UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Grau en Enginyeria Informàtica

Especialitat en Computació

Anàlisi automàtica de tiquets de phishing i malware mitjançant el processament del llenguatge natural

Jaume Casals Vilaplana

Director: Ernest Teniente López

Ponent: Albert Obiols Vives

Tutor de GEP: Carolina Maria Consolación Segura

6 de desembre de 2023

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]

L'impacte d'aquestes ciberamenaces transcendeix les pèrdues financeres meres, impreg-

nant 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. Es posa especial èmfasi en respectar els matisos del sistema de gestió de tiquets emprat per l'**Agència**. La finalitat d'aquesta iniciativa és millorar la capacitat d'aquesta entitat per classificar eficaçment les possibles futures incidències.

La naturalesa intrínsecament sensible del projecte requereix 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 es caracteritza 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 pretén abordar 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 planteja 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 informació 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 és dissenyar i implantar un sistema d'anàlisi i extracció de la informació de tiquets d'incidents que alleugi les deficiències existents. En el projecte també s'inclou el desenvolupament d'una API que permeti utilitzar el sistema de manera senzilla. S'implementa 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 gestió d'incidents proposat 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 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, és 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 és la signatura d'Acords de No Divulgació (NDA) entre les parts implicades, inclosa l'Agència, i2cat i InLab. Aquests acords 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 se sotmetrà 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 d'una eina que aconsegueixi 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.
- Implementar un sistema d'extracció de la informació d'un tiquet basat en l'objectiu anterior per aconseguir un resultat satisfactori.
- Dissenyar i desenvolupar una *pipeline* que extregui els tiquets, extregui i processi les dades, anonimitzi les necessàries i ho emmagatzemi a una altra base de dades.
- Posar en funcionament una API que permeti accedir i utilitzar aquest sistema de manera senzilla.

1.2.2 Requeriments funcionals

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

1. L'usuari introdueix els **paràmetres d'entrada a l'API**, entre els quals s'inclou l'identificador del tiquet a analitzar.
2. Es **comprova si hi ha cap error greu** i es retorna un missatge d'error en aquest cas.
3. S'envia l'ordre a la primera base de dades on, mitjançant l'identificador abans mencionat, es **retorna el tiquet especificat**.
4. **S'executa l'algorisme principal** localment utilitzant el tiquet obtingut i retorna les dades dels camps especificats.
5. El resultat és tractat i es passa per un **algorisme d'anonimització**.
6. **S'emmagatzema** en la segona base de dades escollida.

Aquestes dades que proporcioni l'algorisme seran tractades i passades per un algorisme d'anonimització i, finalment, s'emmagatzemaran a una segona base de dades. Tot aquest procés serà invisible per l'usuari de l'API

1.2.3 Requeriments no funcionals

- **Adaptabilitat:** El sistema ha de permetre l'extracció d'informació de qualsevol seqüència de tiquets independentment de la manera en la qual s'ha escrit. Ha d'aconseguir comprendre el significat dels texts i arribar a conclusions equivalents, fins i tot amb variacions a la redacció. quadern de càrregues que presenti un fabricant independentment de la manera en la qual s'ha escrit.
- **Usabilitat:** L'eina ha de ser fàcil d'usar per facilitar-ne la integració en el flux de treball actual 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 obtenir un temps d'espera curt per evitar que aquest pas esdevingui un coll d'ampolla en el procés o causi molèsties durant el seu ús.
- **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 és 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.

- **La resposta està separada o es troba en articles diferents:** 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, ja que no és factible proporcionar al model una conversa completa d'un tiquet. 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, ja que certes categories d'aquest àmbit no afavoreixen el nivell de flexibilitat desitjat.

- **Escassetat 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 progrés 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.

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 són l'elecció per excel·lència pel desenvolupament d'aquest projecte, més concretament s'ha usat *Scrum*.

