# Guió TFG

## Inici

Bon dia, us presentaré el meu Treball de Final de Grau: Anàlisi automàtica de tiquets de phishing mitjançant el processament del llenguatge natural. Aquest projecte s’ha realitzat dins d’un Conveni de Cooperació Educativa com a part d’un projecte dut a terme per inLab FIB.

## Índex

Parlarem sobre:

* La contextualització i els objectius del projecte
* S’explicarà una mica alguns conceptes rellevants
* La part principal serà el desenvolupament de les parts del sistema
* Després es mostrarà l’avaluació final dels models entrenats
* I finalment, es mostrarà el resultat final amb les conclusions del projecte

## Contextualització

### Estadístiques

L'apogeu de la digitalització i la tecnologia ha introduït una comoditat i una connectivitat sense precedents. Però juntament amb aquests avenços, s'ha vist un augment paral·lel de les ciberamenaces i el frau digital. Les estadístiques revelen que les estafes informàtiques denunciades es disparen a més de 300.000 incidents al país només el 2022.

Aquest augment de la ciberdelinqüència no és una anomalia estadística, sinó més aviat un reflex de l'evolució del panorama tecnològic i les seves vulnerabilitats. Les estafes informàtiques han augmentat un 370% en només sis anys, cosa que subratlla la urgència de la nostra resposta a aquestes amenaces.

Les repercussions van més enllà de les pèrdues econòmiques. La privadesa i la seguretat de les persones es veuen en perill constant, sobretot quan es té en compte les dades confidencials que es confien en empreses que sofreixen els mateixos atacs.

A mesura que ens enfrontem a aquestes creixents amenaces, es fa imperatiu desenvolupar estratègies proactives per mitigar els riscos i protegir-se contra incidents futurs. Aquest projecte pretén abordar aquesta necessitat analitzant els incidents de phishing i extraient informació clau per millorar la preparació i evitar que es repeteixin.

### Objectiu

L'objectiu principal és dissenyar i desplegar un sistema capaç d’extreure certs camps dels tiquets, cosa que permetrà lluitar de manera proactiva contra les ciberamenaces.

Per aconseguir aquest objectiu general, s'ha marcat els següents objectius:

* Realitzar un estudi de l'estat de l'art investigant les solucions existents per a reptes similars.
* Configurar la base de dades i desenvolupar una eina de descàrrega de tiquets.
* Preprocessar automàticament el text que elimini tota la informació redundant possible i afegeixi al text amb dades rellevants.
* Entrenar un model del processament del llenguatge natural.
* Anonimitzar les dades de sortida, salvaguardant la informació confidencial.
* Configurar i gestionar l'emmagatzematge de les dades preprocessades i anonimitzades en la base de dades de sortida.
* Integrant tots els elements anteriors, construir una pipeline que processi els tiquets dins de l'entorn especificat.
* Finalment, desenvolupar una API que simplifiqui l'accés i la utilització del sistema.

### Actors implicats

Per explicar resumidament els actors implicats en aquest projecte, l’Agència de ciberseguretat ha encarregat a i2CAT el desenvolupament del projecte global anomenat “TDA”. i2CAT per la seva part, ha confiat en inLab FIB per a dur a terme una part anomenada “TSC” o “Ticket System Collector”.

### Cronologia

Aquest projecte ha tingut una línia temporal complexa. Primerament, es va iniciar el projecte TSC al gener de 2023, un any després, es va permetre l’accés a les dades reals amb les quals s’ha de treballar. Addicionalment, es va tardar un mes en obtenir accés a llibreries necessàries per al entrenament dels models que s’ha fet servir. Gràcies a l’ampliació del projecte, s’ha pogut redactar i explicar sobre el treball amb els tiquets reals i el desplegament de l’entorn.

## Exploració teòrica

### Anàlisi d’un tiquet

### OTRS i Elasticsearch

## Desenvolupament

### Arquitectura

### Pipeline: Extracció de dades

### Pipeline: Preprocessament

### Preprocessament: Eliminació de signatures

### Pipeline: Execució del model

### Pipeline: Anonimització

### Pipeline: Elasticsearch

### API

### Dades sintètiques: explicació

### Dades sintètiques: Models destacats

### Dades reals: explicació

### Dades reals: Models destacats

## Avaluació dels models

### Resultats dels models

### Percentatge de “NotFound”

## Conclusions