



UNIVERSIDADE DA CORUÑA

Facultade de Informática

Máster Universitario en Enxeñaría en Informática

TRABALLO FIN DE MÁSTER

**MindWell: un *chatbot open source* como
axuda para a detección e o seguemento de
trastornos depresivos**

Estudante: Eliseo Bao Souto

Dirección: Miguel Anxo Pérez Vila

Javier Parapar López

A Coruña, febreiro de 2024.

Agradecementos

Este traballo naceu en setembro do 2021. Un correo electrónico que se convertiu en TFG, que se convertiu en relación laboral, que se convertiu en TFM, que agora se converte nun novo propósito de vida. No contexto deste TFM quero agradecer a Javi e a Anxo a súa orientación e os seus consellos pero, dunha maneira máis ampla, quero agradecer tamén a todos os demais membros do IRLab a súa amizade: David, Manu, Jorge, Juan, Paloma, Antón, Edu, Gilberto, Alfonso, César, Brais e Adrián. Teño a sorte de compartir o meu día a día con vós. Grazas a todos vós e tamén a Álvaro que foi quen me deu a oportunidade en primeiro lugar.

No ámbito persoal síntome tamén moi afortunado por ter ao meu carón as persoas que me queren. Debo recoñecerlle a meus pais todo o seu apoio durante estes anos e a educación que recibín pola súa parte, dous aspectos fundamentais para a consecución dos meus logros. Por último, unha mención especial a Lucía, en quen vexo unha persoa coa que seguir mellorando día a día e construíndo futuro.

Resumo

As persoas usuarias de redes sociais adoitan percibir estas plataformas como espazos de apoio nos que expoñer, comentar e desafogarse respecto dos seus problemas do día a día, de xeito que a súa actividade nestas canles pode conter evidencia sobre o estado da súa saúde mental. Recentemente, investigacións dos campos da *Information Retrieval* (IR), *Natural Language Processing* (NLP) e *Machine Learning* (ML) explotaron esta información en liña para desenvolver modelos de detección de depresión, cuxo obxectivo é identificar as persoas en situación de risco en plataformas como Twitter, Reddit ou Facebook. Relacionado con isto, outras investigacións recentes sinalaron a importancia de utilizar marcadores clínicos, como o uso de síntomas validados, para mellorar a confianza nos modelos computacionais por parte dos profesionais sanitarios. Neste traballo preséntase un *chatbot open source* concebido como un asistente que proporcione explicacións aliñadas con marcadores clínicos sobre a presenza de síntomas de depresión validados nas publicacións en redes sociais, tendo en conta tamén a temporalidade dos mesmos. O obxectivo é desenvolver unha ferramenta que inclúa as funcionalidades necesarias para proporcionar os beneficios comentados anteriormente. Deste xeito, dotaríase aos profesionais dunha ferramenta de apoio que os descargue dunha labor tediosa e moi intensiva en tempo como é o feito de revisar a man o histórico de publicacións dun suxeito. Avaliamos a nosa proposta utilizando coñecemento experto para poder medir a calidade e a aplicabilidade das explicacións do *chatbot* para o ámbito clínico real, e os resultados demostraron a utilidade do sistema para xerar análises baseadas nos síntomas.

Abstract

Social media users often perceive these platforms as supportive spaces in which to expose, comment and disclose their daily problems, and thus their activity can provide clues to their mental health status. Research on *Information Retrieval* (IR), *Natural Language Processing* (NLP) and *Machine Learning* (ML) has recently used this online information to develop screening models that aim to identify at-risk individuals on platforms such as Twitter, Reddit or Facebook. Recently, research has highlighted the importance of using clinical markers, such as the use of validated symptoms, to improve health professionals' confidence in computational models. This work presents a *open source chatbot* designed as an assistant that provides explanations, aligned with validated clinical markers, for the presence of depressive symptoms in social media posts, taking into account the temporality of these symptoms. The aim is to develop a tool that includes the necessary functionalities to provide the above-mentioned benefits. Following this approach, it is possible to provide professionals with a

support tool that relieves them of the tedious and time-consuming task of manually reviewing a subject's posting history. We evaluated our proposal using expert knowledge to measure the quality and applicability of the chatbot's explanations to the real clinical setting, and the results demonstrated the usefulness of the system for generating analyses based on the subject's feelings.

Palabras chave:

- Asistente conversacional
- Procesamento da Linguaxe Natural
- Gran Modelo de Linguaxe
- IA Conversacional
- IA Explicable
- Xeración Aumentada con Recuperación
- Detección de depresión
- Software Libre

Keywords:

- Chatbot
- Natural Language Processing
- Large Language Model
- Conversational AI
- Explainable AI
- Retrieval Augmented Generation
- Depression detection
- Open Source

Índice Xeral

1	Introdución	1
1.1	Motivación	1
1.2	Obxectivos	2
1.3	Estrutura da memoria	3
1.4	Plan de traballo	4
2	Fundamentos e conceptos básicos	7
2.1	<i>Open source</i>	7
2.2	<i>Large Language Models (LLMs)</i>	8
2.3	Intelixencia Artificial conversacional	10
2.4	Intelixencia Artificial explicable	12
2.5	<i>Retrieval Augmented Generation (RAG)</i>	12
3	Tecnoloxías e ferramentas	15
3.1	Linguaxes de programación	15
3.1.1	Python	15
3.1.2	JavaScript	16
3.2	Librarías e compoñentes	16
3.2.1	Transformers	16
3.2.2	DeepSpeed	17
3.2.3	FastChat	17
3.2.4	FastAPI	17
3.2.5	Elasticsearch	17
3.2.6	LangChain	18
3.2.7	Django	18
3.3	Ferramentas de desenvolvemento	19
3.3.1	VSCodium	19
3.4	Ferramentas de metodoloxía	19

3.4.1	Xestión de paquetes	19
3.4.2	Docker	20
3.4.3	Xestión do proxecto	20
3.4.4	Control de versións	20
4	Metodoloxía e xestión do proxecto	23
4.1	Metodoloxía	23
4.1.1	Escolla da metodoloxía	23
4.1.2	Scrum	24
4.1.3	Adaptación de Scrum a este traballo	27
4.2	Xestión do proxecto	28
4.2.1	Estimación	28
4.2.2	Recursos	28
4.2.3	Custes	29
4.2.4	Xestión de riscos	31
5	Proposta	33
5.1	Arquitectura	33
5.1.1	Aplicación base	33
5.1.2	Incremento	35
5.2	Fontes de datos e conxunto creado	36
5.3	<i>Fine-tuning</i> do modelo	39
6	Desenvolvemento	43
6.1	Análise de requisitos	43
6.1.1	Requisitos funcionais	43
6.1.2	Requisitos non funcionais	45
6.1.3	Historias de persoa usuaria	45
6.2	<i>Sprints</i>	47
6.2.1	<i>Sprint</i> 1 (21/09/23 ao 08/10/23): Formación teórica e escolla tecnolóxica	47
6.2.2	<i>Sprint</i> 2 (09/10/23 ao 25/10/23): Elaboración do conxunto de datos	50
6.2.3	<i>Sprint</i> 3 (26/10/23 ao 12/11/23): <i>Fine-tuning</i> do LLM	51
6.2.4	<i>Sprint</i> 4 (13/11/23 ao 30/11/23): Interacción co LLM (1/2)	53
6.2.5	<i>Sprint</i> 5 (01/12/23 ao 17/12/23): Interacción co LLM (2/2)	55
6.2.6	<i>Sprint</i> 6 (18/12/23 ao 03/01/24): Incorporación do contexto	56
6.2.7	<i>Sprint</i> 7 (04/01/24 ao 21/01/24): Avaliación e peche do proxecto	58
6.2.8	Balance final	59

7	Produto Final	63
7.1	Resumo da aplicación	63
7.2	<i>Software</i> liberado	68
8	Conclusións, relación co <i>Master</i> e traballo futuro	71
8.1	Conclusións	71
8.2	Relación co <i>Master</i>	72
8.2.1	Relación con asignaturas	72
8.2.2	Relación con competencias do título	73
8.3	Traballo futuro	75
9	Compromiso ético	77
9.1	Supervisión dun perfil clínico experto	77
9.2	Cumprimento dos acordos de uso dos conxuntos de datos	77
9.3	Emisións de CO ₂ relacionadas co proxecto	78
A	Licenzas <i>software</i>	81
B	Cuestionario BDI-II	83
C	Equivalencia entre DSM-5 e BDI-II	89
D	Equivalencia entre <i>PsySym</i> e BDI-II	91
	Relación de Acrónimos	93
	Glosario	95
	Bibliografía	97

Índice de Figuras

2.1	Vista xeral das compoñentes <i>encoder</i> e <i>decoder</i> da arquitectura Transformer.	9
2.2	Vista xeral dun sistema completo de IA Conversacional.	11
2.3	Vista xeral dun sistema de <i>Retrieval Augmented Generation</i> (RAG).	13
3.1	Índice TIOBE de popularidade con data de Novembro do 2023.	16
4.1	Detalle do proceso de <i>Scrum</i> . Diagrama recollido da guía oficial.	25
5.1	Arquitectura da aplicación. En verde o incremento asociado a este TFM. En azul as modificacións substanciais sobre o proxecto base.	34
5.2	Plantillas definidas para os exemplos tanto positivos como negativos dos turnos.	39
5.3	Evolución da <i>loss</i> obtida durante o proceso de <i>fine-tuning</i> en función das épocas consideradas.	41
5.4	Consumo de GPU do modelo durante o <i>fine-tuning</i>	41
6.1	Planificación inicial do proxecto.	47
6.2	Planificación inicial do <i>Sprint</i> 1.	48
6.3	Historias finalmente completadas no <i>Sprint</i> 1.	49
6.4	Progreso do proxecto ao peche do <i>Sprint</i> 1.	49
6.5	Planificación inicial do <i>Sprint</i> 2.	51
6.6	Historias finalmente completadas no <i>Sprint</i> 2.	51
6.7	Progreso do proxecto ao peche do <i>Sprint</i> 2.	51
6.8	Planificación inicial do <i>Sprint</i> 3.	52
6.9	Historias finalmente completadas no <i>Sprint</i> 3.	53
6.10	Progreso do proxecto ao peche do <i>Sprint</i> 3.	53
6.11	Planificación inicial do <i>Sprint</i> 4.	54
6.12	Historias finalmente completadas no <i>Sprint</i> 4.	54
6.13	Progreso do proxecto ao peche do <i>Sprint</i> 4.	55

6.14	Planificación inicial do <i>Sprint</i> 5.	56
6.15	Historias finalmente completadas no <i>Sprint</i> 5.	56
6.16	Progreso do proxecto ao peche do <i>Sprint</i> 5.	56
6.17	Planificación inicial do <i>Sprint</i> 6.	57
6.18	Historias finalmente completadas no <i>Sprint</i> 6.	58
6.19	Progreso do proxecto ao peche do <i>Sprint</i> 6.	58
6.20	Planificación inicial do <i>Sprint</i> 7.	58
6.21	Historias finalmente completadas no <i>Sprint</i> 7.	59
6.22	Progreso do proxecto ao peche do <i>Sprint</i> 7.	59
7.1	Vista da pantalla de <i>login</i>	64
7.2	Vista do panel de control da aplicación.	64
7.3	Creación dun <i>Corpus</i> no que almacenar a información.	65
7.4	Obtendo os comentarios de múltiples perfís dun determinado <i>subreddit</i>	65
7.5	Obtendo os comentarios dun perfil (suxeito) concreto de Reddit.	66
7.6	Vista da pantalla detalle dun perfil (suxeito).	66
7.7	Vista da pantalla que amosa os perfís (suxeitos) cos que se está a traballar.	67
7.8	Vista do <i>chat</i> con MindWell sobre un perfil (suxeito) determinado.	69

Índice de Táboas

4.1	Estimación salarial para os recursos humanos do proxecto.	30
4.2	Detalle dos custes totais dos recursos humanos.	30
4.3	Detalle dos custes totais dos recursos materiais e de <i>software</i>	30
4.4	Detalle da identificación e clasificación dos riscos presentes no proxecto.	31
5.1	Cada fila presenta un exemplo das fontes de datos. Cabe destacar que estas mostras están parafraseadas en aras da privacidade e acurtadas para maior claridade. En negriña represéntase a evidencia do síntoma de depresión no texto.	37
5.2	Estatísticas dos exemplos positivos do conxunto de datos elaborado. As lonxitudes medias exprésanse en <i>tokens</i>	37
5.3	Táboa resumo dos parámetros e configuracións utilizados para o proceso de <i>fine-tuning</i> de Vicuna-7B-1.5.	40
6.1	Historias de Persoa Usuaria e puntos estimados.	46
6.2	Historias de Persoa Usuaria e puntos reais.	60
6.3	Relación entre <i>Sprints</i> , Tarefas e Historias.	61
A.1	Relación entre os distintos <i>softwares</i> empregados e as súas licenzas.	81
C.1	Equivalencia entre DSM-5 e BDI-II.	89
D.1	Equivalencia entre <i>PsySym</i> e BDI-II.	91

Introdución

O propósito deste Capítulo é servir como preámbulo dos que a continuación o acompañan. Así pois, as Seccións 1.1, 1.2, 1.3 e 1.4 presentan respectivamente a Motivación, os Obxectivos, a Estrutura da memoria e o Plan de traballo deste *Traballo de Fin de Master (TFM)*. A idea detrás desta conceptualización é proporcionar un contexto que dote de cohesión o resto do traballo e lle dea cabida dentro do actual panorama social e informático.

1.1 Motivación

A saúde mental é unha compoñente crucial da saúde e o benestar xeral. As cifras máis recentes estiman que a prevalencia dos trastornos mentais en persoas adultas supera o 20% [1]. A *Organización Mundial da Saúde (OMS)* estima que, aproximadamente, 332 millóns de persoas en todo o mundo están afectadas por trastornos de tipo depresivo [2]. A intervención temperá nos trastornos mentais é crucial para mitigar o seu impacto, especialmente en individuos novos [3]. Con todo, debido ao estigma asociado á depresión, máis do 60% das persoas que padecen este tipo de trastornos non buscan apoio profesional [4]. Para facer fronte a este problema, existe unha interese crecente en axudar na detección temperá e o diagnóstico da depresión, mitigando así o seu impacto social [5].

Neste escenario, as publicacións nas redes sociais eríxense como recursos extremadamente valiosos para buscar signos temperáns de depresión [6, 7, 8, 9, 10, 11], dado que as persoas que experimentan depresión adoitan sentirse máis cómodas expresando os seus pensamentos e emocións nestas plataformas que na vida real motivadas por factores como a privacidade ou o anonimato [12]. Por isto, e dado que a cantidade de datos procedentes das redes sociais é inxente, o uso de técnicas computacionais apoiadas na lingüística deu lugar a importantes avances na detección de signos de depresión [13]. Non obstante, os avances neste eido non pretenden substituír aos profesionais da saúde mental, senón ofrecer apoio á súa labor [14]. Así pois, calquera tipo de *software* derivado deste tipo de investigacións só debe ser usado

para ampliar o alcance dos profesionais e facilitar o seu fluxo de traballo ou, complementariamente, para investigar e analizar patróns dentro das redes sociais no marco de traballos analíticos. Con todo, aínda existe marxe de mellora neste tipo de sistemas para que poidan ser usados en escenarios reais, polo que gañar a confianza dos expertos e aínda un reto de cara á implantación nos dominios clínicos [15]. Unha forma de *conseguir* confianza é deseñar métodos que se fundamenten no uso de síntomas recollidos en cuestionarios clínicos validados [16, 17, 18, 19, 20]. A maioría destas propostas, no campo da depresión, utilizan os marcadores do *Beck Depression Inventory, Second Version* (BDI-II) [21], que engloba unha serie de síntomas como a irritabilidade, o pesimismo ou os problemas de sono. A modo ilustrativo, o Apéndice B recolle unha transcripción literal deste cuestionario.

Con esta motivación en mente, identificouse como área de potencial innovación a creación dun *chatbot* como ferramenta de apoio á revisión do histórico de publicacións dun suxeito na procura de evidencias de síntomas de depresión, unha tarefa tediosa e moi intensiva en tempo. Como alternativa, valéndose do *chatbot*, os equipos clínicos poden aforrar parte dese tempo e adicalo a outros pacientes ou ás tarefas que consideren oportunas. Isto desenvólvese de xeito máis detallado na seguinte Sección, onde se presentan os Obxectivos.

1.2 Obxectivos

O principal obxectivo deste TFM é desenvolver un *chatbot open source* capaz de analizar publicacións no foro en liña Reddit¹ na procura dos síntomas de depresión reflexados no inventario BDI-II (ver Apéndice B), proporcionando con isto información útil no marco da labor clínica acometida polas persoas que exercen a psicoloxía. Este obxectivo pode descompoñerse nas seguintes tarefas específicas:

Obxectivo 1: Recompilar e *curar* un conxunto de datos que permita o *fine-tuning* dun *Large Language Model* (LLM), neste caso orientado á análise de texto no ámbito dos trastornos depresivos. A día de hoxe o fluxo de traballo habitual, e tamén o que mellores resultados proporciona, é partir dun gran modelo *pre-trained* nun conxunto de datos inxente para logo realizar *fine-tuning* de cara a especializar o modelo nunha tarefa determinada. É por isto que o feito de contar cun conxunto de datos de calidade resulta un aspecto clave no resultado final do proxecto.

Obxectivo 2: Realizar *fine-tuning* do LLM Vicuna-7B-1.5² [22], un modelo conversacional *open source* creado a partir de Llama-2³ [23]. Para este proceso de *fine-*

¹ <https://www.reddit.com/>

² <https://huggingface.co/lmsys/vicuna-7b-v1.5>

³ <https://huggingface.co/meta-llama/Llama-2-7b-hf>

tuning empregárase *Low-Rank Adaptation of Large Language Models* (LoRA), unha técnica que permite acelerar o proceso de adestramento ao tempo que se reduce o consumo de recursos. Desde a súa publicación a comezos do ano 2023, as diferentes versións de Llama creadas por Meta AI⁴ eríxíronse como firmes alternativas a outros modelos privativos coma ChatGPT [24]. Así pois, mediante esta tarefa preténdese ensinarlle os síntomas de depresión ao LLM para que logo este sexa capaz de detectalos e explicalos en conxuntos de texto.

Obxectivo 3: Crear índices densos usando o *framework* Elasticsearch, é dicir, representacións vectoriais en forma de *embeddings* que garden o histórico de publicacións en Reddit dos suxeitos estudados. Estes *embeddings* son tamén computados co LLM descrito no anterior obxectivo.

