UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

ANÀLISI AUTOMÀTICA DE TIQUETS DE PHISHING I MALWARE MITJANÇANT EL PROCESSAMENT DEL LLENGUATGE NATURAL

JAUME CASALS VILAPLANA

Director/a: ALBERT OBIOLS VIVES (UNIVERSITAT POLITÈCNICA DE CATALUNYA)

Ponent: ERNEST TENIENTE LOPEZ (Departament d'Enginyeria de Serveis i Sistemes d'Informació)

Titulació: Grau en Enginyeria Informàtica (Computació)

Memòria del treball de fi de grau

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Resum

Aquest informe descriu un projecte desenvolupat per millorar la gestió d'incidents en una empresa de ciberseguretat mitjançant l'ús d'un model de Processament del Llenguatge Natural (NLP). L'objectiu principal és l'extracció de certs camps dels tiquets d'incidents, completant un sistema automàtic per la millor experiència d'usuari. El projecte aborda els reptes del volum creixent i complexitat dels incidents de ciberseguretat, utilitzant un model de NLP que ha estat perfeccionat per comprendre els matisos lingüístics i contextuals dels textos d'entrada, tenint en compte les limitacions tècniques presents. La solució implementada inclou un procés complet de l'extracció de la informació, que inclou la recuperació dels tiquets, les tècniques de preprocessament adaptades i l'extracció de característiques clau. S'emfatitza la seguretat i la confidencialitat de la informació delicada durant tot el procés. La implementació d'aquest sistema no només millora l'eficiència operativa, sinó que també contribueix a millorar la resposta a les amenaces cibernètiques en constant evolució i enforteix la postura de ciberseguretat de l'empresa i la comunitat en general.

Resumen

Este informe describe un proyecto desarrollado para mejorar la gestión de incidentes en una empresa de ciberseguridad mediante el uso de un modelo de Procesamiento del Lenguaje Natural (NLP). El objetivo principal es la extracción de ciertos campos de los tiques de incidentes, completando un sistema automático para la mejor experiencia de usuario. El proyecto aborda los retos del creciente volumen y complejidad de los incidentes de ciberseguridad, utilizando un modelo de NLP que ha sido perfeccionado para comprender los matices lingüísticos y contextuales de los textos de entrada, teniendo en cuenta las limitaciones técnicas presentes. La solución implementada incluye un proceso completo de la extracción de la información, que incluye la recuperación de los tiques, las técnicas de preprocesamiento adaptadas y la extracción de características clave. Se enfatiza la seguridad y confidencialidad de la información delicada durante todo el proceso. La implementación de este sistema no solo mejora la eficiencia operativa, sino que también contribuye a mejorar la respuesta a las amenazas cibernéticas en constante evolución y fortalece la postura de ciberseguridad de la empresa y la comunidad en general.

Abstract

This report describes a project developed to improve incident management in a cybersecurity company by using a Natural Language Processing (NLP) model. The main objective is the extraction of certain fields from incident tickets, completing an automatic system for the best user experience. The project addresses the challenges of the increasing volume and complexity of cybersecurity incidents, using an NLP model that has been refined to understand the linguistic and contextual nuances of the input texts, taking into account the technical limitations present. The implemented solution includes a complete process of information extraction, including ticket retrieval, adapted preprocessing techniques and the extraction of certain fields. The security and confidentiality of sensitive information are emphasized throughout the process. The implementation of this system not only improves operational efficiency, but also contributes to an improved response to evolving cyber threats and strengthens the cybersecurity posture of the company and the community as a whole.

Índex

Índex de taules

Índex de figures

Capítol 1

Contextualització i abast

Aquest és un treball de final de grau del Grau d'Enginyeria Informàtica que s'imparteix a la Facultat d'Informàtica de Barcelona (FIB), que forma part de la Universitat Politècnica de Catalunya (UPC). El treball actual s'ha realitzat dintre d'un Conveni de Cooperació Educativa com a part d'un projecte dut a terme pel laboratori d'innovació i recerca inLab FIB, pertanyent a la Facultat d'Informàtica de Barcelona.

1.1 Contextualització

1.1.1 Context

En una era dominada per la digitalització, en la qual la tecnologia ha revolucionat innegablement les nostres vides i les operacions empresarials, ha sorgit un adversari formidable: una onada de ciberamenaces i frau digital. En endinsar-nos en l'àmbit de la ciberseguretat, ens enfrontem a la necessitat urgent d'adoptar mesures sòlides per contrarestar les ramificacions del frau digital, que van molt més enllà de l'àmbit virtual i s'infiltren en la nostra societat.

Aquesta revolució digital ha donat lloc a un panorama cada cop més desafiador per a la ciberseguretat. A mesura que més persones i empreses depenen de plataformes i serveis en línia, els riscos i costos dels ciberatacs i fraus es disparen a nivells sense precedents. Les estafes informàtiques el 2023 (que representen quasi el 90% de tota la cibercriminalitat) presenten un increment més del 20% sobre el mateix període (de gener a juny) del 2022. Per comprendre millor encara l'evolució de la cibercriminalitat, i el seu impacte sobre el conjunt de la criminalitat, les estafes informàtiques van representar la quantitat anual de 335.995 delictes el 2022 i 70.178 al 2016. Això implica que, en només sis anys, les estafes informàtiques l'any 2022 van créixer més del 370% sobre les del 2016. [1] Aquesta tendència es pot veure el els gràfics que mostren algunes estadístiques a la secció A.

L'impacte d'aquestes ciberamenaces transcendeix les pèrdues financeres meres, impregnant les vides d'individus i organitzacions per igual. Les persones ja no són mers espectadors, sinó que es troben a primera línia d'una guerra cibernètica, en què la informació personal, abans considerada sagrada, s'ha convertit en un bé preuat per als actors maliciosos que aprofites totes les vulnerabilitats. Alhora, les empreses s'enfronten a una allau d'atacs que posen en perill no només la seva estabilitat financera, sinó que també erosionen la confiança i comprometen informació delicada.

El frau digital, més enllà de les implicacions financeres, és un repte social. Soscava la confiança dels consumidors i les empreses en l'entorn en línia, amb grans conseqüències a la privadesa, la seguretat i el benestar general. El robatori d'identitat, el phishing i l'apropiació de comptes exposen informació personal i financera sensible, comprometent comptes i transaccions en línia. La reputació i la credibilitat d'individus i organitzacions estan en joc.

Una de les formes d'engany digital més freqüents són els ciberatacs dirigits al sector sanitari. Aquest àmbit és especialment susceptible de patir atacs de ransomware que bloquegen dades i exigeixen un pagament a canvi del seu alliberament. Aquests atacs poden tenir greus conseqüències tant per als professionals sanitaris com per als pacients, com ara posar en perill la seguretat dels pacients, interrompre els serveis mèdics i comprometre informació confidencial.

Diversos factors contribueixen a fer que el sector sanitari sigui més vulnerable als ciberatacs que altres sectors. Un problema clau és la dependència de programari i sistemes antiquats dins de la sanitat. Aquests programes funcionen sovint amb sistemes operatius obsolets, que ja no reben manteniment ni actualitzacions dels seus fabricants. Aquests sistemes són susceptibles de ser vulnerats a través de falles conegudes que els permeten accedir als dispositius connectats. Un altre factor és el gran valor i importància de les dades de les organitzacions sanitàries que es conserven i manipulen. Aquestes dades inclouen historials mèdics, receptes i altra informació confidencial que els ciberdelinqüents i els competidors poden explotar per a activitats fraudulentes. La corrupció de dades per delictes cibernètics pot tenir greus conseqüències, com ara retards en el diagnòstic, el tractament i els errors de prescripció.

1.1.2 Justificació

A l'àmbit de la ciberseguretat contemporània, la gestió de tiquets d'incidents constitueix un component operatiu crític per a les agències dedicades a salvaguardar les infraestructures digitals. Reconeixent la importància primordial d'una gestió eficient dels tiquets d'incidents, una agència de ciberseguretat destacada s'ha embarcat en un projecte confidencial destinat a extreure i analitzar informació relativa als tiquets de ciberseguretat que reben i posteriorment resolen. Aquesta iniciativa, orquestrada per **l'Agència de**

ciberseguretat, va requerir la contractació d'un intermediari, **i2CAT**, per facilitar els processos d'extracció i anàlisi. Alhora, **i2CAT** va confiar l'execució d'aquesta intrincada tasca a una altra entitat, **inLab FIB**, on treballa l'autor.

Dins d'aquest marc, l'objectiu general del projecte és la recuperació, anàlisi i emmagatzematge segur de la informació delicada relacionada amb aquests tiquets de ciberseguretat. S'ha posat especial èmfasi en respectar els matisos del sistema de gestió de tiquets emprat per l'**Agència**. La finalitat d'aquesta iniciativa ha sigut millorar la capacitat d'aquesta entitat per classificar eficaçment les possibles futures incidències.

La naturalesa intrínsecament sensible del projecte ha requerit un enfocament rigorós i altament confidencial, com subratllen els múltiples acords de confidencialitat (*NDA*) que regeixen les interaccions entre totes les parts implicades: **l'Agència de ciberseguretat**, **i2CAT** i **inLab FIB**. En conseqüència, el projecte s'ha caracteritzat per unes mesures de seguretat estrictes de les dades per garantir que tota la informació relativa als tiquets romangui segura dins dels servidors de l'**Agència**.

1.1.3 Problema a resoldre

El panorama actual de la ciberseguretat està marcat per una sèrie d'amenaces digitals en evolució constant, que requereixen la millora contínua de les mesures defensives i les estratègies de resposta. Un aspecte fonamental d'aquesta postura defensiva gira al voltant de la gestió eficient dels tiquets d'incidents de ciberseguretat. Aquests tiquets serveixen per documentar i fer un seguiment dels incidents notificats, siguin correus de *phishing*, anomalies o programari maliciós. Un sistema de gestió de tiquets d'incidents ben estructurat és indispensable per permetre una ràpida resolució de les amenaces. Aquest projecte ha abordat un problema específic associat a aquesta faceta crucial de la gestió de la ciberseguretat.

L'**Agència** depèn en gran manera d'un sistema de gestió de tiquets d'incidents per gestionar i resoldre incidents de ciberseguretat. Tot i això, el sistema existent emprat per l'**Agència** ha mostrat certes deficiències que necessiten rectificació. Una d'aquestes, és l'absència d'un mecanisme per analitzar les dades contingudes als tiquets. Això ha plantejat reptes importants per a la capacitat de l'**Agència** d'obtenir informació pràctica a partir de les dades històriques dels tiquets i aplicar mesures proactives per frustrar les amenaces recurrents.

Per exemple, considerem un escenari on l'**Agència** s'ha trobat prèviament amb un sofisticat atac de *phishing* que utilitzava un mètode d'atac novell. El tiquet d'incident associat a aquest atac conté informació molt valuosa sobre el modus operandi de l'atac, el punt d'origen de l'atac, les accions de mitigació de l'atac o els usuaris afectats. L'actual sistema d'incidents no permet l'extracció i posterior anàlisi sistemàtica d'aquesta infor-

mació valuosa. En conseqüència, quan torni a sorgir un mètode similar, la capacitat de l'**Agència** per accelerar-ne la resposta i mitigar els possibles danys es veu obstaculitzada per la manca d'informació històrica.

L'objectiu principal d'aquest projecte ha sigut dissenyar i implementar un sistema d'anàlisi i extracció de la informació de tiquets d'incidents que alleugi les deficiències existents. En el projecte també s'ha inclòs el desenvolupament d'una API que permeti utilitzar el sistema de manera senzilla. S'ha implementat també una funcionalitat que permet l'arxiu segur de les dades dins dels seus propis servidors per tal de proveir una anonimització de la informació més sensible. Es preveu que aquest sistema doti l'**Agència** de la capacitat d'extreure informació de les dades històriques dels tiquets, facilitant la ràpida detecció i la resposta a amenaces recurrents i el desenvolupament de mesures preventives.

Amb el compliment d'aquest objectiu, el projecte aspira a satisfer la bretxa existent entre la notificació i l'anàlisi d'incidents, permetent així a l'**Agència** aprofitar tot el potencial de les dades de tiquets d'incidents. Aquest sistema millorat de gestió de tiquets d'incidents garanteix que les dades crítiques romanguin accessibles, confidencials i en compliment dels protocols de seguretat.

En essència, el projecte soluciona una deficiència de l'actual sistema de gestió d'incidents de l'**Agència**, facilitant l'extracció sistemàtica, l'anàlisi i l'emmagatzematge segur de les dades dels incidents, reforçant en darrer terme la capacitat de l'**Agència** per detectar, respondre i prevenir les amenaces a la ciberseguretat amb més eficàcia.

1.1.4 Actors implicats

Són actors totes aquelles parts que, o bé els seus interessos es poden veure afectats positivament o negativament pels resultats d'aquests, o bé estan implicades de forma directa en el projecte. Aquests són els següents:

- **L'agència de ciberseguretat:** D'ara endavant, l'**Agència**, la principal part interessada i beneficiària del sistema de d'anàlisi automàtica d'incidents és la mateixa agència de ciberseguretat. L'objectiu del sistema és millorar la seva eficiència operativa general, proporcionant-los una eina per a l'extracció, anàlisi i emmagatzematge segur de dades d'incidents. L'**Agència** utilitza aquest sistema com a eina vital per a una detecció, resposta i prevenció d'incidents més eficaç.
- **i2CAT:** És l'intermediari contractat per l'**Agència** per executar el projecte. És responsable que el sistema de gestió de tiquets d'incidents arribi a bon port. **i2CAT** és una part interessada en l'èxit del projecte i utilitzarà el sistema durant el desplegament per satisfer les necessitats de l'**Agència**. Es beneficia del compliment de les obligacions contractuals i, potencialment, de l'èxit del desplegament del sistema en altres projectes o contractes.

- **inLab FIB:** És el subcontractista contractat per **i2CAT** per implantar el sistema de gestió de tiquets d'incidències. Són els responsables directes del desenvolupament de la solució tècnica i de garantir-ne la funcionalitat. Els interessos d'**InLab FIB** resideixen a lliurar un producte funcional que satisfaci els requisits de l'**Agència**, així com complir les obligacions amb el soci contractual, **i2CAT**.

1.1.5 Identificació de lleis i regulacions

En la realització d'aquest projecte, ha sigut essencial considerar el marc legal i reglamentari que regeix el tractament de dades confidencials i l'execució de les obligacions contractuals. La base de la confidencialitat i el compliment legal d'aquest projecte són les lleis de *LOPDGDD*, *GDPR* i la signatura d'*Acords de No Divulgació* (*NDA*) entre les parts implicades, inclosa l'**Agència**, **i2CAT** i **InLab FIB**.

La llei *LOPDGDD* regula la protecció de dades personals i els drets digitals al context espanyol. Estableix principis com ara la necessitat d'obtenir consentiment per processar dades, la limitació en la recopilació de dades i l'obligació d'implementar mesures de seguretat.

El *GDPR* és una regulació de la Unió Europea que harmonitza les lleis de protecció de dades a tots els estats membres. Proporciona un marc legal robust per al tractament de dades personals, amb èmfasi en el respecte a la privadesa i els drets individuals.

Els *NDA* són fonamentals per restringir la difusió d'informació més enllà de les persones i entitats designades que participen directament al projecte. L'incompliment d'un acord de confidencialitat pot comportar conseqüències jurídiques, incloent-hi possibles litigis civils i danys i perjudicis. Per conseqüència, aquest projecte s'ha sotmès als convenis establerts per l'acord signat per tots els integrants d'aquest projecte, que conté la restricció de no compartir informació confidencial amb individus externs al projecte.

1.2 Abast

1.2.1 Objectius

El principal objectiu d'aquest projecte és el desenvolupament i implementació d'un sistema eina capaç d'extreure, analitzar i emmagatzemar la informació trobada en els tiquets proveïts per l'**Agència**. A continuació es llisten els objectius:

- Fer un estudi de l'estat de l'art per tal d'identificar solucions ja existents a problemes similars i adaptar-ne una al problema presentat.

- Configurar una base de dades OTRS i crear una eina de descàrrega de tiquets de les fonts de dades proporcionades utilitzant *PyOTRS*.
- Fer un preprocessamnet del text que en elimini tot el text redundant possible, afegixi metadades i informació rellevant i formategi el text segons les especificacions requerides.
- Desenvolupar i entrenar un model de *Deep Learning* per a l'extracció dels camps especificats del text preprocessat.
- Anonimització de les dades de sortida mitjançant un filtre de *Logstash*.
- Configurar i gestionar l'emmagatzematge de les dades preprocessades i anonimitzades en *Elasticsearch*.
- Combinar els elements anteriors i implementar una *pipeline* que descarregui, processi, extregui, anonimitzi i emmagatzemi les dades en l'entorn especificat.
- Posar en funcionament una API que permeti accedir i utilitzar aquest sistema de manera senzilla.

1.2.2 Requisits funcionals

El funcionament de l'API és invisible per l'usuari, però darrere hi ha tot el sistema en funcionament. El funcionament del sistema permet el següent:

- L'usuari accedeix a l'API que està emmagatzemanda a l'entorn cedit per **i2CAT** i especifica l'identificador del tiquet a analitzar. Ha de comprovar els possibles següents errors, els quals interrompen l'execució normal del programa:
  - No existeix el tiquet especificat.
  - No es pot connectar amb la base de dades d'on s'extreuen els tiquets.
  - No es pot accedir a la base de dades on s'emmagatzema el resultat.
  - Un error d'execució degut al contingut del tiquet.
- S'obté un tiquet de la base de dades *OTRS* utilitzant la configuració i llibreries preparades per a fer-ho.
- Es preprocessa el tiquet per tal d'afegir tota aquella informació rellevant al procés d'extracció d'informació propera i extreure tota aquella que sigui repetida i no aporti dades noves. Aquests són els processos pels quals passa:
  - Extreure en text simple tot el text del tiquet en format HTML.

  - Addició del text dels fitxers adjunts.
  - Addició del text dels tiquets als quals es fa referència.
  - Addició de tots els articles del tiquet.
  - Eliminació del text duplicat detectat.
  - Afegir el *system prompt* i reordenar el text en el format necessari per l'entrada del model.

- S'executa el model emprant el tiquet aconseguit i extreu la informació dels camps especificats.
- El resultat es passa per un algorisme d'anonimització implementat amb un filtre de *Logstash*.
- El resultat, ja anonimitzat, es redirigeix i s'emmagatzema en la segona base de dades *Elasticsearch*.

1.2.3 Requisits no funcionals

- **Adaptabilitat:** El sistema ha de permetre l'extracció d'informació de qualsevol tiquet independentment de la manera en la qual s'ha escrit. Ha d'aconseguir comprendre el significat dels texts i extreure els camps correctament, fins i tot amb variacions a la redacció.
- **Usabilitat:** L'eina ha de ser fàcil d'usar per facilitar-ne la integració en el flux de treball actual i amb eines futures amb les mínimes dificultats.
- **Eficiència:** Aquest projecte no prioritza el desenvolupament d'un sistema crític on el temps sigui una preocupació primordial. Tot i això, s'ha de processar una gran quantitat de dades i és crucial evitar un temps d'espera llarg per evitar que aquest pas esdevingui un coll d'ampolla en el procés.
- **Escalabilitat:** Els tiquets a processar varien en mida, tant pel que fa a la longitud dels mateixos articles com al nombre d'articles inclosos en un tiquet. Per obtenir un rendiment òptim, l'eina ha de tenir un rang d'acceptació ampli, que doni cabuda a la màxima quantitat de tiquets i garanteixi al mateix temps una funcionalitat correcta amb tots ells.
- **Confidencialitat:** Els tiquets que es processen estan subjectes a contractes de confidencialitat estrictes. Aquest fet implica que les dades no es poden retirar dels servidors designats i s'han de tractar amb cura, adoptant les mesures d'anonimització adequades. Aquests contractes també imposen limitacions als tipus de models i tècniques que es poden utilitzar durant el projecte.

1.2.4 Obstacles i riscos potencials

- **L'eina no entén correctament el llenguatge:** Comprendre el llenguatge natural és una tasca difícil que evoluciona contínuament i, sobretot, és molt lluny de ser perfecta. Una preocupació important ha estat la possible inadequació dels models disponibles per comprendre eficaçment determinats textos. És una tasca difícil, sobretot en català, trobar models de NLP que tinguin la capacitat de comprendre textos extensos i que extreguin la informació desitjada. A més a més, dependre únicament de models en local pot augmentar aquest risc en impedir que el sistema millori i s'ajusti constantment amb nous models lingüístics i dades, cosa que podria impedir el rendiment sostingut del programari.

- **Disgregació de la resposta:** Aquest repte sorgeix perquè els models de NLP depenen sovint del context i la proximitat per establir connexions entre paraules i frases. Quan els detalls clau estan dispersos o són incoherents, el model pot tenir dificultats per reunir la informació necessària, cosa que dona lloc a respostes incompletes o errònies a les consultes dels usuaris. Això també s'aplica a situacions en què la informació està repartida en diversos articles o tiquets, ja que no és factible proporcionar al model una conversa completa d'un tiquet amb tot el context necessari per comprendre la situació i trobar correctament tots els . En conseqüència, si la resposta es fragmenta i no és analitzada correctament, es pot perdre part de la informació. A més a més, aquesta limitació restringeix la varietat de models disponibles, pel fet que certes categories d'aquest àmbit no afavoreixen el nivell de flexibilitat desitjat.

- **Escassetat i varietat de dades d'entrenament:** L'èxit de l'entrenament del model depèn en gran manera d'un conjunt de dades ampli i variat. Tot i això, l'adquisició d'aquestes dades és lenta i hi ha una llarga demora per aconseguir-les. Aquest estancament impedeix l'avenç del projecte i també limita la capacitat per perfeccionar i optimitzar eficaçment el model. En cas que fos necessari, es buscarien dades sintètiques per compensar aquestes limitacions, encara que estiguessin en llengües diferents.

- **Potència insuficient per executar el model:** Aquests models són reconeguts per la seva complexitat i mida, cosa que exigeix considerables recursos informàtics. És possible que l'**Agència** no tingui la infraestructura necessària per suportar els models d'ús intensiu de recursos. Aquesta circumstància té el potencial de dificultar l'execució exitosa del projecte i donar lloc a problemes de rendiment que poden requerir la reavaluació del model seleccionat.

- **Poca experiència amb les tecnologies necessàries:** Aquesta manca de coneixements podria provocar problemes durant el desenvolupament, com ara un pro-

grés més lent, possibles errors i una corba d'aprenentatge més pronunciada. Per reduir aquest risc, es compta amb orientació, formació addicional programada i col·laboració amb experts als camps pertinents per a una execució del projecte més fluida i satisfactòria.

- **Accés restringit:** La limitació de recursos de programari i maquinari imposa un desavantatge significatiu a la gamma de models que es poden provar i als mètodes disponibles per a l'entrenament. El projecte es pot veure restringit a solucions menys òptimes, cosa que obstaculitza la capacitat d'assolir els nivells de rendiment desitjats. Tots els canvis i imprevistos que hagin sortit en el moment triguen un temps en ser avaluats i admesos (o denegats) per l'**Agència** o **i2CAT**.

1.3 Metodologia i rigor

1.3.1 Metodologia

Per maximitzar la productivitat d'un equip de desenvolupadors, és important tenir una bona metodologia. Així, s'evita que la feina d'un membre de l'equip col·lideixi, endarrereixi o impedeixi la d'un altre. Per aquest motiu, les metodologies Àgils han sigut l'elecció per excel·lència pel desenvolupament d'aquest projecte, més concretament s'ha usat *Scrum*.

Seguint la metodologia *Scrum*, la feina s'ha organitzat de manera que es puguin realitzar *Sprints*. Els *Sprints* són iteracions de dues o tres setmanes durant les quals s'implementen funcionalitats noves que s'afegeixen al producte intentant mantenir-lo sempre usable. Els *Sprints* es finalitzen amb una reunió on s'avalua el progrés i es decideix què implementar durant la següent iteració. Les tasques que s'han decidit implementar han intentat ser d'una durada igual o menor a la durada del *Sprint*, per tant, la feina s'ha dividit en subtasques per arribar a aquesta quota.

A més a més de les reunions anteriorment mencionades, l'equip també s'ha reunit de manera diària (*Daily Scrum* o *Daily Standup*) i setmanal (*Weeklys*). Aquestes reunions més breus han servit per mantenir als desenvolupadors actualitzats i col·laborant mútuament, poder detectar a temps qualsevol problema que pugui sorgir. Independentment d'aquestes reunions, l'equip ha estat comunicat mitjançant un programa de missatgeria instantània.

Per evitar errors al codi, s'ha utilitzat un mètode de desenvolupament conegut com a *Test Driven Development* (TDD), que consisteix a convertir els requisits de programari en casos de prova abans de crear el mateix codi. D'aquesta manera, es crea una gran quantitat de proves al llarg del desenvolupament que verifiquen constantment que es compleixen tots els requisits, cosa que garanteix que el codi funcioni correctament.

A continuació es defineixen els diferents rols dins del marc de la metodologia *Scrum* [2] i quines persones hi han format part de cada rol en el projecte.

Propietari del producte (Product Owner)

El propietari del producte representa el negoci i és responsable d'assegurar que l'equip ofereix el màxim valor, cosa que requereix una relació de confiança amb l'equip de desenvolupament.