Seguint la metodologia *Scrum*, la feina s'organitza 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 es decideixin implementar han de ser d'una durada igual o menor a la durada del *Sprint*, per tant, la feina es divideix en subtasques per arribar a aquesta quota.

A més a més de les reunions anteriorment mencionades, l'equip també es reuneix de manera diària (*dailys*) i setmanal (*weeklys*). Aquestes reunions més breus serveixen per mantenir als desenvolupadors actualitzats i col·laborant mútuament, poden detectar a temps qualsevol problema que pugui sorgir. Independentment d'aquestes reunions, l'equip està 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.

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.
- **Models NLP[2][3][4]**: 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'utilitza OTRS, una eina de gestió i emmagatzematge de tiquets. A continuació, es mostra un tiquet d'exemple d'ORTS i s'explica en deteniment les seves parts.

En el tiquet es pot 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 compatible amb les mateixes llibreries que OTRS i es comporta de manera idèntica. S'utilitzarà 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 fet. 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.

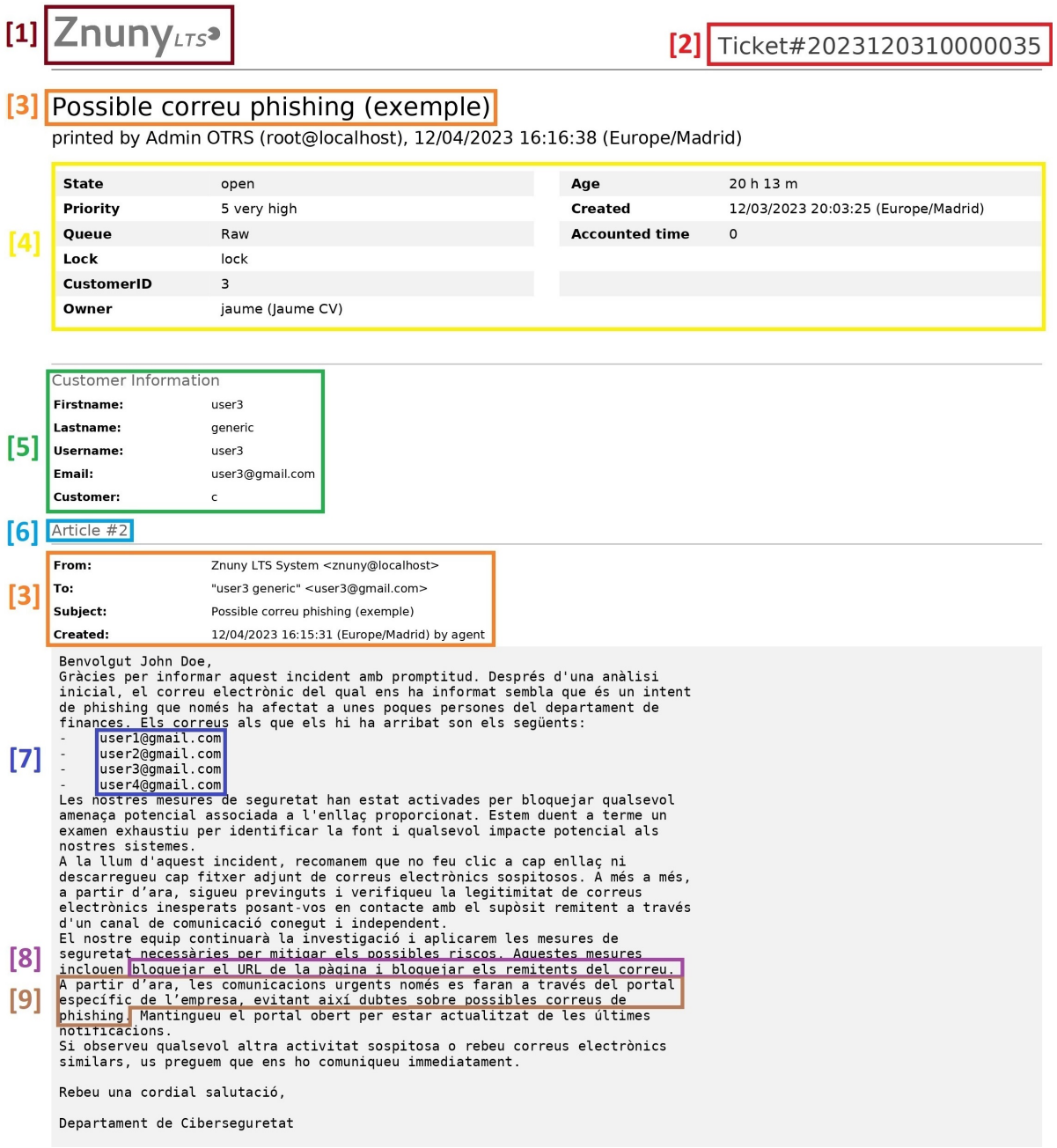

Figura 1: Primera meitat del tiquet d'exemple d'ORTS amb totes les parts indicades. *(Creació pròpia)*