Obxectivo 4: Conectar o LLM resultante cos índices descritos no anterior obxectivo, para así garantir un sistema que, dado que realiza as súas explicacións a partir dun almacén de datos xa existente, sexa máis robusto e fiable en canto ás súas predicións.

Obxectivo 5: Implementar unha *Application Programming Interface* (API) que sirva o modelo resultante e atenda o envío de mensaxes a través do *chat*.

Obxectivo 6: Desenvolver unha interface gráfica para a interacción co *chatbot*.

Ademais das anteriores tarefas, son tamén obxectivos deste proxecto aplicar correctamente a metodoloxía de desenvolvemento plantexada, realizar unha planificación adecuada e manter un control de versións estrito.

1.3 Estrutura da memoria

Esta Sección trata de dar unha visión global da propia memoria, describindo brevemente cada unha das partes que a compoñen:

Capítulo 1: Introduce o presente proxecto, contextualizándoo e definindo a motivación e os obxectivos. Así mesmo, contén a estrutura da memoria e o plan de traballo a seguir.

Capítulo 2: Trata de introducir as bases teóricas sobre as que se sustenta este traballo, presentando os fundamentos e conceptos básicos necesarios.

Capítulo 3: Describe e analiza as linguaxes de programación, librerías e ferramentas empregadas no contexto do proxecto, xustificando para cada unha delas a súa utilización.

⁴<https://ai.meta.com/>

Capítulo 4: Detalla a metodoloxía seguida durante o transcurso do proxecto, así como a xestión do mesmo. Para isto, xustifícase e preséntase a propia metodoloxía, así como tamén a adaptación realizada sobre a mesma.

Capítulo 5: Expón ao completo a proposta deste proxecto, aportando os detalles necesarios para comprendela no seu conxunto.

Capítulo 6: Expón os detalles de deseño e implementación do *software* realizado para o proxecto.

Capítulo 7: Presenta o produto final desenvolvido, analizando as súas particularidades e aplicacións.

Capítulo 8: Analiza o cumprimento ou non dos obxectivos inicialmente establecidos en forma de conclusións, detallando tamén aspectos positivos e negativos. A maiores, introdúcense posibles liñas futuras e ampliacións para o traballo.

Capítulo 9: Reflexiona sobre o traballo cun profesional clínico, o cumprimento dos acordos de uso dos conxuntos de datos e as emisións de CO₂ relacionadas co proxecto.

Apéndice A: Presenta unha transcripción literal do cuestionario psicolóxico *Beck Depression Inventory, Second Version* (BDI-II).

Apéndice B: Presenta a equivalencia definida entre *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) e BDI-II.

Apéndice C: Presenta a equivalencia definida entre *PsySym* e BDI-II.

1.4 Plan de traballo

Para poder levar a cabo o proxecto seguíronse as seguintes fases relacionadas coa metodoloxía empregada:

- Fase 1:** Estudo da propia metodoloxía, así como dos conceptos de base necesarios para levar a cabo o proxecto, centrándose principalmente no *State-of-the-Art* (SOTA) en canto ao ámbito dos LLMs.
- Fase 2:** Análise dos requisitos funcionais e non funcionais para definir as tarefas correspondentes.
- Fase 3:** Estudo das tecnoloxías e ferramentas dispoñibles, seleccionando aquelas que se consideren óptimas en cada momento.

Fase 4: Dado que a metodoloxía queda encadrada dentro das denominadas “áxiles” e é incremental, en cada unha das iteracións realizáronse os seguintes pasos:

Paso 1: Xestión das tarefas identificadas.

Paso 2: Desenvolvemento e implementación das mesmas.

Paso 3: Avaliación dos resultados obtidos.

Fase 5: Documentación do traballo e elaboración da memoria.

Fundamentos e conceptos básicos

ESTE Capítulo debe entenderse como unha introdución ás bases teóricas sobre as que se sustenta todo o traballo realizado. Así pois, o seu propósito é presentarlle ao lector unha serie de conceptos que fagan desta memoria un documento autocontido. Para iso, e dado que é un compromiso firme adquirido desde o primeiro momento deste TFM, a Sección 2.1 expón brevemente a idea de *open source* e a motivación para tomar esta decisión. Logo, na Sección 2.2, introdúcense os *Large Language Model* (LLM). Para continuar, nas Seccións 2.3 e 2.4, preséntanse os conceptos de Intelixencia Artificial Conversacional e Intelixencia Artificial Explicable, respectivamente. Por último, a Sección 2.5 detalla un aspecto clave para este TFM como é a *Retrieval Augmented Generation* (RAG).

2.1 *Open source*

A *Open Source Initiative* (OSI)¹ é unha organización sen ánimo de lucro adicada a promover e protexer o *software* de código aberto. Fundada en 1998, a OSI desempeña un papel central no dictame e o mantemento da definición de *open source*², dado que *open source* non significa unicamente ter acceso ao código. A definición está conformada por un conxunto de criterios que deben cumprirse para que un *software* se considere verdadeiramente de código aberto. Estes inclúen a libre redistribución, a inclusión do código fonte, o permiso para obras derivadas, a integridade do código fonte, a non discriminación de persoas ou grupos, a ausencia de restriccións en canto aos campos de uso, a igualdade de aplicación dos dereitos da licenza, a independencia das licenzas respecto dos produtos, a ausencia de restriccións sobre outro *software* e a neutralidade tecnolóxica das licenzas.

Desenvolver un proxecto baixo unha licenza *open source* ofrece unha serie de vantaxes que levaron ao autor deste traballo a tomar esta decisión. Primeiramente, promove a transparencia e a confianza ao permitir o acceso ao código fonte. Isto, ademais, facilita a identificación

¹ <https://opensource.org/about/>

² <https://opensource.org/osd/>

de erros e fortalece a base de persoas usuarias. Así pois, a colaboración global facilita a diversidade de ideas e contribucións, o que acelera o desenvolvemento e mellora a calidade do final do *software*. Desde o punto de vista ético, ser partícipe do movemento *open source* implica comprometerse cos valores de liberdade e igualdade de acceso ao coñecemento, contribuíndo así á construción dun ambiente dixital máis igualitario e inclusivo. Cabe tamén destacar, como nota de peche, que o compromiso *open source* deste proxecto non se limita ao código implementado, senón que tamén se opta por modelos neuronais que seguen esta filosofía.

A este efecto, o Apéndice A recolle unha Táboa coa relación entre os distintos *softwares* empregados e as súas licenzas.

2.2 Large Language Models (LLMs)

Nos últimos anos o campo do *Natural Language Processing* (NLP) experimentou un cambio de paradigma coa aparición dos *Large Language Model* (LLM). Estes modelos, caracterizados polo seu inmenso tamaño en canto ao número de parámetros e o seu rendemento nunha ampla gama de tarefas relacionadas co ámbito lingüístico, revolucionaron a forma en que os sistemas *comprenden* e xeran texto de xeito similar a como o fan os humanos. As raíces dos LLMs remóntanse aos primeiros avances no modelado estatístico da linguaxe e a aprendizaxe automática. Os enfoques tradicionais, como os modelos de n-gramas, baseábanse en información contextual limitada e tiñan dificultades para captar as complexidades da linguaxe humana. A chegada do *deep learning*, unido á dispoñibilidade de grandes cantidades de datos textuais (algo para o que *Internet* foi un factor chave), o éxito das *Recurrent Neural Network* (RNN) [25] e o concepto de *Long Short-Term Memory* (LSTM) [26] supuxo un avance na procura de modelos lingüísticos máis sofisticados. Máis próximo no tempo, desenvolvementos como Word2Vec [27] e GloVe [28] centráronse en representar as palabras como vectores dentro de espazos vectoriais continuos, o que permite capturar relacións de similaridade semántica entre termos. Non obstante, este tipo de modelos carecen da profundidade e a complexidade necesarias para comprender o contexto en pasaxes máis longas. O gran *punto de inflexión* tivo lugar co desenvolvemento da arquitectura *Transformer* [29]. Este traballo de Vaswani et al. [29] introduce un mecanismo denominado *self-attention*, que se representa segundo a Ecuación 2.1 e permite capturar dependencias de longo alcance de maneira paralelizable, poñendo así solución a un problema clásico.

$$\text{attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.1)$$

Nesta ecuación Q , K e V son matrices que representan *queries*, claves e valores, respectivamente. Estas matrices derivanse da secuencia de entrada e empréganse para computar os pesos da atención e producir a secuencia de saída. d_k representa a dimensionalidade dos

vectores de claves, e permite escalar os produtos das queries e as claves. Nótese que se aplica unha función *softmax* fila por fila para normalizar os produtos a través de todas as claves. Isto resulta en pesos de atención que representan a importancia de cada clave con respecto á *query* inicial.

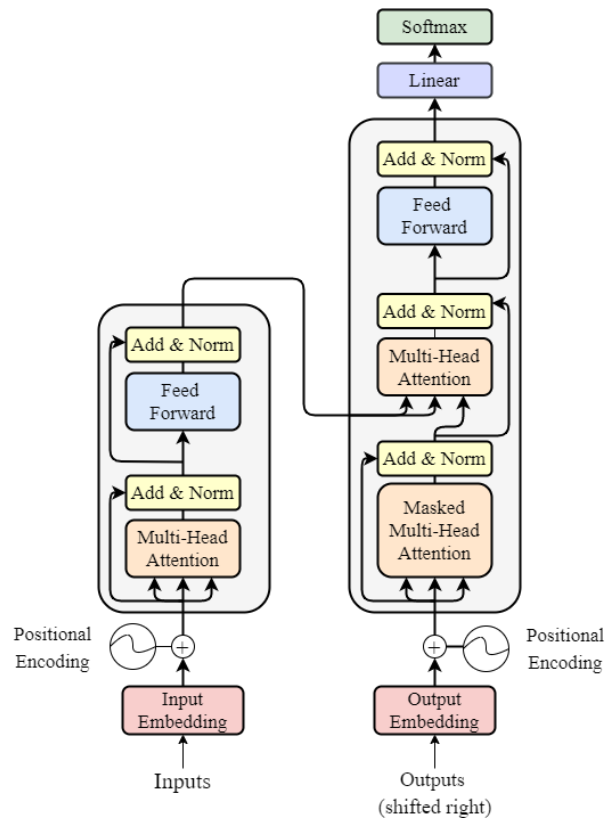


Figura 2.1: Vista xeral das compoñentes *encoder* e *decoder* da arquitectura Transformer.

É así que a arquitectura *Transformer* senta as bases de todos os modelos de linguaxe *State-of-the-Art* (SOTA) a día de hoxe. Deste xeito, e de maneira simplificada, os LLM operan recibindo unha entrada, codificándoa e, a continuación, descodificándoa para producir unha predición de saída segundo o ilustrado na Figura 2.1. Pero antes disto, o modelo require: *i*) *pre-training* para poder cumprir funcións xerais e *ii*) *fine-tuning* para acometer tarefas específicas. A fase de *pre-training* consiste no adestramento non supervisado do modelo utilizando grandes conxuntos de datos textuais de sitios como Wikipedia, GitHub ou outros para que o modelo *aprenda* o significado das palabras e as relacións entre as mesmas. O *fine-tuning*, mediante un axuste de pesos, optimiza o rendemento do modelo base para unha tarefa específica como pode ser, por exemplo, a tradución. Moi ligado ao *fine-tuning*, o *prompt-tuning* defínese como o axuste do modelo base para realizar tarefas específicas pero sen que teña lugar unha

modificación dos pesos do mesmo. Existen dúas aproximacións: o *few-shot prompting* e o *zero-shot prompting*. En ambos casos a *prompt* (instrución) que recibe como entrada o modelo indica a tarefa a realizar e o formato da mesma. A diferenza radica en que, cando se considera a aproximación *few-shot*, na *prompt* tamén se proporcionan unha serie de exemplos para *guiar* o modelo.

Os LLM, non obstante, teñen aínda por diante diversos retos. O máis coñecido son as *alucinacións*, un fenómeno que se manifesta cando este tipo de modelos producen saídas falsas ou que non reflicten a intención da persoa usuaria. Por outra banda, os riscos de seguridade tamén son significativos, xa que os LLM poden filtrar información privada, participar en estafas e contribuír á difusión de información errónea se non se xestionan adecuadamente, o que podería ter repercusións devastadoras a escala global. Ademais, se os datos utilizados para o adestramento só representa un grupo demográfico ou carecen de diversidade, os resultados producidos polo modelo de linguaxe grande tamén carecerán de diversidade, propagando así ao mundo computacional sesgos existentes na sociedade. O consentimento tamén expón problemas éticos, xa que os conxuntos de datos a miúdo extraen datos da *Internet* sen consentimento, o que pode dar lugar a problemas de infracción de dereitos de autor e violacións da privacidade. Como nota final, destacar que gran parte do potencial dos LLM radica na súa capacidade para ser aplicados a múltiples casos de usos e con diversos fins. Entre eles, destacan a recuperación de información, o análise de sentimentos, a xeración de texto e código ou, destacadamente, os *chatbot* e as IA conversacionais. É precisamente este caso de uso o que posibilita a realización deste TFM. Na seguinte Sección coméntase con detalle os fundamentos básicos a coñecer sobre a Intelixencia Artificial Conversacional.

2.3 Intelixencia Artificial conversacional

O concepto de IA Conversacional refírese a aqueles sistemas informáticos deseñados para interactuar e comunicarse con persoas a través da linguaxe natural humana [30]. Estes sistemas buscan comprender as consultas e preguntas das persoas, así como xerar respostas significativas para as mesmas. A aplicación máis destacada da IA Conversacional resultou na proliferación de *chatbots*, que se converteron en omnipresentes en diversos ámbitos como a atención ao cliente, o soporte técnico ou a educación [31]. Estes sistemas están deseñados non só para entender o significado literal das consultas das persoas usuarias, senón tamén para captar a intención e o contexto subxacentes, permitindo así interaccións máis similares ás humanas. Ademais, os sistemas de IA Conversacional tamén teñen a capacidade de *aprender* e adaptarse co tempo, perfeccionando as súas respostas en función dos comentarios das persoas usuarias e da evolución dos patróns lingüísticos. A este efecto, a Figura 2.2 amosa un esquema xeral das partes que compoñen este tipo de sistemas. Por unha parte, no fluxo de

entrada, o primeiro paso é a operación de *Automatic Speech Recognition* (ASR), que consiste en converter a linguaxe oral na súa correspondente representación escrita. Logo, ten lugar a fase de NLU, onde se realiza unha determinada operación sobre a representación textual previamente obtida. No caso da Figura 2.2, trátase dunha tradución. Como último paso da *pipeline*, efectúase a operación de *Text to Speech* (TTS), de modo que a representación escrita xerada como resposta se convirta novamente a linguaxe oral.

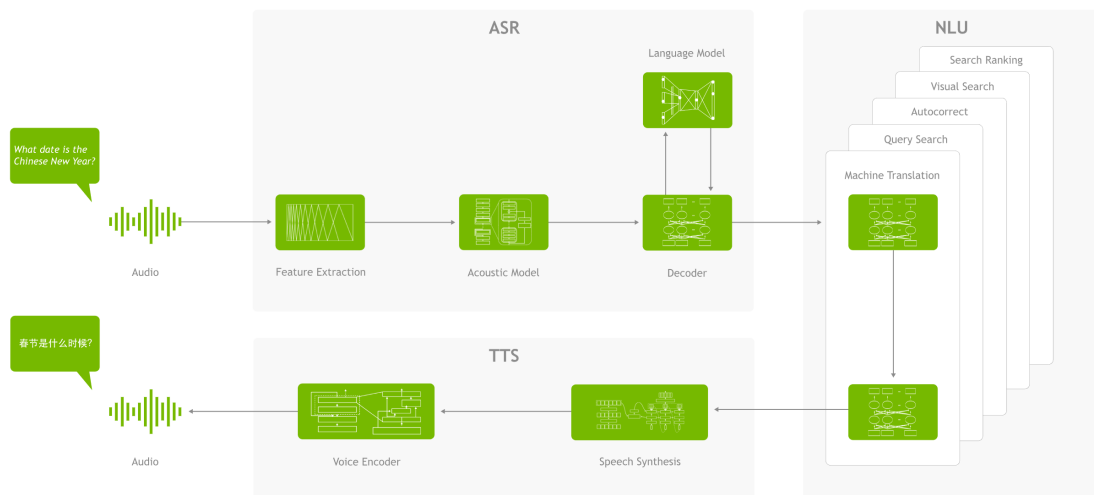


Figura 2.2: Vista xeral dun sistema completo de IA Conversacional.

Neste contexto, os LLM están a mellorar significativamente os sistemas de IA Conversacional debido á súa excelente comprensión da linguaxe natural, a súa capacidade para manexar grandes contextos e a súa habilidade para xerar respostas que se asemellen en forma e contido ás dunha persoa humana. Non obstante, esta capacidade non é inherente aos modelos base, senón que é resultado dun proceso máis complexo. Os LLM, en esencia, *simplemente* ven relacións entre as palabras. Por exemplo, se unha frase empeza con “Eu” é moi probable que a seguinte palabra sexa “son”. É dicir, o comportamento básico é similar a un autocompletado. Non obstante, os modelos que recentemente gañaron popularidade e recoñecemento mediático como ChatGPT [24] non autocompletan, senón que responden. Esta resposta é posible grazas a un proceso de *fine-tuning* con datos de conversas no que o modelo *aprende* a responder instrucións. Así pois, o factor disruptivo está ligado a que os LLM máis modernos admiten o diálogo multiturno, o que os dota de capacidade para manter un historial da conversa. Con isto, as respostas teñen certa personalización en función das preferencias e interaccións previas da persoa usuaria. Como nota aclaratoria final, no contexto deste proxecto, pensar en IA Conversacional é sinónimo de pensar nun LLM axustado para actuar segundo as características que se acaban de expoñer.