El propietari del producte ha de prioritzar la feina basant-se en diversos factors, incloses les necessitats del client i els requisits de les parts interessades (*stakeholders*). Si entressin les prioritats en conflicte, podrien obstaculitzar l'eficàcia de l'equip. Les responsabilitats del *Product Owner* són la gestió del backlog, la supervisió dels desplegaments i el maneig dels *stakeholders*.

En general, el propietari del producte és fonamental per alinear l'equip amb els objectius empresarials i facilitar una comunicació i una col·laboració eficaces entre els *stakeholders*.

En aquest projecte, el propietari del producte o *Product Owner* ha estat **i2CAT**.

Facilitador (Scrum Master)

El facilitador és el paper responsable de garantir l'aplicació efectiva de les pràctiques *Scrum*. Actuant com a líder servidor, facilita la comunicació i la col·laboració dins de l'equip. Se centren a fer visible el treball de l'equip, el foment d'una cultura d'aprenentatge i el foment d'un entorn alineat amb els valors de *Scrum*.

Garanteix la transparència mitjançant la creació de mapes d'històries i l'actualització de la documentació; entrena l'equip en el repartiment del treball i millorar amb la revisió dels resultats; promou l'autoorganització encoratjant els membres de l'equip a provar nous enfocaments; i posa èmfasi en la importància dels valors *Scrum* en la creació d'un ambient de confiança.

Dins de l'àmbit del projecte, el Facilitador o *Scrum Master* ha sigut el **cap del projecte**.

Desenvolupadors (Development Team)

Contràriament a la percepció comuna, el terme "desenvolupador" abasta diverses funcions com a dissenyadors, escriptors i programadors, no només enginyers. Les responsabilitats principals dels desenvolupadors són l'autoorganització, el disseny, el desenvolupament, les proves i el desplegament. Tenen autonomia per prendre decisions, ja que l'autoorganització no consisteix a desafiar l'organització, sinó a capacitar els qui estan més a prop de la feina.

Es comprometen a lliurar la feina dins d'un esprint i garanteix la transparència mitjançant reunions diàries de *Scrum* (*Daily Scrum*). Tot i que el *Scrum Master* pot facilitar el *Scrum* diari, en última instància és responsabilitat de l'equip de desenvolupament dirigir la reunió, fomentant la col·laboració i la millora contínua a la feina.

En aquest projecte, l'equip de desenvolupadors o *Development Team* han sigut els **dos desenvolupadors júniors**.

1.3.2 Eines

- **Git:** És una eina que és utilitzada per controlar i gestionar les versions del codi. També s'utilitza per compartir el codi amb el client.
- **Python:** llenguatge de programació per acomplir les tasques de *Machine Learning* (ML) i consensuat amb l'empresa.
- **Hugging Face:** Principal font d'investigació sobre models i facilitadora d'eines per la seva prova i execució.
- **OTRS:** Sistema lliure que s'utilitza per assignar identificadors únics a sol·licituds de servei o informació. És el sistema utilitzat a la primera base de dades d'on s'extreuen els tiquets.
- **Elasticsearch:** Servidor que proveeix un motor de cerca de text complet, distribuït i amb una interfície web. És publicat com a codi obert i s'utilitza per a la segona base de dades on es guarden els tiquets.
- **Logstash:** És un pipeline de processament de dades que transforma les dades abans de ser emmagatzemades a Elasticsearch. S'ha utilitzat per pseudoanonimitzar els camps abans d'emmagatzemar-los.
- **Models NLP[3]**: Escollit després de fer l'estudi corresponent.
- **Slack:** El servei de missatgeria instantània usat per comunicar-se amb l'equip.
- **Google Meet:** Es fa servir per celebrar les reunions digitals.

Capítol 2

Exploració teòrica

En aquesta secció es repassen totes aquelles idees necessàries per a una comprensió completa del treball. Es comença definint els conceptes fonamentals i explicant un tiquet d'incidències completament. Més endavant, es defineix l'aprenentatge autònom i com funciona, amb una atenció especial als models NLP. Per finalitzar, s'explora l'estat de l'art actual i com se solucionen els problemes de NLP avui en dia. En última instància, s'escull el model que serà usat per resoldre el problema plantejat.

2.1 Definició de conceptes

Anàlisi d'un tiquet

Un tiquet d'incidències és un informe de qualsevol problema o dubte que hagi pogut sorgir, normalment, dins d'una empresa. Aquests tiquets serveixen per comunicar del problema mencionat al tiquet i s'espera obtenir una contestació detallant quins són els passos a seguir per solucionar el problema o una resposta resolent el dubte. Per aquest projecte, s'ha usat OTRS, una eina de gestió i emmagatzematge de tiquets. Gràcies a la reducció de l'abast mencionat a l'apartat 5.5, només es poden trobar tiquets pertenyents a dues categories: *phishing* i *malware*, sent la primera més abundant. A continuació, es mostra un tiquet d'exemple d'ORTS i s'explica en deteniment les seves parts.

En el tiquet es poden veure els següents camps:

1. **Logo Znuny:** Tot i que a l'**Agència** s'utilitza OTRS, aquest tiquet d'exemple, i les proves que s'han dut a terme han sigut realitzades amb Znuny. Znuny és la continuació d'OTRS, ja que a partir en un cert punt, la versió gratuita d'OTRS (OTRS Community Edition) va deixar de rebre actualitzacions de manteniment. A efectes pràctics, Znuny és es comporta de manera idèntica i és compatible amb les mateixes llibreries que OTRS. S'ha utilitzat per a totes les proves locals que

requereixin el servei d'un gestor de tiquets.

2. **Nombre del tiquet:** Nombre identificador únic del tiquet. És l'identificador principal per obtenir el tiquet de la base de dades.

3. **Capçalera:** Els tiquets tenen un assumpte, remitent, destinatari i informació sobre la data d'enviament, igual que un correu electrònic. En aquest cas, està dividida en diversos llocs del tiquet.

4. **Informació del tiquet (metadata):** Hi ha informació inclosa en el tiquet que permet establir alguns camps rellevants relatius a les condicions actuals del tiquet, tals com: l'estat del tiquet, la seva prioritat, a quina cua pertany, quant de temps fa que s'ha creat, etc.

5. **Informació del client:** Conté informació sobre l'empresa que ha patit l'incident, així com l'usuari de l'empresa que ha escrit el tiquet.

6. **Nombre de l'article:** Un tiquet es divideix en diferents articles. Cada article representa un missatge o correu que una de les parts ha enviat. En aquest exemple hi ha dos articles: El primer (2) informant del problema i el segon (1) explicant les mesures preses a causa de l'incident. És important tenir en compte que, tal com es veu a la Figura 1, el text de tots els anteriors tiquets es reescriu sota d'aquest.

Tal i com s'ha explicat, la feina d'extracció dels camps succeeix majoritariament dins del tiquet, analitzant i extraient informació dels diferents articles. Aquests són els camps que es busquen, el seu identificador otorgat per l'**Agència** i la seva definició:

7. **Usuaris afectats (*usuarios_afectados*):** Els usuaris afectats són tots aquells usuaris que han rebut el correu referent a la incidència. Aquests usuaris han de ser molt específics, es requereix que sigui els emails que han rebut aquest correu. No és correcte, per exemple, identificar pel seu nom als usuaris afectats, ni incloure tot un grup, com un departament sencer.

8. **Accions de mitigació (*acciones_mitigacio*):** Les accions de mitigació són les accions preses per evitar l'expansió de l'incident. Normalment, les accions de mitigació són preses per l'equip de ciberseguretat encarregat de solucionar el tiquet, però pot arribar pels dos llocs. Són mesures immediates que només afecten la incidència concreta que s'intenta combatre, i aquest és el punt que més les diferencia de les accions de control.

9. **Accions de control (*acciones_control*):** Les accions de control, a diferència de les accions de mitigació, són aquelles que s'utilitzen per prevenir més incidents del mateix estil. Són aquelles mesures que, estudiant el cas actual, es poden implementar i evitar així un rang més ampli d'aquests esdeveniments.

10. **URL de l'incident (*URL_ mail_ incident*):** És l'adreça electrònica que s'ha usat en el correu de l'incident per enganyar a la víctima i fer creure que estan en una pàgina web coneguda. Es fa ús d'una còpia exacta de la pàgina web per evitar sospites i així robar la informació de la víctima.

11. **Mail de l'atacant (*from_ mail_ incident*):** És l'adreça electrònica des de la qual es van enviar els correus maliciosos. Normalment, pertanyen a un domini estranger, o a un domini de correus temporals per evitar ser rastrejats. El nom d'usuari, però, és esperable que sigui un que simula ser una entitat real, per intentar enganyar a les víctimes.

12. **Mail de la víctima (*recipient_ mail_ incident*):** És un dels usuaris que ha rebut el correu maliciós (*affected_ users*) ha sigut el primer a reportar o a informar sobre aquest succés.

13. **Assumpte correu de l'incident (*subject_ mail_ incident*):** És l'assumpte que tenia el correu de l'atacant. Solen ser assumptes molt cridaners amb paraules que inciten a actuar ràpidament, per aconseguir que la víctima entri sense pensar-s'ho dues vegades.

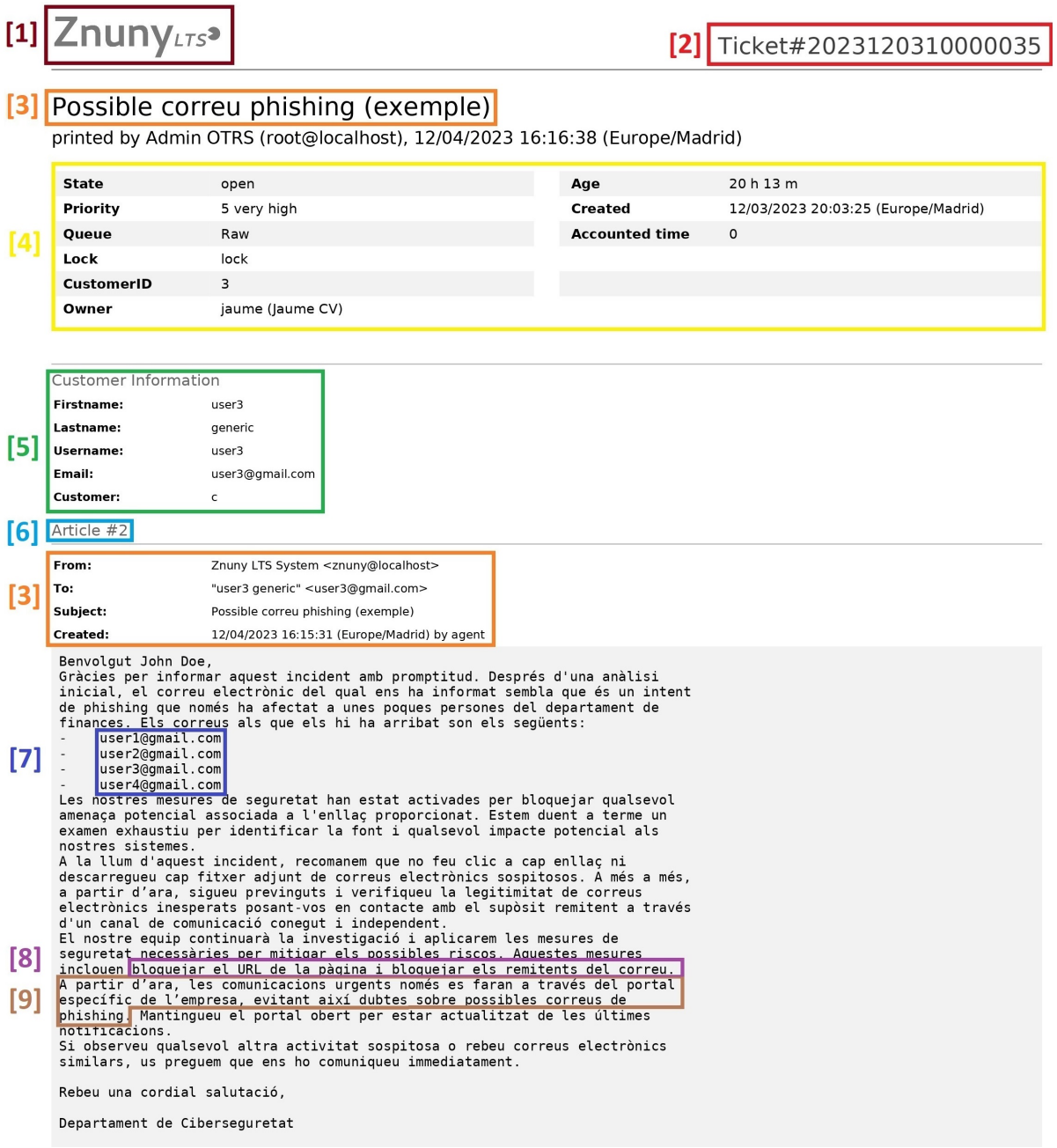


Figura 1: Primera meitat del tiquet d'exemple d'ORTS amb totes les parts indicades. (Creació pròpia)

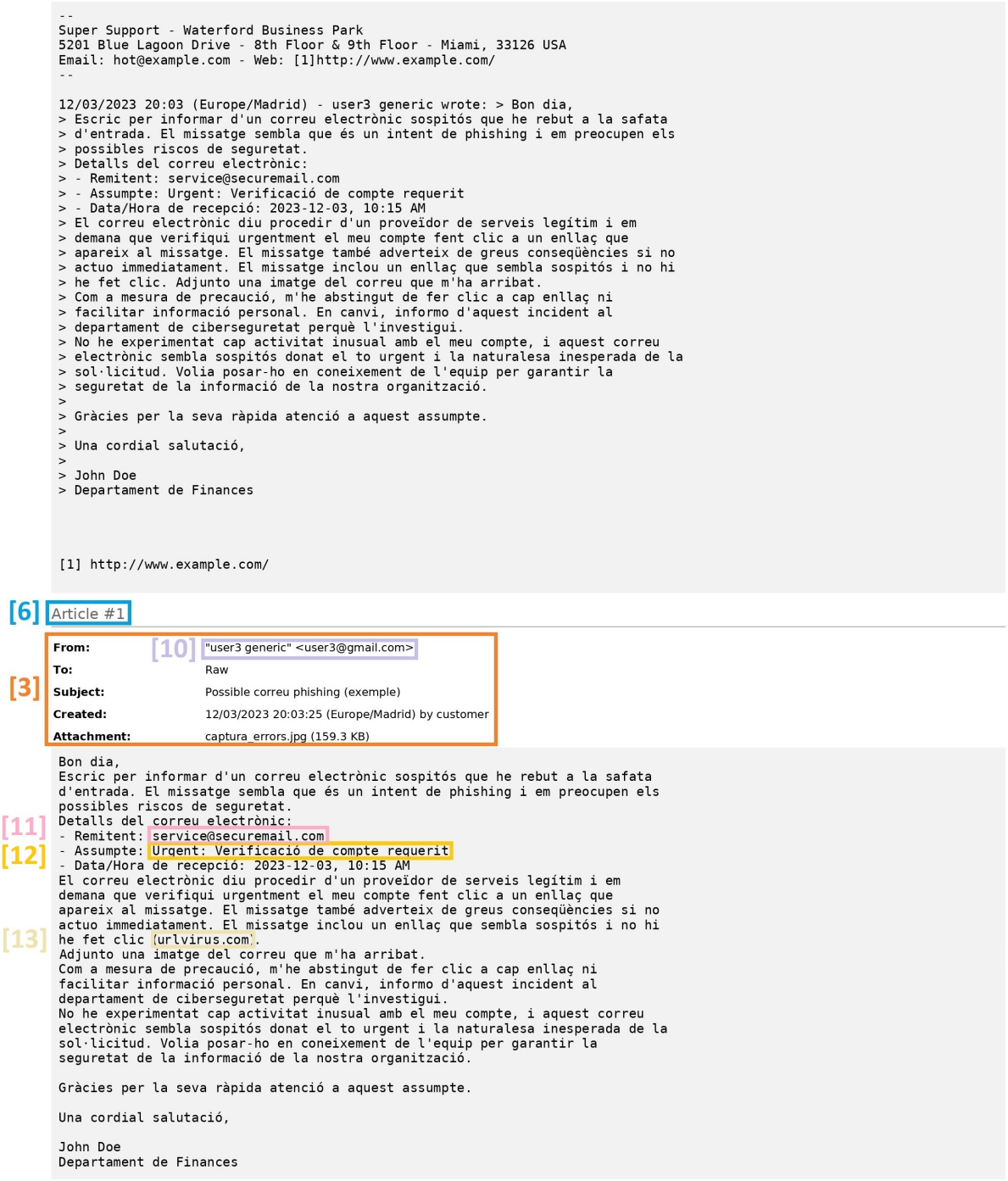


Figura 2: Segona meitat del tiquet d'exemple d'ORTS amb totes les parts indicades. (Creació pròpia)

Non-Disclosure Agreement (NDA)

Un acord de no divulgació (NDA), també conegut com a acord de confidencialitat, és un contracte legal que estableix una relació confidencial entre dues o més parts que acorden protegir qualsevol informació confidencial que puguin compartir o rebre.

Els elements clau d'un acord de confidencialitat solen incloure la identificació de la in-

formació confidencial, la descripció de les obligacions i les responsabilitats de les parts reveladora i receptora, l'especificació de la durada i l'abast de les obligacions de confidencialitat, i la definició de les excepcions o exclusions de l'acord. En cas d'incompliment d'un acord de confidencialitat per una de les parts, l'altra part o les altres parts poden sol·licitar recursos legals, com ara mesures cautelars o danys i perjudicis, per evitar noves divulgacions i compensar qualsevol pèrdua.

En aquest context, l'acord de confidencialitat que s'ha signat per aquest projecte, inclou la protecció de totes les dades que estiguin dins dels servidors de l'**Agència** (en especial els tiquets d'incidències), així com tota l'informació relativa al projecte que contingui informació crítica. Això implica no poder usar cap servei que impliqui moure les dades fora dels servidors.

Phishing

El *phishing* és un tipus de ciberatac que utilitza correus electrònics fraudulents o altres mètodes de comunicació per enganyar els destinataris per tal que revelin informació confidencial o instal·lin *malware* als seus dispositius. Sovint, els correus electrònics de *phishing* suplanten la identitat d'entitats o persones legítimes i redirigeixen els usuaris a pàgines web falses dissenyades per assemblar-se als reals. El *phishing* pot donar lloc a robatoris d'identitat, pèrdues financeres o comptes compromesos.

Malware

El *malware* o programari maliciós es refereix a qualsevol programa dissenyat per interrompre o danyar un sistema informàtic, una xarxa o un dispositiu. El *malware* pot dur a terme una sèrie d'accions malicioses, entre les quals s'inclouen robar, xifrar o esborrar dades i monitorar l'activitat de l'usuari. També es pot infiltrar en els sistemes a través de diversos canals, com ara correus electrònics de *phishing*, descàrregues de fonts no segures, mitjans extraïbles o connexions de xarxa. Un cop dins, el *malware* pot suposar greus amenaces per a la seguretat del sistema, i, fins i tot, pot permetre ciberatacs addicionals.

Dataset

Un *dataset* o conjunt de dades són les dades utilitzades per entrenar, validar i provar els models d'aprenentatge automàtic. És un element essencial de l'aprenentatge automàtic, ja que proporciona les dades necessàries perquè el model adquireixi coneixements i generi prediccions. Depenent del problema que s'abordi, pot incloure dades en diferents formats, com ara text, imatges o àudio. La informació d'un *dataset* sol tenir etiquetes, cosa que indica que cada entrada sol tenir una sortida esperada. Aquestes etiquetes són útils per entrenar el model a distingir patrons a la informació rebuda i generar prediccions precises

per a dades mai vistes abans. Per crear un *dataset* de la màxima qualitat cal considerar meticulosament el procés de recopilació, neteja i etiquetatge. És crucial garantir que el conjunt de dades representi amb precisió el problema que s'aborda i contingui prou informació per entrenar eficaçment el model.

2.2 Aprenentatge autònom

2.2.1 Models d'aprenentatge autònom

L'aprenentatge autònom (*Machine learning* en anglès) és un subcamp de la intel·ligència artificial que permet als ordinadors aprendre els patrons i regles implícites en grans conjunts de dades. Aquests després, es poden fer sevir per fer prediccions o decisions i millorar-ne el rendiment amb l'experiència sense necessitat de programació explícita. Les solucions d'aprenentatge autònom s'apliquen àmpliament a diversos sectors, com ara el comerç electrònic, la sanitat, les finances o la indústria.

A continuació es mostra com hi ha diversos tipus d'algorismes d'aprenentatge autònom segons el resultat desitjat i el tipus de dades que es disposin.

Classificació segons la naturalesa de les dades

Segons la naturalesa de les dades, l'aprenentatge autònom es pot dividir en tres tipus principals:

- **Supervisat:** Aquests algoritmes s'entrenen amb conjunts de dades etiquetades que contenen les variables d'entrada i sortida. L'objectiu principal és aprendre una funció que relacioni les dades d'entrada amb les de sortida i, a continuació, aplicar-la per fer prediccions sobre dades noves o desconegudes. La regressió, classificació i detecció d'anomalies són exemples d'aprenentatge autònom supervisat.
- **No supervisat:** En aquest tipus, els algoritmes s'entrenen amb conjunts de dades no etiquetades que només contenen les variables d'entrada. El seu objectiu és descobrir l'estructura o la distribució subjacent de les dades per agrupar-les o segmentar-les en categories significatives. Alguns exemples d'aprenentatge autònom no supervisat són l'agrupació i la reducció de la dimensionalitat.
- **Per reforç:** L'aprenentatge autònom per reforç implica algorismes que no depenen de conjunts de dades externes, sinó que aprenen de les mateixes accions i de la retroalimentació rebuda de l'entorn. L'objectiu és determinar la política o estratègia òptima que maximitzi la recompensa o minimitzi el cost al llarg del temps. Nor-

malment, s'utilitza en situacions on hi ha un món virtual i un agent que el pugui explorar.

Classificació segons la tasca a resoldre

Segons la tasca a resoldre, l'aprenentatge autònom es pot dividir principalment en quatre tipus, encara que hi ha altres categories i moltes subcategories dins de cada tasca.

- **Multimodal:** Són les tasques entre les quals s'inclou el processament i integració de diverses modalitats, incloent-hi imatges, àudio, text i més. El propòsit és establir una representació cohesiva de les dades independentment del seu format i s'usa en una àmplia gamma d'aplicacions, com creació i edició d'imatges, extracció d'informació i descripció d'imatges.
- **Visió per Computador:** Aquesta tasca consisteix a analitzar i comprendre les dades visuals, tals com imatges i vídeos amb l'objectiu d'extreure informació o característiques de les dades per utilitzar-les en detecció d'imatges, el reconeixement de cares o la segmentació d'escenes, entre altres.
- **Processament del Llenguatge Natural:** Consisteix en l'anàlisi i generació del llenguatge natural, és a dir, text. El seu objectiu és entendre el significat i propòsit de les paraules que rep. Les dades llavors poden ser aplicades a altres tasques com el resum de textos, la traducció o la generació de text. Aquesta és la tasca dels models que es fan servir en aquest projecte, i en l'apartat 2.2.3 es parlarà en més en profunditat.
- **Àudio:** El processament d'àudio s'ocupa de la manipulació i generació d'àudio, tant veu, com música o efectes de soroll. L'objectiu és extreure informació o característiques de la pista d'àudio pel reconeixement de veu, la generació de música o la recomanació de música.

2.2.2 Deep learning

El *deep learning* (aprenentatge profund en català) és un subcamp del *machine learning* que aprofita les xarxes neuronals per analitzar dades.

Les xarxes neuronals consisteixen en capes de neurones artificials que adquireixen patrons i característiques complexos a partir de les dades d'entrada. Fa servir aquestes xarxes neuronals amb múltiples capes per extreure'n característiques i patrons complexos de les dades. Cada capa neuronal obté informació de la capa anterior, fa càlculs i transmet el resultat a la següent capa.

Els models de *deep learning* poden aprendre de grans quantitats de dades i assolir una precisió i eficiència elevades. Destaquen sobretot, perquè són capaços de fer tasques complexes que els algoritmes d'aprenentatge autònom tradicionals consideren difícils, com ara la visió per ordinador i el processament del llenguatge natural. Tot i això, l'aprenentatge profund requereix més recursos i temps per l'entrenament i execució que el *machine learning*.

Fine-Tuning

El *Fine-Tuning* o afinament en el context del *deep learning* és el procés d'ajustar una xarxa neuronal preentrenada per a una tasca específica. Aquest enfocament evita entrenar una xarxa neuronal des de zero, adaptant un model que ja ha estat entrenat en un conjunt de dades ampli i general. Aquest enfocament és especialment útil quan el conjunt de dades pel *Fine-Tuning* és més petit i pot no ser adequat per entrenar un model d'alt rendiment des del principi.

La tècnica consisteix a ajustar els pesos i paràmetres del model preentrenat durant l'entrenament al conjunt de dades específiques. Aquest procés permet que el model aprengui matisos i informació específics de la tasca, alhora que conserva els valuosos coneixements adquirits a la fase de preentrenament.

Few-shot learning

El *Few-shot learning* o aprenentatge d'uns pocs cops, inclòs el *One-shot learning* o aprenentatge d'un sol cop, aborda escenaris en què un model s'entrena per fer prediccions basant-se en un nombre limitat d'exemples. En els entorns tradicionals, els models solen requerir grans quantitats de dades d'entrenament per aconseguir un rendiment satisfactori. En contraposició, en el *Few-shot learning*, l'objectiu és que un model funcioni bé fins i tot quan se li presenten només unes poques instàncies d'exemple durant la inferència, evitant així un reentrenament.