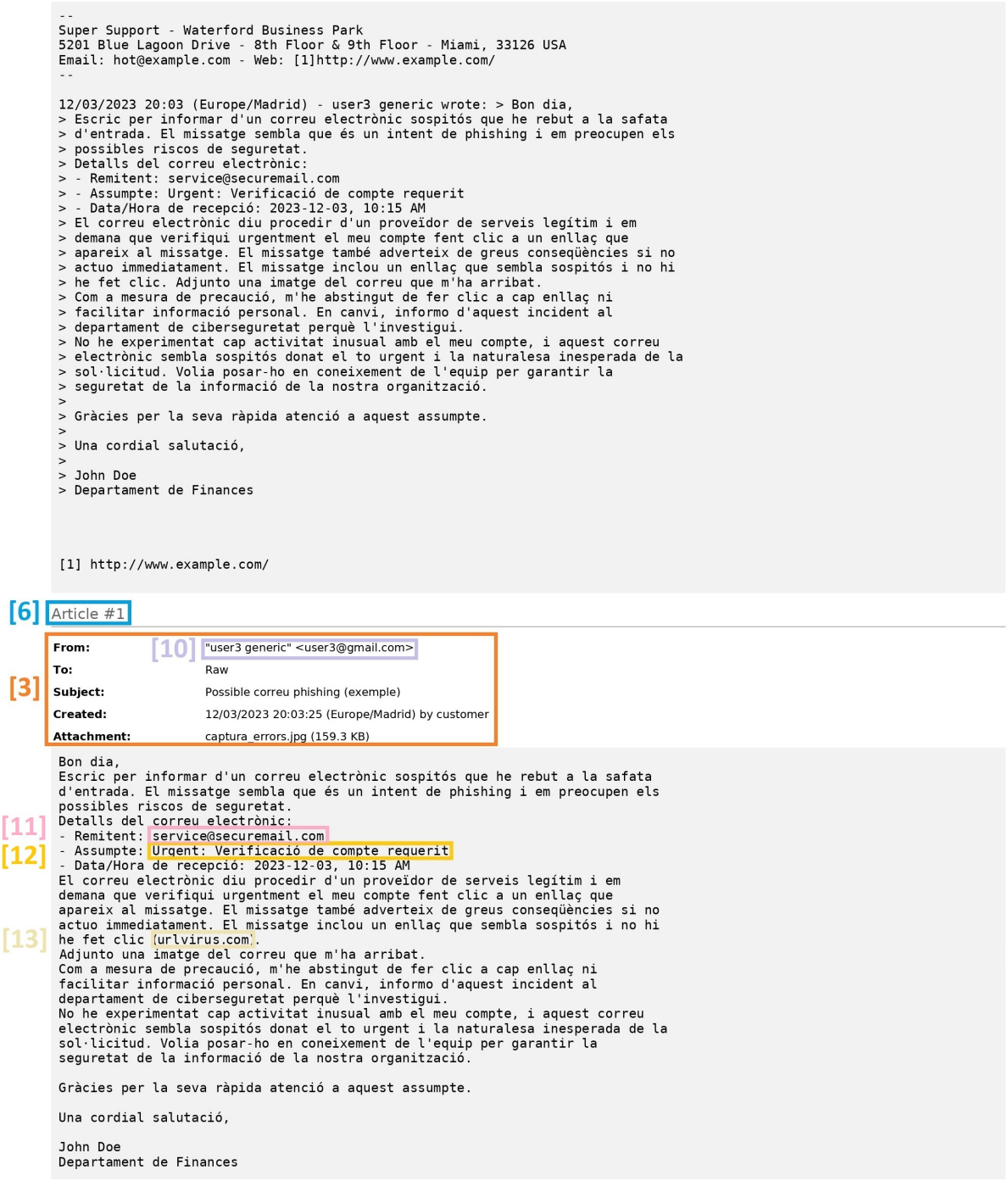

Figura 2: Segona meitat del tiquet d'exemple d'ORTS amb totes les parts indicades. *(Creació pròpia)*

7. Usuaris afectats
8. Accions de mitigació
9. Accions de control
10. Mail de la víctima que informa
11. Mail de l'atacant

12. Assumpte correu de l'incident
13. URL de l'incident
14. Mail de la víctima

NLP

El *Natural Language Processing* també conegut per les seves sigles en anglès *NLP*, és una branca de la intel·ligència artificial que utilitza algorismes i models per a la comprensió i generació de text.

Non-Disclosure Agreement (NDA)

Phishing/programari maliciós

dataset

2.2 Aprenentatge autònom

2.2.1 Models d'aprenentatge autònom

2.2.2 Models del processament del llenguatge natural

2.3 Estat de l'art

2.3.1 Aplicacions comercials

2.3.2 Propostes descartades

2.3.3 Comprovació models disponibles

2.3.4 Models destacats

2.3.5 Justificació de la tria

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 encadenada a les màquines que disposem de l'agència i al sistema que tenen ja muntat fins ara. El sistema consta d'una sèrie de scripts de Python que 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, 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 proporcionada per l'Agència, que substitueix les dades personals o sensibles 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 3 mostra un diagrama de flux que il·lustra el funcionament del sistema.