2.4 Intelixencia Artificial explicable

A IA Explicable refírese á capacidade dos sistemas de intelixencia artificial para proporcionar explicacións claras e comprensibles sobre as súas decisións e procesos internos. A medida que este tipo de modelos se volven máis complexos, como é o caso dos *Large Language Model (LLM)*, a necesidade de comprender como chegan ás súas predicións vólvese crucial, especialmente en dominios críticos como a atención médica, o sistema xudicial ou a toma de decisións automatizada. Existen dous enfoques comúns á hora de explicar as decisións dos modelos. Por unha banda, é posible utilizar modelos intrinsecamente explicables, como árbores de decisión ou modelos lineais [32]. Doutra banda, tamén se pode optar por métodos de explicabilidade agnósticos ao modelo, tipicamente *post hoc* [33, 34], que poden aplicarse a calquera modelo de aprendizaxe automática supervisada independentemente da súa arquitectura. Ademais, as explicacións poden ser locais ou globais. As explicacións locais, a destacar *Local Interpretable Model-Agnostic Explanations (LIME)* [35] e *Shapley Additive exPlanations (SHAP)* [36], céntranse en decisións individuais. Deste xeito, permiten entender por que o modelo produce unha decisión particular [35, 37]. Alternativamente, as explicacións globais consideran todo o comportamento do modelo e as súas predicións en conxunto [38, 39, 40]. Como se viu con anterioridade, a arquitectura *encoder-decoder* dos Transformer captura eficazmente a información contextual das secuencias de entrada e xera as correspondentes secuencias de saída. A este respecto, existen traballos que exploran o mecanismo de atención como ferramenta de explicabilidade [41, 42, 43]. A pesar da súa utilidade, o mecanismo de atención tamén ten limitacións, particularmente na interpretación dos pesos asignados ás características de entrada. Hai discrepancias na literatura sobre a utilidade dos pesos de atención para fins de interpretabilidade, con algunhas investigacións defendendo os seus beneficios [44, 45], mentres que outros sinalan as súas limitacións [46, 47].

Más recentemente, xurdiu unha estratexia alternativa para construír modelos interpretables: o uso de explicacións xenerativas en linguaxe natural. Esta técnica presenta varias vantaxes: *i)* son facilmente comprensibles para os usuarios finais, *ii)* os anotadores humanos poden traballar máis facilmente coa linguaxe natural, o que simplifica a recompilación de datos, e *iii)* pode ser factible extraer explicacións en linguaxe natural de grandes conxuntos de datos de expertos no dominio [48].

2.5 Retrieval Augmented Generation (RAG)

Como se viu nas anteriores Seccións, os LLM de uso xeral poden axustarse para realizar diferentes tarefas como a análise de sentimentos ou o recoñecemento de entidades. En xeral, estas tarefas non adoitan requirir coñecementos previos adicionais. Para tarefas máis comple-

xas e que requiren máis coñecementos específicos, sobretudo cando se consideran dominions concretos, é posible crear sistemas movidos por un LLM que acceda a fontes de coñecemento externas para efectuar as súas predicións. Isto permite unha maior coherencia factual, mellora a fiabilidade das respostas xeradas [49] e axuda a mitigar o problema das “alucinacións” [50]. Os investigadores de Meta AI³ introduciron un método chamado *Retrieval Augmented Generation* (RAG) [51] para abordar este tipo de tarefas intensivas en coñecemento. Fundamentalmente, RAG combina unha compoñente de recuperación de información cun modelo xerador de texto, de modo que o seu coñecemento interno pode modificarse de maneira eficiente e sen necesidade de volver adestrar o todo o modelo.

En esencia, e segundo o ilustrado na Figura 2.3, os sistemas de RAG toman unha entrada e recuperan un conxunto de documentos relevantes ou de apoio dada unha fonte (por exemplo, un índice de Elasticsearch). Os documentos concaténanse como contexto para a entrada orixinal e son considerados polo modelo para producir o resultado final. Deste xeito, 2.3 adáptase a situacións nas que os feitos poden evolucionar co tempo. Isto é moi útil, xa que o coñecemento paramétrico dos LLM é estático. 2.3 permite que os LLM eviten o reentrenamento, permitindo o acceso á información máis recente para xerar saídas fiables mediante a xeración baseada na recuperación.

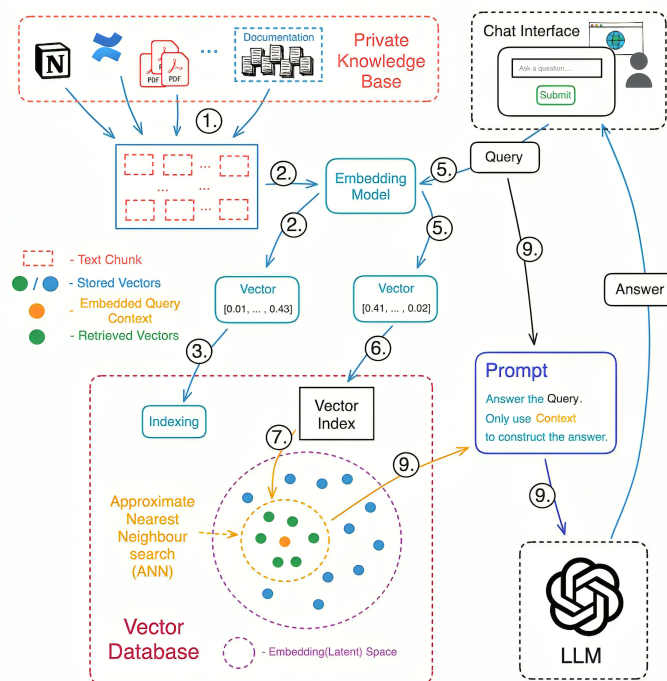


Figura 2.3: Vista xeral dun sistema de *Retrieval Augmented Generation* (RAG).

³ <https://ai.meta.com/>

Tecnoloxías e ferramentas

CANDO a obtención de resultados de alta calidade e a realización exitosa das tarefas son prioridades, é crucial investir tempo na selección e análise das tecnoloxías e ferramentas dispoñibles para que o encaixe no proxecto sexa o mellor posible. Na Sección 3.1 proporciónase unha descrición das linguaxes de programación empregadas. Posteriormente, na Sección 3.2, abórdanse as librerías e compoñentes escollidas. Ademais, ofrécese información detallada sobre as ferramentas de desenvolvemento elixidas na Sección 3.3. Por último, a Sección 3.4 adícase a explicar as ferramentas relacionadas co enfoque metodolóxico do proxecto.

3.1 Linguaxes de programación

Nesta Sección describíense as linguaxes de programación empregadas neste proxecto.

3.1.1 Python

A linguaxe de programación Python¹ destácase como unha linguaxe interpretada de alto nivel que integra módulos, excepcións, tipado dinámico e clases. A súa versatilidade maniféstase na capacidade para soportar diversos paradigmas de programación, incluíndo a orientación a obxectos, programación funcional e imperativa, entre outros. A súa fortaleza radica na potencia que ofrece, combinada cunha sintaxe clara e sinxela. Python tamén proporciona interfaces para realizar chamadas ao sistema e conta cunha ampla variedade de bibliotecas. É importante destacar a súa portabilidade, xa que pode executarse tanto en sistemas operativos Windows coma en variantes de Unix incluíndo Linux e macOS.

Un aspecto a considerar é que ao ser unha linguaxe interpretada, o rendemento do código pode ser máis lento en comparación con linguaxes compiladas coma C. Con todo, tal e como se ilustra na Figura 3.1, Python é na actualidade a linguaxe de programación máis popular segundo o índice TIOBE [52]. Isto implica que a maioría de librerías e desenvolvementos

¹<https://docs.python.org/3/faq/general.html>

son implementados ou compatibles con esta linguaxe, e máis especialmente no ámbito da intelixencia artificial e a aprendizaxe máquina [53].

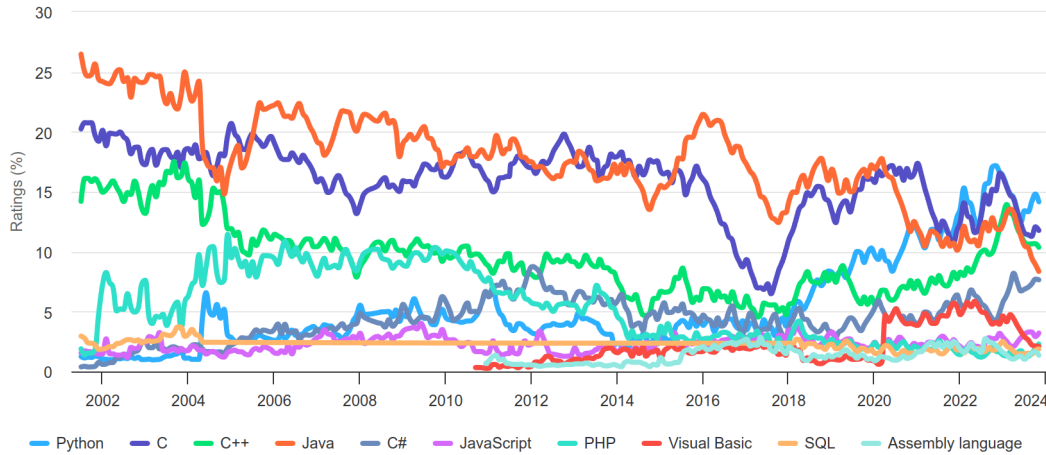


Figura 3.1: Índice TIOBE de popularidade con data de Novembro do 2023.

3.1.2 JavaScript

JavaScript é unha linguaxe de programación de alto nivel, interpretada e orientada a obxectos. Xunto coa parte estrutural definida con *HyperText Markup Language* (HTML) e os estilos configurados con *Cascading Style Sheets* (CSS) permite a creación de páxinas *web* interactivas e dinámicas. En concreto, cando se fala de JavaScript como compoñente do *frontend* dunha aplicación, esta tecnoloxía xestiona a interacción das persoas usuarias coa interface e a actualización dinámica da páxina sen necesidade de recargar o sitio. Ligado a isto, é habitual utilizar JavaScript en conxunto con *Asynchronous JavaScript and XML* (AJAX), posibilitando o intercambio asíncrono de datos.

3.2 Librerías e compoñentes

Esta Sección aborda as librerías e compoñentes escollidas para o proxecto.

3.2.1 Transformers

A librería *Transformers*² convertiuse nunha ferramenta esencial no ámbito da intelixencia artificial e a aprendizaxe máquina. Desenvolvida por Hugging Face, esta librería proporciona tanto implementacións predestradas de modelos de *Natural Language Processing* (NLP), *Computer Vision* (CV), *Automatic Speech Recognition* (ASR), etc., coma tamén ferramentas para adestrar novos modelos, avaliar o seu rendemento e utilidades para integrar de xeito doado

² <https://huggingface.co/docs/transformers/index>

os modelos existentes nas aplicacións. A librería está dispoñible baixo a licenza Apache³, o que a fai accesible e utilizada amplamente entre a comunidade.

3.2.2 DeepSpeed

DeepSpeed⁴ é unha biblioteca *open source* desenvolvida por Microsoft baixo licenza Apache⁵ para mellorar a eficiencia e a velocidade de adestramento de modelos de aprendizaxe máquina. A característica principal da biblioteca é o uso de técnicas como a compresión de modelos, o uso eficiente da memoria e a paralelización do adestramento para aproveitar ao máximo os recursos dispoñibles.

3.2.3 FastChat

FastChat⁶ é unha plataforma *open source* dispoñible baixo licenza Apache⁷ para adestrar, servir e avaliar *chatbots* baseados en LLMs. Esta librería baséase na noutras xa introducidas en esta Sección, nomeadamente Transformers (ver 3.2.1) e DeepSpeed (ver 3.2.2). Tamén se integra con solucións de almacenamento de datos, incluíndo Elasticsearch (ver Sección3.2.5), para facilitar a recuperación de información e contexto durante a conversación.

3.2.4 FastAPI

FastAPI⁸ é un *framework* dispoñible baixo licenza MIT⁹ orientado ao desenvolvemento de *Application Programming Interface (API)s REST* de xeito áxil e, ademais, proporcionando unha documentación automática grazas á integración con Swagger UI¹⁰. Por outra banda, FastAPI tamén permite a validación de datos de entrada e saída mediante o uso de modelos Pydantic¹¹, o que fai que as APIs sexan máis seguras e fáciles de manter.

3.2.5 Elasticsearch

Elasticsearch¹² é un motor deseñado para o procesamento, a procura, e a análise de diferentes tipos de datos a gran escala. Un punto a destacar é que os datos poden ser tanto estruturados coma non estruturados, podendo ir desde ficheiros JavaScript Object Notation

³ <https://github.com/huggingface/transformers/blob/main/LICENSE>

⁴ <https://www.deepspeed.ai/getting-started/>

⁵ <https://github.com/microsoft/DeepSpeed/blob/master/LICENSE>

⁶ <https://github.com/lm-sys/FastChat>

⁷ <https://github.com/lm-sys/FastChat/blob/main/LICENSE>

⁸ <https://fastapi.tiangolo.com/>

⁹ <https://github.com/tyangolo/fastapi/blob/master/LICENSE>

¹⁰ <https://swagger.io/tools/swagger-ui/>

¹¹ <https://docs.pydantic.dev/latest/>

¹² <https://www.elastic.co/elasticsearch>

(JSON) ata datos xeoespaciais. O motor está construído sobre Apache Lucene¹³ e utiliza unha arquitectura distribuída, o que lle permite escalar horizontalmente a través de múltiples nodos. Isto garante unha alta dispoñibilidade, tolerancia a erros e capacidade para manexar grandes volumes de datos. Ademais, Elasticsearch admite procuras en tempo real, o que permite aos usuarios consultar e analizar os datos a medida que se reciben. Elasticsearch utiliza a *Server Side Public License* (SSPL)¹⁴, o que xerou debates na comunidade *open source* debido á súa incompatibilidade percibida coas licenzas *GPL*.

3.2.6 LangChain

LangChain¹⁵ é un proxecto *open source* distribuído baixo licenza MIT¹⁶ que foi deseñado para simplificar a creación de aplicacións que integran algún *Large Language Model* (LLM). Pode empregarse en forma de biblioteca Python e proporciona unha serie de abstraccións que permiten que poida ser usada para a implementación de *chatbots* e/ou asistentes virtuais. Exemplos destas abstraccións son a xestión do estado do diálogo entre as diferentes interaccións ou a memoria contextual, o que permite xerar texto máis preciso e coherente. Outro aspecto fundamental é que LangChain facilita o acceso a datos externos, de xeito que os LLMs poden comunicarse e ler información de bases de datos ou índices de Elasticsearch (ver 3.2.5).

3.2.7 Django

Django¹⁷ é un *framework* escrito en Python que busca facilitar un desenvolvemento limpo, veloz e pragmático de aplicacións *web*. Desta maneira, Django permite abstraer aspectos máis complexos de baixo nivel de maneira que o desenvolvedor se poida centrar na lóxica da aplicación. Este *framework* está baseado na reutilización de compoñentes, e incorpora unha gran cantidade de padróns con funcionalidades básicas como a autenticación, a conexión con base de datos ou as operacións de creación, lectura, actualización e borrado de elementos. No marco concreto deste proxecto, Django utilízase para o desenvolvemento da interface a través da cal se interactúa co *chatbot*. Relacionado con isto, cabe destacar a existencia das *views*, que permiten implementar de xeito moi doado os *endpoints* que recuperan datos do servizo *Representational State Transfer* (REST) que serve a lóxica dos *Large Language Model* (LLM).

¹³ <https://lucene.apache.org/>

¹⁴ <https://github.com/elastic/elasticsearch/blob/main/LICENSE.txt>

¹⁵ https://python.langchain.com/docs/get_started/introduction

¹⁶ <https://github.com/langchain-ai/langchain/blob/master/LICENSE>

¹⁷ <https://www.djangoproject.com/start/overview/>

3.3 Ferramentas de desenvolvemento

Esta Sección detalla as ferramentas de desenvolvemento elixidas para o proxecto.

3.3.1 VSCodium

VSCodium¹⁸ é unha distribución totalmente *open source* dispoñible baixo licenza MIT¹⁹ do editor Visual Studio Code (VS Code)²⁰. Aínda que VS Code é gratuito e o seu código fonte *open source*, a versión distribuída por Microsoft está suxeita a unha licenza privativa que inclúe funcións de telemetría. VSCodium, pola súa banda, toma os binarios fonte de VS Code e ofrece unha versión que elimina a devandita telemetría. Cómpre destacar que tanto VSCodium como VS Code comparten o mesmo núcleo, polo que todas as extensións do *marketplace* de VS Code son plenamente funcionais.

Así pois, VSCodium é un editor de ficheiros lixeiro e dispoñible para Windows, Linux e macOS. Entre as súas características inclúese a posibilidade de depurar o código, o autocompletado, as macros, facilidades para refactorización, etc. Ademais é moi personalizable tanto en aparencia coma en comportamento. Por todo isto, entre outros motivos, actualmente sitúase como a ferramenta de desenvolvemento máis popular, cunha cota de uso do 70% [54]. Polos motivos anteriormente expostos, ademais de ser unha ferramenta xa coñecida polo autor, é o editor empregado para todo o desenvolvemento deste proxecto.

3.4 Ferramentas de metodoloxía

Esta Sección explica as ferramentas relacionadas co enfoque metodolóxico do proxecto.

3.4.1 Xestión de paquetes

Pip

Pip²¹ é un sistema de xestión de paquetes empregado para instalar e administrar paquetes de Python, moitos deles dispoñibles no *Python Package Index*. Unha vantaxe de Pip é que permite xestionar ditos paquetes de maneira moi sinxela mediante a súa interfaz de liña de comandos. Outra característica de interese é que permite controlar listas de paquetes e as súas versións mediante un ficheiro de requisitos, permitindo unha recreación eficaz dun conxunto de paquetes nun novo entorno. Por último, destacar que está distribuído baixo licenza MIT²² de *software* libre.

¹⁸ <https://vscodium.com/>

¹⁹ <https://github.com/VSCodium/vscodium/blob/master/LICENSE>

²⁰ <https://code.visualstudio.com/>

²¹ <https://pip.pypa.io/en/stable/>

²² <https://github.com/pypa/pip/blob/main/LICENSE.txt>

3.4.2 Docker

Docker²³ é unha plataforma que facilita a creación, implementación e execución de aplicacións en contedores. Estes contedores permiten empacar todas as dependencias dunha aplicación xunto co seu código nunha única unidade, o que asegura que a aplicación se execute de maneira consistente en calquera contorno, asegurando a reproducibilidade en diferentes máquinas. O motor de Docker atópase dispoñible baixo licenza Apache²⁴ de *software* libre e código aberto.

3.4.3 Xestión do proxecto

Taiga

Taiga²⁵ é unha ferramenta para a xestión de proxectos encadrados dentro das metodoloxías denominadas áxiles: Scrum e Kanban. A súa principal funcionalidade é a xestión de tarefas, xa que permite crear as historias de usuario, a estimación de puntos de historia e a asignación das mesmas aos diferentes *sprints*. Todo isto facilita o seguemento do proxecto e dá unha visión global tanto dos puntos completados coma dos restantes ata a finalización en cada un dos *sprints*. A maiores, ofrece integración con GitLab (ver 3.4.4), o que unido a que está deseñada para ser empregada coa metodoloxía seguida neste proxecto (ver 4.1) fixo que fora a ferramenta seleccionada para a xestión do proxecto. Destacar, finalmente, que está distribuída baixo licenza Mozilla²⁶ de *software* libre e código aberto.