El funcionament és tan senzill com exposar un exemple o diversos exemples del problema ja solucionat amb la seva resposta i fer contestar al model en el mateix missatge un problema similar. El raonament és que el model entén el que s'espera que hagi de retornar i produeix així una inferència molt més satisfactòria.

Avaluació dels models

Per avaluar i millorar l'entrenament d'un model, s'utilitzen certes mètriques que permeten quantificar com s'assembla la generació del model amb la resposta correcta de referència. S'ha fet ús de les següents mètriques:

- *Train_Loss*: Error durant l'entrenament. Calculat automàticament per la funció *Trainer* de la llibreria *transformers*. La funció pel còmput d'aquest error s'escull automàticament i depèn de la tasca amb la qual s'està entrenant, en aquest cas, fa servir la funció de `CrossEntropyLoss` (Error d'entropia creuada).
- *Val_Loss* Error durant la validació. Es calcula també automàticament i fa servir la funció de `CrossEntropyLoss`.
- *ROUGE-1*: Mesura la superposició d'unigrames (cada paraula) entre el text generat i el text de referència.
- *ROUGE-2*: Mesura el solapament d'unigrames (dues paraules seguides) entre el text generat i el text de referència.
- *ROUGE-L*: Mesura la subseqüència comuna més llarga (LCS) de les paraules entre el text generat i el text de referència.
- *ROUGE-Lsum*: Mesura la LCS de les paraules entre el text generat i el text de referència, però també inclou un factor de normalització de longitud que premia les LCS més llargues.

2.2.3 Processament del Llenguatge Natural

El processament del llenguatge natural (*NLP* per les seves sigles en anglès) es refereix a la disciplina de dissenyar màquines capaçes de comprendre el llenguatge humà, o dades semblants al llenguatge humà, de la manera com s'escriu, es parla i s'organitza. Els models d'aprenentatge profund poden aprendre a representar el llenguatge natural d'una manera significativa i utilitzar aquestes representacions per realitzar diverses tasques de *NLP* [3]. Per cadascuna d'elles es descriu i s'indica si és adequada pel projecte que es desenvolupa.

- **Classificació de text:** És la tasca d'assignar una etiqueta o classe a un text donat. Aquesta etiqueta normalment es proporciona amb la probabilitat de que el text tingui certa etiqueta. Alguns dels seus usos són l'anàlisi de sentiments, la inferència del llenguatge natural i comprovar la correctesa del llenguatge. Aquesta tasca no es pot aplicar de manera directa al cas d'ús d'aquest projecte, però es va intentar fer servir per l'eliminació de les signatures dels tiquets.
- **Resum de text:** Aquesta tasca consisteix a generar un breu resum d'un text més llarg, com ara articles de notícies, ressenyes i informes, preservant a la vegada la informació important. Alguns models extreuen frases senceres del text d'entrada mentre que d'altres generen text completament nou. Aquesta tasca no ha sigut implementada en el projecte encara pot ser una tasca útil en una situació ideal

ja que permetria eliminar tota aquella part del tiquet que no fos rellevant i així permetre que el model principal extregui els camps requerits.

- **Traducció automàtica:** Aquesta tasca consisteix a traduir text d'un idioma a un altre, com ara l'anglès a l'espanyol. Aquesta tasca és particularment imprecisa perquè hi ha moltes paraules o expresions que no es poden traduir d'un idioma a un altre. Pel projecte pot ser vital, ja que els models en anglès tenen un rendiment molt superior als models en català, i traduïr els tiquets, mantenint el significat, millora molt el rendiment.

- **Resposta a preguntes:** Aquesta tasca consisteix a trobar la resposta a una pregunta de llenguatge natural a partir d'un text o una base de coneixements determinats. Depenent del model, no és necessari subministrar un context per contestar preguntes, es pot contestar només amb l'informació que ha après. És una de les principals propostes per aquest projecte, ja que soluciona tots els problemes ingerint el tiquet com a "context" i permetent fer preguntes per trobar els camps buscats.

- **Generació de text:** Serveix per produir text coherent i fluid a partir d'una entrada determinada, com ara una instrucció o una paraula clau a substituïr. Poden ser molt versàtils, ja que generen el text més probable que bé a continuació, sent ideals per resoldre moltes tasques de *NLP*. Té una aplicació directe al projecte, ja que es pot formular frases o preguntes juntament amb el tiquet a analitzar, de manera que la continuació d'aquestes sigui la resposta que es busca.

- **Reconeixement d'entitats:** És la tasca de trobar entitats a un text. Aquestes entitats poden ser noms de persones, llocs o organitzacions, entre altres. Cada paraula de l'entrada és assignada una etiqueta, basat en la probabilitat que la paraula pertanyi a la classe en qüestió. Tot i que normalment està dissenyat per reconeixer entitats senzilles i de poques paraules, es podria arribar a plantejar com una solució pel projecte, ja que podria aprendre a trobar els camps buscats al text com si fóssin entitats.

2.3 Estat de l'art

En aquesta secció es revisa l'estat de l'art dels models que poden solucionar el problema plantejat. Per aconseguir la millor solució, s'ha proposat mantenir un rang de cerca ampli, i considerar moltes propostes que en un principi no havien sigut plantejades.

2.3.1 Aplicacions comercials

Lògicament, un model de generació del llenguatge natural pot ser una eina molt potent si es busca el propòsit adequat. És per això, que moltes empreses han decidit invertir en aquest sector per assolir aquests models tan potents. Aquests models són, generalment, l'estat de l'art en aquest àmbit, ja que disposen de molts recursos per obtenir el millor resultat possible. A continuació es presenten els models comercialitzats més importants en l'àmbit global:

- **GPT-4:** *GPT-4* és un gran model multimodal que pot agafar entrades d'imatge i text i produir sortides de text. Representa el darrer avenç en la investigació d'*OpenAI* sobre l'ampliació del *deep learning*. El seu principal ús és per crear xatbots (bots conversacionals) generals, que poden conversar sobre temes molt diversos i, fins i tot, cercar a la xarxa per aportar informació extra. Alguns exemples són *ChatGPT* i *Bing Chat*, tots dos són serveis en línia i, per tant, no aplicables per al projecte.

- **LaMDA:** *LaMDA* és una col·lecció de grans models lingüístics desenvolupats per Google que treballen conjuntament per resoldre diverses tasques de generació de text. *LaMDA* no està disponible públicament i només es pot accedir utilitzant la seva plataforma. Google va integrar aquest model en alguns dels seus productes, per millorar la seva funcionalitat i també va desenvolupar una primera versió d'un xatbot (*Google Bard*) que durant un temps va usar aquest model.

- **PaLM 2:** *PaLM 2* és un gran model lingüístic (LLM) que amplia el llegat de Google en matèria d'aprenentatge automàtic i de recerca responsable de la IA. S'ha entrenat en un corpus divers i multilingüe de text, codi, matemàtiques i pàgines web i va ser usat durant un temps com a xatbot amb *Google Bard*. Tot i que no és de codi obert, es pot accedir a PaLM 2 a través d'una API.

- **Gemini:** *Gemini* de Google és un ambiciós projecte d'intel·ligència artificial. És gran model de llenguatge (LLM) que tingui la capacitat de fer una gran varietat de tasques, incloses les relacionades amb text, imatges i àudio. *Gemini* compta amb un disseny multimodal que permet la integració de diversos tipus d'informació i dona lloc a una àmplia gamma de formats de sortida. No és una plataforma de codi obert, per la qual cosa el codi i les dades no són accessibles al públic. Per altra banda, s'integra amb eines d'IA per a la generació de continguts, com ara la versió més nova de *Google Bard*, que permet la interacció de l'usuari amb aquest model.

- **Claude 2:** *Claude 2* és un model de llenguatge d'intel·ligència artificial avançat desenvolupat per *Anthropic* amb un enfocament a IA segura i beneficiosa. Permet

generar textos de primera qualitat i contextualment apropiats per a diverses aplicacions, com ara la generació de codi, l'anàlisi de textos i la redacció de composicions. És accessible al públic mitjançant un lloc web beta (en forma de xatbot), però el codi i les dades utilitzades per la formació i el funcionament no són accessibles pel públic.

- **Inflection-1:** *Inflection-1* és un gran model de llenguatge (LLM) desenvolupat per *Inflection AI*. És la base de *Pi*, un assistent d'IA que pretén ser un xatbot de companyia, que continua les converses més llargues amb els usuaris i s'adapta a cadascun d'ells en funció d'aquestes converses. *Inflection-1* no està disponible públicament i només és accessible en línia a través del xatbot que fa servir.

- **Grok-1:** *Grok-1* va ser creat per *xAI*. Es tracta del motor que impulsa *Grok*, un xatbot capaç de respondre gairebé qualsevol dubte i fins i tot d'oferir suggeriments sobre les preguntes que s'han de formular. *Grok* està configurat per assistir als usuaris amb la xarxa social que del que és propietari el mateix fundador de l'empresa. *Grok-1* no és un model de codi obert i no es pot emprar de manera lliure a excepció d'una beta oberta.

2.3.2 Propostes descartades

El procés de selecció del model va exigir una avaluació acurada de múltiples candidats per garantir la seva adequació als objectius del projecte. El fet de considerar tants models, cadascun amb els seus punts forts i febles, ha implicat inevitablement un rebuig de molts d'ells per trobar la solució que millor s'adeqüi al problema.

Seguint aquestes directrius, s'ha intentat aconseguir una avaluació exhaustiva i imparcial dels models candidats. En última instància, s'han triat les opcions més adequades per complir els objectius del projecte, alhora que es treballa dins de les limitacions imposades per la sensibilitat de les dades i els recursos computacionals. La secció següent aprofundeix en les raons que van portar a descartar determinats models durant la fase de recerca.

Resolució inadequada d'extreure informació

Un dels principals criteris de selecció d'un model ha sigut la seva capacitat per extreure les paraules adequades del text que se li ofereix. Els models que no han complert una prova base que representa l'ús que se li donarà al model són exclosos immediatament.

La prova en qüestió es basa en, donat un tiquet d'exemple, fer una pregunta o formular una frase, de manera que la continuació sempre sigui una frase que estigui continguda en el text. Aquesta prova s'ha modificat depenent de per quina tasca està entrenat cada model. Això ha servit per descartar directament els models menys adequats i quedar-se

amb el que són capaços de contestar correctament, permetent entrenar només aquells que ja anaven encaminats.

Accessibilitat en línia i restriccions dels acords de confidencialitat

Alguns models han sigut eliminats a causa de la seva accessibilitat només en línia, cosa que no s'ajusta al marc del projecte, que requereix un funcionament *in situ* sense dependència de servidors externs. És inviable emprar models que exigeixin la transferència de dades fora dels servidors controlats, atesa la sensibilitat de les dades i les restrictives disposicions dels acords de confidencialitat vigents. Aquesta restricció ha sigut imperativa per mantenir la confidencialitat i la integritat de la informació confidencial.

Limitació de recursos

Les exigències computacionals dels algorismes de *deep learning* són considerables, per la qual cosa requereixen grans recursos de maquinari per agilitzar els processos d'entrenament. En concret, l'ús d'unitats de processament gràfic (GPU) d'alt rendiment s'ha tornat essencial per accelerar aquest tipus de tasques. A diferència de les CPU tradicionals, l'arquitectura paral·lela de les GPU millora l'eficiència de les operacions d'alta càrrega computacional inherents als algoritmes de *deep learning*. Per tant, és necessari disposar de GPU amb una potència computacional per poder entrenar i afinar la majoria dels models disponibles.

Tot i que alguns models resulten eficaços per abordar la tasca, han hagut de ser descartats a causa dels seus requisits poc pràctics en matèria de recursos tecnològics. En el context del projecte, la viabilitat de l'aplicació d'un model dins de la infraestructura disponible és molt limitant. Els models que exigien una potència de càlcul o uns excessius recursos de memòria es consideren inviables per a l'entorn d'execució del projecte.

2.3.3 Comprovació dels models disponibles

Avaluem cada model per valorar-ne l'eficàcia a l'hora de respondre les proves que han sigut seleccionades per cobrir els casos més típics. Aquesta secció presenta una descripció dels models que han sigut considerats, incloent-hi els punts forts i febles de cadascun. A través d'una anàlisi dels avantatges i desavantatges de cada model, l'objectiu és trobar raons per a la seva selecció o eliminació.

Per cada model, s'ha descarregat la versió més gran (i, en conseqüència, s'espera que la més efectiva), que ha permès l'equip actual. S'ha executat el model amb una sèrie de tiquets de prova depenent del tipus de tasca per la qual ha sigut entrenat.

Models de *Question Answering*		
Model	**Params.**	**Observacions**
BERT	340M	És un model antic | Contesta meitat de la resposta
BERT (fine-tuned)	340M	Ja està afinat | Contesta meitat de la resposta
RoBERTa	355M	Millora de BERT | Contesta meitat de la resposta
DeBERTaV3	304M	Millora de RoBERTa | Contesta meitat de la resposta
GPT-2	774M	Massa general | Escriu altres coses
BART	406M	Més complicat d'afinar | Serveix per resumir
Prometheus	7000M	Requereix massa recursos
LUKE	483M	Resposta buida
XLNet	110M	Escriu altres coses
XLNet (large)	340M	Escriu altres coses

Taula 1: Comparació dels models de *Question Answering* (Resposta a preguntes)

Models de *Text Generation*		
Model	**Params.**	**Observacions**
GPT-2	774M	Massa general | Escriu altres coses
BART	406M	Més complicat d'afinar | Serveix per resumir
Prometheus	7000M	Requereix massa recursos
GoLLIE-7B	7000M	Dissenyat per seguir instruccions | Requereix massa recursos
GPT-J	6053M	Té una versió en català | Requereix massa recursos
GPT-Neo	2700M	Té una versió en català | Entèn el significat però no contesta amb el format esperat
Llama 2	7000M	És pitjor, en general, que Mistral 7B | Requereix massa recursos
Mistral 7B	7300M	Requereix massa recursos
Mistral 7B (Instruct)	7300M	Dissenyat per seguir instruccions | Ha sigut reduït (Quantificació) | Respon correctament
Zephyr 7B β	7300M	Reentrenament de Mistral 7B | Requereix massa recursos
Flor-Instructed	6300M	Està dissenyat per la comprensió i escriptura de text en català, no comprèn el format de sortida
Qwen1.5-Instructed	7000M	Comprèn el català i segueix les instruccions a la perfecció. Es requereix una GPU per executar-se.

Taula 2: Comparació dels models de *Text Generation* (Generació de text)

Models de *Text to Text Generation*		
Model	**Params.**	**Observacions**
Flan-T5-Small	80M	No entèn el significat, Serveix per tasques més petites
Flan-T5-Base	250M	És relativament petit, Contesta correctament
FLAN-T5 Large	780M	Contesta correctament
FLAN-T5 XL	3B	Massa gran com per ser entrenat, Contesta correctament
M2M100	418M	Massa complicat d'afinar, Serveix per traduir
Flan-T5-Base (QA)	300M	Reentrenat específicament per contestar preguntes, Contesta meitat de la resposta

Taula 3: Comparació dels models de *Text to Text Generation* (Generació de text a text)

Capítol 3

Desenvolupament del sistema

3.1 Arquitectura del sistema (pipeline)

Aquest apartat descriu l'arquitectura del sistema que s'ha desenvolupat per a l'anàlisi automàtica dels tiquets d'incidències de ciberseguretat de l'**Agència**. L'arquitectura del sistema està directament lligat a les màquines des de les quals s'han disposat propietat de l'**Agència** i al sistema que tenien en producció. L'**Agència** va posar a disposició de l'equip tres servidors idèntics per cada secció on estaran les dades i dos portàtils amb accés als servidors a través d'una VPN privada.

- **Servidor 1 (OTRS):** Servidor on s'emmagatzema la base de dades d'OTRS amb els tiquets al seu interior.
- **Servidor 2 (Pipeline):** Servidor encarregat d'accedir als altres servidors per llegir i guardar informació i passar-la pel programa desenvolupat per resoldre el problema.
- **Servidor 3 (Elasticsearch):** Servidor que recull les dades processades i anonimitzades per un futur us.

- **Portàtil 1 (amb GPU):** Portàtil amb la funció principal d'entrenar el model que serà utilitzat per extreure l'informació dels tiquets. Està configurat per evitar cap bretxa de dades.
- **Portàtil 2 (sense GPU):** Portàtil amb la capacitat d'accedir als servidors de l'**Agència** mitjançant una VPN. Té la mateixa configuració que el Portàtil 1.

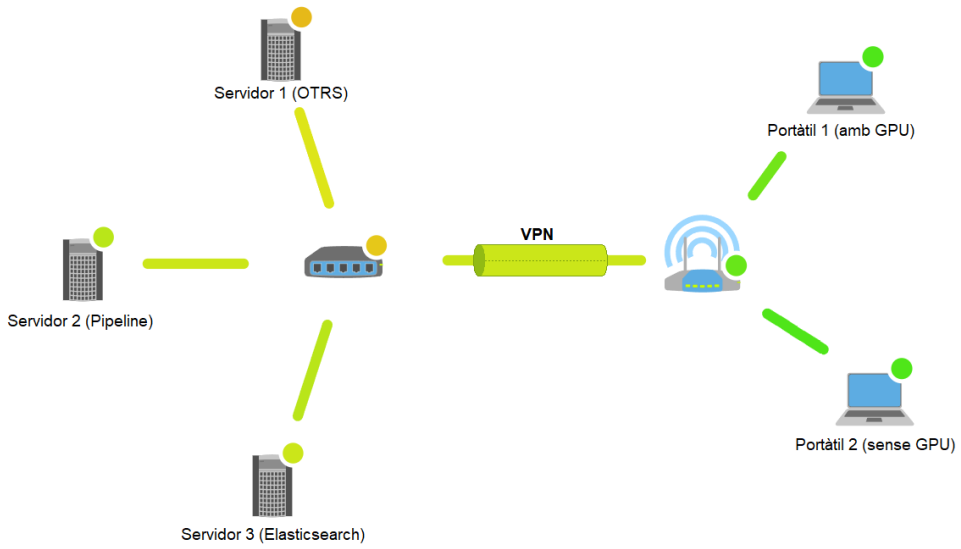


Figura 3: Diagrama de la xarxa del projecte.
(Creació pròpia amb recursos de **yWorks**)

El sistema consta d'una sèrie de scripts de Python que s'executen seqüencialment i duen a terme les tasques següents:

1. Extreure els tiquets de la base de dades OTRS de l'**Agència**.
2. Preprocessar el text de cada tiquet, eliminant el soroll (dades repetides o innecessàries), normalitzant el format i inserint les referències necessàries.
3. Aplicar un model de processament del llenguatge natural (NLP) que detecta i extreu els camps rellevants de cada tiquet.
4. Anonimitzar els camps extrets mitjançant una funció a Elasticsearch, que substitueix les dades de sortida per símbols o etiquetes genèriques.
5. Emmagatzemar els camps anonimitzats en una base de dades Elasticsearch, que és un motor de cerca i anàlisi distribuïda.

La figura 4 mostra un diagrama que il·lustra el funcionament del sistema.

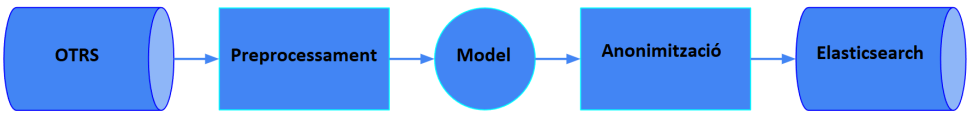


Figura 4: Diagrama de l'arquitectura del sistema.
(Creació pròpia)

En les següents seccions es detallen els components i les tecnologies que han sigut utilitzades a cadascuna de les etapes del *pipeline*.

3.1.1 Extracció de tiquets d'OTRS

Per extreure els tiquets del Servidor OTRS de l'**Agència**, s'ha fa servir *PyOTRS*, que permet accedir a les dades mitjançant peticions a través de la REST API. S'ha implementat un script de Python que utilitza la llibreria mencionada per fer les peticions i obtenir els tiquets. L'script s'encarrega d'autenticar-se amb les credencials cedides per l'**Agència**, extreure els tiquets especificats a un fitxer Excel i fer un postprocessament per aconseguir tota la informació del tiquet en un format estandarditzat i sense repeticions innecessàries. A continuació es detalla el funcionament del codi seguint l'ordre que recorren els tiquets.

1. **Iniciar sessió:** Es fan unes configuracions per assegurar que els tiquets es reben amb el format esperat i mitjançant les funcions creades a OTRS. Després es crea un client amb l'adreça IP del servidor, l'usuari i la contrasenya obtingudes de l'**Agència**.

2. **Iterar pels tiquets:** En cas que s'extreguin molts tiquets seguits es busca i s'obre el fitxer Excel especificat en el qual hi haurà la informació necessària dels tiquets que s'ha de processar. S'itera per la columna en la qual es troben els identificadors dels tiquets. També hi ha una opció per especificar un sol identificador de tiquet.

3. **Llegir ID del tiquet:** Per cada fila, es llegeix l'identificador del tiquet i es crida a la següent funció.

4. **Comprovar la utilització del tiquet:** Abans de començar a extreure els tiquets, es comprova si el tiquet en qüestió ja ha sigut extret abans. Això es fa per evitar cicles infinits amb referències cícliques dins dels tiquets. Es reinicia la llista de tiquets visitats després de cada iteració.

5. **Extreure tiquet:** S'envia la petició al servidor OTRS mitjançant la funció `get_ticket_by_number` de **PyOTRS**. El qual retorna un tiquet de la classe `Ticket` , molt convenientment, pel seu futur processament. En cas que el tiquet amb l'identificador especificat no existeixi, es llançarà un error informant del succés.

3.1.2 Preprocessament de tiquets

El preprocessament dels tiquets té com a objectiu preparar el text per ser analitzat pel model de NLP. S'ha implementat un script de Python que efectua les operacions següents sobre el text de cada tiquet:

1. **Iterar pels articles:** El processament bàsic del tiquet consisteix a iterar per tots els articles del tiquet i, per cadascun d'ells unir el cos principal de l'article amb els arxius adjunts i les referències a altres tiquets. Fent això repetit per cada article fins a arribar a l'últim i evitant sempre repetir text ja mencionat anteriorment.

2. **Processament del cos:** El primer que s'insereix és el cos principal de l'article que s'està tractant en el moment. Gràcies a una opció d'OTRS, el cos està adjuntat com un fitxer adjunt més en format HTML, permetent molta més flexibilitat a l'hora de modificar el seu contingut. Un problema que es va trobar en tots els tiquets, és que per cada article que es contesta, es torna a enviar tota la conversa fins al moment a sota. És per això que es busca l'element HTML que marca el final del tiquet (Una vora esquerra blava) i s'elimina tot el que hi ha dins. Després es converteix a text normal, s'elimina la firma (indica l'hora, lloc i altra informació irrellevant) amb una expressió regular (regex) i s'elimina tots els salts de línia sobrants.

3. **Processament dels adjunts:** Els fitxers adjunts estan tots codificats amb la codificació **Base 64**, per tant, el primer pas és descodificar el fitxer. Després depenent del seu format es fa ús de llibreries diferents per acabar amb el fitxer en format text UTF-8. No tots els fitxers poden ser processats per extreure la seva informació, per exemple, s'ha considerat que analitzar les imatges està fora de l'àmbit d'aquest projecte i, per tant, s'ha decidit tractar els seguents formats:

    - **Fitxer de text (.txt):** Per tractar els fitxers de text, simplement es pot descodificar l'arxiu amb la codificació **Base 64** i codificar el resultat binari amb la codificació UTF-8.
    - **Fitxer de correu electrònic (.msg):** Per processar els fitxers de correu electrònic, s'ha implementat una funció que extreu el contingut de l'adjunt codificat i es descodifica a partir de la codificació **Base 64**. Les dades binàries després es codifiquen com un missatge d'Outlook amb la llibreria `extract_msg` . Amb aquest format, ara es pot llegir en format text el remitent, destinatari, subjecte i cos per recrear el correu original.
    - **Fitxer PDF (.pdf):** Per llegir fitxers PDF, es descodifica el fitxer codificat a **Base 64** que representa un arxiu PDF i extreu el contingut de text del PDF descodificat utilitzant una biblioteca de lectura de PDF anomenada `PyPDF2`. El text extret de cada pàgina es concatena i es torna.
    - **Document de Word (.docx):** Aquesta funció descodifica l'adjunt amb la codificació **Base 64**. El document després es descodifica en dades binàries i fa ús de la llibreria de python `docx` per carregar el fitxer .docx des de l'objecte binari. El contingut del document es pot accedir per paràgrafs i es retorna concatenant-los tots.

- **Antic document de Word (.doc):** Per extraure el contingut del document adjunt, el descodifica fent servir la codificació **Base 64**. Les dades binàries descodificades s'escriuen en un fitxer temporal amb extensió .doc. Després, s'utilitza l'eina externa `antiword` a través d'un subprocés per convertir el fitxer temporal .doc en text pla. Finalment, es torna el text extret com a resultat de la funció.