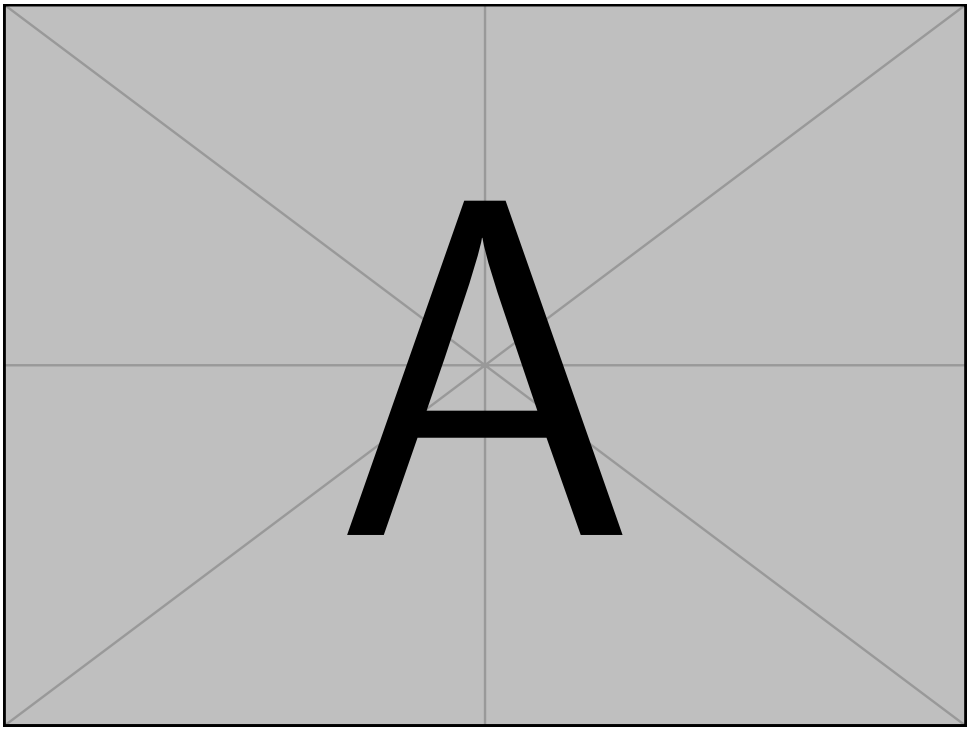

Figura 3: Diagrama de flux del sistema d'anàlisi automàtica de tiquets d'incidències de ciberseguretat.

A continuació es detallen els components i les tecnologies utilitzades a cadascuna de les etapes del pipeline.

3.1.1 Extracció de tiquets d'OTRS

Per extreure els tiquets de la base de dades OTRS de l'Agència, s'ha fet servir l'API REST d'OTRS, que permet accedir a les dades mitjançant peticions HTTP. S'ha implementat un script de Python que utilitza la llibreria requests per fer les peticions i obtenir els tiquets en format JSON. L'script s'encarrega d'autenticar-se amb les credencials de l'Agència, filtrar els tiquets per data, estat i categoria, i emmagatzemar-los en un fitxer local per al processament posterior.

3.1.2 Preprocessament de tiquets

El preprocessament dels tiquets té com a objectiu preparar el text per analitzar-lo pel model de PLN. S'ha implementat un script de Python que realitza les operacions següents sobre el text de cada tiquet:

- Eliminar els elements HTML, les capçaleres, les signatures i els missatges anteriors que puguin contenir els tiquets, ja que no aporten informació rellevant per a l'anàlisi.

- Normalitzar el text, convertint totes les lletres a minúscules, eliminant els signes de puntuació, els números i els caràcters especials, i substituint els espais múltiples per un de sol.
- Tokenitzar el text, és a dir, dividir-lo en unitats mínimes de significat, que en aquest cas són les paraules. S'ha utilitzat la llibreria NLTK per fer aquesta tasca, que ofereix diferents algorismes de tokenització per a diferents idiomes.

El resultat del preprocessament és una llista de tokens per cada tiquet, que s'emmagatzema en un altre fitxer local per a la seva posterior anàlisi.

3.1.3 Aplicació del model de PLN

Per aplicar el model de PLN que detecta i extreu els camps rellevants de cada tiquet, s'ha fet servir la llibreria spaCy, que és un framework de codi obert per al processament del llenguatge natural. SpaCy ofereix diversos models preentrenats per a diferents idiomes, que permeten fer tasques com l'anàlisi sintàctica, l'etiquetació morfològica, l'extracció d'entitats nomenades, la classificació de textos, etc.

En aquest cas, s'ha utilitzat el model preentrenat per a l'espanyol, que s'anomena $es_core\_news\_sm$. Aquest model és capaç de reconèixer les entitats anomenades següents: persona, organització, lloc, producte, esdeveniment, obra d'art, llei, data, hora, percentatge, diners, quantitat, ordinal i cardinal. Tot i això, no totes aquestes entitats són rellevants per a l'anàlisi dels tiquets d'incidències de ciberseguretat, ia més hi ha alguns camps que no es corresponen amb cap d'aquestes entitats, com el tipus d'incidència o el nivell de prioritat.