3.4.4 Control de versións

Git e GitLab

Git²⁷ é un *Source Code Management* (SCM) distribuído deseñado por Linus Torvalds para o manter o desenvolvemento do *kernel* de Linux. Está deseñado para traballar tanto con pequenas aplicacións como con grandes proxectos, sendo a facilidade para o uso e o rendemento dous dos seus piares fundamentais.

O ser estritos no control de versións permite ter unha correcta trazabilidade nos cambios e evolución do proxecto, de modo que é requisito indispensable para un desenvolvemento de calidade. Fronte a outros sistemas de control de versións, coma SVN²⁸, Git ofrece certas vantaxes que o levaron a ser o *SCM* de referencia:

²³ <https://www.docker.com/>

²⁴ <https://github.com/moby/moby/blob/master/LICENSE>

²⁵ <https://www.taiga.io/>

²⁶ <https://community.taiga.io/t/whats-taigas-open-source-license/138>

²⁷ <https://git-scm.com/>

²⁸ <https://subversion.apache.org/>

- Posibilidade de traballar sen conexión a internet: Git funciona con repositorios locais e remotos, o que permite efectuar o control de versións desconectados da rede e conservar en local toda a funcionalidade.
- Desenvolvemento non lineal: Git potencia o uso do **branching**, facilitando o traballo dos equipos de desenvolvemento e maximizando a produtividade.

Para xestionar os repositorios Git, existen dúas plataformas amplamente asentadas: GitLab²⁹ e GitHub³⁰. Para este proxecto, faise uso dunha conta en GitLab facilitada polo IRLab³¹, laboratorio no cal se enmarca o desenvolvemento deste proxecto.

²⁹ <https://about.gitlab.com/>

³⁰ <https://github.com/>

³¹ <https://www.dc.fi.udc.es/irlab/>

Metodoloxía e xestión do proxecto

ESTE Capítulo aborda aspectos vinculados á metodoloxía elixida para a xestión do proxecto. En particular, empregase *Scrum* con diversas adaptacións debido ao feito de que o tipo de proxecto non é de enxeñaría clásica e ás limitacións no equipo de desenvolvemento, xa que se trata dun *TFM*. Detállanse cuestións como a metodoloxía en si, as estimacións de custos e duración, a planificación e a xestión de riscos.

4.1 Metodoloxía

A metodoloxía de desenvolvemento é crucial, pois establece unha estrutura que dirixe todo o proceso do proxecto. Por este motivo, é importante facer unha escolla fundamentada e apropiada. Ao longo do tempo, xurdiron diversas metodoloxías para a xestión de proxectos, formando parte do campo de estudo e desenvolvemento dentro da disciplina da enxeñaría do *software*.

4.1.1 Escolla da metodoloxía

Nun proxecto de desenvolvemento en investigación como é este, a escolla da metodoloxía resulta de igual importancia. Habitualmente, no referido á metodoloxía de investigación, as eleccións adoitan estar baseadas no método científico. Por outra banda, para a xestión do proxecto científico, é preciso escoller unha metodoloxía. Dada a natureza do proxecto, esta escolla non resulta tan clara coma nun proxecto de enxeñaría clásico. A continuación preséntanse unha serie de características desexables que se deben ter en conta á hora de seleccionar a metodoloxía.

Ciclos de vida curtos

Búscase iterar de xeito constante, para adaptarse aos resultados que se van obtendo nos sucesivos ciclos e definir novos experimentos a partir destes.

Adaptación aos cambios

Se en calquera proxecto é habitual que aparezan cambios, nun proxecto deste tipo máis aínda. As necesidades varían en función dos resultados obtidos.

Xestión de riscos

É necesario ser quen de detectar o antes posible as ameazas que poidan xurdir durante a investigación para poder corrixir o curso do desenvolvemento, de tal modo que se acaden os obxectivos establecidos.

Visión do avance da investigación

Resulta desexable poder estudar os avances conseguidos durante a investigación. Isto faise coa intención de incorporar a información recadada e obter conclusións intermedias que permitan dirixir obxectivos concretos.

Todos estes requisitos apuntan á necesidade dunha metodoloxía áxil¹, amplamente estendidas a día de hoxe na xestión de proxectos. Están baseadas no uso de ciclos de vida iterativos, enfatizando na idea de obter un produto intermedio á finalización de cada un destes ciclos. Isto permite adaptarse mellor aos cambios nos requisitos e detectar o antes posibles os riscos. A maiores, facilita o estudo do avance do proxecto e supón unha motivación para o equipo. En concreto, escolleuse *Scrum*. As súas particularidades explícanse a continuación.

4.1.2 Scrum

Scrum [55] é unha das denominadas metodoloxías áxiles. Preséntase coma un marco deliberadamente incompleto, definindo unicamente as partes necesarias para implementar a súa teoría. Deste xeito, en lugar de proporcionar instrucións detalladas, serán as regras de Scrum as que guíen as relacións e interaccións.

Scrum (ver Figura 4.1) utiliza un enfoque iterativo e incremental. Isto permite, por unha parte, mellorar a comunicación co cliente, xa que ao remate de cada incremento debería ser posible ter un entregable *software*. Ademais, tamén favorece a inspección continua, o que axuda a mellorar o control de riscos.

Para comprender mellor Scrum, cómpre definir o seu equipo (ver 4.1.2), os seus eventos (ver 4.1.2) e os artefactos dos que se dispón (ver 4.1.2). Todo isto fundaméntase nos piares empíricos da propia metodoloxía:

Transparencia

Favorece a inspección, tendo en conta que unha inspección sen transparencia pode levar a engano.

¹<https://agilemanifesto.org/iso/gl/manifesto.html>

Inspección

Permite a adaptación, partindo de que a inspección sen adaptación resulta inútil. Debe ser continua e frecuente, para detectar desviacións e problemas co maior tempo posible. A inspección está, por tanto, moi relacionada cos eventos Scrum (4.1.2), deseñados para provocar cambios.

Adaptación

Indispensable cando se detecta unha desviación. No menor tempo posible, debe realizarse un axuste que permita minimizar a desviación adicional.

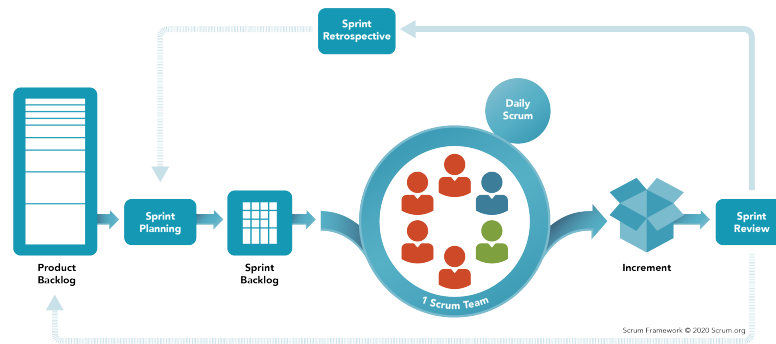


Figura 4.1: Detalle do proceso de *Scrum*. Diagrama recollido da [guía oficial](#).

Equipo

Un equipo de Scrum debe ser multifuncional, de modo que os seus membros teñan as habilidades necesarias para crear valor en cada *Sprint* (4.1.2). É, por tanto, unha unidade cohesionada que elimina os sub-equipos e as xerarquías. Os equipos Scrum, idealmente, deben manterse entornando ás 10 persoas para buscar o equilibrio entre ser áxil e ter a capacidade de completar un traballo significativo dentro dun *Sprint*. Ademais, estarán presentes os seguintes tres roles, coas súas responsabilidades específicas.

Desenvolvedores

Deben estar comprometidos a crear calquera aspecto dun incremento funcional dentro do *Sprint*. Serán sempre responsables de crear un plan para o *Sprint*, inculcar calidade e adaptar o seu traballo diario de cara á consecución do obxectivo do *Sprint*.

Product Owner

Actúa como persoa propietaria do produto, debendo maximizar o seu valor. Tamén é

responsable da xestión do *Product Backlog* polo que, entre outras tarefas, deberá desenvolver e comunicar explicitamente o *Product Goal*, crear e comunicar elementos de traballo pendentes do produto, etc.

Scrum Master

É o responsable de establecer Scrum tal e como se define na súa guía [55]. Buscará, por tanto, axudar a todas as persoas que dalgún modo participan no proxecto a comprender a teoría e a práctica da metodoloxía.

Eventos

O *Sprint* actúa como evento contedor de todos os demais eventos, que deberían ser vistos como oportunidades para a inspección e adaptación dos artefactos (ver 4.1.2). Os *Sprints*, idealmente, deben durar un mes ou menos e comezar un a continuación do anterior. Deste xeito, poderase garantir a inspección e a adaptación cunha periodicidade suficiente para que o *Sprint Goal* non se volva obsoleto, controlando tamén o risco e a complexidade. Todos os eventos que contén o *Sprint* son os seguintes:

Sprint Planning

Reunión que dá comezo ao *Sprint*, establecendo o traballo que se realizará dentro do mesmo e os obxectivos que se pretenden acadar. Debe responder certas cuestións, como son identificar porque o *Sprint* é valioso, que se pode facer nel e como se realizará o traballo escollido. Como máximo, este evento debería durar oito horas para *Sprints* dun mes.

Daily Scrum

Reunión diaria de quince minutos de duración para as persoas desenvolvedoras do equipo *Scrum* (ver 4.1.2). O seu obxectivo debe ser revisar o progreso de cara á consecución do *Sprint Goal*, inspeccionando o traballo realizado. Para isto, resulta útil poñer en común o traballo realizado no día anterior, definir o traballo que se realizará no propio día e comentar posibles problemas ou obstáculos que apareceran no proceso.

Sprint Review

Reunión coas partes interesadas ao remate do *Sprint*, de modo que sexa posible revisar e inspeccionar o traballo realizado durante o incremento e determinar futuras adaptacións. Permite ver que se logrou durante o *Sprint* e decidir que facer a continuación. Como máximo, este evento debería durar catro horas para *Sprints* dun mes.

Sprint Retrospective

Reunión posterior e complementaria á *Sprint Review* co propósito de planificar formas

de aumentar a calidade e a eficacia. O equipo Scrum debe analizar que foi ben durante o *Sprint*, que problemas se atoparon e como estes foron ou non foron resoltos. Como máximo, este evento debería durar tres horas para *Sprints* dun mes.

Artefactos

Ao longo do desenvolvemento do produto elabóranse unha serie de elementos denominados *artefactos*. Os artefactos de *Scrum* están deseñados para maximizar a transparencia da información clave. Son os seguintes:

Product Backlog

Traballo pendente do produto. Organízase como unha lista emerxente e ordenada do necesario para mellorar o produto, sendo os seus elementos implementados durante os *Sprints*. Exprésase en forma de Historias de Persoa Usuaría e é a única fonte de requisitos do produto. Dado que é dinámica e evoluciona conforme o fai o produto, nunca está completa.

En calquera TFM, e en especial se se trata dun proxecto de desenvolvemento en investigación, o *Product Backlog* debe estar practicamente completo ao comezo. Isto é debido a que se deben ter claros os obxectivos da investigación, as ferramentas das que se dispón e o plan de traballo, que foi especificado no anteprojecto. Non obstante, os requisitos serán refinados en función dos resultados que se vaian obtendo, quedando este aspecto cuberto polas facilidades para a adaptación que aporta *Scrum*.

Sprint Backlog

É o subconxunto de tarefas do *Product Backlog* que o equipo selecciona e incorpora ao *Sprint*. Elabórase durante o *Sprint Planning*. É posible engadir novos elementos durante a realización do *Sprint* en caso de consideralo necesario para cumprir co obxectivo establecido.

Incremento

Correspóndese co resultado do *Sprint*, sendo a suma de todas as tarefas, casos de uso, historias e elementos completados durante o *Sprint*. É posto en disposición da persoa usuaria en forma de software, e resulta un aspecto clave para o carácter iterativo e incremental de *Scrum*.

4.1.3 Adaptación de Scrum a este traballo

Como se comentou con anterioridade, existen unha serie de aspectos que imposibilitan unha fiel adopción da metodoloxía. Por exemplo, o equipo de desenvolvemento está formado

por unha única persoa, a dedicación non se mantén constante, etc. A continuación, preséntanse as adaptacións realizadas sobre *Scrum* para poder utilizar esta metodoloxía no proxecto:

- O rol de *Product Owner* está desempeñado polos directores do proxecto Miguel Anxo Pérez Vila e Javier Parapar López.
- O rol de *Scrum Master* está desempeñado polo alumno Eliseo Bao Souto.
- O *Equipo de Desenvolvemento* está composto por unicamente unha persoa, o alumno Eliseo Bao Souto.
- Dado que o *Equipo de Desenvolvemento* é unipersoal, prescínlese das reunións diarias. Non obstante, o alumno debe avaliar as cuestións que se trataría nestas reunións.
- A duración dos *Sprints* é de 2/3 semanas, cun esforzo de 50 puntos de historia.

4.2 Xestión do proxecto

A xestión do proxecto busca satisfacer unha serie de obxectivos que permitan maximizar o éxito do mesmo. É desexable cumprir coa planificación e os custes estimados, así como tamén acadar os termos de calidade desexados. Por outra parte, preténdese optimizar o uso de recursos e tamén ter constancia en todo momento do estado do proxecto.

4.2.1 Estimación

Para realizar unha estimación entendible e non demasiado complexa das historias de persoa usuaria, establécese que 1 hora de traballo equivale a 1 punto de historia. Isto permite estimar o custe de realizar unha historia de persoa usuaria en horas por persoa. A maiores, fíxanse 50 puntos de historia por *Sprint*, sendo a duración destes 2/3 semanas. Por tanto, en condicións ideais, o tempo aproximado de traballo é de 4 horas ao día.

4.2.2 Recursos

Os recursos empregados para un proxecto destas características poden dividirse nas tres categorías que a continuación se detallan.

Recursos humanos

Tal e como se introduciu na Sección 4.1.3, un total de tres persoas participan activamente neste proxecto no referente á metodoloxía e xestión do mesmo. Os directores Miguel Anxo Pérez Vila e Javier Parapar López exercen a función de *Product Owners*, tomando parte activa na planificación e definición de requisitos. O resto de roles, tanto o de *Scrum Master* coma o

de *Equipo de Desenvolvemento*, están representados polo alumno Eliseo Bao Souto. Isto implica que o alumno terá que, por un lado, asegurar o cumprimento do marco metodolóxico, así como tamén exercer de analista, deseñador e desenvolvedor. Por outra parte, unha persoa experta con formación en psicoloxía participa tamén no proxecto fornecendo validación dos resultados finais e, por tanto, axudando a avaliar a utilidade e usabilidade do produto desenvolvido nun marco clínico real. Así pois, son un total de 4 persoas as que conforman os recursos humanos necesarios para o desenvolvemento do proxecto.

Recursos materiais

Os recursos materiais son dous equipos facilitados polo grupo de investigación no cal se desenvolve este proxecto. Por unha parte, para o desenvolvemento do *software*, tarefas relacionadas cos aspectos metodolóxicos, etc. é abondo un equipo básico sen especiais requisitos mínimos. Por outra banda, para o *fine-tuning*, o servido do modelo e o aloxamento da parte *web*, dispónse dunha máquina que ofrece como principais características 128 GB de *Random Access Memory (RAM)* e unha *Graphics Processing Unit (GPU)* NVIDIA RTX 6000 Ada Generation 48 GB. Respecto disto, a Sección 9.3 ofrece unha estimación das emisións de CO₂ relacionadas co proxecto.

Recursos *software*

Son todos os mencionados no Capítulo 3. Todos eles son de uso gratuito.

4.2.3 Custes

Para definir os custes dos recursos humanos faise uso do informe *Salary Guide 2023 Spain - Technology* elaborado pola compañía Michael Page [56]. Na Táboa 4.1 recóllese a estimación salarial para os diferentes perfís identificados. A maiores, deben terse en conta os seguintes aspectos:

- En total, realízanse 6 *Sprints*.
- Tal e como se define na Subsección 4.1.3, os *Sprints* teñen unha duración de 2/3 semanas.
- Os *Product Owner* (directores) adican cada un 5 horas de traballo por *Sprint*.
- O equipo de desenvolvemento (alumno) adica 20 horas semanais de traballo: 2 como analista e 18 como desenvolvedor.
- Jesús Adrián Pérez Reales, a persoa con coñecemento experto, é accesible a través do laboratorio de investigación no que se desenvolve este proxecto e non imputa custe, xa que se contan con recursos deste tipo no marco de proxectos coma o PLEC2021-007662

(*Big-eRisk: Early Prediction of Personal Risks on Massive Data*). É por iso que na Táboa 4.1 se indica que o seu custe *non aplica*.

Así, o detalle completo dos custes dos recursos humanos queda especificado na Táboa 4.2.

Perfil	Custe
<i>Director</i>	50€/hora
<i>Analista</i>	30€/hora
<i>Desenvolvedor</i>	25€/hora
<i>Psicólogo</i>	<i>Non aplica</i>

Táboa 4.1: Estimación salarial para os recursos humanos do proxecto.

Perfil	Tempo (h/Sprint)	Dedicación (nº Sprints)	Custe (€/h)	Total (€)
<i>Director</i>	5 x 2	6	50	3 000
<i>Analista</i>	5	6	30	900
<i>Desenvolvedor</i>	45	6	25	6 750
<i>Psicólogo</i>				0
Total				10 650

Táboa 4.2: Detalle dos custes totais dos recursos humanos.

A Táboa 4.3 recolle os custes asociados aos recursos materiais e de *software*.

Recurso	Uds.	Custe (€/ud)	Vida (meses)	Uso (meses)	Total (€)
<i>Máquina básica</i>	1	1000	60	5	83.34
<i>Máquina GPU</i>	1	10000	60	5	833.34
<i>software</i>					0
Total					916.68

Táboa 4.3: Detalle dos custes totais dos recursos materiais e de *software*.

O custe combinado dos recursos humanos, materiais e *software* suma 11 566.68€.

4.2.4 Xestión de riscos

En todo proxecto, de cara a evitar as ameazas e mitigar as súas consecuencias, resulta necesario realizar unha adecuada identificación e xestión dos riscos aos que se está exposto. Por este motivo, en primeiro lugar, debe levarse a cabo unha fase de identificación e clasificación dos mesmos, obtendo a exposición do proxecto a cada risco en función da probabilidade de aparición e o impacto en caso de ocorrer o risco. Na Táboa 4.4 pode verse o resultado deste proceso de identificación e clasificación.

Código	Descrición	Probabilidade	Impacto	Exposición
R1	Planificación inadecuada	Media	Alto	Alta
R2	Mal deseño	Media	Medio	Media
R3	Mala implementación	Baixa	Moi alto	Alta
R4	Baixa eficiencia	Media	Alto	Media
R5	Falta de recursos	Baixa	Alto	Baixa
R6	Perda de información	Baixa	Moi alto	Alta

Táboa 4.4: Detalle da identificación e clasificación dos riscos presentes no proxecto.