4. **Processament de les referències:** Aquesta funció intenta buscar en el text totes aquelles referències a incidències passades per intentar recrear tot el context necessari per entendre el tiquet. Es defineix un patró d'expressió regular que cerca exactament 16 dígits, que representen un número de tiquet, i cerca totes les aparicions d'aquest patró dins de l'article. Per a cada número de tiquet identificat, es recupera informació detallada sobre el tiquet referenciat utilitzant la funció `get_ticket_by_number` i indicant, en cas que sigui necessari si ja ha sigut inserida aquesta referència anteriorment.

5. **Unir i retornar:** Finalment, s'uneixen totes les peces en un mateix text, de manera que quedarien els articles ordenats temporalment i, sota de cada article, el text dels fitxers adjunts (si han pogut ser descodificats) i els tiquets als quals es fan referència.

3.1.3 Eliminació de signatures i peus de pàgina

La majoria dels articles d'un tiquet contenien un últim paràgraf que contenia informació genèrica sobre ciberseguretat, telèfons o correus de contacte, advertències, etc. A més a més, a sobre d'aquest peu de pàgina hi havia una signatura automàtica informant de la seva posició a l'empresa, ubicació, etc.

Combinant les signatures i els peus de pàgina, hi havia articles on la majoria de text era aquesta part irrellevant del correu. Es va considerar que aquesta informació no era rellevant per a l'entrenament del model principal, ja que el fet que un correu fos confidencial o el número de telèfon d'una persona no hauria d'influir en el resultat, pel fet que no era cap dels camps que es buscava extreure. L'objectiu d'aquest model de "preprocessament"dels articles era:

- Alleugerir la càrrega computacional del model principal.
- Mostrar al model només la informació rellevant.
- Permetre entrenar amb més informació d'entrada com els adjunts dels tiquets o les referències.

Per a l'entrenament d'aquest model es va decidir treballar amb articles individuals en comptes de amb tot el tiquet, ja que les signatures són individuals de cada article i no

en general al tiquet. Això també va permetre la possibilitat d'utilitzar models amb un context més reduït.

Conjunt de dades

L'objectiu del model és rebre el text d'un article directament com surten del procés de preprocessament de tiquets i obtenir com a resultat un tiquet sense les signatures ni peus de pàgina mencionats anteriorment. S'ha plantejat diverses maneres en què un model pot aprendre aquest comportament desitjat:

- **Detectar el text:** El resultat d'aquest model és directament el text rellevant que es vol. No és necessari cap processament després d'obtenir el resultat.
- **Detectar la signatura:** El resultat del model serà la signatura present a l'article, en cas que no n'hi hagi cap, no es retornaria res. Després es pot detectar aquesta signatura en l'article d'entrada i eliminar-la. El problema que presenta aquest mètode és que en cas que el model escrigui una sola lletra incorrectament del resultat ja no es podria trobar al text original.
- **Detectar caràcter d'inici de la signatura:** S'ha vist que la majoria de les correccions que es poden realitzar a un article es troben al final. Aprofitant aquest fet, es pot entrenar un model que indiqui a partir de quina alçada s'inicia la signatura i així poder-la retallar. No obstant hi ha casos on informació rellevant com el departament de la persona que escriu l'article pot ser esborrat.
- **Classificació de frases:** Inspirat en altres projectes amb objectius similars, s'ha plantejat un conjunt de dades amb elements més reduïts i simples en forma de frase i que l'objectiu sigui classificar-les segons si formen part de la signatura o no. Utilitzant un altre model del llenguatge natural, es pot dividir un text en frases i, reutilitzant intel·ligentment les dades dels conjunts de dades anteriors, es pot evitar tornar a etiquetar. Per tenir més flexibilitat, s'ha afegit la posició de la frase dins de l'article juntament amb el nombre de frases totals de l'article. Totes aquestes dades ens poden ajudar a classificar millor si una frase pertany a una signatura.

Entrenament

A continuació es presenten els models utilitzats i quin ha sigut el conjunt de dades en el que s'ha entrenat.

Flan-T5-Base

Aquest model s'ha entrenat amb l'objectiu de detectar el text dels articles, descartant en el procés les signatures. Utilitzant les tècniques de Gradient Checkpointing i Gradient

Accumulation, s'ha aconseguit entrenar aquest model dins de la GPU de 8GB sacrificant en el procés el temps d'entrenament. S'ha entrenat durant 10 epochs que ha sigut l'equivalent a més de 24 hores d'entrenament continuat. Tot i el bon historial d'aquest model, el resultat d'aquest finetuning ha sigut un fracàs. El model no ha sigut capaç de generalitzar la tasca d'eliminar la signatura, ni en els casos on més es repetia. La causa d'aquest mal entrenament queda desconeguda, ja que aquesta tasca es considera teòricament senzilla de resoldre per un model del llenguatge natural.

FLOR-760M CONTINUAR AQUI!!!!! A causa del fet que el model **Flan-T5-Base** no va demostrar bons resultats en eliminar les signatures dels correus, es va decidr canviar de model. El model **FLOR** és un model de llenguatge causal basat en l'arquitectura transformers per a català, castellà i anglès. És el resultat d'una tècnica d'adaptació lingüística realitzada al model **BLOOM**, que implica modificar el vocabulari del model i la capa d'*embedding*, i entrenar el model contínuament amb fitxes de 26B en les llengües objectiu. S'ha fet l'entrenament amb el model de 760M per 5 epochs. En observar els resultats de l'entrenament, s'esperaven resultats acceptables pel nostre cas d'ús. Això no obstant, en fer l'avaluació en les dades de testatge, hem vist que el model no generalitza bé. Addicionalment, un altre desavantatge del model és el temps d'inferència. Hem observat que en netejar un article tarda entorn d'un minut, la qual cosa fa inviable netejar les dades amb un model de generació de text. A causa d'aquests problemes, s'ha decidit canviar el mètode per detectar les signatures en els articles dels tiquets.

Modificació RoBERTa

Com que els models generatius triguen molt a executar-se i no hem trobat resultats bons, hem decidit canviar de mètode per resoldre el problema. Hem decidit canviar el problema d'un problema de generació de text a un problema de classificació de frases. Per a classificar les frases el que es vol és convertir cada frase en un embeddings de mida fixa que mantingui el significat de la frase i utilitzar-lo per classificar-la com a una de les dues classes resultat: "pertany a una signatura"o "no pertany a una signatura". Per a fer-ho, s'ha fet servir el model RoBERTa per a generar l'embedding de la frase i s'ha emprat una sèrie de capes denses que reben aquest embedding i el processen iterativament fins a arribar a una capa de classificació amb dues neurones on la neurona més activada indica quina classificació té. Addicionalment, s'ha afegit una capa que donada la posició de la frase i el nombre de frases per arribar al final (ambdues normalitzades entre 0 i 1), augmenta la seva mida fins a arribar a la mateixa que els embeddings, així tenen una representació igual i no són menyspreats. Aquests nombres es concatenen amb els embeddings i es posen com a input a les capes de classificació per a aconseguir més informació sobre la frase. Després de moltes iteracions (més de 10) sobre l'arquitectura del model, la naturalesa de les dades i la manera d'entrenament, s'ha arribat a la conclusió que la millor precisió a la qual pot arribar és a un 87

Spacy Pipeline

En observar que els resultats del model ROBERTA no milloraven i pel fet que una classificació incorrecta, ens podria perdre informació important, vam provar un últim mètode. Vam decidir fer una solució més tradicional fent servir la llibreria Spacy. Spacy és una llibreria de processament del llenguatge natural (NLP) desenvolupada en Python que ofereix eines eficients per analitzar i processar text de manera precisa i eficaç. Amb Spacy, pots realitzar diverses tasques de NLP, com ara etiquetar de manera automàtica el tipus de paraules en un text (POS tagging), reconeixement d'entitats amb models preentrenats, anàlisi de dependències sintàctiques i molt més. Una de les funcionalitats destacades de Spacy és la seva capacitat per a la classificació de textos, que permet etiquetar automàticament els documents o fragments de text en categories prèviament definides, facilitant així la seva organització i anàlisi en aplicacions de processament de llenguatge natural. Nosaltres hem fet servir la pipeline tradicional que fa servir tok2vec en lloc de Transformers per crear els embeddings de les frases. S'ha fet servir tok2vec, ja que no requereix fer ús de GPU. En cas que haguéssim de fer servir Transformers per millorar els resultats, faria falta baixar la versió de cuda toolkit de 12.1 a 11.8. La causa d'això és el fet que encara hi ha llibreries com Cupy (essencial per entrenar Spacy amb models de Transformers) que requereixen la versió Cuda toolkit 11.8. I pel fet que per baixar la versió del Cuda Toolkit ens fa falta permisos d'administrador, hem decidit no fer servir GPU. Per això s'ha procedit a entrenar fent servir la CPU. Encara que hàgim fet servir un mètode més tradicional, s'han millorat els resultats respecte al model de Transformers i hem aconseguit una f1 score de 90. Podem observar que els resultats són impressionants considerant que no fem servir un mètode estat de l'art com poden ser els transformers a l'hora de trobar embeddings. Si en el futur considerem que hauríem de millorar el model de filtratge de signatures, crearem un tiquet al lloc de treballar per baixar la versió de Cuda toolkit de 12.1 a 11.8.

Conclusió eliminació de signatures i peus de pàgina

En conclusió, el viatge per desenvolupar un model d'eliminació de signatures eficaç ha estat ple de reptes i contratemps. Els models inicials, inclosos el Flan-T5-Base i el FLOR-760M, no van assolir les expectatives establertes, lluitant amb la generalització i els temps d'inferència poc pràctics. El posterior canvi a la classificació de frases mitjançant el model RoBERTa també va quedar curt, aconseguint una precisió que, tot i que lloable, no era suficient per al desplegament en un entorn de producció a causa del risc d'omissió informació crítica.

L'exploració de mètodes tradicionals amb el pipeline tok2vec de Spacy ha donat resultats prometedors, superant els dels models Transformer més avançats. Tanmateix, la puntuació f1 de 90, tot i ser un assoliment important, no proporciona el nivell de confiança necessari per garantir que no es perdi cap informació important durant el procés d'eliminació de la signatura. Aquesta desconfiança prové del fet que hi ha articles on la

informació més rellevant sol estar al final de l'article en una mena de resum fet per la persona que ha resolt el tiquet. La pèrdua d'aquesta informació pot reduir l'efectivitat del model d'extracció posterior. Per tant, s'ha pres la decisió de no utilitzar aquest model en el seu estat actual. A més a més, a causa de les limitacions de temps, l'enfocament s'ha desplaçat cap a la millora del model principal, que és fonamental per al projecte.

3.1.4 Aplicació del model

Per aplicar el model NLP que detecta i extreu els camps rellevants de cada tiquet, s'ha fet servir la llibreria *Transformers*.

Transformers és una biblioteca de codi obert desenvolupada per *Hugging Face* [3] que ofereix models preentrenats i eines per a tasques de processament del llenguatge natural. Aquesta llibreria facilita molt l'aplicació del model i, després d'instalar totes les dependències requerides, s'utilitza la següent commanda per inicialitzar un *pipeline*, que amaga tot el procés d'inferència:

```
t2t_gen_pipe = pipeline(
     "text2text-generation",
     model='./trained_model/',
     tokenizer='google/flan-t5-base',
     device='cuda'
     )
output_text = t2t_gen_pipe(input_ticket_str, max_new_tokens=300, beams=
                                    5)[0]['generated_text']
```

Aquest codi inicialitza un *pipeline* amb la funció de generació de text a text (la tasca del model *Flan-T5*), el model afinat anteriorment, el *tokenitzador* per defecte del model i s'indica que es faci ús dels drivers de cuda, per aprofitar tota la potència computacional de la GPU. Amb aquest *pipeline*, s'indica l'entrada, el nombre màxim de *tokens* que pot generar el model (el model té la capacitat per parar sol, però s'indica un màxim com a mesura de seguretat) i el nombre de *beams* (possibles solucions candidates, diferents unes de les altres) durant el *beam search*.

3.1.5 Anonimització de camps

L'anonimització dels camps extrets té com a objectiu protegir la privadesa i la confidencialitat de les dades personals o sensibles que puguin contenir els tiquets d'incidències de ciberseguretat. Per fer-ho, s'ha utilitzat una funció proporcionada per **i2CAT**, que rep com a paràmetre el text d'un camp i torna un text anonimitzat, que substitueix les dades per símbols o etiquetes genèriques. Per exemple, si el text del camp és "Marc Vila", la funció torna "NOM COGNOM", i si el text és "192.168.1.1", la funció torna "IP PRIVADA".

La funció d'anonimització s'ha implementat en un script de Python que rep com a entrada el fitxer amb els camps extrets pel model de NLP, i torna com a sortida un altre fitxer amb els camps anonimitzats.

3.1.6 Emmagatzematge del resultat

A l'etapa final de la cadena de processament dels tiquets, l'**Agència** ha fet servir *Elasticsearch* com a solució d'emmagatzematge per als camps extrets. *Elasticsearch* és conegut per les seves capacitats de cerca distribuïda i en temps real, cosa que el fa eficient per gestionar i consultar grans volums de dades estructurades.

El procés comença establint una connexió amb el *Servidor Elasticsearch* mitjançant una funció d'inici de sessió. Aquesta funció empra la llibreria de Python `Elasticsearch` per verificar la connectivitat amb l'adreça IP del **Servidor 2** especificada.

Un cop s'estableix una connexió amb èxit, es crida la funció per pujar els camps extrets d'un tiquet a la base de dades. La funció pren com a paràmetres un diccionari que representa els camps extrets del tiquet, la connexió Elasticsearch establerta i alguns camps necessaris com el nombre del tiquet i l'identificador.

A continuació, utilitza una funció pròpia de la llibreria `Elasticsearch` per inserir el tiquet a l'índex especificat, amb l'identificador del tiquet com a clau primària.

3.2 Creació del dataset sintètic

Durant el desenvolupament i millora dels models del projecte, es va plantejar un repte important: la manca d'un conjunt complet de dades del món real. Les dades del món real són crucials per entrenar i avaluar models que puguin funcionar en aplicacions pràctiques. L'obtenció d'un *dataset* prou ampli i divers per als models que es desenvolupen va ser un repte, ja que, com que no es disposava de les dades originals amb les quals s'hauria de treballar, es va haver de buscar un *dataset* similar. La naturalesa sensible de les dades i restriccions de propietat associades als informes d'incidències dificulten la cerca d'un conjunt de dades similar.

Per tant, es va prendre la decisió d'utilitzar un *dataset* sintètic, que és una col·lecció simulada de dades generades per imitar escenaris del món real. Tot i que no reprodueix les complexitats i matisos de les dades reals, les dades sintètiques serveixen com una valuosa eina per provar i refinar models quan les dades reals són inaccessibles o limitades.

3.2.1 Descripció general

El *dataset* escollit és *Named Entity Recognition in Indian court judgments* [4] (Reconeixement d'Entitats Nomenades a Sentències Judicials de l'Índia). Aquest *dataset* sintètic és una col·lecció seleccionada de sentències judicials que han sigut anotades centrant-se en entitats amb més rellevància dins del context judicial. Aquesta elecció ha estat guiada per les següents consideracions que l'han convertit en un candidat adequat per avaluar el rendiment dels models:

- **Entitats complexes:** Les sentències dels tribunals indis solen contenir terminologia jurídica complexa, noms diversos i algunes entitats llargues que suposen un repte per als sistemes NER. Aquesta complexitat proporciona un terreny de proves per als models, cosa que permet avaluar la seva capacitat per extreure amb precisió entitats intricades i variades, tal com apareixen als tiquets reals.
- **Sensibilitat al context:** El conjunt de dades escollit fa especial èmfasi en la importància del context en el reconeixement d'entitats. En els documents jurídics, els noms es poden referir a múltiples entitats en funció del context, una característica que reflecteix els reptes s'esperava trobar al conjunt de tiquets proporcionats per l'**Agència**. Aquesta consideració s'alineava amb les complexitats contextuals que preveiem en escenaris d'incidències on una única entitat (per exemple, una adreça de correu electrònic) pot representar diferents papers (atacant, destinatari, usuaris afectats) en funció del context. Comprendre i abordar aquests matisos específics del context ha sigut crucial per al desenvolupament de models precisos.

3.2.2 Visió general de les dades

Per com funcionen els judicis a l'Índia, una sentència típica d'un tribunal es pot dividir en dues parts: el preàmbul i la sentència. La separació entre el preàmbul i la sentència sol estar marcat per una paraula clau com "JUDGMENT" o "ORDER".

- **Preàmbul:** El preàmbul d'una sentència conté les dades importants de la sentència tals com els noms de les parts, jutges, advocats, etc. Aquestes dades no solen tenir frases gramaticalment completes, sinó que estan formatades.
- **Sentència:** Per altra banda, el text que segueix el preàmbul fins al final de la sentència es diu simplement "sentència".

Per a l'ús que se li ha donat, s'ha considerat que els preàmbuls i les sentències poden ser combinades en el mateix conjunt de dades per aconseguir més dades i més diversitat.

El *dataset* conté les següents entitats extretes:

- **TRIBUNAL:** Nom de qualsevol tribunal esmentat.
- **SOL·LICITANT:** Nom dels sol·licitants del cas actual.
- **DEMANDAT:** Nom dels demandats/acusats/oposició del cas actual.
- **JUTGE:** Nom dels jutges del cas actual i dels casos anteriors.
- **ADVOCAT:** Nom dels advocats d'ambdues parts.
- **DATA:** Qualsevol data esmentada a la sentència.
- **ORG:** Nom de les organitzacions esmentades al text a part del tribunal.
- **GPE:** Llocs geopolítics que inclouen noms d'estats, ciutats i pobles.
- **ESTATUT:** Nom de la llei esmentada a la sentència.
- **DISPOSICIÓ:** Seccions, subseccions, articles, lleis, normes d'un estatut.
- **PRECEDENT:** Tots els casos judicials anteriors referits a la sentència com a precedent.
- **NÚMERO_CAS:** Tots els altres números de casos esmentats a la sentència (a part del precedent).
- **TESTIMONI:** Nom dels testimonis de la sentència actual.
- **ALTRES_PERSONES:** Nom de totes les persones no catalogades com a sol·licitant, demandat, jutge, advocat o testimoni.

3.2.3 Preprocessament del dataset

El *dataset* originalment està en format `.json`, un format que pot ser còmode per la tasca de NER, però per generació de text és més útil tenir-ho en altres formats. Per transformar el conjunt de dades en format `.csv` s'han executat els següents passos:

1. **Llegir el *dataset*:** Per començar, s'ha carregat el *dataset* des dels fitxers JSON on s'emmagatzemen les dades utilitzant la llibreria `json`.
2. **Unir totes les sentències:** Les sentències i els preàmbuls que estaven en arxius diferents s'han unit en la mateixa variable, pel motiu mencionat anteriorment.
3. **Escriure la capçalera:** La capçalera d'un arxiu `.csv` és el títol de les columnes. En aquest cas ha sigut "sentence" i "output" pel text d'entrada i les entitats de sortida, respectivament.

4. **Iterar per les sentències:** Cada sentència conté un camp identificador, un d'anotacions, un de dades i un de metadades.

5. **Escriure el text:** S'ha escrit el text en el fitxer de sortida amb el format esperat per un `.csv`.

6. **Recórrer les anotacions:** S'ha iterat per cada anotació d'una sentència on hi ha les entitats extretes del text.

7. **Escriure les entitats:** Cada entitat, ha sigut escrita a la seva categoria separada per comes. En cas que un tipus d'entitat no tingui cap entrada, s'ha escrit `'# not_found#'`.

Aquest procés s'ha repetit pels arxius de *train* i de *test*, que estaven separats. Addicionalment, s'ha separat 1000 entrades de l'arxiu *train* en un anomenat *validation* per tenir tots els arxius necessaris per a l'entrenament.

3.2.4 Exploració de les dades

Recompte d'entitats per tipus

A partir del nombre d'entitats recopilades mostrades a la taula 4, es va veure que cap entitat domina el conjunt de dades, i que la distribució està relativament equilibrada entre les diferents classes. Encara que hi ha algunes variacions, cap dels tipus d'entitat té una majoria ni està gaire infrarepresentat.

Entitat	Recompte Sentències	Recompte Preàmbuls	Total
TRIBUNAL	1293	1074	**2367**
SOL·LICITANT	464	2604	**3068**
DEMANDAT	324	3538	**3862**
JUTGE	567	1758	**2325**
ADVOCAT	NA	3505	**3505**
DATA	1885	NA	**1885**
ORG	1441	NA	**1441**
GPE	1398	NA	**1398**
ESTATUT	1804	NA	**1804**
DISPOSICIÓ	2384	NA	**2384**
PRECEDENT	1351	NA	**1351**
NÚMERO_CAS	1040	NA	**1040**
TESTIMONI	881	NA	**881**
ALTRES_PERSONES	2653	NA	**2653**
Total	17485	12479	**29964**

Taula 4: Recompte de les entitats recopilades en les sentències, els preàmbuls i en total.

L'equilibri entre les classes d'entitats es pot veure més clarament en la figura 5 on es compara percentualment la proporció d'entitats al *dataset*.

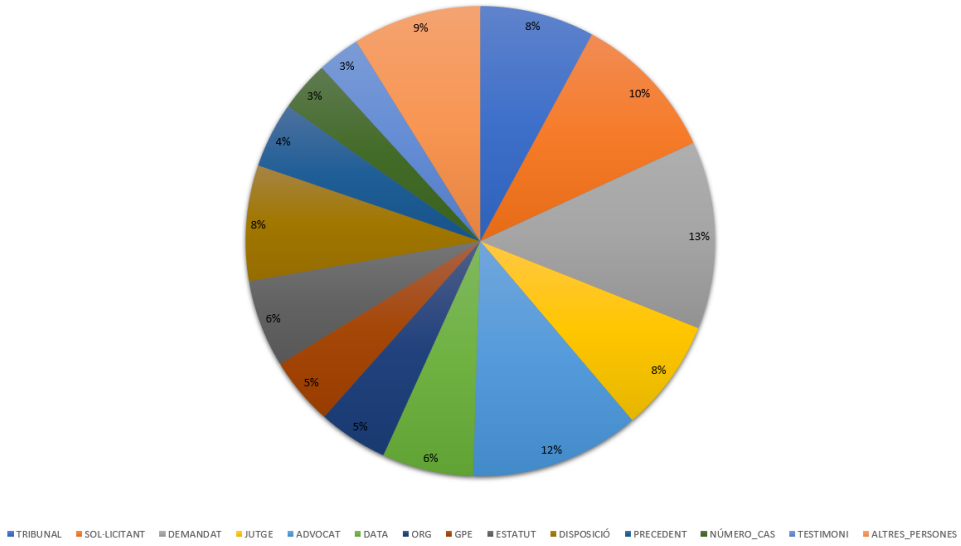


Figura 5: Diagrama de sectors comparant el nombre d'entitats a les sentències. (Creació pròpia)

Recompte de sentències per tipus de cas i per tribunal

A l'hora de recopilar les sentències, és probable que les sentències més esmentades siguin les més importants. Però limitar-se a prendre les sentències més citades d'un determinat tribunal donaria lloc a un biaix en determinats tipus de casos (per exemple, casos criminals). Per tant, cal controlar els tipus de casos per tenir en compte la varietat de sentències. Així doncs, al *dataset* hi ha els 8 tipus de casos més freqüents: civils, constitucionals, criminals, financers, industrials i laborals, de terreny i propietat, de vehicles i d'impostos.

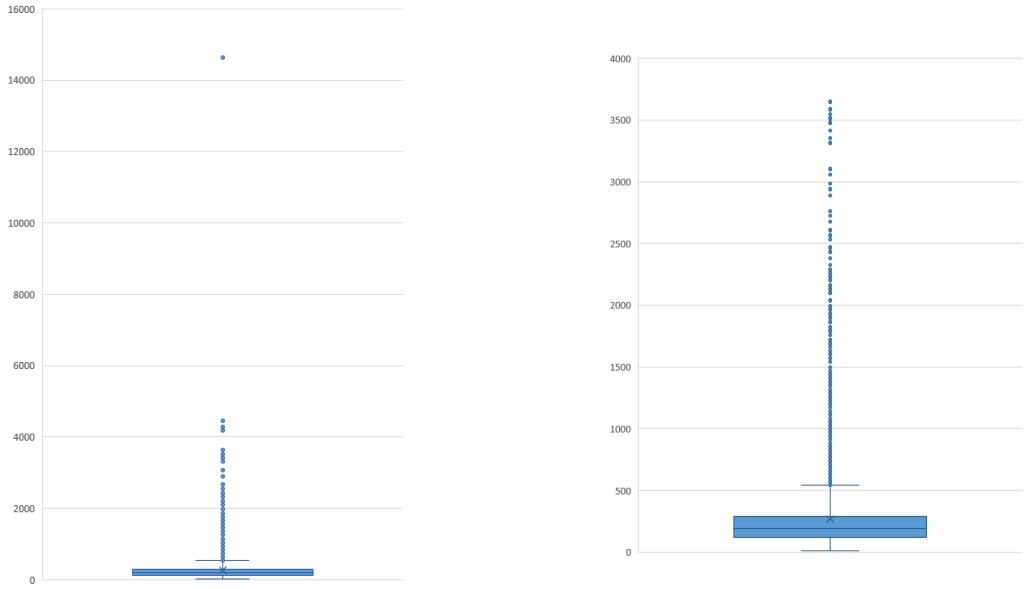
Per mostrar que aquesta distribució és equilibrada, es pot veure el recompte fet a la taula 23, on es mostra el recompte de les sentències per tipus de cas i per tribunal de les sentències.