Per tant, s'ha realitzat un procés d'adaptació del model preentrenat, que consisteix a afegir noves entitats i reentrenar el model amb un conjunt de dades etiquetat específicament per al domini dels tiquets de ciberseguretat. El conjunt de dades etiquetatge s'ha obtingut a partir d'una mostra de tiquets reals de l'Agència, que s'han anotat manualment amb les entitats següents:

- TIPUS: el tipus d'incidència, que pot ser phishing, malware, ransomware, atac DDoS, etc.
- PRIORITAT: el nivell de prioritat assignat al tiquet, que pot ser baixa, mitjana, alta o crítica.
- USUARI: nom o àlies de l'usuari afectat per la incidència.
- DISPOSITIU: el nom o l'adreça IP del dispositiu compromès per la incidència.
- DATA: la data en què es va produir o es va reportar la incidència.

- HORA: l'hora en què es va produir o es va reportar la incidència.
- ORG: l'organització a què pertany l'usuari o dispositiu afectat, que es manté com una entitat preexistent del model.
- LOC: el lloc des del qual es va originar o detectar la incidència, que es manté com una entitat preexistent del model.

El procés d'adaptació del model s'ha realitzat seguint els passos descrits a la documentació de spaCy, que consisteixen en:

1. Crear un objecte de la classe Language a partir del model preentrenat.
2. Afegir les noves entitats al component de reconeixement d'entitats nomenades (NER) del model, mitjançant el mètode add_label.
3. Deshabilitar els altres components del model, com l'analitzador sintàctic o l'etiquetador morfològic, ja que no es faran servir i poden interferir amb l'entrenament del NER.
4. Crear un objecte de classe EntityRecognizer a partir del component NER del model, i assignar-li un optimitzador, una funció de pèrdua i unes mètriques d'avaluació.
5. Entrenar l'objecte EntityRecognizer amb el conjunt de dades etiquetatge, mitjançant el mètode update, que rep com a paràmetres els textos i les anotacions de les entitats, i torna la pèrdua i les mètriques a cada iteració.
6. Desar el model adaptat en un fitxer local, mitjançant el mètode to_disk.

El resultat de l'entrenament és un model de PLN capaç de detectar i extreure els camps rellevants dels tiquets d'incidències de ciberseguretat, que s'aplica a cada tiquet preprocessat i s'emmagatzema el resultat en un altre fitxer local per a la seva posterior anonimització.

3.1.4 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 l'Agència, 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 "Juan Pérez", la funció torna "NOM COGNOM", i si el text és "192.168.1.1", la funció torna ÏP 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 PLN, i torna com a sortida un altre fitxer amb els camps anonimitzats, que s'emmagatzemen en un diccionari amb la següent estructura:

3.2 Creació dataset sintètic

3.3 Creació de tests

3.4 Finetune sintètic

3.5 Dataset i finetune amb dades reals

3.5.1 Estadístiques descriptives

3.5.2 Tendències i patrons

3.5.3 Problemes i incidències comunes

3.5.4 Anomalies i valors atípics

3.6 Desplegament API

Capítol 4

Avaluació amb resultats reals

4.1 Eficàcia de la solució i anàlisi de resultats

Capítol 5

Planificació temporal

Segons el conveni de cooperació educativa acordat, aquest projecte s'inicia el dia 16 de setembre de 2023 i acaba 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. Es treballen 5 hores al dia, tot i que, puntualment, es treballà 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, es dediquen unes 600 hores en aquest projecte. Es dediquen 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 1: Taula de tasques amb les dependències, els recursos i l'estimació de les hores (Elaboració pròpia)