Prevenición

Aqueles riscos aos que se considera que se está altamente exposto deben ser tratados dun xeito especial, xa que de ocorrer poden supoñer grandes cambios na raíz do proxecto ou incluso o seu fracaso. É por iso que son analizados nunha fase temperá do proxecto e se definen os correspondentes plans de continxencia. A continuación detállanse os plans de continxencia para os riscos de alta exposición:

- **R1 - Planificación inadecuada.** Dado o corte innovador do proxecto, pode ser habitual e incluso tolerable que existan certas desviacións na planificación do proxecto dadas as características do mesmo. En parte, este risco queda minimizado de xeito natural pola propia metodoloxía pero, aínda así, faise un exhaustivo control para evitar retrasos.
- **R3 - Mala implementación.** Unha mala implementación daría lugar á obtención de resultados erróneos no funcionamento do sistema, mermando a confianza dos expertos no mesmo. Aínda que isto é pouco probable que ocorra, xa que a complexidade do *software* a implementar non é elevada, este suposto sería catastrófico. Para evitalo,

establécense controis rutinarios con perfís expertos, encarnados polos directores do proxecto, que se encargarán de validar a implementación.

- **R6 - Perda de información.** De novo, trátase dun risco pouco probable pero de gran impacto en caso de suceder. A perda de información pode darse por múltiples causas: erro humano, fallo hardware, roubo... En calquera caso, perder o propio código fonte, os resultados dos experimentos ou a documentación suporía un grave retraso no proxecto. Por iso, establécese en todo momento unha política de copias de seguridade e control de versións na nube.

Seguemento

Para os riscos de exposición media realízase un seguemento constante durante o proxecto para controlar que non se convirtan en riscos de alta exposición. En concreto:

- **R2 - Mal deseño.** Os ciclos de vida curtos e o carácter iterativo da metodoloxía permiten detectar erros no deseño e corríxilos antes de que se propaguen.
- **R4 - Baixa eficiencia.** De comezo, é posible que a proposta inicial non sexa tan eficiente como podería chegar a ser. Non obstante, o carácter iterativo da metodoloxía permite refinar a solución gradualmente.

Asimilación

Os riscos con pouca exposición son simplemente aceptados, xa que é pouco probable que ocorran e ademais o seu impacto en caso de que se dea este suposto é pouco elevado. Así, realizar plans de continxencia ou manter un seguemento exhaustivo sería máis custoso que realizar a propia xestión de problemas en caso de que se chegaran a dar. Son riscos de baixa exposición os seguintes:

- **R5 - Falta de recursos.** Os recursos son xa coñecidos de antemán, polo que a planificación se fai acorde a eles e sempre con certas folguras para paliar posibles situacións de baixa dispoñibilidade dos recursos.

Proposta

O obxectivo deste Capítulo é detallar todo o relacionado coa proposta do proxecto. Así pois, na Sección 5.1 detállase a nivel visual e explícase a arquitectura definida para o sistema. A continuación, na Sección 5.2 preséntanse as fontes de datos identificadas, as súas características e como foron aproveitadas para construír o conxunto de datos necesario para levar a cabo este traballo.

5.1 Arquitectura

O *chatbot* que se desenvolve no marco deste TFM intégrase dentro dunha solución *software* xa existente¹. Na Figura 5.1 ilústrase a arquitectura do sistema final, debendo ter en conta as seguintes consideracións: a agrupación lóxica dentro do cadrado verde representa a parte da plataforma implementada desde cero. Como é de agardar, outras partes da solución orixinal foron modificadas para dar soporte ás novas funcionalidades, estando estas coloreadas en azul. A continuación analízanse os aspectos máis destacados:

5.1.1 Aplicación base

A aplicación base segue unha arquitectura cliente-servidor, onde por unha parte a persoa usuaria interacciona cos distintos elementos por medio do cliente e por outra banda o servidor obtén os datos, prepáraos, realiza a lóxica de negocio necesaria e envíaos de volta ao cliente. Así pois, o cliente *web* impleméntase a través de *templates* de Django, que conteñen partes estáticas de HTML e tamén unha sintaxe especial que describe como insertar contido dinámico. En canto ao servidor, as compoñentes máis destacadas son as *views*, que actúan como controladores comunicando as *templates* e facendo chamadas a outros servizos, e as *tasks*, que lanzan tarefas pesadas de manexo de datos en segundo plano.

¹<https://gitlab.irilab.org/anxo.pvila/psyprof>

Redis² emprégase para almacenar mensaxes producidas polo código da aplicación, que describen o traballo para realizar na cola de tarefas de Celery. A compoñente de Celery³ encárgase entón de procesar de maneira asíncrona os datos, de xeito que cando se solicitan grandes cantidades de datos estas encólanse para ser despois atendidas. Isto permítelle á aplicación poder procesar as respostas da API de Reddit sen ter que agardar a que se recolla todo o histórico dun suxeito. PostgreSQL⁴ utilízase para o almacenamento persistente dos datos.

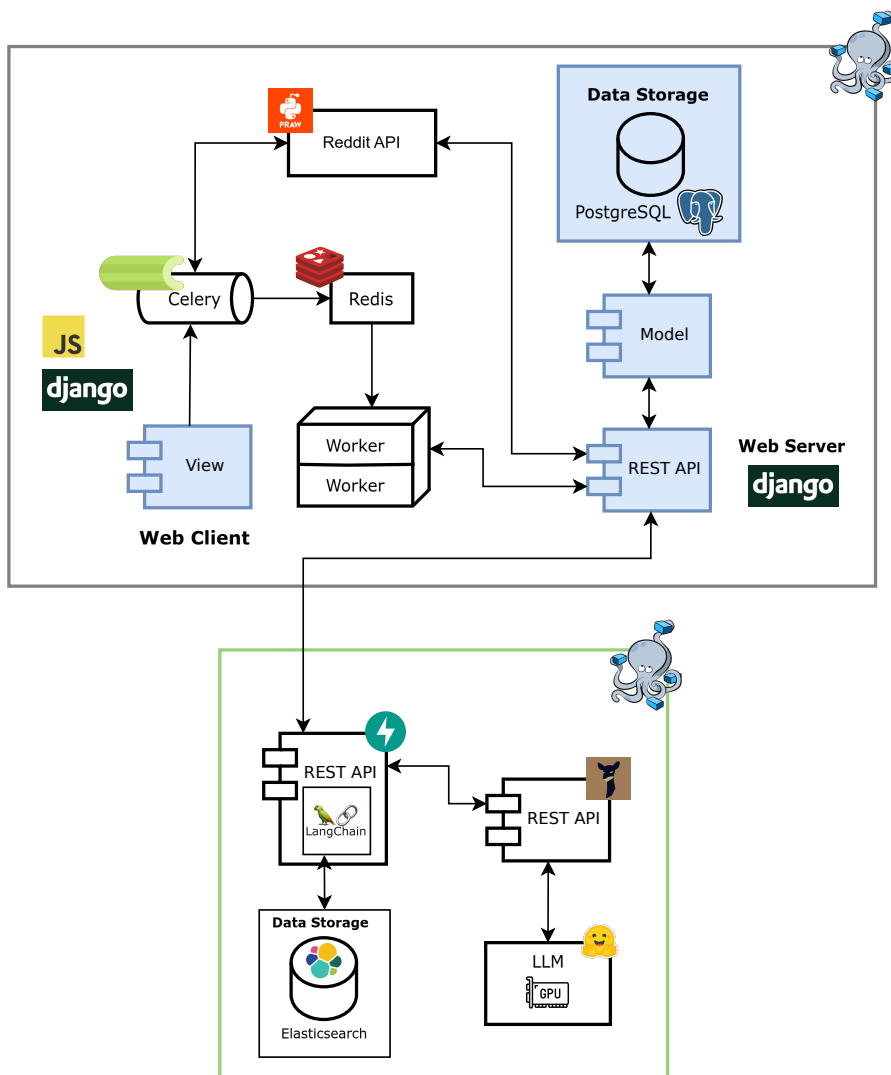


Figura 5.1: Arquitectura da aplicación. En verde o incremento asociado a este TFM. En azul as modificacións substanciais sobre o proxecto base.

² <https://redis.io/>

³ <https://docs.celeryproject.org/en/stable/>

⁴ <https://www.postgresql.org/>

5.1.2 Incremento

Cando o usuario escribe unha mensaxe no *chat* envíase unha petición co contido da propia mensaxe, o identificador do suxeito de análise e o posible historial da conversa. Esta petición é recollida por unha nova [REST API](#) que implementa toda a lóxica de Langchain. Esta lóxica debe comunicarse así mesmo co [LLM](#), que se enmascara tamén detrás dunha [API](#), e co índice de Elasticsearch para obter o histórico de publicacións mediante o identificador do suxeito de análise. Todo isto orquéstase co seguinte *docker compose*:

```
1 version: '2.3'
2 services:
3
4   mindwell_elasticsearch:
5     image: docker.elastic.co/elasticsearch/elasticsearch
6     container_name: mindwell_elasticsearch
7     ports:
8       - "9200:9200"
9     environment:
10      - discovery.type=single-node
11      - xpack.security.enabled=${ELASTICSEARCH_SECURITY}
12      - xpack.security.http.ssl.enabled=${ELASTICSEARCH_SECURITY}
13     volumes:
14      - es-data:/usr/share/elasticsearch/data
15     network_mode: host
16
17   mindwell_model:
18     image: ${MINDWELL_IMAGE}
19     container_name: mindwell_model
20     volumes:
21      - ${MODELS_PATH}:/app/models
22     runtime: nvidia
23     command: >
24       /bin/bash -c '
25         python3 -m fastchat.serve.controller &
26         sleep 10s &&
27         python3 -m fastchat.serve.model_worker --model-names
28         "gpt-3.5-turbo,text-embedding-ada-002" --model-path
29         /app/models/${MINDWELL_MODEL} --num-gpus 1 &
30         sleep 30s &&
31         python3 -m fastchat.serve.openai_api_server --host 0.0.0.0
32         --port ${MODEL_PORT} &
33         sleep infinity
34         '
35     network_mode: host
36
37   mindwell_api:
```

```

35 image: gitlab.irlab.org:5050/irlab/slurm_images:mindwell-api
36 container_name: mindwell_api
37 volumes:
38   - ${PROJECT_PATH}:/app
39   - ${MODELS_PATH}:/app/models
40 environment:
41   - OPENAI_API_BASE=http://${SLURM_NODE}:${MODEL_PORT}/v1
42   - OPENAI_API_KEY=EMPTY
43   - CONTEXT_API_TOKEN=EMPTY
44 command: >
45   /bin/bash -c 'cd /app && uvicorn app:app --host 0.0.0.0
46   --port ${API_PORT} --reload '
47 depends_on:
48   - mindwell_model
49 network_mode: host
50 volumes:
51   es-data:
52     driver: local
53     driver_opts:
54       o: bind
55       type: none
56       device: ${ELASTICSEARCH_DATA}

```

5.2 Fontes de datos e conxunto creado

Para este proxecto considéranse os síntomas do cuestionario clínico *Beck Depression Inventory, Second Version* (BDI-II) [21]. Este formulario cubre 21 síntomas recoñecidos, tales coma a tristeza, o pesimismo ou as alteracións nos ciclos de sono. A decisión para seguir esta formulación radica na súa ampla adopción na práctica clínica [57] e a súa presenza na literatura previa relacionada coa detección de depresión na *Internet* [16, 19, 58, 59]. Baixo estes criterios, selecciónanse as seguintes dúas fontes de datos: BDI-Sen [60] e PsySym [18]. Ambas as dúas coleccións consisten en sentenzas extraídas de Reddit⁵ anotadas a nivel de síntoma de depresión e foron publicadas en conferencias de alto prestixio nos campos da IR e o NLP (SIGIR 2023⁶ e EMNLP 2022⁷, respectivamente). A modo ilustrativo, na Táboa 5.1 recóllense varios exemplos destas fontes de datos. Complementariamente, a Táboa 5.2 amosa as estatísticas de sentenzas positivas, é dicir, que denotan evidencia da presenza dalgún dos síntomas considerados. É importante saber que BDI-Sen representa os 21 síntomas do BDI-II, mentres que PsySym se basea nos síntomas cubertos no *Diagnostic and Statistical Manual of*

⁵ <https://www.reddit.com/>

⁶ <https://sigir.org/sigir2023/>

⁷ <https://2022.emnlp.org/>

Mental Disorders, Fifth Edition (DSM-5) [61]. A este efecto, cabe destacar que se realizou en colaboración cunha persoa con formación en psicoloxía a conversión dos síntomas recollidos no DSM-5 aos síntomas do BDI-II. Esta equivalencia, que é posible dado que existen síntomas en común entre os cuestionarios, recóllese no Apéndice C. As clases etiquetadas en PsySym, se ben se basean no DSM-5, reciben denominacións lixeiramente diferentes aos nomes dos síntomas. A este efecto, realizouse tamén en colaboración coa persoa experta a correspondente equivalencia, que se pode consultar no Apéndice D.

Exemplos do conxunto de datos	
(. . .) I want to start a business one moment, then pay out my IRA and travel throughout Europe. I do not comprehend who I am. My short-term memory is terrible, and I can not concentrate. I'm unsure of what to do. You guys are going to advise some really fantastic actions for me to pursue, but ultimately I lack the willpower or energy to carry out your advice.	
Recently, I have been having a lot of difficulty with this. I have been depressed, worried, or ill since I was a child. Like my youth has been taken from me. As a last-ditch effort to feel like a person once more, I am actually considering seeing a naturopathic physician. Anyway, I hope you soon feel better.	
I abhor myself to the core. Even reading back through Reddit postings I posted a few days ago makes me want to commit suicide. I am such a disgusting waste of life – useless, unproductive, and with a future that is already in uncertainty. And the fact that Im feeling this way on spring break is something I detest so fiercely. (. . .)	

Táboa 5.1: Cada fila presenta un exemplo das fontes de datos. Cabe destacar que estas mostras están parafraseadas en aras da privacidade e acurtadas para maior claridade. En negraña represéntase a evidencia do síntoma de depresión no texto.

	BDI-Sen	PsySym
<i>Número de posts</i>	357	752
<i>Número de explicacións de síntomas</i>	546	746
.....		
<i>Número medio de explicacións por post</i>	1,53	1,02
<i>Lonxitude media dos posts</i>	127	515
<i>Lonxitude media das explicacións</i>	13,76	13,44

Táboa 5.2: Estatísticas dos exemplos positivos do conxunto de datos elaborado. As lonxitudes medias exprésanse en *tokens*.

No *listing* que se amosa a continuación ilústrase o aspecto do conxunto de datos final-

mente construído e consumido polo modelo⁸. Dado que se trata dun modelo conversacional, o conxunto de datos representa conversas entre un humano e o propio modelo, onde o humano pregunta sobre a evidencia de síntomas nun determinado texto e o modelo responde en función de se detecta esa evidencia ou non. Por brevidade, só se amosan un caso positivo e un caso negativo⁹, pero o total de conversas (é dicir, exemplos), son 3008 unha vez efectuadas as limpezas e postprocesados pertinentes dos exemplos dos conxuntos de datos que se toman como fonte. Para dotar de riqueza léxica e naturalidade os datos, definíronse unha serie de plantillas que se ecollen de xeito aleatorio para as preguntas e respostas, de acordo co ilustrado na Figura 5.2.

```

1 [
2   {
3     "id": 0,
4     "conversations": [
5       {
6         "from": "human",
7         "value": "I'd like to inquire whether there are any
8         indications of depressive symptoms in the following passage:\n\n\ughhhh
9         i'm in this position too and i am really pissed off, i feel your pain so
10        much i'm very disappointed in how such an academically great university
11        is unprepared to meet the needs of its new students i'm in a double at
12        west mesa, but at least we only have to pay the triple rate!"
13      },
14      {
15        "from": "gpt",
16        "value": "This scenario is concerning. The text strongly
17        suggests the presence of severe depressive symptoms. For example,
18        'ughhhh i'm in this position too and i am really pissed off, i feel your
19        pain so much i'm very disappointed' clearly reflects the symptom
20        'sadness'."
21      }
22    ]
23  },
24  {
25    "id": 1,
26    "conversations": [
27      {
28        "from": "human",
29        "value": "I'm concerned about someone. Can you evaluate this
30        text for any signs of depression?\n\n\nbabies crying i used to work in
31        old mansion that was converted into a residential treatment facility for
32        teenagers in the juvenile justice system the building was really creepy
33        and i had several experiences i'd describe as paranormal but this is my
34        favorite this took place in late 2001 or early 2002 our residents
35        (court-ordered teenage girls on parole with the juvenile court) had
36        their bedrooms on the second floor of the building a kid came out of her
37        bedroom one night to complain that she heard babies crying and she
38        couldn't sleep she said the cries were coming from directly above her

```

⁸ Dado que a sintaxe se herda da API de OpenAI, nomes como “gpt” son usados de maneira explícita en FastChat.