Recompte de sentències per llargada

A l'exploració del conjunt de dades, un aspecte que ha cobrat importància és la distribució de les longituds de les sentències. Aquesta anàlisi proporciona informació sobre la variabilitat de les longituds d'entrada i ha ajudat a descobrir l'estructura general i els possibles *outliers* o valors atípics del conjunt de dades.

En primer lloc, es va generar un *boxplot* (diagrama de caixa) per visualitzar les longituds dels texts d'entrada del *dataset* com es pot veure a la figura 6a. El diagrama de caixa va

revelar un valor atípic: una sentència que era significativament més llarga que totes les altres. Es va descobrir que aquesta frase era un error, i que no tenia sentit. A més a més, es van identificar quatre sentències atípiques les quals superaven els 4.000 caràcters. Es va decidir eliminar aquests valors atípics per evitar anàlisis esbiaixades i garantir que el model s'entreni amb dades representatives i significatives.



(a) Amb totes les sentències.

(b) Sense *outliers*.

Figura 6: Boxplot de la longitud de les sentències, abans i després d'eliminar els *outliers*. (Creació pròpia)

Després d'eliminar els valors atípics identificats, es va generar un nou *boxplot* per visualitzar la nova distribució de la longitud de les frases 6b.

Addicionalment, a la figura 7 es pot veure que s'ha creat un histograma per oferir una visió més detallada de la distribució de freqüències dins dels diferents rangs de longitud.

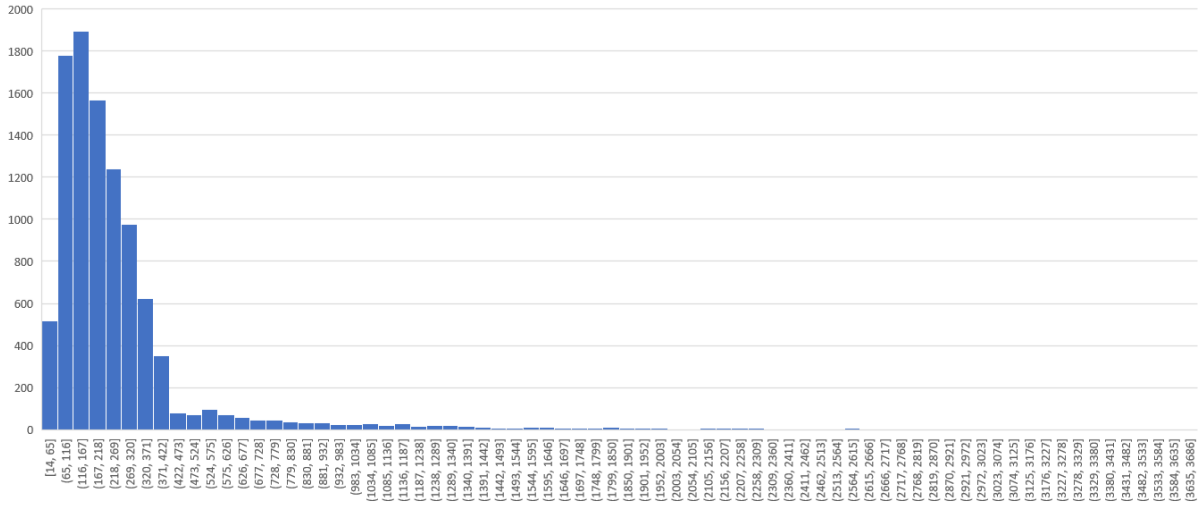


Figura 7: Histograma de la longitud de les sentències, sense *outliers*.
(Creació pròpia)

Recompte d'entitats per llargada

L'anàlisi de la longitud de les entitats proporciona informació sobre l'estructura i la composició del *dataset*, revelant possibles patrons i anomalies. L'histograma a la figura 8 il·lustra la distribució de les longituds de les entitats entre els diferents tipus. Cada barra representa un interval de longituds de dos caràcters per un tipus concret d'entitat, i l'alçada de la barra indica el recompte d'entitats compreses en aquest interval de longitud.

Observant la figura 8 es pot veure que la categoria "DEMANDAT" ha destacat per mostrar un nombre significatiu d'entitats amb longituds superiors o iguals a 100 caràcters.

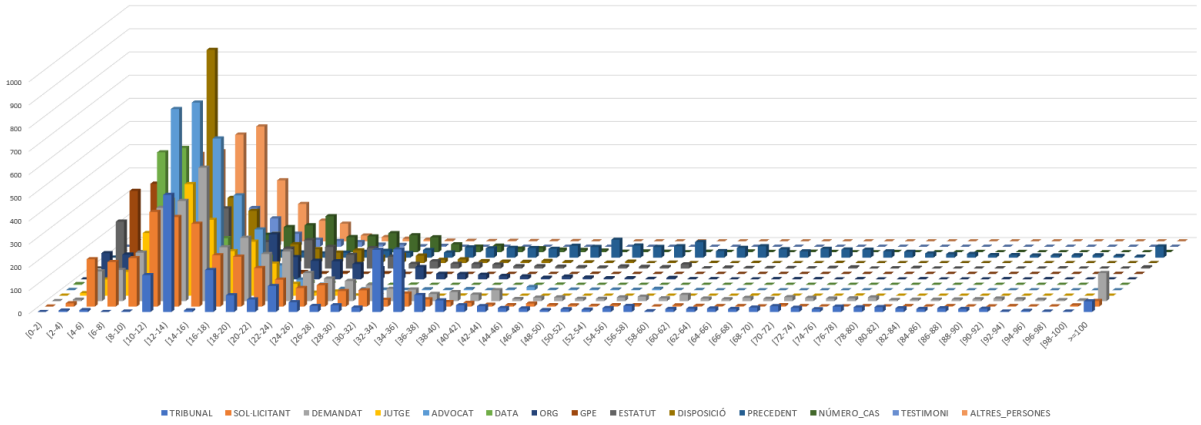


Figura 8: Histograma de la longitud de les entitats, per categoria.
(Creació pròpia)

Tal com es veu la figura 9, s'ha generat també un histograma complet combinant tots els tipus d'entitats. L'histograma unificat proporciona una visió global de la distribució

de la longitud a tot el *dataset*. Les entitats de diferents categories tendeixen a tenir una longitud semblant. Més concretament, el 55% de les entitats tenen entre 8 i 18 caràcters i el 90% de les entitats en tenen menys de 36. Aquesta uniformitat en la distribució de la longitud suggereix que les entitats es representen de forma coherent en el *dataset*.

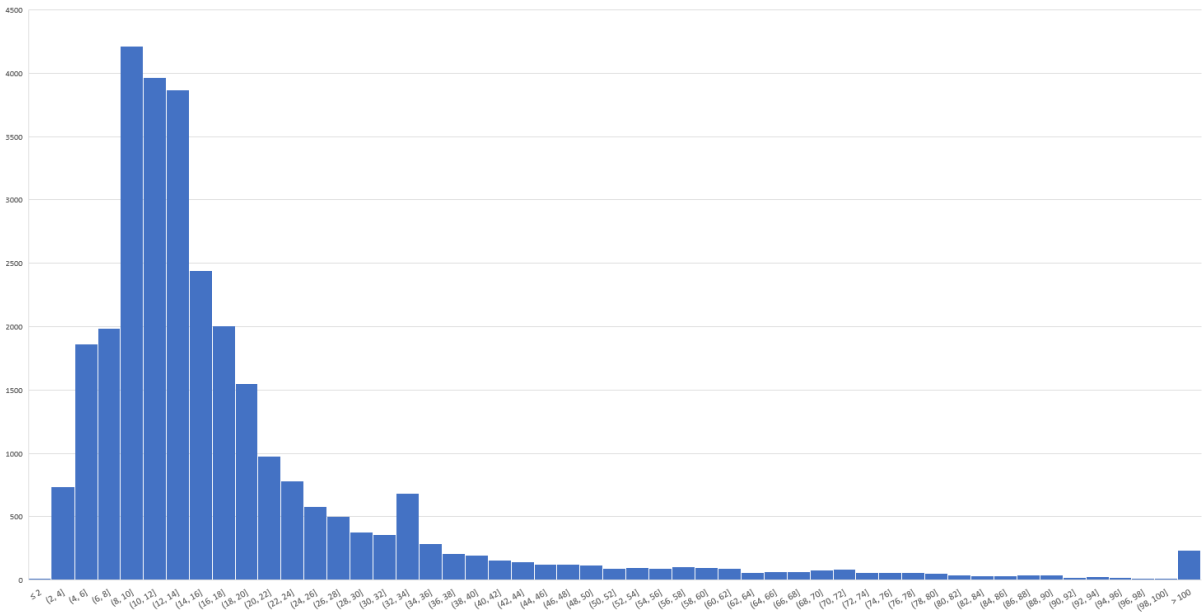


Figura 9: Histograma de la longitud de totes les entitats.
(Creació pròpia)

3.3 Models destacats

De la comparació dels models a l'apartat 2.3.3 n'hi ha uns quants que han destacat per sobre dels altres. Aquests models han continuat per les proves de selecció i han sigut sotmesos a més testos i a afinaments per quantificar el seu rendiment. D'aquesta manera, es pot escollir el model més adequat pel projecte de manera acurada.

A causa del retard en l'arribada de les dades, aquest apartat es divideix en dues seccions diferents. La primera secció presenta els models comparats utilitzant dades de prova utilitzant el hardware disponible a l'oficina en el moment el qual només disposa de CPU. La segona secció avalua uns nous models amb les dades reals, aprofitant el portàtil equipat amb una GPU procedent de l'Agència.

3.3.1 Comparació de models amb dades sintètiques

Mistral 7B (Instruct)

Mistral 7B Instruct [5] és un model del llenguatge natural que està afinat a partir del model *Mistral 7B*. Aquest reentrenament es va realitzar, segons expliquen els creadors, per demostrar la facilitat que té el model original per adaptar-se a qualsevol tasca. Aquesta és una de les raons que va cridar molt l'atenció aquest model.

Un altre motiu pel qual aquest model és tan interessant és que afirma ser el millor model de la seva mida. Això ha sigut comprovar comparant aquest model amb els més populars de la seva categoria en els *datasets* més populars.

El problema que té aquest model és que té massa paràmetres i tardaria massa temps a ser afinat correctament. De fet, el model sencer no es pot fer servir a les màquines disponibles, havent de recórrer a la versió quantificada del model per poder fer la inferència.

Aquest model s'ha executat més concretament utilitzant el programa *localGPT* [6]. El programa permet, en essència, parlar a un assistent sobre un document indicat. *LocalGPT* primer ingereix un document i després deixa fer preguntes sobre el document, tenint l'habilitat de comprendre'l, llegir-lo i extreure informació per contestar tots els dubtes. A més a més, pot recordar el context de la conversació, el que significa que es poden fer preguntes encadenades fent referència a les anteriors respostes.

En les proves, ha donat molt bons resultats, encertant en gairebé totes les preguntes. El seu temps d'execució és bastant lent, però acceptable pels objectius que es plantegen. Finalment, aquest model s'ha descartat, ja que no es pot assegurar que les màquines de l'**Agència** tinguin la capacitat per executar aquest model.

Llama 2

Llama 2 [7] és un model NLP de codi obert desenvolupat per *Meta*. Aquest model va ser publicat afirmant ser el millor model durant el seu temps en moltes proves de rendiment. Té $7000M$ paràmetres, fent impossible la seva execució localment amb els dispositius disponibles. Per aquest motiu, s'executa la seva versió quantificada. Dona resultats acceptables, però lluny de ser la solució del problema i, molt inferior comparat amb *Mistral 7B*, el seu principal competidor.

Aquest model també és compatible amb el programa *LocalGPT* [6] i, és per això i per la seva mida similar que és tan comparat amb el model *Mistral 7B*. *Llama 2* és el model que està seleccionat per defecte al programa *localGPT*, i per aquesta raó va ser el primer a ser provat. En veure la millora significativa comparat amb el seu model competidor, va ser ràpidament substituït per aquesta versió millorada seva.

Flan-T5-Base

El model *Flan-T5-Base* [8] és un model de processament del llenguatge natural que presenta millores en tots els aspectes respecte al seu model previ: *T5 Base*. Els seus desenvolupadors van escriure:

> "Amb T5, proposem reformular totes les tasques de NLP en un format unificat de text a text on l'entrada i la sortida són sempre cadenes de text, a diferència dels models d'estil BERT que només poden donar com a sortida una etiqueta de classe o un tram de l'entrada. El nostre marc de text a text ens permet utilitzar el mateix model, funció de pèrdua i hiperparàmetres en qualsevol tasca NLP."

Un punt important d'aquest model és que està entrenat amb 60 llenguatges. Això significa que té una comprensió molt gran de tots els texts que se li puguin donar. A la vegada, no interessa tant que entengui gaires idiomes, ja que el format d'entrada estarà molt centrat en un idioma.

Durant els experiments, ha demostrat una capacitat de comprensió del llenguatge i del tiquet. El principal problema és que, en la majoria dels casos, no compleix correctament l'objectiu especificat, a excepció d'una tasca en concret. Gràcies al potencial que s'ha vist en el model, es continua el procés, afinant el model amb una tasca de NER.

L'afinament ha finalitzat satisfactòriament aportant un *rouge1* de $0,96$. Fent proves amb inferència s'ha vist que aquest model contesta correctament a la majoria, però ha fallat sobretot perquè repeteix més d'una vegada les respostes.

Flan-T5-Small

Flan-T5-Small [8] és el germà més petit de la família *Flan-T5*. Tenint això en compte, és lògic pensar que donaria un pitjor rendiment que els altres models. El raonament darrere de la decisió de provar aquest model a fons és aconseguir un model que seria més ràpid, tant en entrenament com en inferència, a canvi de perdre poca precisió en les respostes.

Aquest model, tal com s'havia vist, té la capacitat de formular frases senzilles, però en una primera instància no ha sigut capaç de respondre cap pregunta correctament. Després de l'afinament, s'ha comprovat que no ha assolit trobar els patrons més bàsics de les dades, donant així un resultat negatiu i descartant el model per futurs passos.

Flan-T5-Base (LaMini)

Repetint el mateix procés que amb la seva versió inferior, *Flan-T5-Base (LaMini)* [8] [9] és la versió afinada del model *Flan-T5-Base* amb el *dataset* per a seguir instruccions

LaMini. A causa dels bons resultats reportats pels mateixos creadors, amb un augment del rendiment en totes les proves comparatives, es decideix afinar aquest model, en cerca d'un rendiment superior als resultats obtinguts al *Flan-T5-Base*.

S'ha de destacar que aquest model en les seves primeres proves no va destacar per la seva capacitat pràctica de resoldre els reptes plantejats. De fet, va ser més aviat les promeses teòriques que van impulsar a continuar amb el desenvolupament d'aquest model, ja que, en un principi va demostrar un rendiment similar o lleugerament inferior que el model del qual parteix, *Flan-T5-Base*.

Flan-T5-Small (LaMini)

El model *Flan-T5-Small (LaMini)* [8] [9] és una versió afinada del model *Flan-T5-Small* amb el *dataset* per a seguir instruccions *LaMini*. Amb aquest afinament, el model és capaç de millorar respecte al model del qual ha partit. Els creadors afirmen que té el millor rendiment general donat la seva mida.

Malauradament, no s'ha notat cap millora respecte al model anterior. No només no ha sigut capaç de generar text relacionat amb el text donat sinó que no ha contestat a les preguntes seguint el format amb el qual ha sigut entrenat. Deguts aquests motius, i veient que és el model més potent de la seva categoria, ha quedat clar que és necessari més paràmetres per poder completar aquesta tasca correctament. El descens sorprenent del rendiment observat durant la prova del model *Flan-T5-Small (LaMini)* per a la resolució d'una tasca específica, planteja diverses causes potencials per a aquest resultat:

- **Qualitat del model:** El conjunt de dades *LaMini* pot tenir característiques perjudicials per al rendiment a la tasca entrenada. Atès que el *LaMini dataset* ha sigut generat sintèticament, és possible que no capturi amb precisió la complexitat i els matisos dels escenaris del món real per a dur a terme correctament les proves en qüestió. La naturalesa sintètica de les dades podria introduir biaixos o patrons poc realistes que afectin negativament la capacitat del model per generalitzar ara la tasca específica.
- **Rendiments decreixents a l'afinament:** L'ajust repetitiu del model pot haver resultat en un ajust excessiu, disminuint-ne l'adaptabilitat durant les proves. Els múltiples cicles d'afinament podrien haver trobat inicialment els patrons més rellevants, de manera que quan s'ha ajustat per última vegada, oferiria millores limitades i introduiria soroll potencialment. Tot i destacar en tasques generals de seguiment d'instruccions, el conjunt de dades *LaMini* pot no ajustar-se als requisits específics de la tasca.

3.3.2 Comparació de models amb dades reals

FLOR-6.3B-Instructed

A causa del fet que el camp de la IA generativa és un camp que evoluciona molt ràpidament, encara que se seleccioni un model pel seu entrenament, es continua observant els diferents models que van sortir. Com s'explica en l'apartat de riscos potencials 1.2.4, es va demanar permisos per la instal·lació de WSL (Subsistema de Windows per Linux) i va trigar un mes fins a donar permisos.

A causa d'això, s'ha buscat models alternatius que es puguin entrenar amb llibreries comunes disponibles a Windows. Aquesta cerca és bastant peculiar, ja que requereix els pesos del model seleccionat càpiga en una GPU de 8 GB i a més a més que pugui entendre el català, castellà i, en certa manera, l'anglès. A part del model *Flan-T5*, s'ha trobat el model *FLOR*, desenvolupat pel projecte AINA. Aquest model té diverses mides que són de 780M, 1.3B i 6.3B.

S'ha optat per entrenar el model FLOR 1.3B per l'extracció dels camps que s'han definit en el projecte. Es creu que aquest model hauria de comportar-se millor que el QWEN-1.5, ja que se centra específicament en els 3 idiomes que són rellevants pel problema. En canvi, el QWEN-1.5 té molts més idiomes i podria ser pitjor en idiomes com el català.

Tal com s'havia comprovat a les primeres proves, el model no entén el format de sortida, el qual és crucial per poder completar el procés d'anonimització i emmagatzematge. Fins i tot després d'un entrenament de 10 hores no arriba als estàndards mínims requerits.

Qwen1.5-7B

El model *Qwen1.5* s'ha provat després de dur a terme les proves d'entrenament en la màquina proporcionada per l'Agència. Aquesta màquina té 8 GB de GPU, la qual cosa és una gran limitació per entrenar models LLM. Per tant, després de fer proves amb els models anteriors, es va observar que el model *Mistral* és un model massa gran per a poder-lo entrenar en la GPU d'aquesta màquina. Els models de la família *Flan-T5* tampoc tenen suport GPTQ, cosa que ens permet escalar els pesos dels models per poder-los entrenar en màquines amb menys recursos. Per tant, el màxim que es va poder entrenar va ser el model *Flan-T5* utilitzant els adaptadors LoRa i PEFT. Aquest model té 250M paràmetres, la qual cosa va demostrar no ser suficient per a aquest projecte.

Amb el ràpid creixement d'aquest camp, nous models apareixen cada setmana. Entre aquests models, es va trobar el model *Qwen1.5*, un model que s'assembla al *Flan-T5* en la seva filosofia multilingüe i multitasca, però promet resultats molt millors. Aquest model té 6 versions: 0.5B, 1.8B, 4B, 7B, 14B i 72B. A més a més, té suport per molts idiomes, incloent-hi el català. Després de fer proves, es creu que la versió 7B quantitzada i

entrenat amb *QLoRA* pot ser idònia per entrenar en aquesta màquina. La raó per la qual es van poder entrenar models molt més grans que el *Flan-T5* base és perquè té versions GPTQ, cosa que el *Flan-T5* no disposa.

Finalment, cal esmentar que aquest model permet un context de fins a trenta-dos mil tokens, la qual cosa és un avantatge, ja que podrem entrenar el model amb el tiquet sencer. A diferència del T5, on hauríem creat el model per article del tiquet, cosa que hauria provocat la pèrdua del context global del tiquet.

3.3.3 Justificació de la tria

L'entrenament del model final va començar amb *FLOR 6.3B*, ja que era el millor candidat pel fet d'estar entrenat en els 3 idiomes que es requereixen en aquest projecte. Una conseqüència positiva de només estar centrat en aquests tres idiomes és que el model és lleugerament inferior en mida que el model original.

L'entrenament del model es va realitzar durant unes 10 hores pel fet que la versió GPTQ no estava disponible. Addicionalment, cal mencionar que les dades es van preparar de forma que l'output sigui un JSON. Això no obstant, a l'hora d'avaluar els resultats del model entrenat, el model només continuava el tiquet de l'input amb dades addicionals que no existeixen en el tiquet original. Per tant, aquest model no va treure els resultats i la qualitat que s'esperava. En conseqüència, es va decidir descartar aquest model, ja que més es valora és que els resultats surtin de forma organitzada com un JSON.

Pel fet que el *FLOR 6.3B* no ha demostrat uns resultats molt aptes per aquest projecte, es va decidir centrar l'atenció en l'últim model *Qwen1.5*. Es va entrenar el model de 7B i l'entrenament d'aquest model va ser més ràpid que els altres gràcies al fet que s'utilitzava la versió quantitzada i entrenant només un nombre petit de paràmetres. En analitzar els resultats, s'observa que el model és capaç de treure un JSON en bon format i a més és capaç de treure algun dels camps que hi ha en el tiquet. És per això que es decideix escollir aquest model per l'entrenament i futura implementació al resultat final.

3.4 Fine-tune del model

En aquesta secció s'analitza el procés de *fine-tuning* del model *Flan-T5* (i les seves diferents versions) per millorar la seva capacitat d'extracció d'entitats. Aquesta secció explica com s'ha preprocessat el conjunt de dades, postprocessats els resultats i com s'ha configurat i entrenat el model amb el *dataset* escollit.

3.4.1 Preparació de les dades

A continuació es mostren els primers passos pel *fine-tuning* del model:

1. **Carregar *dataset*:** Es llegeixen els fitxers on s'ha emmagatzemat les dades en format `.csv`.

2. **Carregar *tokenitzador*:** S'utilitza la funció `from_pretrained` de la llibreria `transformers` per carregar el *tokenitzador* del model oficial de Google: "google/flan-t5-small". Aquest model ha canviat segons el que s'entreni en cada moment, agafant inclús models no oficials de l'empresa però compatibles igualment amb l'estructura emprada.

Preprocés del *dataset*

Aquest apartat explora el procés de preparació final del conjunt de dades abans del *fine-tuning*. Els objectius principals són determinar les longituds de les entrades i sortides posteriors a la *tokenització* i articular els passos necessaris per elaborar la funció de preprocessament. Les tasques pertanyents a aquest apartat de preprocessament són les següents:

3. **Calcular la longitud de les dades:** S'ha hagut de determinar la llargada màxima de l'entrada, ja que serà necessària per a la pròxima tasca. Això és perquè s'ha d'establir una llargada màxima pel model, pel fet que no pot llegir una entrada indefinidament llarga. Aquest càlcul s'ha fet dues vegades, per les dades d'entrada i per les de sortida, agafant totes les dades, *tokenitzant-la* amb el *tokenitzador* i agafant el mínim entre la seva llargada i el màxim acceptat pel model. S'ha hagut de retallar la mida de les dades ja que

4. **Funció de preprocés:** Abans de passar l'entrada pel model, s'ha de modificar les dades perquè el model les pugui entendre. Primer s'ha afegit una instrucció anomenada *system_prompt*, que indica al model la tasca que ha de fer. Després, s'ha *tokenitzat* les sentències i els resultats truncant a la llargada màxima calculada anteriorment. Aquesta mateixa funció de *tokenització* també afegeix un *padding* a tots aquells textos que són menors que la llargada màxima, fent que siguin tots de la mateixa mida.

Postprocés i avaluació del resultat

En aquesta secció s'analitza la fase de postprocessament del nostre model. L'objectiu és millorar la precisió mitjançant l'aplicació de dues funcions diferents de postprocessament

en diferents fases de l'entrenament. Per avaluar objectivament l'eficàcia del model, s'ha utilitzat la mètrica *ROUGE*, una mesura del processament del llenguatge natural per avaluar la similitud entre dos textos. Aquests han sigut els passos:

5. **Carregar el model:** El model ha sigut carregat fent servir el mateix mètode per carregar el *tokenitzador*, i és fent ús de la funció `from_pretrained` i s'ha obtingut de la mateixa font "google/flan-t5-small" (o la que es faci servir en el moment).

6. **Primera funció de postprocessament:** La primera funció de postprocessament senzillament descodifica els *tokens* generats pel model en text, neteja qualsevol excedent d'espai que hi hagi i el divideix en frases per poder ser avaluades.

7. **Segona funció de postprocessament:** La segona funció és pràcticament idèntica a la primera, però elimina les etiquetes que escriu el model abans de donar la resposta. Això s'ha fet perquè experimentalment s'ha vist que si el model simplement escriu les etiquetes, ja aconsegueix un resultat elevat. Eliminant aquest factor de l'avaluació permet assolir un resultat quantificat més representatiu de la realitat. Aquesta segona funció només es pot aplicar quan el model ja ha sigut entrenat per suficient temps i sempre genera, com a mínim les etiquetes esperades.

8. **Funció d'avaluació:** Per a avaluar el rendiment del model s'ha usat la mètrica *ROUGE* a través de la llibreria *evaluate*. Aquesta funció retorna les mètriques de *ROUGE-1*, *ROUGE-2*, *ROUGE-L* i *ROUGE-Lsum*.