5.4 Diagrama de Gantt

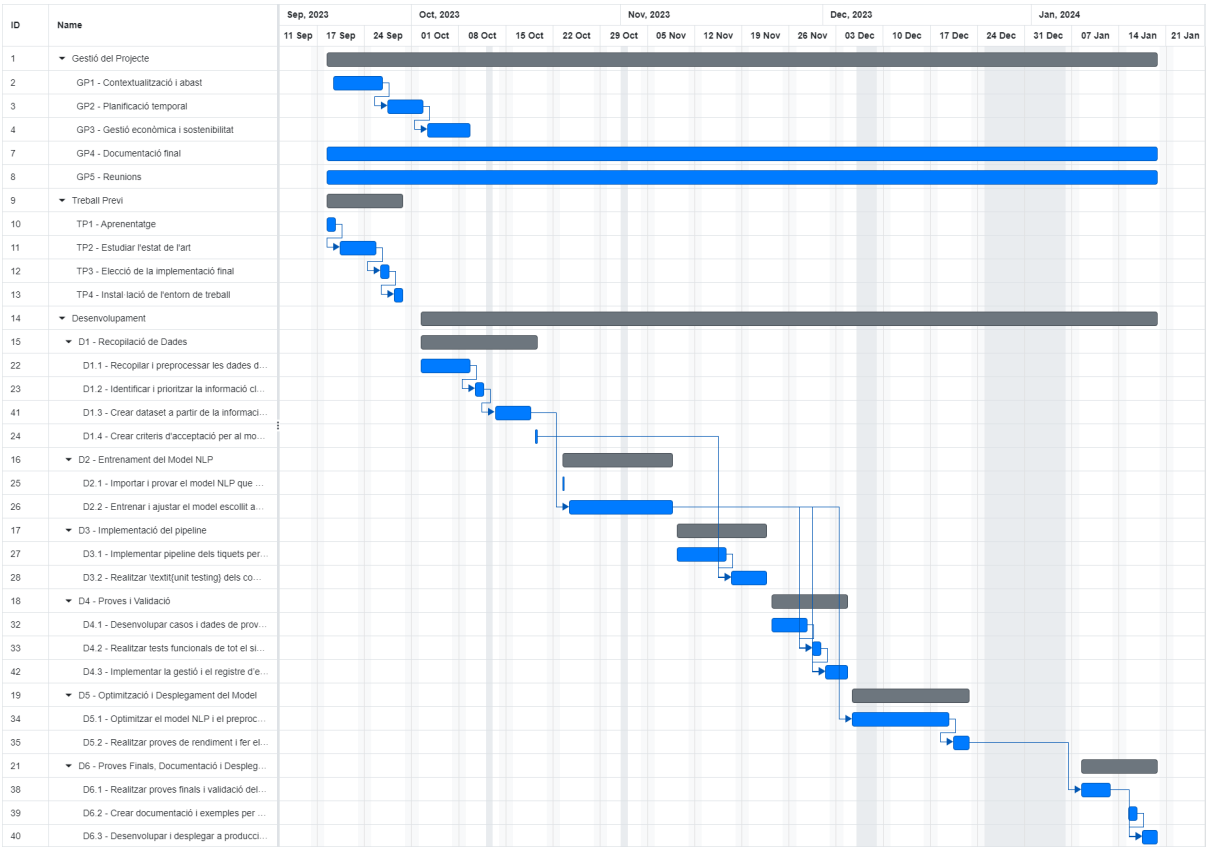

Figura 4: Diagrama de Gantt del projecte
(Elaboració pròpia amb *onlinegantt.com*)

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, ara es poden anticipar els possibles retards i dificultats que provoquin aquests punts. Naturalment, la gestió de riscs és un procés continu i aquests busquen 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

es treballa sigui confidencial i l'alta potència que requereixen les solucions amb els quals es treballen, crea una barrera entre tots els models disponibles i els que podem utilitzar. És important mencionar, que el problema amb el qual es treballa 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 podria provocar 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 fa 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 garanteix una anàlisi més coherent, sinó que també ofereix 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 que no sigui així, 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 planteja 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 molt 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).