⁹ Parafraseados para preservar o anonimato dos suxeitos de acordo cos *data agreements* das fontes de datos.

```

21     },
22     {
23         "from": "gpt",
24         "value": "There don't seem to be any signs of depression or
related symptoms in the text."
25     }
26 ]
27 }
28 ]

```

```

HUMAN_TEMPLATES = [
    "Could you please tell me if there is any evidence of the presence of symptoms of depression in the following text?\n\n{post}",
    "Please analyze the following text and tell me if you notice any signs of depression:\n\n{post}",
    "Could you provide your opinion on whether the given text contains indications of depression?\n\n{post}",
    "Hey, can you check if there are any signs of depression in this text?\n\n{post}",
    "I'd like to know if there is any evidence of depressive symptoms in the following passage:\n\n{post}",
    "I'm worried about someone. Can you analyze this text for any signs of depression?\n\n{post}",
    "I'm trying to learn how to identify depression symptoms. Can you help me by analyzing this text?\n\n{post}",
    "As part of a formal assessment, I need to determine if this text exhibits any indications of depression:\n\n{post}",
    "My friend has been acting strangely, and I'm concerned. Can you check if there are any signs of depression in this text?\n\n{post}",
    "I came across this text and was wondering if it shows any signs of depression. What do you think?\n\n{post}",
    "In my professional capacity, I need to assess if there's evidence of depressive symptoms in this text:\n\n{post}",
    "I'm studying psychology, and I'd like to practice identifying depression symptoms. Can you help me by analyzing this text?\n\n{post}",
]

GPT_TEMPLATES = {
    "positive": [
        "Yes, the text indicates the presence of symptoms of depression. For example, fragment '{evidence}' reveals the symptom known as '{symptom}'.",
        "Indeed, the text suggests the presence of depression-related symptoms. One instance is '{evidence}', which appears to be a manifestation of the symptom called '{symptom}'.",
        "Yes, I've identified signs of depressive symptoms in the text. For instance, '{evidence}' is indicative of the symptom known as '{symptom}'.",
        "The text does contain subtle indications of depression. For example, the mention of '{evidence}' might be related to the symptom of '{symptom}'.",
        "While not immediately evident, the text does contain subtle clues that may point to the presence of depressive symptoms. For instance, '{evidence}' could be linked to the symptom called '{symptom}'.",
        "This is a concerning case. The text strongly indicates the presence of severe depression symptoms. For instance, '{evidence}' is a clear manifestation of the symptom '{symptom}'."
    ],
    "negative": [
        "I have found no evidence of the presence of depressive symptoms in the text you have provided.",
        "I couldn't detect any evidence of depressive symptoms in the provided text.",
        "It appears that the text does not exhibit any signs of depression or related symptoms.",
        "I found no indications of depressive symptoms in the text you supplied.",
        "I've carefully analyzed the text, but I couldn't find any evidence of depressive symptoms.",
        "The text is clear of any indicators related to depression or its symptoms. There's no evidence of depressive symptoms here.",
        "I want to reassure you that I couldn't identify any signs of depressive symptoms in the text you provided. If you have concerns, consider seeking professional advice.",
        "I've carefully reviewed the text, and I couldn't find any evidence of depressive symptoms."
    ]
}

```

Figura 5.2: Plantillas definidas para os exemplos tanto positivos como negativos dos turnos.

5.3 *Fine-tuning* do modelo

A Táboa presenta un resumo de todos os parámetros e configuracións utilizados para o proceso de *fine-tuning* de Vicuna-7B-1.5¹⁰ [22]. Para mellorar a eficacia do modelo, o adestramento incorpora a atención de longo alcance *Low-Rank Adaptation of Large Language Mo-*

¹⁰ <https://huggingface.co/lmsys/vicuna-7b-v1.5>

dels (LoRA). O adestramento estendeuse ao longo de 1000 épocas e emprega precisión mixta con representación de punto flotante de 16 bits. O tamaño dos *batches* tanto para o adestramento como para a avaliación foi configurado a 2. Os pasos de acumulación de gradientes foron establecidos a 1. As estratexias de avaliación e gardado do modelo baseáronse en intervalos de pasos, efectuando a avaliación cada 100 pasos e gardando os *checkpoints* intermedios cada 200. O número total de modelos gardados foi limitado a 2, asegurándose de sempre preservar o mellor ata o momento. A taxa de aprendizaxe foi establecida en $2e - 5$ e a lonxitude máxima do modelo fixouse en 2048 *tokens*. O proceso foi optimizado con DeepSpeed.

Parámetro	Valor
--lora_r	8
--lora_alpha	16
--lora_dropout	0.05
--num_train_epochs	1000
--fp16	True
--per_device_train_batch_size	2
--per_device_eval_batch_size	2
--gradient_accumulation_steps	1
--eval_steps	100
--save_steps	200
--save_total_limit	2
--learning_rate	$2e - 5$
--model_max_length	2048

Táboa 5.3: Táboa resumo dos parámetros e configuracións utilizados para o proceso de *fine-tuning* de Vicuna-7B-1.5.

A Figura 5.3 correspóndese con unha gráfica que ilustra os valores da función de *loss* en adestramento e avaliación durante o proceso de *fine-tuning* do LLM. En xeral, unha función de *loss* (función de perda) é unha medida que avalía a diferenza entre as predicións do modelo e as etiquetas verdadeiras do conxunto de datos. O obxectivo durante o adestramento é minimizar a función de perda para mellorar a precisión do modelo na tarefa desexada mediante a variación dos pesos que se están a modificar. Por outra parte, o propósito de vixiar a función de perda nun conxunto de avaliación é evitar o sobreadestramento do modelo e, de acordo coa Figura que se comenta, neste caso é xusto afirmar que o adestramento podería ter rematado

antes das épocas totais definidas. Como nota complementaria, destacar que ao axustar modelos conversacionais do tipo de Vicuna-7B-1.5, o método común de cálculo adoita ser a suma da perda de entropía cruzada de todos os tokens de cada secuencia e dividir este valor pola lonxitude da propia secuencia. Tamén se ilustra coa Figura 5.4 o uso de GPU (ver Subsección 4.2.2 durante todo este proceso.

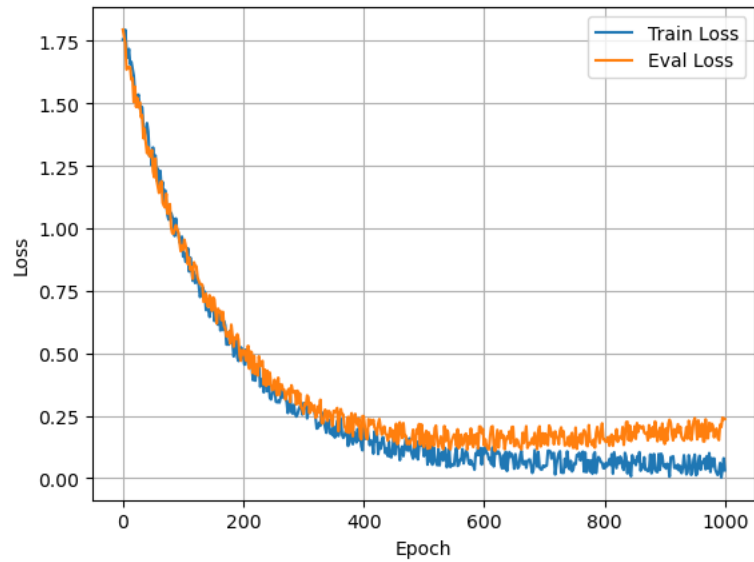


Figura 5.3: Evolución da *loss* obtida durante o proceso de *fine-tuning* en función das épocas consideradas.



Figura 5.4: Consumo de GPU do modelo durante o *fine-tuning*.

Desenvolvemento

O capítulo que segue subdivídese en dúas Seccións. En primeiro lugar, a Sección 6.1 aborda a fase de análise do proxecto, onde se introducen tanto os requisitos funcionais coma os non funcionais. Esta etapa é de vital importancia para o traballo e debe ser realizada nas primeiras etapas do proxecto. Ademais, describe a súa correspondente transformación en Historia de persoa usuaria. A continuación, a Sección 6.2 detalla o traballo realizado en cada un dos *Sprints*, mostrando o progreso, os problemas e as solucións que ocorreron durante o desenvolvemento do proxecto.

6.1 Análise de requisitos

A Tarefa de análise de requisitos desenvólvese, habitualmente, a través de encontros cos clientes nos que estes deben especificar todas as súas necesidades. Aquelas necesidades que sexan viables (no sentido de ser implementables e satisfacibles) conformarán os requisitos funcionais. No caso deste proxecto, o proceso de elicitación de requisitos tivo lugar nas primeiras reunións entre o alumno e os directores, que asumen o papel de *Product Owners*.

6.1.1 Requisitos funcionais

En liña co explicado anteriormente, os requisitos funcionais poden definirse tamén como as características e/ou funcionalidades que debe ter o sistema. Os requisitos identificados a un nivel alto son os seguintes:

Recompilar e curar un conxunto de datos

Este requisito non queda satisfeito polo sistema que se desenvolve, se non que debe ser cumprido para que o sistema poida chegar a funcionar cos valores de calidade agardados. Así pois, o primeiro paso é ter dispoñible unha cantidade suficiente de datos que permita o *fine-tuning* do LLM escollido. Xustifícase como requisito en tanto que é unha demanda dos *Product Owners*.

Realizar o *fine-tuning* do LLM

Novamente, este requisito non queda satisfeito polo sistema en si, se non que é un requisito para que o propio sistema (o *chatbot*) poida chegar a funcionar cos valores de calidade agardados. Xustifícase como requisito en tanto que é unha demanda dos *Product Owners*.

Desenvolver unha interface gráfica para o *chatbot*

O sistema final será interactuable a través dunha plataforma *web*. Neste sentido, dotar ao modelo dunha interface gráfica amigable é un dos primeiros requisitos elicitados durante a reunión cos *Product Owners*.

Conseguir unha plataforma profesional na que integrar o sistema

Os *Product Owners* deixan claro que a intención é que a ferramenta sexa usada por profesionais da psicoloxía clínica, de modo que a plataforma debe soportar a xestión das contas dos mesmos. Por outra parte, en tanto que o obxecto das análises son persoas usuarias de Reddit, a plataforma ten que permitir recoller información sobre este tipo de perfís.

Manter un índice de Elasticsearch coa actividade dos pacientes

Durante a reunión inicial os *Product Owners* poñen especial énfase en que queren manter a actividade dos suxeitos analizados nun índice de Elasticsearch. Este é un almacén denso de representacións vectoriais en forma de *embeddings*, que a maiores ofrece tamén metadatos sobre os mesmos como pode ser a data de publicación o *subreddit*.

Implementar un mecanismo para que o LLM acceda ao índice

O propósito que pretende satisfacer este requisito é que o LLM poida *razonar* sobre a actividade dos suxeitos, de modo que as súas respostas estean baseadas en evidencias extractivas. Con isto preténdese que todas as respostas estean baseadas na histórico dos suxeitos e así non *inventar*.

Implementar un mecanismo de confiabilidade nas respostas do LLM

Requisito moi ligado co anterior e que, neste caso, pretende que as respostas sexan o máis confiables posibles, reducindo o coñecido problema de *hallucination* destes modelos.

Validar o comportamento da solución desenvolvida con persoal experto

É imprescindible asegurarse de que o produto final sexa usable, confiable e de utilidade para as persoas usuarias finais. Neste sentido, deberase levar a cabo unha fase de avaliación por parte deste tipo de persoal experto.

6.1.2 Requisitos non funcionais

Os requisitos non funcionais establecen os criterios que o sistema debe cumprir. Máis que simplemente as súas características específicas, definen as cualidades, atributos e restricións do sistema. Desta maneira, están directamente ligados ás decisións de deseño que se toman, pois establecen os estándares para o rendemento, a seguridade e a usabilidade do sistema. En particular, para este proxecto, identificáronse os seguintes requisitos non funcionais:

Eficacia

É necesario que o *chatbot* se comporte segundo o esperado. En concreto, dado o seu propósito e a natureza dos datos que manexa, é imprescindible que as súas respostas sexan confiables e teñan unha base sólida nos feitos.

Escalabilidade

O sistema debe ser quen de manexar un número variable (nos dous sentidos) de clientes. É dicir, debería ser *elástico*.

Extensibilidade e manteñibilidade

Resulta desexable poder engadir novas funcionalidades e modificacións ao longo do ciclo de vida completo do produto. Tamén se persegue o poder integrar o sistema dentro doutras plataformas *web* xa existentes.

Velocidade de resposta asumible

A inferencia con este tipo de modelos é un proceso moi intensivo en recursos, cómputo e, en consecuencia, tempo. Non obstante, ao interactuar co *chatbot*, a experiencia que cabe esperar é rápida e fluída. Se ademais sumamos o feito de que o sistema debe ser *elástico* en canto ao número de clientes que atende, este requisito vólvese máis crítico se cabe.

6.1.3 Historias de persoa usuaria

Partindo dos requisitos funcionais descritos a alto nivel (ver Sección 6.1.1), é factible especificar a totalidade dos requisitos funcionais mediante a definición das Historias de persoa usuaria. Todas as Historias definidas para este proxecto están recollidas na Táboa 6.1, cada unha delas coa súa estimación de Puntos de Historia. É importante notar que esta lista de Historias constitúe o chamado *Product Backlog*, desde onde se seleccionan as Historias para incorporar ao *Sprint Backlog* de cada *Sprint*.

Táboa 6.1: Historias de Persoa Usuaria e puntos estimados.

ID	Definición	Puntos
1	Como <i>developer</i> necesito formarme respecto dos fundamentos teóricos do proxecto.	30
2	Como <i>analyst</i> e <i>developer</i> necesito estudar e seleccionar as tecnoloxías a empregar.	20
3	Como <i>product owner</i> quero construír un conxunto de datos apropiado para esta tarefa.	30
4	Como <i>developer</i> necesito coñecer o <i>SOTA</i> en canto aos conxuntos de datos.	10
5	Como <i>product owner</i> quero facer <i>fine-tuning</i> dun <i>LLM</i> para a tarefa definida.	50
6	Como <i>developer</i> necesito formarme e coñecer en detalle os aspectos relacionados co <i>fine-tuning</i> .	20
7	Como <i>developer</i> quero ser quen de interactuar co <i>LLM</i> por unha <i>API</i> .	30
8	Como <i>user</i> quero comunicarme co <i>LLM</i> mediante unha <i>Graphical User Interface (GUI)</i> .	25
9	Como <i>product owner</i> quero conseguir unha plataforma profesional na que integrar o sistema.	25
10	Como <i>user</i> quero conversar co <i>LLM</i> sobre as publicacións dos suxeitos.	20
11	Como <i>product owner</i> quero manter nun índice de Elasticsearch as publicacións dos suxeitos.	25
12	Como <i>user</i> quero que o sistema incorpore un mecanismo de confiabilidade para as respostas do <i>LLM</i> .	25
13	Como <i>developer</i> e <i>product owner</i> quero validar o comportamento da solución desenvolvida con persoal experto.	50

6.2 Sprints

Esta Sección detalla o proceso de desenvolvemento levado a cabo ao longo do proxecto, describindo as Historias de persoa usuaria realizadas en cada *Sprint* e facendo un breve balance ao rematar cada un deles. Despois da planificación inicial do proxecto, a suma total do *Product Backlog* ascende a 350 Puntos de Historia. Considerando que a duración do proxecto será de 7 *Sprints*, a meta é completar 50 Puntos de Historia en cada *Sprint* idealmente. Na Figura 6.1 amósase a estimación inicial do proxecto, onde se poden observar varios detalles:

Project points

Representan a totalidade dos puntos asignados para o proxecto.

Defined points

Representan a cantidade de puntos asignados ás Historias de Persoa Usuaria.

Closed points

Representa a suma dos Puntos de Historia xa completados.

Points/Sprint

Representa a media de Puntos completados en cada *Sprint*. Na terminoloxía de Scrum 4.1 coñécese como *velocity* e, de acordo co explicado previamente, debe manterse sempre arredor dos 50 Puntos de Historia por *Sprint*.



Figura 6.1: Planificación inicial do proxecto.

6.2.1 *Sprint* 1 (21/09/23 ao 08/10/23): Formación teórica e escolla tecnolóxica

O primeiro *Sprint* do proxecto supón a base sobre a que se levantará todo o desenvolvemento posterior. En concreto, selecciónanse as Historias 1 e 2, tal e como ilustra a Figura 6.2. Ambas as dúas son Historias de formación e configuración, o cal é esperable en tanto que se trata dunha fase temperá do proxecto. Construír unha base sólida sobre a que despois iterar é un aspecto fundamental para un resultado final de calidade. Como parte do proceso, realizouse unha división das Historias en Tarefas, identificando as seguintes:

T1 - Estudo dos fundamentos teóricos.

T1.1 - Estudo e asimilación do relacionado con *Large Language Models*.

T1.2 - Estudo e asimilación do relacionado con *IA* conversacional.

T1.3 - Estudo e asimilación do relacionado con *IA* explicable.

T1.4 - Estudo e asimilación do relacionado con *Retrieval Augmented Generation*.

T2 - Análise das tecnoloxías.

T2.1 - Busca e comparativa de librarías e compoñentes.

T2.2 - Estudo da documentación das librarías escollidas.

T2.3 - Configuración do contorno de desenvolvemento.

T2.4 - Configuración do contorno de xestión do proxecto.

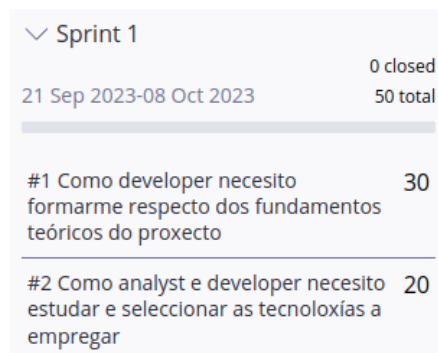


Figura 6.2: Planificación inicial do *Sprint* 1.

As Historias seleccionadas inicialmente completáronse antes da finalización do *Sprint*, é dicir, a estimación dos Puntos de Historia foi demasiado conservadora. Non obstante, cabe destacar que este resultado estaba contemplado, xa que se preferiu pecar de conservadurismo e non de optimismo en canto á asignación de Puntos de Historia nas fases temperás do proxecto. Con isto preténdese evitar a toda costa retrasos xa desde fases iniciais e poder ter unha primeira proba do ritmo de avance do proxecto. Así mesmo, a experiencia previa do autor con temática relacionada á deste traballo fixo que se avanzase máis rápido do previsto estudando os fundamentos e seleccionando as tecnoloxías. Deste xeito, para tratar de manter constante a dedicación dos *Sprints*, modificouse a estimación de Puntos para as Historias 1 e 2 de acordo co sucedido e incorporouse a Historia 4 ao *Sprint Backlog*. Para esta nova Historia, fíxose a correspondente descomposición en Tarefas:

T3 - Estudo do SOTA en canto aos conxuntos de datos.

T3.1 - Busca de conxuntos de datos relacionados coa Tarefa de *fine-tuning* para o noso escenario concreto.

T3.2 - Familiarización cos conxuntos de datos seleccionados.

Sprint Review

A estimación inicial foi demasiado conservadora en canto aos Puntos de Historia. Por iso, as Historias 1 e 2 completáronse antes da finalización do *Sprint*. Para manter a *velocity* no valor desexado, incorporouse a Historia 4 ao *Sprint Backlog*, tendo feitas as correspondentes modificacións nas estimacións de Puntos. Deste xeito, a Figura 6.3 amosa as Historias finalmente completadas no *Sprint 1*, así como os Puntos de Historia reais. A Figura 6.4 ilustra o progreso do proxecto ao remate desta iteración, onde pode ver que ao remate do primeiro *Sprint* a *velocity* do proxecto é a adecuada, manténdose nos 50 Puntos de Historia por *Sprint*. Como conclusión ao peche deste *Sprint*, o resultado considerase exitoso, xa que aínda que houbo que facer un pequeno axuste, o equipo respondeu ben e a *velocity* puido manterse.



Figura 6.3: Historias finalmente completadas no *Sprint 1*.