3.4.2 Configuració i entrenament

Aquesta última part se centra en la configuració i l'execució del *fine-tuning*. A continuació es mostren els passos que s'han pres:

9. **Configuració per l'entrenament:** Les configuracions més rellevants utilitzades han sigut:

    - **Batch size:** Ha sigut igual tant per entrenament com per inferència per evitar incrementar l'ús de la memòria el màxim possible. En els models més petits s'ha arribat a usar 4 però en la resta, només 1.
    - **fp16:** La precisió *fp16* hauria permès utilitzar altres configuracions més adequades, però per un error de compatibilitat, ha hagut d'estar desactivat (`False`).
    - **Learning rate:** La taxa d'aprenentatge ha estat igual que en l'entrenament original del model [8]: $5e-4$.

- **Epochs:** El nombre de vegades que les dades passen per l'algorisme d'entrenament. Ha estat 3 per la primera funció de postprocessament i 2 més per la segona, donant un total de 5.
- **Optimitzador *Adam*:** S'ha emprat l'optimitzador *Adam* per a calcular el descens del gradient, amb els paràmetres per defecte.
- **Màxima generació de *tokens*:** S'ha incrementat fins a 300 la generació de *tokens* del model per permetre que en l'avaluació pugui fer ús de tot el seu potencial.

10. **Buidar la memòria:** Per assegurar que el model cap en la limitada memòria disponible, abans de cada entrenament, s'ha buidat la memòria de la GPU eliminant la memòria cau (memòria *cache*) i forçant la recollida de la memòria brossa (*garbage collection*).

11. **Iniciar el *fine-tuning*:** Es comença el *fine-tuning* especificant el model, els paràmetres d'entrenament, les dades d'entrenament i la funció d'avaluació. El procés triga diverses hores, en aquest cas, han sigut unes tres hores i mitja per cada model petit i set hores i mitja per cadascun dels grans.

Capítol 4

Avaluació dels models

4.1 Anàlisi dels resultats

En aquesta secció s'ofereix una anàlisi dels resultats obtinguts amb els models que han sigut entrenats anteriorment amb el procediment de *fine-tuning* explicat a la secció 3.4.

Alguns resultats s'han perdut durant el procés d'entrenament. Aquesta pèrdua de dades és deguda a les característiques de la plataforma utilitzada per executar el codi, els documents *Jupyter Notebook*, que tant en un entorn local o mitjançant *Google Colab* és possible no assolir aquests resultats si no es prenen les mesures adequades. Malauradament, aquests resultats d'entrenament perduts no es poden recuperar sense tornar a entrenar els models. Aquest desafiament inesperat ha demostrat la importància de la transparència en el procés de recerca, tot reconeixent tant els èxits com els reptes trobats durant l'experimentació.

En cada secció es poden veure dos tipus de taules:

- **Taula d'entrenament:** Es mostra el progrés que s'ha aconseguit guardar de cada entrenament. Al final de cada *Epoch* s'indiquen les mètriques de *Train_ Loss* (escurçat a *Train*), *Validation_ Loss* (escurçat a Val), *ROUGE-1*, *ROUGE-2*, *ROUGE-L* i *ROUGE-Lsum*.
- **Taula d'avaluació:** Es fa una recopilació dels resultats obtinguts amb cada *checkpoint* o punt de control que s'ha recuperat. A part de les mètriques de *ROUGE*, es va afegir la mètrica de "Coincidència exacta" (escurçat a "Exacte") que simplement indica el percentatge de resultats generats que han coincidit exactament amb el resultat real.

4.1.1 Flan-T5-Base

L'entrenament inicial de la versió base del model *Flan-T5* es resumeix a la taula següent, que mostra les mètriques clau durant les primeres *Epochs*.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	0,058	0,045	94,662	92,672	93,843	93,955
2	0,037	0,036	96,162	94,540	95,625	95,725
3	...	...	...	...	...	...

Taula 5: Primer entrenament del model *Flan-T5-Base* (amb etiquetes)

Després d'aquest entrenament inicial, el model es va sotmetre a dos cicles addicionals. Aquest segon entrenament s'ha realitzat amb la segona funció de postprocessament, la qual no té en compte les etiquetes a l'hora de calcular l'error amb les diferents mètriques.

Els resultats de l'avaluació després d'aquest període d'entrenament estès es detallen a la taula següent.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Exacte
Model base	0,00	0,00	0,00	0,00	0,00%
Checkpoint 2	95,84	94,27	95,19	95,20	61,82%
Checkpoint 3	96,70	95,21	96,20	96,21	64,99%
Checkpoint 4	95,97	94,50	95,37	95,38	64,24%
Checkpoint 5	96,57	95,17	96,05	96,09	66,01%

Taula 6: Avaluació del model *Flan-T5-Base* (sense etiquetes)

Observant el resultat, es va poder veure que el model va mostrar una millora del rendiment necessària, com indica l'augment de les puntuacions *ROUGE* als punts de control següents. La puntuació de coincidència exacta, una mesura de la capacitat del model per proporcionar respostes literals, també mostra una tendència positiva, assolint el seu punt màxim al *Checkpoint 5*. La raó per la qual es creu que el model va disminuir el seu rendiment en finalitzar el *Checkpoint 4*, és perquè va ser la primera *epoch* on el model s'avaluava sense les etiquetes i això pot haver causat una disminució de la puntuació, però no necessàriament de la qualitat del model.

4.1.2 Flan-T5-Base (LaMini)

L'entrenament inicial, com es mostra a la taula següent, dona resultats prometedors després de només la segona *epoch*.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	...	...	...	...	...	...
2	0,013	0,022	97,800	96,553	97,248	97,350

Taula 7: Primer entrenament del model *Flan-T5-Base (LaMini)* (amb etiquetes)

L'entrenament posterior sense etiquetes mostra un refinament més gran de les capacitats del model, amb la quarta i cinquena *epochs* mostrant les següents tendències. Es pot observar un lleuger empitjorament de les mètriques, que segurament és degut al canvi a no utilitzar les etiquetes pel càlcul d'aquestes.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
3	...	...	...	...	...	...
4	0,012	0,029	96,594	95,267	96,109	96,136
5	0,005	0,029	96,794	95,538	96,257	96,290

Taula 8: Segon entrenament del model *Flan-T5-Base (LaMini)* (sense etiquetes)

La taula següent resumeix els resultats de l'avaluació, que demostren la millora contínua del model a través dels diversos *checkpoints*.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Exacte
Model base	0,00	0,00	0,00	0,00	0,00%
Checkpoint 1	95,31	93,48	94,76	94,75	58,28%
Checkpoint 2	96,74	95,37	96,18	96,18	67,87%
Checkpoint 3	96,39	95,03	95,90	95,90	66,66%
Checkpoint 4	96,55	95,22	96,08	96,06	66,75%
Checkpoint 5	96,75	95,45	96,19	96,19	69,36%

Taula 9: Avaluació del model *Flan-T5-Base (LaMini)* (sense etiquetes)

Els resultats mostren que tot i que algunes mètriques no han progressat com s'esperava, hi ha hagut un increment en la valoració de coincidències exactes que asseguren que, finalment, el *fine-tuning* ha sigut exitós.

4.1.3 Flan-T5-Small

A la taula següent es resumeix el procés d'entrenament del model *Flan-T5-Small* al llarg de cinc *epochs*.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	0,084	0,041	94,86	92,50	93,70	93,84
2	0,055	0,036	96,13	93,95	94,94	95,06
3	0,043	0,033	96,61	94,74	95,59	95,73
4	0,040	0,032	96,62	94,81	95,74	95,89
5	0,037	0,032	96,58	94,78	95,71	95,86

Taula 10: Entrenament del model *Flan-T5-Small* (amb etiquetes)

El progrés del model durant l'entrenament indica una disminució dels errors tant a *Train* com a *Val* al llarg de les cinc *epochs*, cosa que suggereix que ha après eficaçment de les dades etiquetades. Per altre banda, hi ha hagut una lleugera disminució de les mètriques de *ROUGE* durant l'última *epoch*, que podrien donar a entendre que el penúltim *checkpoint* pot tenir millor rendiment que l'últim. Aquesta última afirmació va ser contrastada comprovant l'avaluació dels punts de control mostrats a la següent taula.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Exacte
Model base	0,00	0,00	0,00	0,00	0,00%
Checkpoint 4	95,07	93,20	94,40	94,40	54,09%
Checkpoint 5	95,07	93,17	94,39	94,39	54,74%

Taula 11: Avaluació del model *Flan-T5-Small* (sense etiquetes)

Després de l'avaluació sense etiquetes, s'ha pogut comprovar que el model ha decidit comprometre una lleguera part de puntuació *ROUGE* per aconseguir una última millora de les coincidències exactes.

4.1.4 Flan-T5-Small (LaMini)

La taula següent presenta el progrés de l'entrenament del model Flan-T5-Small amb el *fine-tuning LaMini*, incloses les mètriques clau de l'error de *Train* i *Val*, així com les puntuacions *ROUGE*.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	...	...	...	...	...	...
2	...	...	...	...	...	...
3	0,066	0,043	95,55	92,75	93,87	94,12
4	0,064	0,042	95,61	92,88	93,93	94,19
5	...	...	...	...	...	...

Taula 12: Entrenament del model *Flan-T5-Small (LaMini)* (amb etiquetes)

Després de l'entrenament continuat, es va avaluar el model sense etiquetes. Els resultats es resumeixen a la taula següent.

Taula 13: Evaluació (sense etiquetes)

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Exacte
Model base	0,00	0,00	0,00	0,00	0,00%
Checkpoint 1	90,27	86,79	88,56	88,59	31,09%
Checkpoint 2	92,51	89,48	91,03	91,05	37,80%
Checkpoint 3	92,83	89,96	91,39	91,41	40,31%
Checkpoint 4	93,36	90,62	92,07	92,08	43,01%
Checkpoint 5	93,48	90,77	92,16	92,16	43,38%

Taula 14: Avaluació del model *Flan-T5-Small (LaMini)* (sense etiquetes)

El model mostra una millora progressiva al llarg dels *checkpoints* d'entrenament, amb puntuacions ROUGE i percentatges de "coincidència exacta" creixents.

4.1.5 Flan-T5-Base (LoRA)

El model Flan-T5-Base es va ajustar mitjançant el mètode LoRA (*Low-rank Adaptation of LLMs*), que es va centrar a entrenar únicament les matrius "q", "v", "k" i "o" del mòdul d'atenció del model. Aquests paràmentres entrenables només representen l'1,41% de tots els paràmetres del model, reduïnt en gran mesura la càrrega computacional.

Epoch	Train	Val	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	0,114	0,042	20,847	16,240	20,715	20,733
2	0,057	0,036	20,757	16,301	20,672	20,673
3	0,048	0,034	20,748	16,288	20,655	20,660

Taula 15: Entrenament del model *Flan-T5-Base* amb LoRA (amb etiquetes)

El *fine-tuning* amb el mètode *LoRA* mostra un patró clar en els errors de *Train* i *Val*, cosa que indica una millora gradual del rendiment del model al llarg de les *epochs* especificades. No obstant això, les puntuacions ROUGE mostren una tendència diferent, cosa que suggereix que la capacitat del model per captar el matisos del *dataset* i generar resultats rellevants es pot veure qüestionada en aquest entrenament. És per aquest motiu que es va decidir cancelar el entrenament i no continuar amb el procés.

4.2 Eficàcia de la solució

4.2.1 Comparació dels models

El model que millors resultats ha donat ha sigut, *FLan-T5-Base (La Mini)*. Comparant les mètriques de l'últim *checkpoint*, es pot veure que supera en totes les categories al *FLan-T5-Base*.

Per altra banda, si es compara amb la mida del model, el model *FLan-T5-Small* ha donat el millor resultat tenint en compte que és 3 vegades més petit i només ha disminuït la seva mètrica de coincidència exacta 14,62%. A més a més, és l'única mida de model que es preveu que es podria entrenar a les màquines de l'**Agència**, donant poc marge de millora.

Ha sigut notable durant l'entrenament la diferència dels models *LaMini*, donant un increment del rendiment amb el model base i una disminució amb el model petit, comparat amb el seu model oficial. Per assegurar que no ha sigut un error d'entrenament, s'hauria de tornar a entrenar el model per eliminar el component estocàstic de l'entrenament, però independentment, es creu que simplement per la mida del model, no pot ser entrenat dos cops sense perdre la capacitat de completar algunes tasques.

Pel que fa al model *FLan-T5-Base (LoRA)*, ha sigut entrenat amb aquest mètode que optimitza l'entrenament centrant-se en mòduls o paràmetres específics, cosa que permet un procés més eficient en termes de recursos. Tot i això, els resultats obtinguts no han coincidit amb les expectatives inicials. En aquest cas, ha donat un rendiment inferior als models més petits en totes les mètriques, a excepció de la mètrica de "Coincidència exacta", en la qual no s'ha provat. La raó per la qual ha pogut donar mals resultats pot ser perquè la selecció dels paràmetres entrenables no ha pogut captar del tot els matisos del *dataset*, cosa que ha pogut conduir a un àmbit de millora del model més reduït. A més a més, algunes tasques o conjunts de dades poden ser intrínsecament complexes i requerir una perspectiva més àmplia durant l'entrenament.

Finalment, s'ha vist que la mètrica *ROUGE* pot ser útil per moltes aplicacions, però en una última instància s'ha optat per fer servir una mètrica com la de "Coincidència exacta" que doni una informació més clara sobre com progressa el model. El problema principal amb la mètrica *ROUGE* és que s'ha vist que es pot buscar una optimització d'aquesta mètrica sense que el resultat generat sigui exactament el que s'esperava, permetent que es pugui generar text de manera desordenada i que no es penalitzi negativament. Per aquesta raó, aquesta mètrica ha sigut utilitzada només com a referència i no per a entrenar el model.

4.2.2 Solució final

En conclusió, el *fine-tuning* de diversos models per al *dataset* escollit ha mostrat les capacitats potencials d'aquests models a l'hora d'abordar els reptes que plantegen els requisits del projecte. Tot i això, continua sent difícil determinar quin seria el model més eficaç en un desplegament real a causa de la manca de dades oficials que ha proporcionat l'**Agència**.

Tot i les seves limitacions, el model *FLan-T5-Base (La Mini)* s'ha posicionat primer en termes de mètriques. El seu rendiment en l'entorn simulat que s'ha plantejat el converteix en un candidat prometedor teòric per a la seva implementació.

A més a més, la solució és completa, ja que abasta no només el model ajustat, sinó també la *pipeline*, la configuració del servidor OTRS i la recuperació i el preprocessament dels tiquets d'incidents. En teoria, el disseny garanteix una funcionalitat sense fissures, però la veritable prova està en l'aplicació al món real. Malauradament, l'absència de dades crucials impedeix una avaluació concloent de l'eficàcia de la solució en un entorn operatiu real.

Capítol 5

Planificació temporal

Segons el conveni de cooperació educativa acordat, aquest projecte es va iniciar el dia 16 de setembre de 2023 i va concloure el 19 de gener de 2024, amb una duració de setze setmanes i dos dies (en un espai temporal de quatre mesos i tres dies). Aquestes dates han sigut escollides per començar amb l'inici de GEP i finalitzar amb la lectura del Treball de Fi de Grau. S'han treballat 5 hores al dia, tot i que, puntualment, s'ha treballat fora de l'horari laboral. S'han extret els dies festius tals com les setmanes del 25 de desembre fins al 7 de gener per vacances de Nadal.

En total, s'ha dedicat unes 600 hores a aquest projecte. S'ha dedicat 460 hores al treball previ i desenvolupament del projecte i 140 per la documentació i redacció d'aquesta memòria. Aquesta planificació temporal només és una estimació de les hores dedicades per l'autor, tot i que a l'equip hi participi més persones.

5.1 Descripció de les tasques

5.1.1 Gestió de Projecte [GP] (180 hores)

- **GP1 (25 hores) - Contextualització i abast.** Contextualitzar i descriure l'abast del projecte on es justifica la solució proposada al problema. És necessari estar sincronitzat amb el ponent del TFG.
- **GP2 (15 hores) - Planificació temporal.** Planificar temporalment el projecte sencer, amb una descripció detallada de les tasques que cal completar, inclosa una estimació de la durada i els recursos necessaris. A més a més, també inclou una anàlisi de riscos relacionats amb el projecte.
- **GP3 (20 hores) - Gestió econòmica i sostenibilitat.** Crear un pressupost que identifiqui i estimi els costos del projecte i la seva gestió, així com un informe de

sostenibilitat del projecte en termes econòmics, socials i ambientals.

- **GP4 (80 hores) - Documentació final.** Elaborar i redactar una memòria completa que inclou una descripció dels aspectes tècnics i de gestió del projecte. Es realitza periòdicament durant el desenvolupament del projecte.
- **GP5 (40 hores) - Reunions.** Celebrar reunions *Sprint* amb els membres de l'equip per discutir els objectius i fer trobades de seguiment periòdiques. També inclou les reunions amb el ponent o director del TFG. Aquesta tasca es realitzarà constantment durant l'elaboració del projecte.

5.1.2 Treball Previ [TP] (80 hores)

- **TP1 (10 hores) - Aprenentatge.** Formació per obtenir els coneixements previs necessaris per iniciar correctament el projecte, tals com: l'aprenentatge autònom, el processament del llenguatge natural (NLP), bases de dades basades en OTRS, sistemes d'emmagatzematge utilitzant Elasticsearch, etc.
- **TP2 (40 hores) - Estudiar l'estat de l'art.** Recerca bibliogràfica per explorar possibles solucions des de diversos enfocaments. Això inclou, tant projectes similars que s'han dut a terme, com articles teòrics.
- **TP3 (20 hores) - Elecció de la implementació final.** Comparació de manera pràctica les diverses alternatives i revisar i discutir els avantatges, possibles millores i inconvenients. Inclou la planificació de tot el projecte, és a dir, del procés que seguiran les dades. També se seleccionarà les tecnologies utilitzades per implementar la solució.
- **TP4 (10 hores) - Instal·lació de l'entorn de treball.** Instal·lació i configuració del programari necessari per al desenvolupament, així com, creació del repositoris pertinents.

5.1.3 Desenvolupament [D] (340 hores)

D1 (65 hores) - Recopilació de Dades

- D1.1 (25 hores) - Recopilació i preprocessament de les dades d'entrenament.
- D1.2 (10 hores) - Identificació i priorització la informació clau a extreure dels tiquets.
- D1.3 (20 hores) - Creació *dataset* a partir de la informació processada.
- D1.4 (10 hores) - Creació criteris d'acceptació per al model.

D2 (75 hores) - Entrenament del Model NLP

- D2.1 (5 hores) - Primeres proves amb el model NLP que es farà servir.
- D2.2 (70 hores) - Entrenament i ajustament del model escollit amb les dades d'entrenament preprocessades.

D3 (50 hores) - Implementació del *pipeline*

- D3.1 (30 hores) - Implementació *pipeline* dels tiquets per a la inferència del model.
- D3.2 (20 hores) - Desenvolupament *unit testing* dels components del *pipeline*.

D4 (40 hores) - Proves i Validació

- D4.1 (20 hores) - Desenvolupament casos i dades de proves per a tests funcionals.
- D4.2 (10 hores) - Creació tests funcionals de tot el sistema.
- D4.3 (10 hores) - Implementació la gestió i el registre d'errors.

D5 (60 hores) - Optimització i Desplegament del Model

- D5.1 (40 hores) - Optimització del model NLP i el preprocessament dels tiquets.
- D5.2 (20 hores) - Realització proves de rendiment i els ajustos finals a l'entorn d'assaig.

D6 (50 hores) - Proves Finals, Documentació i Desplegament Final

- D6.1 (25 hores) - Realització proves finals i validació del sistema.
- D6.2 (10 hores) - Redacció documentació i exemples per a l'usuari.
- D6.3 (15 hores) - Desenvolupament i desplegament a producció l'API del sistema.

5.2 Recursos

5.2.1 Recursos humans

Es defineixen els següents rols dins de l'equip:

- **Cap del projecte:** Responsable de supervisar tot el cicle de vida del projecte. Respecta les guies d'estil i la coherència. Entre les seves funcions s'inclouen la planificació, l'organització de reunions d'equip i la garantia que el projecte avança segons els terminis establerts.
- **Desenvolupador júnior:** S'encarrega d'implementar la solució escollida. Les seves responsabilitats inclouen codificar i posar a prova el sistema per desenvolupar les característiques i les funcionalitats especificades. Hi participen dos desenvolupadors júniors en el projecte.
- **Expert científic:** Proporciona orientació experta sobre tècniques, algorismes i metodologies de processament del llenguatge natural. La seva funció és consultiva i contribueix assessorant sobre les millors pràctiques i aportant idees per millorar els aspectes computacionals del projecte. Es disposarà d'un únic

5.2.2 Recursos materials

Aquests són els recursos materials que s'estima que seran necessaris per al correcte desenvolupament del projecte:

- **Ordinador** amb els seus perifèrics.
- **Sala de reunions** per celebrar les reunions setmanals.
- **PyCharm** serà l'entorn de desenvolupament predilecte.
- **Git** per sistematitzar el control de les versions del codi.
- **Overleaf** per redactar la memòria.
- **onlinegantt.com** per l'elaboració del diagrama Gantt.
- **VPN** per accedir a les bases de dades i als servidors cedits per l'Agència.
- **Portàtil** cedit per l'Agència amb l'entorn configurat per ser el més confidencial possible.
- **Google Meet** per les reunions no presencials.
- **Google Drive** per l'emmagatzematge de documents relacionats amb el projecte.

5.3 Taula de tasques

Nom	Dependències	Recursos	Duració (hores)
Gestió de Projectes (GP)			**180**
GP1		Overleaf, Reunions	25
GP2	GP1	Overleaf, onlinegantt	15
GP3	GP2	Overleaf	20
GP4		Overleaf	80
GP5		Meet, Reunions	40
Treball Previ (TP)			**80**
TP1			10
TP2	TP1	PyCharm	40
TP3	TP2		20
TP4	TP3	PyCharm, Git	10
Desenvolupament (D)			**340**
Recopilació de dades (D1)			**65**
D1.1		PyCharm, Portàtil, VPN	25
D1.2	D1.1	Meet, Reunions	10
D1.3	D1.2	PyCharm, Portàtil, VPN	20
D1.4		PyCharm, Git	10
Desenvolupament del model NLP (D2)			**75**
D2.1		Pycharm, Git	5
D2.2	D1.3	Pycharm, Git, VPN	70
Implementació del *pipeline* (D3)			**50**
D3.1		PyCharm, Git	30
D3.2	D1.4, D3.1	PyCharm, Git	20
Proves i validació (D4)			**40**
D4.1		PyCharm, Git, Reunions	20
D4.2	D2.2, D4.1	PyCharm, Git	10
D4.3	D2.2, D4.2	PyCharm, Git	10
Optimització i desplegament del model (D5)			**60**
D5.1	D2.2	PyCharm, Git	40
D5.2	D5.1	PyCharm, Git	20
Proves finals, documentació i desplegament final (D6)			**50**
D6.1	D5.2	PyCharm, Git	25
D6.2		Overleaf, Drive	10
D6.3	D6.1, D6.2	PyCharm, Git, Portàtil, VPN	15
TOTAL			**600**

Taula 16: Taula de tasques amb les dependències, els recursos i l'estimació de les hores (Elaboració pròpia)

5.4 Diagrama de Gantt

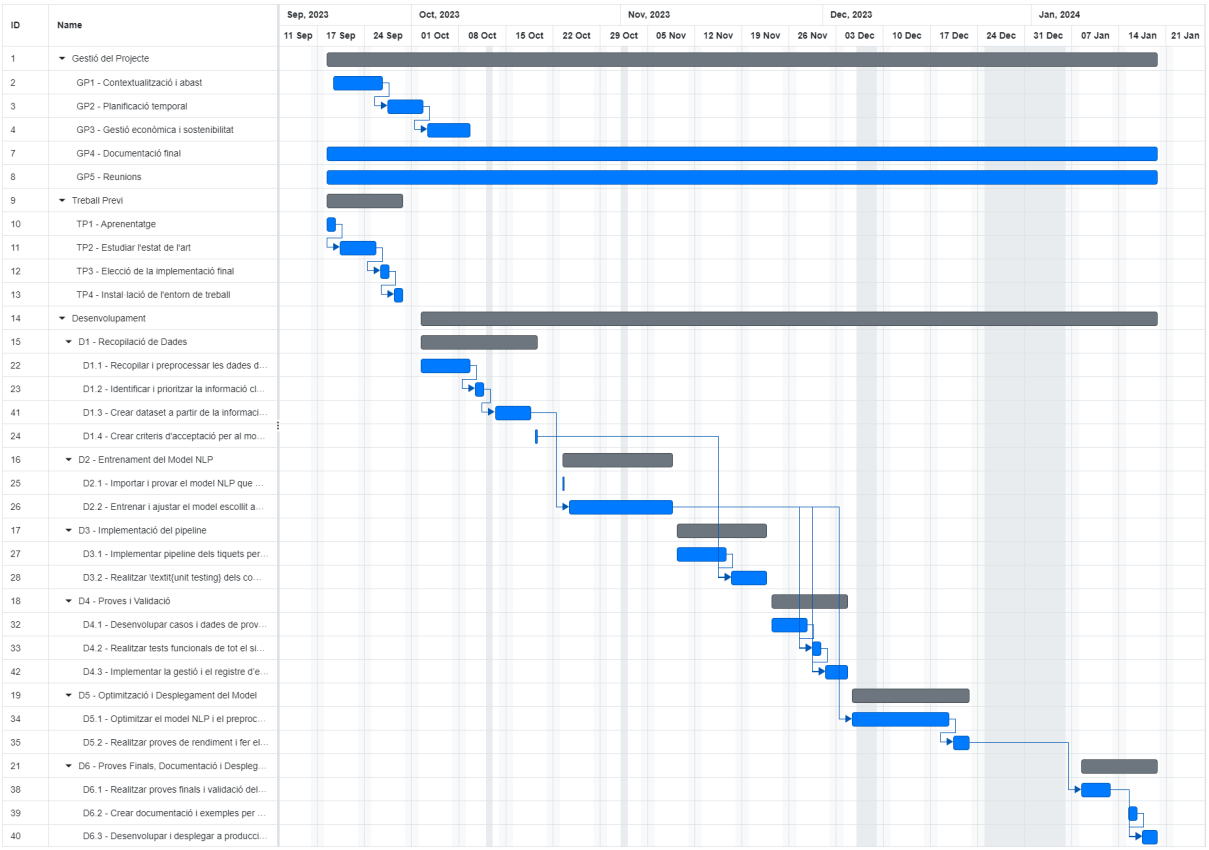


Figura 10: Diagrama de Gantt del projecte
(Elaboració pròpia amb *onlinegantt.com*)

5.5 Gestió del risc

A l'apartat d'obstacles i riscos potencials 1.2.4 s'ha mencionat alguns dels obstacles potencials que poden sorgir. Gràcies a aquestes prediccions, ha sigut possible anticipar els possibles retards i dificultats que provoquin aquests punts. Naturalment, la gestió de riscs s'ha treballat com un procés continu i han sigut buscats en cada part de totes les fases durant el desenvolupament del projecte.

