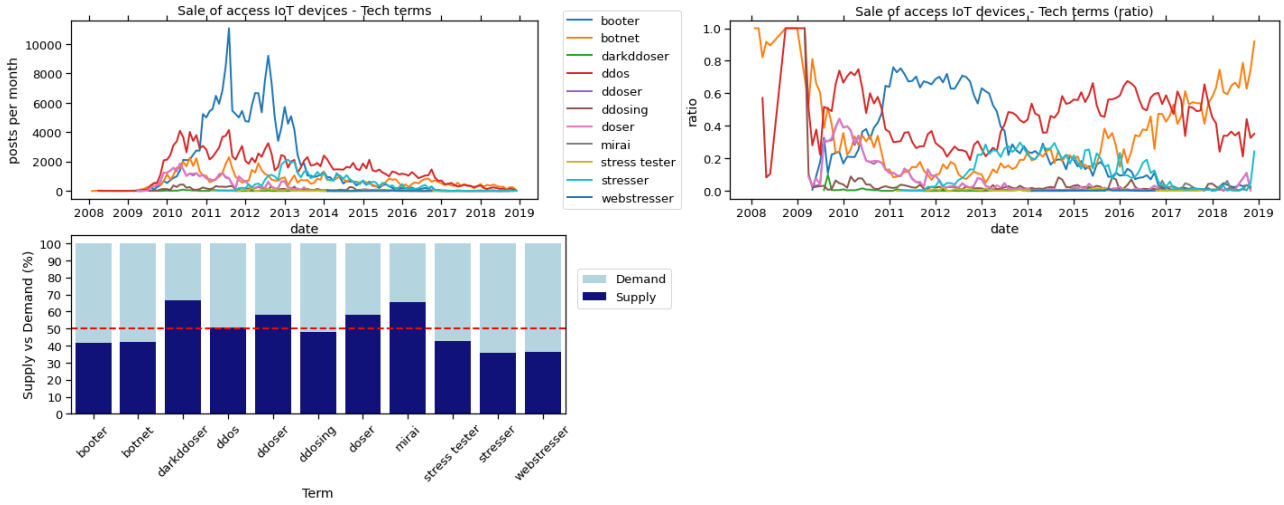

Figure 3.10: Sale of access IoT devices - DDoS Tech terms

3.5.3 Preferred payment methods

Fig. 3.11 shows how preferred payment methods, used for the Sale of access IoT devices, have evolved in last years. It is a key point in order to understand economics behind that type of criminal services.

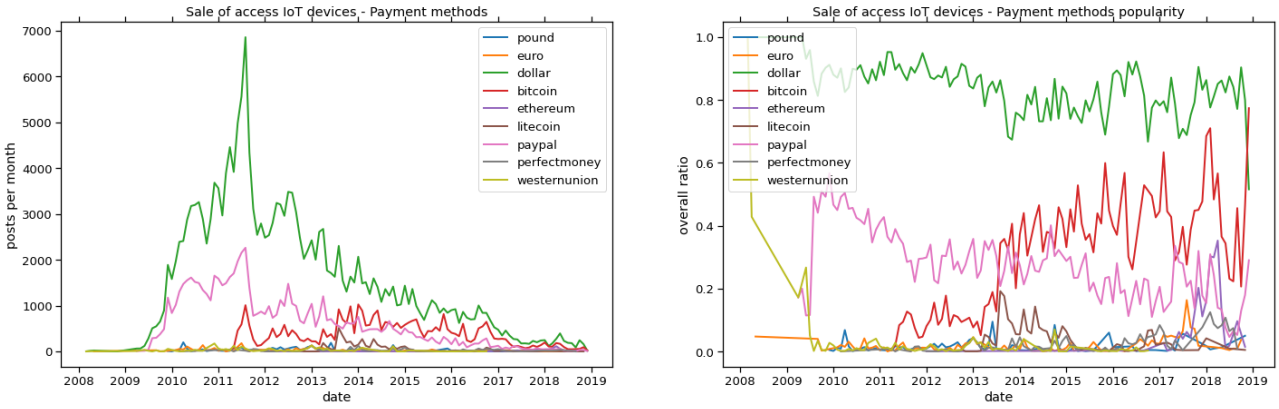

Figure 3.11: Preferred payment methods

The use of bitcoin as a preferred payment method has grown strongly in recent years.

Chapter 4

Conclusions

Data obtained in *Data Analysis* section 3.5 reveals that underground forums as HackForums are clearly lossing market activity over last years. Sale of access IoT devices services for DDoS has been a very important topic in the market section, but the current trend is downward and it is no longer a topic of great relevance.