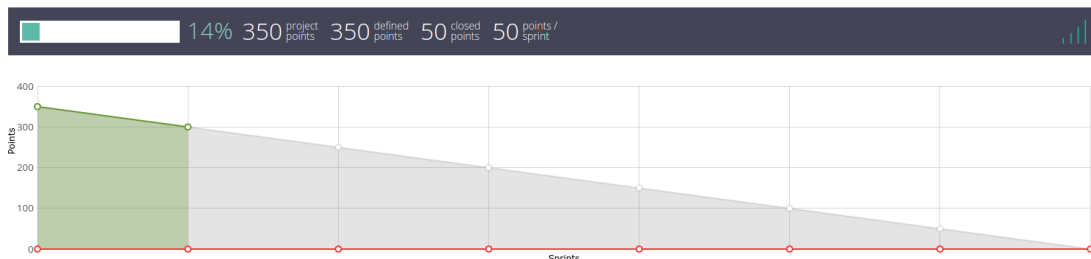


Figura 6.4: Progreso do proxecto ao peche do *Sprint 1*.

6.2.2 *Sprint* 2 (09/10/23 ao 25/10/23): Elaboración do conxunto de datos

Superado o proceso inicial de formación teórica, escolla de tecnoloxías e análise dos conxuntos de datos relacionados, o *Sprint* 2 centrouse na elaboración do conxunto de datos propio para este proxecto, co obxectivo principal de facer *fine-tuning* dun LLM fundacional¹. Debe terse en conta que, dado que no anterior *Sprint* se completaron as Historias do *Sprint Backlog* antes da finalización do mesmo, decidiuse acometer tamén a Historia 4 dentro do *Sprint* 1. Esta Historia, a nivel conceptual, debería completarse neste segundo *Sprint*, polo que ao xa estar finalizada outorga unha certa marxe para adiantar traballo correspondente a *Sprints* posteriores. Así pois, á hora de seleccionar no *Product Backlog* os elementos para levar a cabo neste *Sprint* seleccionáronse as Historias 3 e 6 (ver Figura 6.5). Deste xeito, de acordo coa metodoloxía autoimposta, efectuouse a correspondente descomposición das Historias en Tarefas, identificando as seguintes:

T4 - Construcción dos conxuntos de datos.

T4.1 - Preprocesado da información.

T4.1.1 - Limpeza de datos.

T4.1.2 - Estandarización.

T4.2 - Integración das orixes homoxenizando formatos.

T4.3 - Verificación da integridade e consistencia do conxunto resultante.

T5 - Adquisición de habilidades prácticas para *fine-tuning*.

T5.1 - Identificación de ferramentas/métodos.

T5.2 - Realización de titoriais, identificando casos de uso e exemplos prácticos.

T5.3 - Familiarización con parámetros e opcións de configuración habituais.

Sprint Review

A planificación inicial do *Sprint* resultou ser acertada, de modo que ambas as dúas Historias seleccionadas se completaron cos Puntos estimados, de acordo co ilustrado na Figura 6.6. De xeito complementario, a Figura 6.7 amosa o estado do proxecto e a *velocity* ao remate do *Sprint* 2, onde se pode observar que os Puntos/*Sprint* se manteñen no valor esperado a esta altura do desenvolvemento.

¹ <https://aws.amazon.com/what-is/foundation-models/>

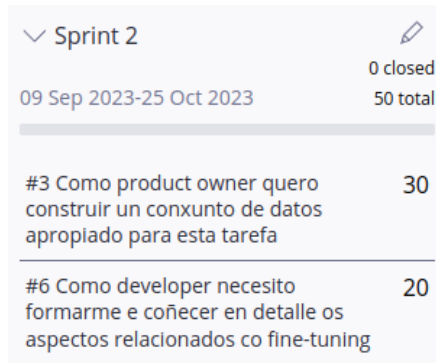


Figura 6.5: Planificación inicial do *Sprint 2*.

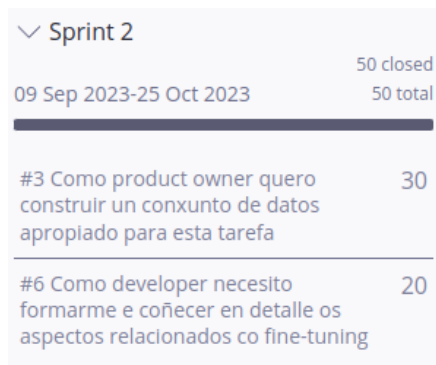


Figura 6.6: Historias finalmente completadas no *Sprint 2*.

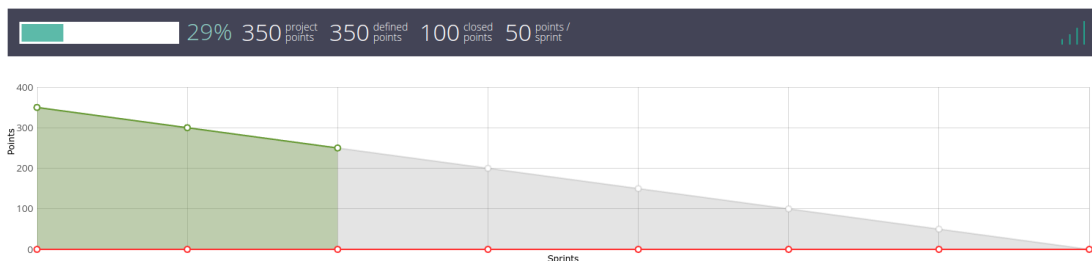


Figura 6.7: Progreso do proxecto ao peche do *Sprint 2*.

6.2.3 *Sprint 3* (26/10/23 ao 12/11/23): *Fine-tuning* do LLM

O incremento asociado a este *Sprint* resulta especialmente importante, xa que o seu entregable constitúe o *corazón* do proxecto entanto que o *LLM* é o elemento sobre o que se orquestra o resto do desenvolvemento. Así pois, é lóxico pensar que a Historia asociada a este obxectivo será moi intensiva en canto ao esforzo (e o tempo), o que xustifica os 50 Puntos asignados para a Historia 5 que constitúe o *Product Backlog* do *Sprint*. Deste xeito, a Figura 6.8

ilustra a estimación inicial de Historias para o *Sprint 3*. Como é habitual, fíxose a pertinente identificación de Tarefas, definindo as seguintes:

T6 - Preparación do contorno de adestramento, incluíndo *hardware* e *software*.

T6.1 - Estudo da plataforma *hardware*.

T6.2 - Preparación da imaxe Docker.

T7 - *Finetuning* do LLM.

T7.1 - Implementación do código para o adestramento.

T7.2 - Selección de hiperparámetros e escolla dos valores a empregar.

T7.3 - Execución do adestramento e revisión dos *logs*.

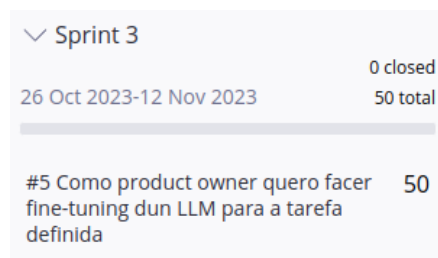


Figura 6.8: Planificación inicial do *Sprint 3*.

Sprint Review

Durante este *Sprint* tivo lugar un imprevisto que supuxo o primeiro problema serio na xestión do proxecto. Ocorreu que, durante dúas semanas deste período, obrigas laborais do autor restáronlle horas de dedicación e impediron que o esforzo se mantese no valor desexado de 50 Puntos de Historia/*Sprint*. Así pois, non se puido completar a Historia 5, polo que se decidiu partila en dúas, pechar a parte completada e devolver ao *Product Backlog* a parte restante. Isto queda reflexado na Figura 6.9 e, complementariamente, a Figura 6.10 amosa o estado do proxecto ao final desta iteración. Pode observarse que a *velocity* sufriu en este punto un descenso moi considerable, ocasionando unha desviación que podería resultar nun retraso na entrega final. Isto tratará de corrixirse nos seguintes *Sprints* por medio de xornadas con dedicación *extra*.

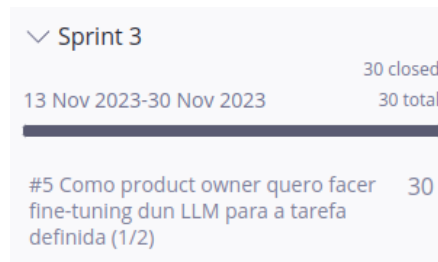


Figura 6.9: Historias finalmente completadas no *Sprint* 3.

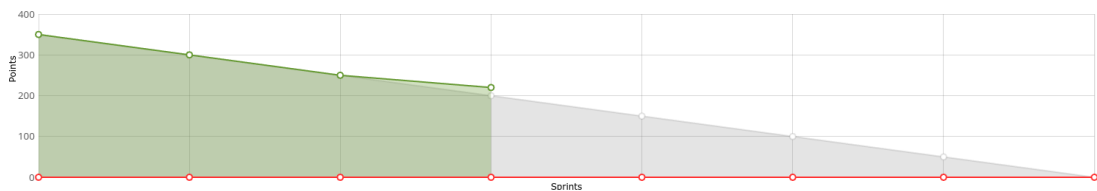


Figura 6.10: Progreso do proxecto ao peche do *Sprint* 3.

6.2.4 *Sprint* 4 (13/11/23 ao 30/11/23): Interacción co LLM (1/2)

Unha vez rematado o *Sprint* anterior, o resultado deste novo incremento debería permitir a interacción co LLM dunha maneira básica a través dunha API REST. Non obstante, hai que recordar que a Historia 5 tivo que ser dividida en dúas partes, xa que non se puido completar na súa totalidade no *Sprint* anterior. Así pois, acometer a segunda parte da mesma é o obxectivo prioritario neste punto do proxecto. É por iso que, tal e como se pode ver na Figura 6.11, se incorpora ao *Sprint Backlog* unha nova Historia froito da división que se acaba de mencionar. Por outra parte, en liña co nome do *Sprint*, selecciónase a Historia 7. Así mesmo, selecciónase a Historia 9 co propósito de *adiantar* traballo e compensar así a desviación do *Sprint* anterior. Tamén, e de acordo co habitual, fíxose a habitual identificación de Tarefas:

T8 - Implementación da API.

T8.1 - Conceptualización dos *endpoints*.

T8.2 - Implementación.

T8.2 - Documentación.

T9 - Procura de plataforma na que integrar o sistema.

T9.1 - Análise de repositorios públicos *open source*.

T9.2 - Probas de concepto.

T9.3 - Decisión final e estudo das adaptacións necesarias.

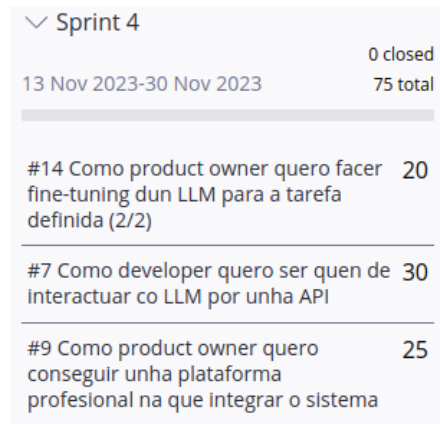


Figura 6.11: Planificación inicial do *Sprint 4*.

Sprint Review

Tratando de corrixir o retraso acumulado, fíxose unha planificación demasiado obxectiva, buscando completar 75 Puntos de Historia neste *Sprint*. Finalmente, isto non puido ser así e só se completaron 50 Puntos de Historia, tal e como ilustra a Figura 6.12. Isto, se ben identifica a *velocity* deste *Sprint* no valor agardado para cada un dos *Sprints*, non permite recuperar a desviación que se arrastra. Así pois, nos seguintes *Sprints* deberase seguir a plantexar estratexias que permitan corrixir o retraso acumulado. Complementariamente, a Figura 6.13 amosa o estado do proxecto e a *velocity* ao remate do *Sprint 4*, onde se pode observar que os Puntos/*Sprint* se sitúan en 45 a esta altura do desenvolvemento.



Figura 6.12: Historias finalmente completadas no *Sprint 4*.

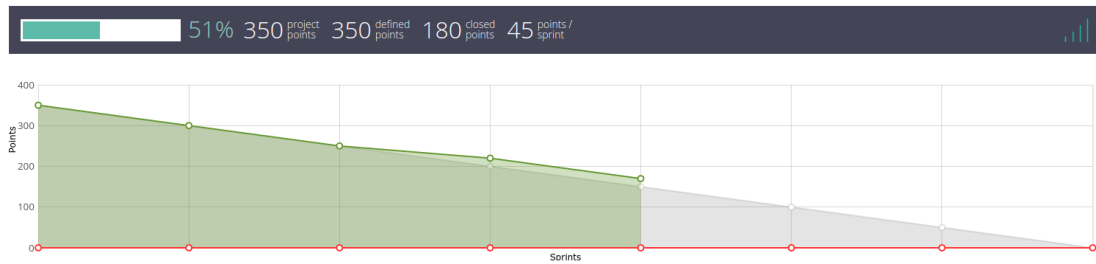


Figura 6.13: Progreso do proxecto ao peche do *Sprint* 4.

6.2.5 *Sprint* 5 (01/12/23 ao 17/12/23): Interacción co LLM (2/2)

Como entregable destacable do anterior *Sprint*, elaborouse unha *API REST* que permite a interacción básica co *LLM*. Durante ese *Sprint*, tratouse de corraxir sen éxito a desviación acumulada, algo que neste *Sprint* 5 se tratou de facer de novo. Non obstante, a estratexia foi máis realista, pois só se planificaron 10 Puntos de Historia extra sobre a *velocity* habitual. Así pois, seleccionáronse as Historias 8 e 9 como propias do *Sprint* e incluíuse tamén a Historia 10 para recuperar esforzo perdido. A estes efectos, a Figura 6.14 ilustra a configuración do *Sprint Backlog* ao comezo da iteración. Como é habitual, realizouse a correspondente descomposición das Historias en Tarefas:

T10 - Comunicación co LLM por medio dunha GUI.

T10.1 - Deseño gráfico da interface.

T10.2 - Implementación do *frontend*.

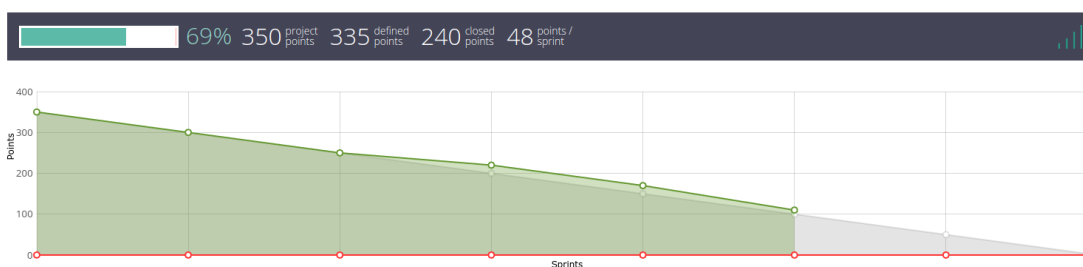
T10.2 - Implementación do *backend*.

T10.3 - Incorporación do historial do propio *chat*.

T11 - Conversa co LLM sobre o contido das publicacións.

Sprint Review

Nesta ocasión a planificación foi exitosa, xa que as condicións permitiron o esforzo extra e se puideron completar os 60 Puntos de Historia, tal e como se ilustra na Figura 6.15. O feito de subir a *velocity* neste *Sprint* permitiu recuperar parte do retraso acumulado, de acordo co amosado na Figura 6.16, que representa o estado do proxecto ao remate desta iteración. Cabe destacar que a *velocity* global se sitúa xa moi próxima aos 50 Puntos por *Sprint* desexados. Neste sentido, asúmese que a desviación está xa corraxida.

Figura 6.14: Planificación inicial do *Sprint 5*.Figura 6.15: Historias finalmente completadas no *Sprint 5*.Figura 6.16: Progreso do proxecto ao peche do *Sprint 5*.

6.2.6 *Sprint 6* (18/12/23 ao 03/01/24): Incorporación do contexto

Para este penúltimo *Sprint* selecciónanse as Historias 11 e 12, co propósito de manter nun índice de Elasticsearch as publicacións dos suxeitos e incorporar un sistema de confiabilidade

para as respostas do LLM, tal e como se ilustra na Figura 6.17. Así pois, a descomposición en Tarefas é a seguinte:

T12 - Creación e mantemento dun índice coas publicacións.

T12.1 - Escolla do conector de LangChain.

T12.2 - *Parsing* dos documentos.

T12.3 - Aseguramento dos metadatos.

T12.4 - Implementación.

T13 - Asegurar confiabilidade nas respostas.

T13.1 - Estudo das *chains* de LangChain.

T13.2 - Escolla da mellor opción.

T13.2 - Implementación.

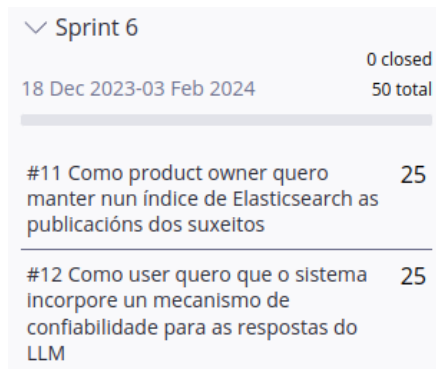


Figura 6.17: Planificación inicial do *Sprint 6*.

Sprint Review

Este penúltimo *Sprint* foi exitoso no seu plantexamento, xa que as Historias seleccionadas se completaron sen ter que sufrir ningunha modificación nas estimacións iniciais segundo o ilustrado na Figura 6.18. A este efecto, a Figura 6.19 representa o estado do proxecto ao remate deste *Sprint*, podendo apreciar unha *velocity* de 48 Puntos por *Sprint*.

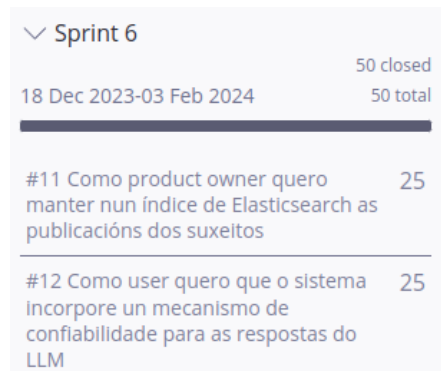


Figura 6.18: Historias finalmente completadas no *Sprint 6*.