No només es mencionaran els riscos, sinó que es detallarà, per cadascun d'ells, quines són les mesures, plans secundaris, efectes en els recursos o bloquejos temporals que podrien provocar. En el cas del temps extra estimat, s'arrodoniran les hores al següent enter. A continuació es detallen els riscos principals:

- **Models insuficients:** A causa de les limitacions per aquest projecte, es va preveure que no hi hauria un gran nombre de models del llenguatge natural que puguin satisfer tots els requisits que es necessiten. El fet que la informació amb la qual

s'ha treballat sigui confidencial i l'alta potència que requereixen les solucions que s'ha plantejat, crea una barrera entre tots els models disponibles i els que es poden utilitzar. És important mencionar, que el problema amb el qual s'ha treballat és d'alta complexitat i, es va plantejar la possibilitat, que els resultats amb el millor model disponible fossin insuficients pels estàndards establerts. En última instància, també es va plantejar una inversió econòmica si fos necessari per accedir a un model privat. Aquest risc recau sobre les tasques TP2 i TP3 amb una extensió estimada del temps del 15% (9 hores).

- **Diversitat de les dades:** La diversitat dels documents disponibles ha fet necessària una decisió estratègica per racionalitzar l'enfocament i refinar l'abast del projecte. En conseqüència, s'ha optat per usar exclusivament tiquets que tractin sobre amenaces de **phishing** o de **malware**. A més a més, es va demanar una reducció del nombre camps que s'extreuen dels tiquets amb el qual es va consensuar l'extracció només els set camps més importants. Aquest enfocament específic no només ha garantit una anàlisi més coherent, sinó que també ha ofert l'oportunitat d'extreure idees més significatives de les dades disponibles, millorant en última instància la qualitat i el rigor del treball. Aquest risc recau sobre les tasques D1.1 i D1.2 amb una extensió estimada del temps del 5% (2 hores).

- **No tenir suficient informació:** Com s'ha observat en els primers passos del projecte, hi ha hagut una mancança de tiquets per poder extreure informació. Des del desconeixement, s'esperava que hi hagués un flux més estable de dades pròximament, però, en cas contrari, es destinaria més temps a recopilar els tiquets necessaris, a generar casos sintètics o aconseguir un *dataset* similar per assegurar la qualitat d'aquests. Aquest risc recau sobre la tasca D1.1 amb una extensió estimada del temps del 10% (3 hores).

- **Falta d'experiència:** La complexitat d'aquest projecte ha plantejat dubtes sobre la capacitat de l'equip per dur a terme tasques tan intrincades. El tractament d'informació confidencial i la necessitat de solucions molt potents requereixen un nivell de competència elevat. Per minimitzar aquest risc, s'han organitzat sessions de formació addicionals i activitats d'intercanvi de coneixements. Això requereix assignar temps addicional perquè els membres de l'equip es familiaritzin amb les eines i tecnologies necessàries per al projecte. A més a més, per mitigar el dèficit d'experiència s'ha decidit consultar un especialista en la matèria per encaminar correctament el projecte. Aquest risc és especialment pertinent per a les tasques TP1 i TP2, amb un augment de temps estimat aproximat del 10% (5 hores).

- **Limitació dels recursos computacionals:** La limitació dels recursos de maquinari, especialment a les GPU, ha plantejat reptes importants en l'àmbit de l'aprenentatge autònom. L'elecció del model adequat per a un projecte té un impacte

significatiu en la qualitat dels resultats i el temps d'entrenament. La manca de recursos limita les opcions, cosa que presenta reptes i riscos. En concret aquest risc afecta les tasques TP2 i D2.2 a causa d'haver d'explorar més models i a haver de dedicar més temps a la inferència i afinament de cadascun d'ells (inclosa l'elecció final) afegint un possible increment del temps del 15%.

Capítol 6

Gestió Econòmica

En aquesta secció, s'aborda l'avaluació dels factors econòmics que intervenen en la realització del projecte. Per aconseguir-ho, es du a terme una anàlisi detallada dels costos associats, amb una especial atenció a la categorització dels mateixos en costos de personal, costos genèrics, costos de contingència i imprevistos. En conclusió, s'analitza la viabilitat econòmica del projecte, estimant els costos en categories específiques i es presenta un pla de contingència per gestionar-los i adaptar-los davant de riscos.

6.1 Costos de personal i activitat

Els costos de personal són els que ocupen el percentatge més gran del cost total d'aquest projecte. S'ha utilitzat els rols especificats en apartats anteriors de Cap de Projecte (CP), Desenvolupador Júnior (DJ) i Expert Científic (EC) per calcular el seu cost. El cost total es pot desengranar en el salari anual brut i la cotització de la Seguretat Social del 30% del salari brut. Els salaris han sigut estimats amb l'ajuda de la pàgina web *payscale.com* creada per ajudar a comprendre als empleats el seu valor en el mercat laboral. A la taula 4 es pot visualitzar els costos de salari, de quotes i el cost total de cada un dels rols que participen en el projecte, incloent-hi els costos per hora.

Posició	**Salari brut anual**	**Cost Seguretat Social**	**Cost total anual**	**Cost total per hora**
Cap de Projecte (**CP**)	50.029€[10]	15.008,7€	65.037€	31,27€
Desenvolupador Júnior (**DJ**)	23.631€[11]	7.089,3€	30.720,3€	14,7€
Expert científic (**EC**)	54.000€[11]	16.200€	70.200€	33,59€

Taula 17: Taula dels costos totals de les posicions del projecte
(Elaboració pròpia amb *payscale.com*)

El cost previst per a cadascuna de les posicions enumerades anteriorment, es mostra detalladament a la Taula 18. El preu de cadascuna es determina multiplicant el nombre d'hores que cada perfil hi dedicarà pel cost per hora determinat a la taula 17.

Identificador	Hores DJ	Hores CP	Hores EC	Hores totals	Cost total
GP	**220**	**140**	**10**	**370**	**7947,70€**
GP1	25	20	0	45	992,90€
GP2	15	10	0	25	533,20€
GP3	20	10	0	30	606,70€
GP4	80	80	0	160	3677,60€
GP5	40 (x2)	20	10	110	2137,30€
TP	**160**	**60**	**40**	**260**	**5571,80€**
TP1	10 (x2)	0	0	20	294,00€
TP2	40 (x2)	25	20	125	2629,55€
TP3	20 (x2)	25	20	85	2041,55€
TP4	10 (x2)	10	0	30	606,70€
D	**680**	**25**	**10**	**715**	**11113,65€**
D1	**130**	**15**	**10**	**155**	**2715,95€**
D1.1	25 (x2)	5	0	55	891,35€
D1.2	10 (x2)	10	5	35	774,65€
D1.3	20 (x2)	0	0	40	588,00€
D1.4	10 (x2)	0	5	25	461,95€
D2	**150**	**0**	**0**	**150**	**2205,00€**
D2.1	5 (x2)	0	0	10	147,00€
D2.2	70 (x2)	0	0	140	2058,00€
D3	**100**	**0**	**0**	**100**	**1470,00€**
D3.1	30 (x2)	0	0	60	882,00€
D3.2	20 (x2)	0	0	40	588,00€
D4	**80**	**10**	**0**	**90**	**1488,70€**
D4.1	20 (x2)	10	0	50	900,70€
D4.2	10 (x2)	0	0	20	294,00€
D4.3	10 (x2)	0	0	20	294,00€
D5	**120**	**0**	**0**	**120**	**1764,00€**
D5.1	40 (x2)	0	0	80	1176,00€
D5.2	20 (x2)	0	0	40	588,00€
D6	**100**	**0**	**0**	**100**	**1470,00€**
D6.1	25 (x2)	0	0	50	735,00€
D6.2	10 (x2)	0	0	20	294,00€
D6.3	15 (x2)	0	0	30	441,00€
Total	**1060**	**225**	**60**	**1345**	**24633,15€**

Taula 18: Taula de les estimacions dels costos per tasca
(Elaboració pròpia)

6.2 Costos genèrics

Les despeses dels costos genèrics no estan vinculades a tasques específiques, però tenen un paper fonamental en l'avaluació financera. Aquesta secció abasta la depreciació dels components de maquinari i programari, així com el consum d'energia, la connexió a la xarxa i l'espai que s'ocupa.

6.2.1 Amortitzacions

Software

Hi ha cert programari, sobretot en l'entorn de desenvolupament controlat per **l'Agència**, que no han estat sota control o són desconeguts i, a més a més, poden variar. En tot cas, aquest programari no ha afectat el pressupost, i, per tant, no s'ha tingut en compte. A continuació es calcula el cost del programari utilitzat:

- **Google Drive:** No s'ha contractat cap espai extra.
- **Overleaf:** No s'ha utilitzat les característiques de la subscripció mensual.
- **Git:** No ha sigut necessari l'ús dels extres que s'ofereixen.
- **PyCharm:** S'ha usat la llicència d'estudiant que dura un any.

En conclusió, el programari no ha afegit cap cost addicional al projecte.

$$\text{Amortització Software} = 0€$$

Hardware

Només ha sigut necessari invertir per un ordinador de sobretaula pels tres integrants de l'equip per la realització d'aquest projecte. S'ha estimat que un ordinador com els fets servir amb els seus dos monitors i perifèrics costa uns 1200€. La vida útil d'aquest ordinador s'ha estimat que oscil·la els cinc anys (seixanta mesos) i aquest projecte té una duració d'uns quatre mesos. Amb aquesta informació es calcula l'amortització dels ordinadors:

$$\text{Amortització Hardware} = 1200€ * 3 \text{ unitats} * \frac{4}{60} \text{ mesos} = 240€$$

6.2.2 Consum elèctric

El cost elèctric que s'ha calculat serà el dels ordinadors de sobretaula que es fan servir amb dos monitors cadascun. El teclat i el ratolí tenen una potència negligible en aquest càlcul. Com s'ha calculat a la taula 18, en total s'han invertit 1345 hores en aquest projecte entre tots els integrants de l'equip. El preu mitjà de la llum de 0,1728€/kWh [12], un ordinador amb un consum de 400 W [13], les dues pantalles de cada monitor són idèntiques amb un consum màxim de 35 W [14].

Producte	Potència	Unitats	Hores d'ús	Consum total	Cost total
Ordinador sobretaula	400 W	3	1345	1614 kWh	278,9€
Pantalla	35 W	6	1345	282,45 kWh	48,81€
Total	**470 W**	**3**	**1345**	**1896,45 kWh**	**327,71€**

Taula 19: Taula del consum elèctric
(Elaboració pròpia)

6.2.3 Connexió internet

S'ha calculat que el preu de la connexió a internet ronda els 50€ per mes. Amb un projecte de quatre mesos de duració, el total ha sigut de:

$$\text{Preu Connexió} = \frac{50€}{1 \text{ mes}} * 4 \text{ mesos} = 200€$$

6.2.4 Espai d'oficina

El cost de l'espai d'oficina s'ha estimat segons el cost mitjà de les oficines disponibles per lloguer en aquesta època [15]. S'ha calculat un preu de $21€/m^2$ amb l'aigua inclosa i han sigut necessaris $25m^2$ d'oficina.

$$\text{Preu Espai} = \frac{21€}{1m^2} * 25m^2 = 525€$$

6.2.5 Total costos genèrics

Concepte	Cost
Amortització *Software*	0€
Amortització *Hardware*	240€
Consum Elèctric	327,71€
Connexió Internet	200€
Espai d'oficina	525€
Total	1292,71€

Taula 20: Suma total dels costos genèrics
(Elaboració pròpia)

6.3 Contingències

Tenir un cost de contingència en un projecte de *software* és crucial perquè permet flexibilitat a l'hora de gestionar reptes i canvis imprevistos que poden sorgir durant el desenvolupament. En assignar un cost de contingència del 15%, és assegurat que hi ha hagut disponibles els recursos financers per abordar problemes inesperats, canvis d'abast o requisits addicionals sense comprometre la qualitat o el calendari del projecte. Aquest enfocament proactiu de la gestió del risc ha sigut implementat per millorar el resultat global del projecte i minimitzar el potencial de retards o compromisos costosos en el producte final.

$$\begin{aligned}\text{Cost contingència} &= (\text{Costos de personal} + \text{Costos genèrics}) * 0,15 = \\ &= (24633,15€ + 1292,71€) * 0,15 = \\ &= 3888,88€\end{aligned}$$

6.4 Imprevistos

Per poder tenir en compte els potencials riscos que han pogut sorgir durant el desenvolupament de les tasques planejades, ha sigut essencial incorporar un sobrecost per les que tenen un major potencial d'error. S'ha assignat un percentatge de sobrecostos basat tant en la probabilitat d'ocurrència com en l'impacte potencial dels quatre riscos identificats anteriorment. Aquest enfocament ha garantit una preparació financera per fer front als riscos que s'han pogut intuir, mantenint l'estabilitat del projecte.

Imprevist	Tasques afectades	Despesa original	Risc	Sobrecost
Models insuficients	TP2, TP3	4671,10€	15%	700,67€
Diversitat de les dades	D1.1, D1.2	1666,00€	5%	83,30€
No tenir suficient informació	D1.1	891,35€	10%	89,14€
Falta d'experiència	TP1, TP2	2923,55€	10%	292,36€
Limitació recursos computacionals	TP2, D2.2	4687,55€	15%	703,13€
Total				**1868,59€**

Taula 21: Taula dels sobrecostos afegits pels imprevistos
(Elaboració pròpia)

6.5 Cost total del projecte

En la Taula 22 es presenten els diferents costos estudiats en aquest capítol i la suma total, donant el cost total estimat del projecte.

Concepte	Cost
Costos de personal	24633,15€
Costos genèrics	1292,71€
Contingències	3888,88€
Imprevistos	1868,59€
Total	**31683,33€**

Taula 22: Taula del cost total del projecte
(Elaboració pròpia)

6.6 Control de gestió

El control de gestió eficaç d'un pressupost per al projecte és essencial per garantir que es mantingui econòmicament estable. L'enfocament escollit ha consistit a aprofitar metodologies ja establertes per controlar i mantenir el pressupost. Les reunions periòdiques de *sprint* han permès avaluar les desviacions i prendre accions per a corregir-ho. Si es produeixen desviacions importants, s'intenta cobrir-les amb els fons assignats per a imprevistos i, si aquests no fossin suficients, es faria ús del cost per a contingències.

Per detectar desviacions, s'han usat indicadors clau de rendiment per identificar l'origen del problema. A més a més, els membres de l'equip han fet un seguiment de les seves hores dedicades a les tasques, permetent comparacions setmanals amb les hores estimades

per completar les tasques. A continuació es mencionen els indicadors utilitzats:

- **Desviació en el cost per tasca**

$$d_c = (c_r - c_e) * h_r$$

- **Desviació de la dedicació d'hores per tasca**

$$d_h = (h_r - h_e) * c_r$$

on c_r és el cost real, c_e és el cost estimat, h_r són les hores dedicades reals i h_e són les hores dedicades estimades.

Capítol 7

Sostenibilitat

7.1 Autoavaluació

És fonamental reflexionar sobre les pròpies conclusions sobre la sostenibilitat i el desenvolupament sostenible abans d'embarcar-se en aquest projecte, especialment en el context de l'àmbit informàtic. Això s'ha aconseguit responent al qüestionari del projecte **EDINSOST2-ODS**.

La integració de coneixements generals, qüestions socials i implicacions ambientals tecnològiques reconeix la importància crítica de la sostenibilitat en aquest camp. Si ens fixem en la tendència de les empreses tecnològiques líders, moltes són ben conscients dels problemes socials, econòmics i ambientals als quals s'enfronta la societat actual, i reconeixen que aquesta disciplina no pot existir aïllada d'aquests reptes globals. A més, s'han d'investigar els mètodes i les eines utilitzades per estimar la viabilitat econòmica del projecte per assegurar-se que són coherents amb els objectius de sostenibilitat. La gestió de recursos és un component crític del desenvolupament a llarg termini i mereix una consideració especial en el context dels projectes informàtics.

En aquesta època de major consciència sobre els problemes socials, econòmics i ambientals, no es pot exagerar el paper dels productes i serveis de la informàtica a l'hora d'exacerbar o mitigar aquests problemes. L'innegable impacte mediambiental del sector, així com les possibles conseqüències per a la salut, la seguretat i la justícia social derivades dels projectes i accions de la informàtica, posen de manifest la necessitat de posar més èmfasi en la sostenibilitat.

Finalment, quan s'aprofundeix en projectes, productes i serveis informàtics, és fonamental reconèixer la complexa xarxa d'interaccions que es produeixen amb altres actors en processos, activitats i projectes més grans, ja que aquestes interaccions tenen un impacte significatiu dels nostres esforços. En essència, aquesta reflexió estableix les bases per a la investigació sobre la sostenibilitat informàtica, destacant la seva naturalesa polifacètica i

les implicacions per a la nostra societat.

Pretenem ampliar aquests punts en els apartats següents, discutint les dimensions econòmica, ambiental i social del projecte. Aquesta investigació pretén demostrar no només una comprensió a fons d'aquests aspectes crítics, sinó també un compromís per incorporar els principis de sostenibilitat al nucli del treball.

7.2 Dimensió econòmica

Has estimat el cost de la realització del projecte (recursos humans i materials)?

Si, s'ha fet una anàlisi de la gestió econòmica del projecte on es tracta, tant el cost de la realització del projecte, com les seves parts individuals i també s'ha tingut en compte el control de gestió per evitar inversions innecessàries.

Com es resol actualment el problema que vols tractar (estat de l'art)? En què millora econòmicament la teva solució a les ja existents?

El mètode actual de resolució d'incidències mitjançant sistemes de tiquets depèn en gran manera de processos manuals d'anàlisi i de resolució, que sovint exigeixen grans quantitats de temps i recursos humans. No obstant això, aquest mètode és intrínsecament limitat a causa de la seva naturalesa reactiva i de la seva incapacitat per aprofitar eficaçment la gran quantitat de dades textuals que contenen els tiquets. La solució que es proposa permet extreure informació valuosa de les incidències. L'automatització de l'anàlisi de les dades permet la identificació proactiva de possibles amenaces. Això agilitza la resolució d'incidències i optimitza la utilització de recursos en abordar els problemes recurrents de manera preventiva.

7.3 Dimensió ambiental

Has estimat l'impacte ambiental que tindrà la realització del projecte? T'has plantejat minimitzar l'impacte, per exemple, reutilitzant recursos?

És important destacar que aquest projecte tindrà un impacte ambiental, encara que aquest pugui ser petit. Gràcies a l'ambient de *co-working* en el que es treballa, molts dels recursos crítics que no es volen malgastar estan sent compartits i aprofitats per més persones que les incloses en aquest projecte.

Com es resol actualment el problema que s'afronta (estat de l'art)? En què millora ambientalment la teva solució respecte a les existents?

És difícil estimar el benefici ambiental de la solució emprada respecte a l'anterior. Tot i que els processos computacionals del processament del llenguatge natural comporten un consum d'electricitat més gran, aquest enfocament ofereix avantatges mediambientals convincents, com ara frustrar amenaces potencials i optimitzar l'assignació de recursos. En identificar i resoldre els problemes de forma proactiva mitjançant una extracció eficaç de la informació, la nova solució redueix la probabilitat que es prolonguin els incidents, cosa que evita el consum innecessari de recursos i repercuteix positivament en la petjada mediambiental dels procediments de gestió d'incidents.

7.4 Dimensió social

Que creus que t'ha aportat en l'àmbit personal la realització d'aquest projecte?

Aquest projecte ha permès demostrar, a mi i a l'empresa, les meves habilitats com a enginyer. També m'ha ajudat a desenvolupar les meves habilitats tècniques i ha aprofundit els meus coneixements de sostenibilitat econòmics, ambientals i socials.

Com es resol actualment el problema que vols afrontar (estat de l'art)? En què millora socialment (qualitat de vida) la teva solució respecte a l'existent?

La nova solució augmenta l'eficàcia operativa i millora la capacitat de l'empresa per fer front de manera proactiva a les amenaces emergents. Això, alhora, contribueix a millorar la qualitat de vida de la societat. Es redueixen els temps de resposta i es millora la precisió en la identificació d'amenaces, cosa que fomenta un entorn més segur i resistent i minimitza els possibles efectes adversos sobre les persones i les comunitats.

Existeix una necessitat real del projecte?

A causa de les tres dimensions d'aquest projecte, es creu que el projecte té una necessitat real i que permetrà a l'empresa identificar millor les amenaces i poder evitar-les d'una manera més ràpida i, en conseqüència, econòmica que anteriorment. També és esperable, que es pugui reutilitzar tant el resultat final com l'experiència adquirida en el futur i més àmpliament en altres àmbits.

Capítol 8

Integració del coneixement

8.1 Competències tècniques del projecte

En aquesta secció es descriuen els aspectes associats a les competències tècniques descrites a la Facultat d'Informàtica de Barcelona dins del projecte i es fa una anàlisi del nivell d'assoliment de cadascuna.

CCO1.1

> Avaluar la complexitat computacional d'un problema, conèixer estratègies algorísmiques que puguin dur a la seva resolució, i recomanar, desenvolupar i implementar la que garanteixi el millor rendiment d'acord amb els requisits establerts.

S'ha estudiat el problema i el camp al qual pertany i s'ha après sobre les tècniques de *deep learning* més recents per la resolució dels objectius plantejats. A més a més, s'ha realitzat un estudi de l'estat de l'art per tal d'aconseguir la solució amb el millor rendiment possible. Es considera que aquesta competència **ha sigut assolida satisfactòriament**.

CCO2.1

> Demostrar coneixement dels fonaments, dels paradigmes i de les tècniques pròpies dels sistemes intel·ligents, i analitzar, dissenyar i construir sistemes, serveis i aplicacions informàtiques que utilitzin aquestes tècniques en qualsevol àmbit d'aplicació.

S'ha comprès els models NLP i el coneixement de les tècniques de *fine-tuning* que després s'ha aplicat per crear l'aplicació al sistema de tiquets d'incidències. Es considera que

aquesta competència **ha sigut assolida satisfactòriament**.

CCO2.2

> Capacitat per a adquirir, obtenir, formalitzar i representar el coneixement humà d'una forma computable per a la resolució de problemes mitjançant un sistema informàtic en qualsevol àmbit d'aplicació, particularment en els que estan relacionats amb aspectes de computació, percepció i actuació en ambients o entorns intel·ligents.

Els models NLP han sigut entrenats a partir de textos generats per humans. El procés ha consistit a formalitzar i representar aquests coneixements de manera computable per resoldre problemes a l'àmbit de la ciberseguretat. Es considera que aquesta competència **ha sigut assolida satisfactòriament**.

CCO2.4

> Demostrar coneixement i desenvolupar tècniques d'aprenentatge computacional; dissenyar i implementar aplicacions i sistemes que les utilitzin, incloent les que es dediquen a l'extracció automàtica d'informació i coneixement a partir de grans volums de dades.

La tèncica d'aprenentatge computacional ha sigut, principalment el *deep learning*. S'ha demostrat el coneixement sobre la tècnica i s'ha implementat un sistema amb la funció principal d'extreure informació automàticament. Aquest sistema ha sigut desenvolupat amb un gran volum de dades d'entrenament per l'aprenentatge del model. Es considera que aquesta competència **ha sigut assolida satisfactòriament**.

CCO3.1

> Implementar codi crític seguint criteris de temps d'execució, eficiència i seguretat.

En treballar amb tiquets d'incidents per a una agència de ciberseguretat, és crucial donar prioritat a la seguretat de tot el sistema. Això implica l'aplicació de pràctiques el maneig adequat de la informació delicada i la protecció davant de possibles vulnerabilitats. Es considera que aquesta competència **ha sigut assolida satisfactòriament**.