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'utilitzen 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€[5]	15.008,7€	65.037€	31,27€
Desenvolupador Júnior (**DJ**)	23.631€[6]	7.089,3€	30.720,3€	14,7€
Expert científic (**EC**)	54.000€[6]	16.200€	70.200€	33,59€

Taula 2: 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 3. El preu de cadascuna es determina multiplicant el nombre d'hores que cada perfil hi dedicarà pel cost per hora determinat a la taula 2.

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 3: 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 estan sota control o són desconeguts i, a més a més, poden variar. En tot cas, aquest programari no afecta al pressupost, i, per tant, no es tindrà en compte. A continuació es calcula el cost del programari utilitzat:

- **Google Drive:** No es contracta cap espai extra.
- **Overleaf:** No s'utilitzen les característiques de la subscripció mensual.
- **Git:** No és necessari l'ús dels extres que s'ofereixen.
- **PyCharm:** S'usa la llicència d'estudiant que dura un any.

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

$$\text{Amortització Software} = 0€$$

Hardware

Només és necessari invertir per un ordinador de sobretaula pels tres integrants de l'equip per la realització d'aquest projecte. S'estima que un ordinador com els fets servir amb els seus dos monitors i perifèrics costa uns 1200€. La vida útil d'aquest ordinador s'estima 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 es calcularà 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 3, 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 [7], un ordinador amb un consum de 400 W [8], les dues pantalles de cada monitor són idèntiques amb un consum màxim de 35 W [9].

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 4: Taula del consum elèctric
(Elaboració pròpia)

6.2.3 Connexió internet

Es calcula que el preu de la connexió a internet ronda els 50€ per mes. Amb un projecte de quatre mesos de duració, el total és 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'estima segons el cost mitjà de les oficines disponibles per lloguer en aquesta època [10]. Es calcula un preu de $21€/m^2$ amb l'aigua inclosa i són 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 5: 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 es tenen 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 millora l'èxit global del projecte i minimitza 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 poden sorgir durant el desenvolupament de les tasques planejades, és 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 garanteix que estiguem preparats financerament per fer front als riscos que s'han pogut intuir, mantenint l'estabilitat del projecte i salvaguardant l'èxit general de l'empresa.

Imprevist	Tasques afectades	Despesa original	Risc	Sobrecost
Models insuficients	TP2, TP3	4671,1€	15%	700,67€
Diversitat de les dades	D1.1, D1.2	1666,0€	5%	83,3€
No tenir suficient informació	D1.1	891,35€	10%	89,14€
Falta d'experiència	TP1, TP2	2923,55€	10%	292,36€
Total				**1165,46€**

Taula 6: Taula dels sobrecostos afegits pels imprevistos
(Elaboració pròpia)

6.5 Cost total del projecte

En la Taula 7 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	1165,46€
Total	**30980,20€**

Taula 7: 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 consisteix a aprofitar metodologies ja establertes per controlar i mantenir el pressupost. Les reunions periòdiques de *sprint* permeten 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'usen indicadors clau de rendiment per identificar l'origen del problema. A més a més, els membres de l'equip fan 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

8.2 Coneixement de les assignatures

Capítol 9

Conclusions

9.1 Assoliment dels objectius

9.1.1 Estudi de l'estat de l'art

9.1.2 Sistema d'extracció d'informació

9.1.3 Implementació del pipeline

9.1.4 Desplegament API

9.2 Treball futur 9.3 Conclusions personals

9.2 Treball futur

9.3 Conclusions personals

9.3 Conclusions personals

Capítol 10

Apèndix

Referències

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

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". A: *North American Chapter of the Association for Computational Linguistics*. Accedit: 21/02/2023. 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[3] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". A: *ArXiv* abs/1907.11692 (2019). Accedit: 21/02/2023. URL: https://api.semanticscholar.org/CorpusID:198953378.

[4] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". A: (2019). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

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

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

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

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

[9] *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.

[10] *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.