However, the data obtained in this study do not allow us to affirm that the sale of this type of services has lost importance. Other forms of communication such as Telegram or Discord channels or other social networks could be gaining users to the detriment of this type of web forums. It could be due to its higher level of anonymity, ease of use and instant messaging benefits.

Regarding supply and demand, it seems that the interest of potential buyers has been declining and there is more activity in the supply than interest in buying. Although as an exception, it seems that services with a stress-as-a-service approach currently continue to have more demand than supply. Which suggests that there is a niche of professional buyers interested in hiring criminal services in order to test the resilience of their network systems.

Preferred payment methods have evolved significantly. The use of cryptocurrencies, more specifically Bitcoin, as a preferred payment method has been growing, being clearly more used nowadays. As we discussed earlier regarding forms of communication, the adoption of cryptocurrencies could be due to their higher level of anonymity, ease of use, and availability around the world.

Chapter 5

Future work

AS stated in 1.3.1 section, underground forums activity collected in crimeBB database is a vast amount of information. A possible future line of work is to analyze the information contained in the rest of the underground forums and analyze the results by comparing them.

In the present study it has not been possible to analyze the evolution of prices regarding sale of access IoT devices services for DDoS. Analising price evolution in last years could be an interesting line of future work.

Chapter 6

Source code

Source code developed for current study is available, under GNU General Public License v3.0, in author's public Github repository:
https://github.com/serfer2/tfm

The purpose of this repository is to allow any interested person to reproduce the results obtained.
Source code is divided in four main areas:

- Infrastructure: It includes all resources needed in order to port and deploy project infrastructure, over docker containers, and for hydrating database.
- Data extraction and preprocess: It is Python helper software meant to assist in datasets creation.
- Data analysis: Jupyter Workbook used for models trainning, data analysis and building graphs.
- Documentation: LaTeX sources for current doc.

Bibliography

[1] U. Akyazi, M. J. G. van Eeten, and C. Hernandez Ganan. Measuring cybercrime as a service (caas) offerings in a cybercrime forum. Technical report, Delft University of Technology, 2021. Paper presented at Workshop on the Economics of Information Security.

[2] Sifer Aseph, Ethan Friedman, Maxwell Aliapoulios, Rasika Bhalerao, Tobias Lauinger, and Damon McCoy. An exploration of cybercriminal discussions of iot-based ddos services. Technical Report FA8750-19-2-0009, New York University, New York, NY, USA, 2021.

[3] Alistair Cockburn. Dr. Alistair Cockburn.

[4] David Cournapeau. CountVectorizer, from Python scikit-learn library.

[5] David Cournapeau. TfidfTransformer, from Python scikit-learn library.

[6] Steve Freeman and Nat Pryce. *Growing Object-Oriented Software Guided by Tests*. Addison Wesley, London, 2009.

[7] Luzia Gonçalves, Ana Subtil, M Rosário Oliveira, and P d Bermudez. Roc curve estimation: An overview. *REVSTAT–Statistical Journal*, 12(1):1–20, 2014.

[8] Stephen Hilt, Vladimir Kropotov, Fernando Mercês, Mayra Rosario, and David Sancho. The internet of things in the cybercrime underground. *Trend Micro Research*, pages 44–44, 2019.

[9] Stephen Hilt, Vladimir Kropotov, Fernando Mercês, Mayra Rosario, and David Sancho. The internet of things in the cybercrime underground. *Trend Micro Research*, pages 27–27, 2019.

[10] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[11] Nikhil Ketkar. Stochastic gradient descent. In *Deep learning with Python*, pages 113–132. Springer, 2017.

[12] Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. In *Australasian Joint Conference on Artificial Intelligence*, pages 488–499. Springer, 2004.

[13] Laszlo Kozma. k nearest neighbors algorithm (knn). *Helsinki University of Technology*, 2008.

[14] Robert C. Martin. Robert C. Martin (Uncle Bob). The Clean Architecture.

[15] Wes McKinney. From Python pandas library. Dataframe is Two-dimensional, size-mutable, potentially heterogeneous tabular data.

[16] Dragan Peraković, Marko Periša, and Ivan Cvitić. Analysis of the iot impact on volume of ddos attacks. *XXXIII Simpozijum o novim tehnologijama u poštanskom i telekomunikacionom saobraćaju–PosTel*, 2015:295–304, 2015.

[17] Alice Hutchings Sergio Pastrana, Daniel R. Thomas and Richard Clayton. Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In Proceedings of The Web Conference 2018 (WWW 2018), Lyon, France. ACM, New York, NY, USA, 10 pages.

[18] Wei T Yue, Qiu-Hong Wang, and Kai-Lung Hui. See no evil, hear no evil? dissecting the impact of online hacker forums. *Mis Quarterly*, 43(1):73, 2019.