CCO3.2

> Programar considerant l'arquitectura hardware, tant en asemblador com en alt nivell.

En aquest projecte, l'arquitectura del sistema ha determinat no només les possibles tècniques a desenvolupar, sinó que també el rendiment final del sistema. Addicionalment, la dependència de les tècniques de *deep leaning* a les GPU, ha obligat a la consideració de l'arquitectura amb la qual es treballa. Es considera que aquesta competència **ha sigut assolida satisfactòriament**.

8.2 Coneixement de les assignatures

Aprenentatge Automàtic

En aquesta assignatura s'estudien diverses tècniques de modelatge, entre els quals s'inclouen les xarxes neuronals artificials i altres sistemes d'aprenentatge autònom. Addicionalment, també ha permès familiaritzar-se amb els ambients de *Jupyter Notebook*.

Intel·ligència Artifical

L'assignatura ofereix una àmplia panoràmica del camp, el caràcter multidisciplinari, el desenvolupament històric i les aplicacions actuals. Ha permès un enteniment general de la intel·ligència artificial i de les seves aplicacions no només a la teoria, sinó que també aplicades de forma pràctica.

Llenguatges de programació

El contingut de l'assignatura no contribueix directament a la tasca de l'anàlisi automàtica de tiquets d'incidents, però atès que *Python* és el llenguatge de programació principal per al desenvolupament del sistema, la informació específica sobre aquest llenguatge es pot aplicar directament al projecte. Els coneixements sobre la sintaxi, les biblioteques i les millors pràctiques de Python han ajudat a implementar el model NLP.

Probabilitat i Estadística

El contingut d'aquesta assignatura ha resultat beneficiosa per al projecte. Comprendre com modelar fenòmens aleatoris ha sigut important a l'hora de tractar els textos d'entra-

da. Una comprensió dels models de probabilitat ha ajudat a analitzar correctament les dades i els resultats dels models.

Capítol 9

Conclusions

9.1 Assoliment dels objectius

9.1.1 Estudi de l'estat de l'art

S'ha realitzat un estudi extens de l'estat de l'art, que ha requerit diverses setmanes de recerca i estudi dels avenços més nous en la matèria. S'ha pogut veure com aquest camp avança a un pas imparable i on pràcticament setmanalment apareixen noves solucions per provar. Tant a l'espai *open source* com al privat, s'ha pogut posar a prova les tecnologies més modernes disponibles i valorar els seus avantatges i desavantatges. Es considera que aquest objectiu **ha sigut assolit satisfactòriament**.

9.1.2 Implementació del pipeline

El *pipeline* ha sigut dissenyat i implementat des de zero, tenint en compte les especificacions i limitacions del projecte. S'ha estudiat i comprès el funcionament intern del sistema de tiquets OTRS i la llibreria PyOTRS i per aconseguir configurar una base de dades amb aquest sistema i extreure els tiquets amb tota la informació necessària. Addicionalment, s'ha escrit una cadena de preprocessament de tiquets per tal de poder obtenir totes les dades en el millor format possible i evitar repeticions o dades irrellevants. S'ha experimentat amb la base de dades *Elasticsearch* per poder inserir informació en una base de dades mitjançant la llibreria requerida. Finalment, no s'ha pogut treballar amb la funció d'anonimització que s'ha mencionat per limitacions externes, però es considera que no és una prioritat que hagi afectat el desenvolupament del projecte. És per això, que es considera que aquest objectiu **ha sigut assolit satisfactòriament**.

9.1.3 Sistema d'extracció d'informació

Aquest sistema d'extracció d'informació ha pres forma d'un model del llenguatge natural. Aquest model preentrenat ha sigut entrenat amb la tècnica de *fine-tuning* per aconseguir el millor rendiment possible. Tot i que el projecte ha assolit fites notables, és essencial reconèixer certes limitacions que han repercutit en la seva capacitat per assolir el màxim potencial. Aquestes limitacions són inherents a l'abast del projecte i factors externs que escapen al control. Notablement, per validar encara més l'eficàcia dels resultats obtinguts, seria valuós fer proves amb dades del món real. Aquest pas proporcionaria una representació més precisa dels resultats i demostraria la seva fiabilitat i l'aplicabilitat. Es considera que aquest objectiu **ha sigut assolit satisfactòriament**.

9.1.4 Desplegament API

El desplegament de l'API s'ha endarrerit principalment a causa de la manca de temps disponible. Es va prendre la decisió de donar prioritat a totes les altres seccions del projecte per tal de garantir que totes les responsabilitats essencials s'abordessin a temps. Tot i aquest retard, l'equip manté el compromís de completar el desplegament de l'API com més aviat millor, una vegada que la càrrega de treball actual estigui més ben equilibrada. Es considera que aquest objectiu **no ha sigut assolit**.

9.2 Treball futur

9.3 Conclusions personals

Aquest projecte m'ha transformat tant personalment com professionalment. Ha abastat diversos aspectes que han contribuït significativament al meu creixement.

Gràcies a aquest projecte, he adquirit una valuosa experiència i habilitat en el camp del *deep learning* i més concretament en el processament del llenguatge natural. Documentar el progrés, les metodologies i els resultats del projecte ha perfeccionat les meves habilitats de redacció tècnica. Aprofundir en la recerca per l'estat de l'art i la documentació del projecte ha garantit una base robusta per als meus futurs projectes. El treball en equip eficaç ha sigut essencial, i col·laborar, compartir idees i afrontar reptes ha fomentat el creixement mutu.

Tot i que el projecte ha comprès conceptes i tecnologies complexes, l'enfocament estructurat i el suport de l'equip l'han convertit en una introducció accessible a aquest camp. El procés d'aprenentatge gradual, juntament amb un entorn de suport, ha permès una transició fluida de la teoria a l'aplicació pràctica.

Aquest projecte ha enriquit les meves habilitats tècniques i m'ha motivat per aprendre més sobre els temes que s'ha treballat, demostrant així, el poder de les experiències pràctiques i la col·laboració en la formació de l'enginyer informàtic.

Referències

[1] *PORTAL ESTADÍSTIC DE CRIMINALITAT*. https://estadisticasdecriminalidad.ses.mir.es/. Accedit: 22/11/2023.

[2] *What is scrum? [+ How to start]*. https://www.atlassian.com/agile/scrum. Accedit: 13/12/2023.

[3] *Hugging Face*. https://huggingface.co/. Accedit: 07/10/2023.

[4] Prathamesh Kalamkar et al. *Named Entity Recognition in Indian court judgments*. 2022. arXiv: 2211.03442 [cs.CL].

[5] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[6] PromptEngineer. *localGPT*. Maig de 2023. URL: https://github.com/PromtEngineer/localGPT.

[7] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[8] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG].

[9] Minghao Wu et al. *LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions*. 2023. arXiv: 2304.14402 [cs.CL].

[10] *Average Information Technology (IT) Manager Salary in Spain*. https://www.payscale.com/research/ES/Job=Information_Technology_(IT)_Manager/Salary. Accedit: 05/10/2023.

[11] *Average Junior Software Engineer Salary in Spain*. https://www.payscale.com/research/ES/Job=Junior_Software_Engineer/Salary/29db7a3a/Barcelona. Accedit: 05/10/2023.

[12] *Precio de la luz por horas: Detalles y Evolución de la tarifa PVPC*. https://tarifaluzhora.es. Accedit: 06/10/2023.

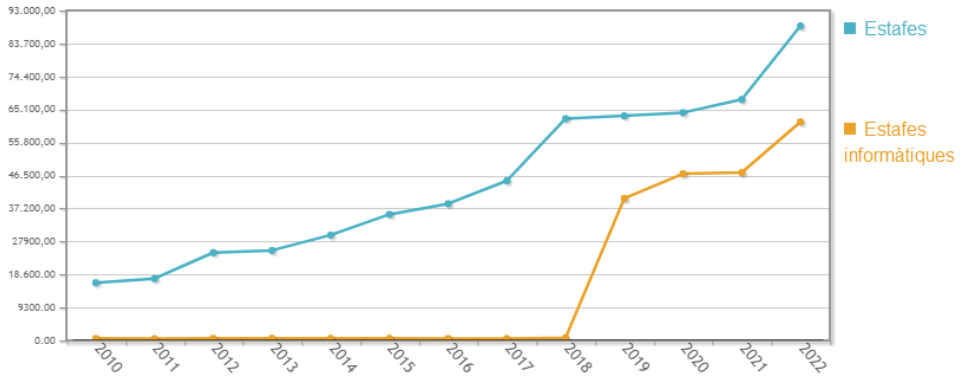
[13] *PC de sobremesa HP Elite Tower 800 G9*. https://www8.hp.com/h20195/V2/GetPDF.aspx/c08086877. Accedit: 13/11/2023.

[14] *Monitor LCD panorámico de 22 pulgadas HP Compaq LA2205wg*. https://support.hp.com/es-es/product/product-specs/hp-compaq-la2205wg-22-inch-widescreen-lcd-monitor/3955309. Accedit: 06/10/2023.

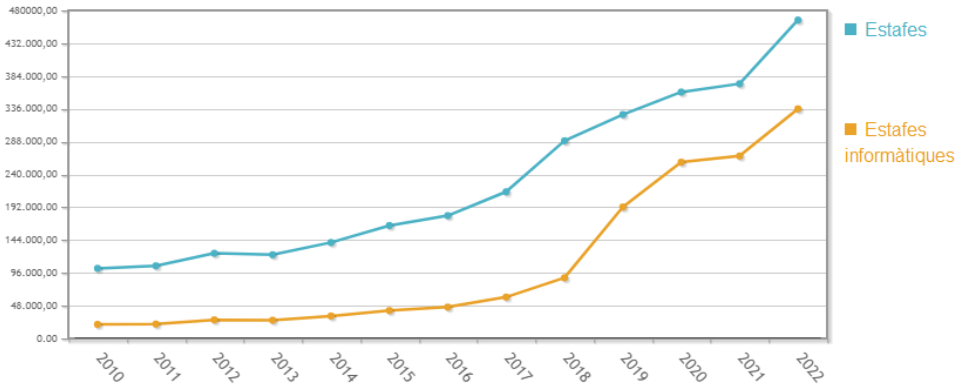
[15] *Oficinas en alquiler*. https://www.bgoficinasbarcelona.com/espacios-oficinas-en-alquiler/?operacion-oficinas=en-alquiler&ciudad=barcelona&zona-oficina=diagonal&oficina-tipo=edificio-de-oficinas. Accedit: 06/10/2023.

Apèndix A

Estadístiques de frau digital

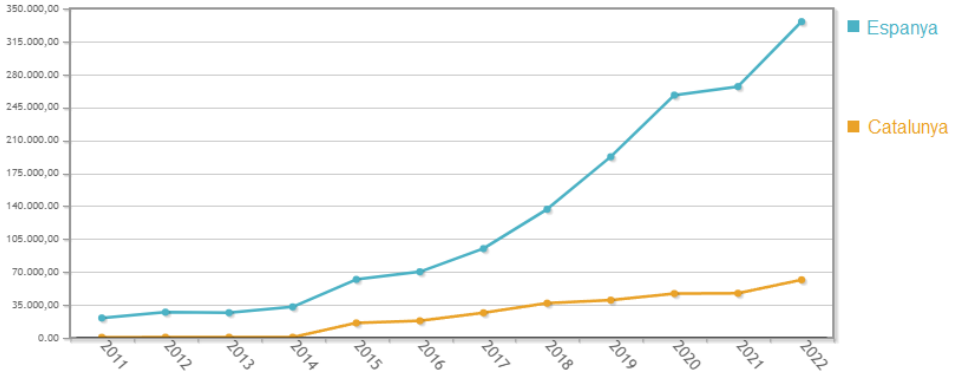


(a) Fets coneguts per període a Catalunya. Nombre de casos d'estafes en comparació amb el nombre de casos d'estafes informàtiques.

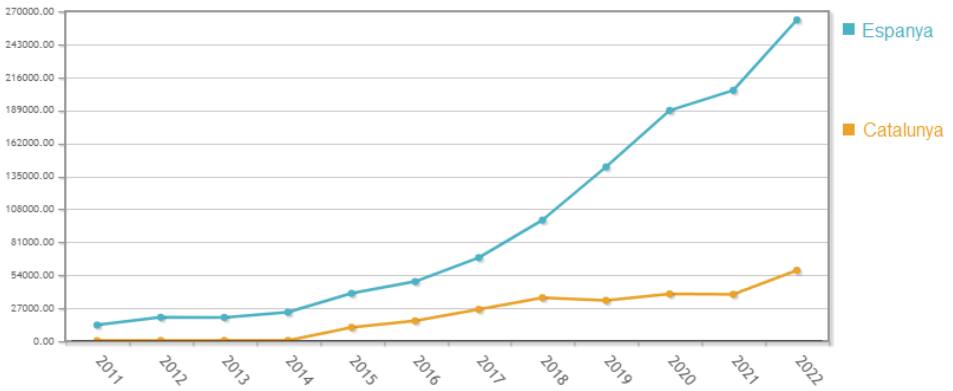


(b) Fets coneguts per període a Espanya. Nombre de casos d'estafes en comparació amb el nombre de casos d'estafes informàtiques.

Figura 11: Comparació de les estafes i les estafes informàtiques durant el període 2011-2022

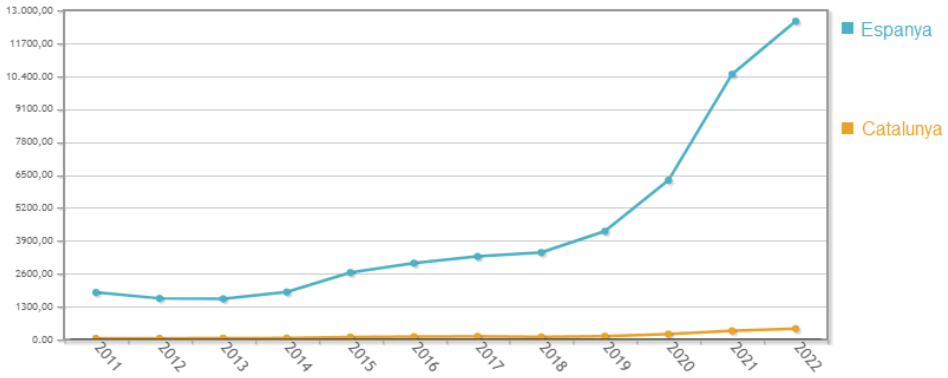


(a) Fets coneguts d'infraccions penals relacionades amb la cibercriminalitat per comunitats autònomes i període. Nombre de casos de frau informàtic.

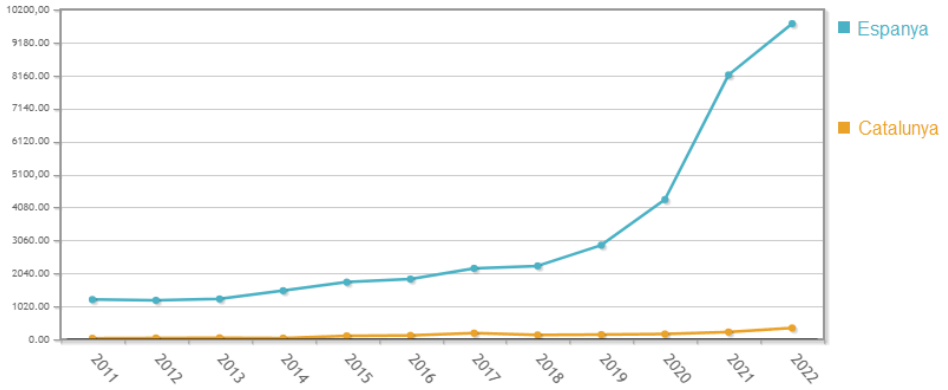


(b) Vicitimitzacions per causes de cibercriminalitat per comunitats autònomes i període. Nombre de víctimes de frau informàtic.

Figura 12: Estadístiques comparant el frau informàtic a Catalunya i Espanya durant el període 2011-2022



(a) Fets coneguts d'infraccions penals relacionades amb la cibercriminalitat per comunitats autònomes i període. Nombre de casos de falsificació informàtica.



(b) Vicitimitzacions per causes de cibercriminalitat per comunitats autònomes i període. Nombre de víctimes de falsificació informàtica.

Figura 13: Estadístiques comparant la falsificació informàtica a Catalunya i Espanya durant el període 2011-2022

Apèndix B

Taules de l'anàlisi del *dataset*

B.1 Distribució de les sentències per tipus de cas i tribunal

Tribunal	Civil
Tribunal Suprem de l'Índia	200
Tribunal Superior de Delhi	100
Tribunal Superior de Bombai	100
Tribunal Superior de Madras	100
Tribunal Superior de Patna	100
Tribunal Superior de Rajasthan	100
Tribunal Superior de Madhya Pradesh	96
Tribunal Superior de Karnataka	95
Tribunal Superior de Calcuta	70
Tribunal Superior de Kerala	0
Tribunal Superior de Panjab-Haryana	100
Tribunal Superior de Gujarat	79
Tribunal de Districte de Delhi	99
Tribunal Superior d'Allahabad	100
Tribunal Superior d'Andhra	100
Tribunal Superior d'Orissa	0
Tribunal d'Apel·lació de l'Impost sobre la Renda	0
Tribunal Superior d'Himachal Pradesh	0
Tribunal Superior de Gauhati	0
Tribunal de Districte de Bangalore	0
Tribunal Superior de Jharkhand	0
Tribunal Superior de Chhattisgarh	0
Tribunal Superior de Jammu i Caixmir	0
Tribunal Superior d'Uttarakhand	0
Tribunal de Duanes, Impostos Especials i Or	0
Tribunal Superior de Sikkim	0
Tribunal Superior de Meghalaya	0
Tribunal Superior de Tripura	0
Tribunal de Duanes, Impostos Especials i Impostos sobre Serveis	0
Total	**1439**

(a) Distribució de les sentències *(Nom Tribunal i Civil)*.

Constitució	Criminal	Financer	Industrial i Laboral
200	200	200	200
100	99	100	100
100	100	87	100
100	100	35	100
100	0	116	100
100	100	110	58
100	100	77	88
99	97	51	90
100	79	72	17
100	100	94	100
100	100	77	53
100	100	98	14
100	91	98	63
100	100	40	22
100	100	75	18
0	0	77	0
0	0	0	0
0	0	18	0
0	0	39	0
0	8	2	36
0	0	24	0
0	0	16	0
0	0	15	0
0	0	7	0
0	0	0	0
0	0	9	0
0	0	5	0
0	0	0	0
0	0	0	0
1599	**1474**	**1542**	**1159**

(b) Distribució de les sentències *(Constitució, Criminal, Financer i Industrial i Laboral)*.

Terrenys i Propietats	Vehicles	Impostos	Total
200	200	200	**1600**
100	100	100	**799**
53	100	100	**740**
100	100	100	**735**
111	100	100	**727**
70	71	100	**709**
91	55	100	**707**
79	96	97	**704**
164	100	100	**702**
100	100	100	**694**
76	83	100	**689**
86	100	100	**677**
100	100	0	**651**
74	100	100	**636**
38	100	86	**617**
89	0	0	**166**
80	0	0	**80**
44	0	0	**62**
16	0	0	**55**
0	0	0	**46**
18	0	0	**42**
16	0	0	**32**
16	0	0	**31**
13	0	0	**20**
19	19	0	**38**
6	0	0	**15**
7	0	0	**12**
2	0	0	**2**
0	1	0	**1**
1768	**1525**	**1483**	**11989**

(c) Distribució de les sentències *(Terrenys i Propietats, Vehicles, Impostos i Total)*.

Taula 23: Distribució de les sentències per tipus de cas i tribunal.

B.2 Llargada de les entitats trobades al *dataset*

Rang	TRIBUNAL	SOL·LICITANT	DEMANDAT	JUTGE	ADVOCAT
[0-2)	0	0	0	0	1
[2-4)	6	14	6	11	7
[4-6)	8	204	128	70	10
[6-8)	0	192	134	103	59
[8-10)	2	209	209	270	271
[10-12)	159	408	402	413	781
[12-14)	505	386	432	481	809
[14-16)	7	357	575	327	654
[16-18)	181	221	232	192	409
[18-20)	72	215	273	233	262
[20-22)	54	166	203	138	107
[22-24)	112	116	214	51	46
[24-26)	42	80	121	13	24
[26-28)	26	92	97	18	6
[28-30)	29	67	87	1	4
[30-32)	19	70	69	1	3
[32-34)	268	29	53	0	2
[34-36)	270	56	50	0	4
[36-38)	73	31	31	0	2
[38-40)	50	19	39	1	5
[40-42)	29	16	28	0	1
[42-44)	25	8	47	0	3
[44-46)	17	6	9	0	14
[46-48)	15	13	15	0	2
[48-50)	8	6	16	0	3
[50-52)	12	5	10	0	3
[52-54)	10	3	10	0	3
[54-56)	17	6	16	0	2
[56-58)	26	6	19	0	6
[58-60)	3	1	12	0	1
[60-62)	13	5	27	0	1
[62-64)	14	3	11	0	0
[64-66)	15	2	8	0	0
[66-68)	13	2	14	1	0
[68-70)	19	1	10	0	0
[70-72)	25	6	17	0	0
[72-74)	18	2	14	0	0
[74-76)	12	1	11	0	0
[76-78)	23	1	12	1	0
[78-80)	20	2	16	0	0
[80-82)	21	1	5	0	0
[82-84)	16	2	6	0	0
[84-86)	13	1	6	0	0
[86-88)	16	2	7	0	0
[88-90)	13	1	8	0	0
[90-92)	14	1	11	0	0
[92-94)	2	2	8	0	0
[94-96)	5	1	6	0	0
[96-98)	2	1	4	0	0
[98-100)	2	5	3	0	0
>=100	46	24	121	0	0
Total	**2367**	**3068**	**3862**	**2325**	**3505**

(a) Llargada de les entitats *(Rang, TRIBUNAL, SOL·LICITANT, DEMANDAT, JUTGE i ADVOCAT).*

DATA	ORG	GPE	ESTATUT	DISPOSICIÓ	PRECEDENT
3	0	0	0	0	0
0	113	12	202	5	0
1	107	358	54	68	0
60	32	389	201	115	0
571	27	339	135	232	1
590	112	120	52	919	0
100	75	53	259	281	1
201	93	50	72	225	31
182	84	40	114	125	26
120	194	17	86	81	22
26	96	8	122	60	18
12	80	3	93	45	21
8	78	4	59	55	25
8	66	1	88	27	29
1	48	3	51	36	20
0	34	0	20	32	33
1	53	0	31	14	22
0	24	0	20	16	44
0	23	0	19	8	40
0	20	0	17	6	42
0	15	0	10	5	41
0	13	0	10	9	38
0	7	0	8	4	51
1	12	0	6	3	46
0	8	1	4	2	77
0	5	0	9	6	52
0	5	0	7	0	46
0	2	0	6	2	49
0	6	0	18	0	67
0	1	0	1	1	28
0	1	0	4	0	42
0	0	0	2	1	49
0	0	0	2	0	36
0	1	0	2	0	27
0	1	0	1	0	38
0	1	0	3	1	35
0	0	0	5	0	34
0	1	0	0	0	27
0	0	0	1	0	25
0	0	0	3	0	18
0	1	0	0	0	15
0	0	0	0	0	16
0	0	0	0	0	11
0	0	0	1	0	11
0	1	0	0	0	9
0	0	0	0	0	9
0	1	0	1	0	10
0	0	0	0	0	10
0	0	0	0	0	7
0	0	0	0	0	4
0	0	0	5	0	48
1885	**1441**	**1398**	**1804**	**2384**	**1351**

(b) Llargada de les entitats *(DATA, ORG, GPE, ESTATUT, DISPOSICIÓ i PRECEDENT)*.

NÚMERO_CAS	TESTIMONI	ALTRES_PERSONES	Total
1	0	0	**5**
0	0	14	**390**
0	57	258	**1323**
3	99	377	**1764**
8	120	388	**2782**
16	162	460	**4594**
39	167	495	**4083**
75	122	263	**3052**
108	56	161	**2131**
116	30	89	**1810**
155	25	76	**1254**
65	20	25	**903**
68	9	18	**604**
82	8	11	**559**
73	2	7	**429**
64	4	3	**352**
34	0	2	**509**
24	0	1	**509**
28	0	2	**257**
15	0	1	**215**
17	0	1	**163**
11	0	1	**165**
10	0	0	**126**
7	0	0	**120**
2	0	0	**127**
3	0	0	**105**
2	0	0	**86**
0	0	0	**100**
2	0	0	**150**
1	0	0	**49**
3	0	0	**96**
1	0	0	**81**
0	0	0	**63**
2	0	0	**62**
1	0	0	**71**
0	0	0	**88**
0	0	0	**73**
1	0	0	**53**
0	0	0	**63**
0	0	0	**59**
0	0	0	**43**
1	0	0	**41**
0	0	0	**31**
1	0	0	**38**
0	0	0	**32**
0	0	0	**35**
0	0	0	**24**
0	0	0	**22**
0	0	0	**14**
0	0	0	**14**
1	0	0	**245**
1040	**881**	**2653**	**29964**

(c) Llargada de les entitats *(NÚMERO_ CAS, TESTIMONI, ALTRES_ PERSONES i Total)*.

Taula 24: Llargada de les entitats trobades al *dataset*.