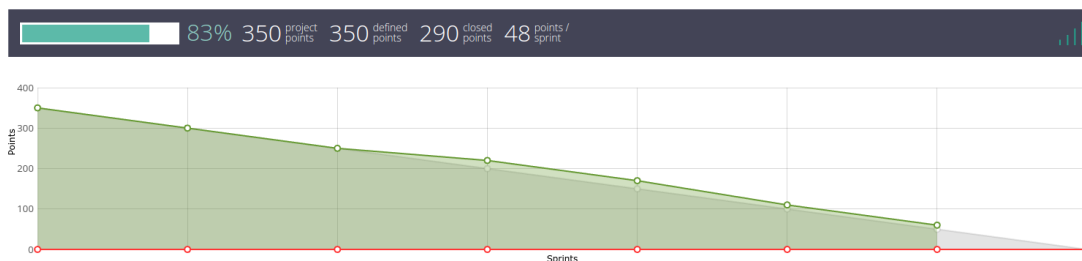


Figura 6.19: Progreso do proxecto ao peche do *Sprint 6*.

6.2.7 *Sprint 7* (04/01/24 ao 21/01/24): Avaliación e peche do proxecto

Este último *Sprint* supón a conclusión do proxecto. Así pois, e unha vez corrixida a desviación que tivo lugar en anteriores etapas do ciclo de vida, a selección das Historias a incluír no *Sprint Backlog* vese unicamente motivada por completar as Historias pendentes. Neste caso, tal e como se mostra na Figura 6.20, só é unha, a Historia 13. O propósito desta Historia é recoller o *feedback* final do perfil clínico experto co que se colaborou durante o desenvolvemento do proxecto, polo que para esta Historia só se identifica unha Tarefa, que se corresponde coa sesión final de *feedback* co perfil experto.

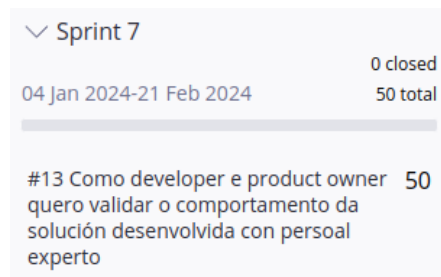


Figura 6.20: Planificación inicial do *Sprint 7*.

Sprint Review

O *Sprint 7* supuxo, de acordo coa planificación, o incremento final do proxecto. É importante saber que este *Sprint* non implicou carga de traballo en canto á implementación, xa que no *Product Backlog* unicamente quedaba por completar a Historia 13. Inicialmente, contempouse unha dedicación maior da que finalmente supuxo esta Historia, polo que se fixeron os axustes pertinentes de acordo co amosado na Figura 6.21. Dado que esta Historia era a derradeira, o *Sprint* pechouse con menos Puntos completados dos que estaban estipulados para cada incremento, pero isto non supón un inconveniente en tanto que o proxecto chegou á súa fin. Complementariamente, a Figura 6.22 ilustra esta situación.

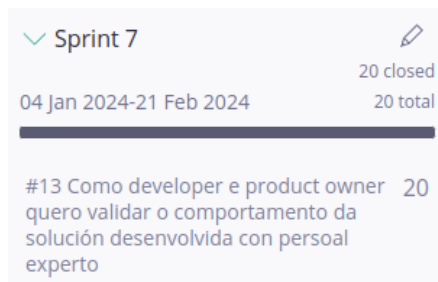


Figura 6.21: Historias finalmente completadas no *Sprint 7*.

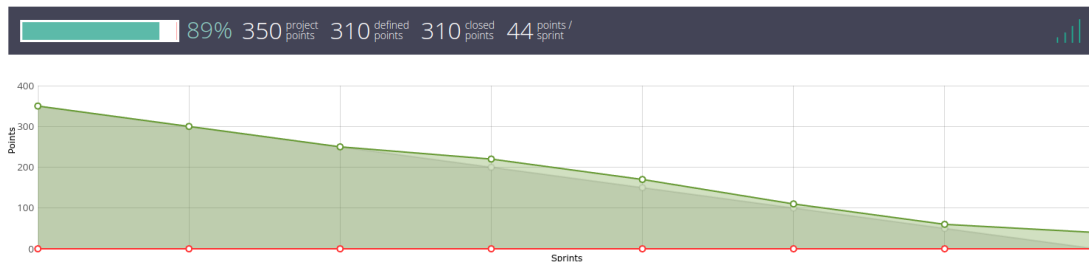


Figura 6.22: Progreso do proxecto ao peche do *Sprint 7*.

6.2.8 Balance final

Na Táboa 6.2 pode verse a asignación real de Puntos para cada unha das Historias de Persoa Usuaria. Complementariamente, a Táboa 6.3 detalla a relación existente as Historias, as Tarefas e os *Sprints*. Reflexionando sobre o desenvolvemento do proxecto, cómpre destacar algúns detalles:

- O uso dunha metodoloxía áxil permitiu corrixir desviacións durante o desenvolvemento.
- O proxecto completouse no tempo establecido e de acordo cos *Sprints* previstos.

- A duración dos *Sprints* mantívose constante nas 2/3 semanas fixadas por iteración.
- O proxecto completouse con menos Puntos de Historia definidos (310) dos inicialmente estimados (350).
- A *velocity* non foi constante, pero mantívose sempre próxima ao valor ideal de 50 (Puntos/*Sprint*).

Táboa 6.2: Historias de Persoa Usuaría e puntos reais.

ID	Definición	Puntos
1	Como <i>developer</i> necesito formarme respecto dos fundamentos teóricos do proxecto.	30 → 25
2	Como <i>analyst</i> e <i>developer</i> necesito estudar e seleccionar as tecnoloxías a empregar.	20 → 15
3	Como <i>product owner</i> quero construír un conxunto de datos apropiado para esta tarefa.	30 → 30
4	Como <i>developer</i> necesito coñecer o <i>SOTA</i> en canto aos conxuntos de datos.	10 → 10
5	Como <i>product owner</i> quero facer <i>fine-tuning</i> dun <i>LLM</i> para a tarefa definida (1/2).	50 → 30
6	Como <i>developer</i> necesito formarme e coñecer en detalle os aspectos relacionados co <i>fine-tuning</i> .	20 → 20
7	Como <i>developer</i> quero ser quen de interactuar co <i>LLM</i> por unha <i>API</i> .	30 → 30
8	Como <i>user</i> quero comunicarme co <i>LLM</i> mediante unha <i>Graphical User Interface (GUI)</i> .	25 → 25
9	Como <i>product owner</i> quero conseguir unha plataforma profesional na que integrar o sistema.	25 → 25
10	Como <i>user</i> quero conversar co <i>LLM</i> sobre as publicacións dos suxeitos.	20 → 20

Continúa na seguinte páxina.

Táboa 6.2 – *Continuación*

ID	Definición	Puntos
11	Como <i>product owner</i> quero manter nun índice de Elasticsearch as publicacións dos suxeitos.	25 → 25
12	Como <i>user</i> quero que o sistema incorpore un mecanismo de confiabilidade para as respostas do LLM.	25 → 25
13	Como <i>developer e product owner</i> quero validar o comportamento da solución desenvolvida con persoal experto.	50 → 20
14	Como <i>product owner</i> quero facer <i>fine-tuning</i> dun LLM para a tarefa definida (2/2).	50 → 20

<i>Sprint</i>	Tarefas	Historias
1 (6.2.1)	T1.1, T1.2, T1.3, T1.4, T2.1, T2.2, T2.3, T2.4, T3.1, T3.2	1, 2, 4
2 (6.2.2)	T4.1.1, T4.1.2, T4.2, T4.3, T5.1, T5.2, T5.3	3, 6
3 (6.2.3)	T6.1, T6.2, T7.1, T7.2, T7.3	5
4 (6.2.4)	T8.1, T8.2, T8.3, T9.1, T9.2, T9.3	7, 14
5 (6.2.5)	T10.1, T10.2, T10.3, T11	8, 9, 10
6 (6.2.6)	T12.1, T12.2, T12.3, T12.4, T13.1, T13.2, T13.3	11, 12
7 (6.2.7)	<i>Feedback final</i>	13

Táboa 6.3: Relación entre *Sprints*, Tarefas e Historias.

Produto Final

O propósito deste Capítulo é amosar o produto finalmente desenvolvido no marco deste TFM. Para isto, a Sección 7.1 ofrece un resumo das vistas e pantallas máis destacadas da aplicación, comentando así as funcionalidades principais da mesma e destacando un exemplo práctico do caso de uso asociado ao *chatbot*. Por outra banda, na Sección 7.2 recóllense as ligazóns a todo o *software* liberado como froito deste proxecto.

7.1 Resumo da aplicación

A aplicación base aporta as características necesarias para considerar este un sistema viable para ser posto eventualmente en produción. Un exemplo disto é o manexo de contas de persoa usuaria. Así pois, na Figura 7.1 amósase a vista asociada á pantalla de *login*, ofrecendo tamén a posibilidade de modificar o contrasinal. Cabe destacar que, dado que se trata dun sistema destinado ao uso por expertos, as novas contas son manualmente creadas por un rol de administración. Por outra banda, a Figura 7.2 ilustra o panel principal da aplicación. Desde el accédese ás demais funcionalidades do sistema e, a efectos deste documento, amósanse as vistas e pantallas correspondentes co fluxo de traballo que debería seguir un profesional usando a aplicación ata poder conversar co *chatbot*. En primeiro lugar, é necesario crear un *Corpus*, operación que se amosa na Figura 7.3. Unha vez creado o *Corpus*, xa se poden obter perfís (suxeitos de estudo) da plataforma Reddit¹. Isto pode facerse de dúas maneiras: ben obtendo os comentarios de múltiples perfís dun determinado *subreddit*, ou ben procesando unha determinada persoa usuaria da rede social. As Figuras 7.4 e 7.5 ilustran ámbalas dúas vías, respectivamente. Unha vez se recolleu a mostra de perfís (suxeitos) desexados, estes amósanse nunha vista resumo que se corresponde coa Figura 7.7. Cada tarxeta é un resumo dun perfil e, facendo *click* nela, accédese ao detalle do perfil, ilustrado tamén na Figura 7.6.

¹<https://www.reddit.com/>

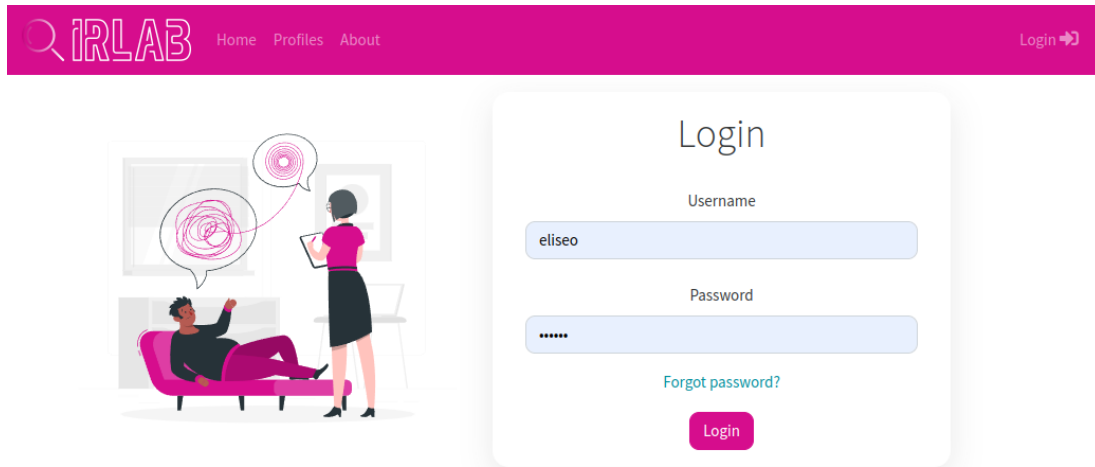


Figura 7.1: Vista da pantalla de *login*.

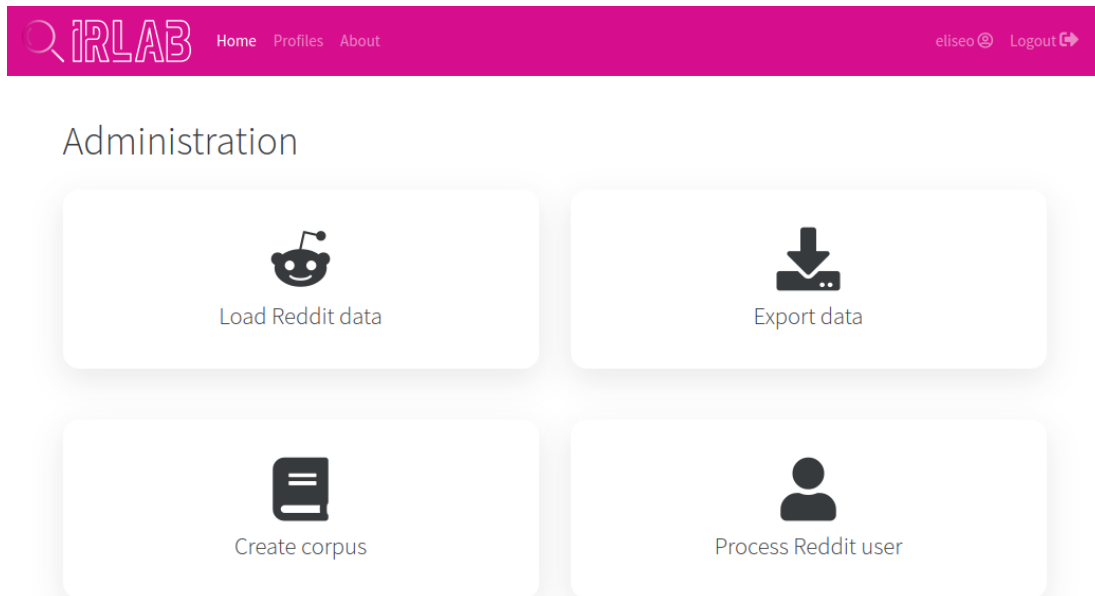


Figura 7.2: Vista do panel de control da aplicación.

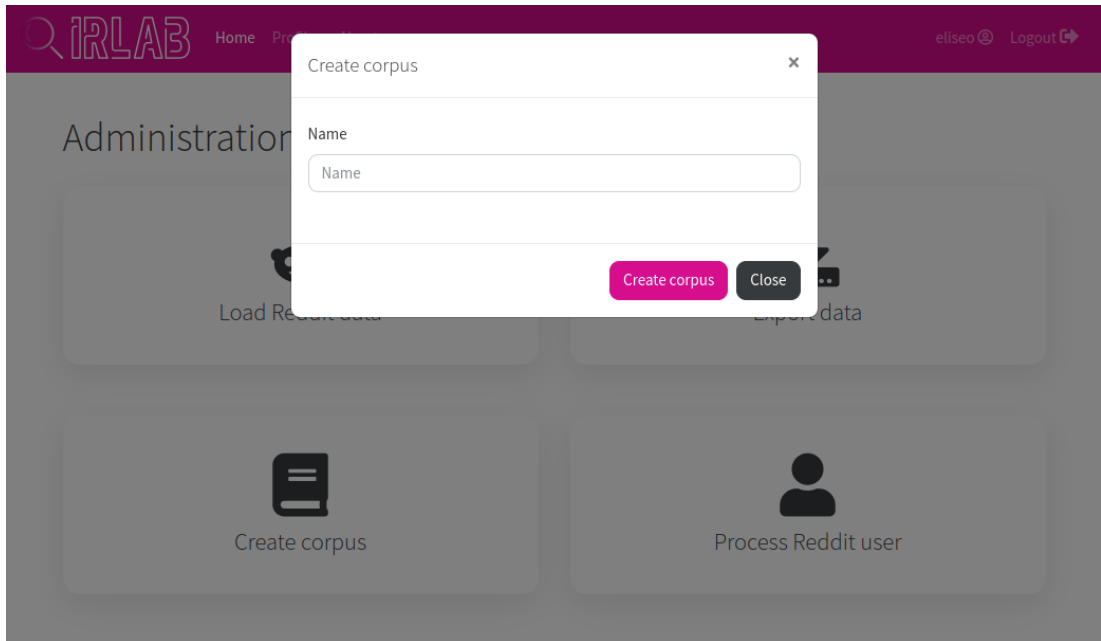


Figura 7.3: Creación dun *Corpus* no que almacenar a información.

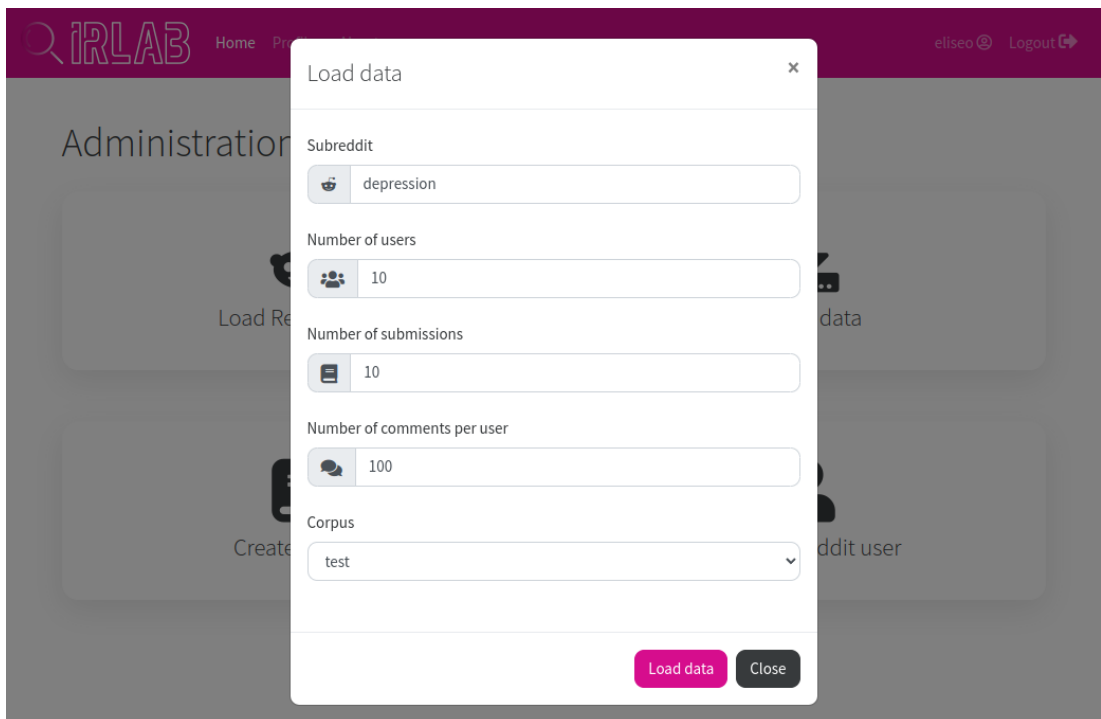


Figura 7.4: Obtendo os comentarios de múltiples perfís dun determinado *subreddit*.

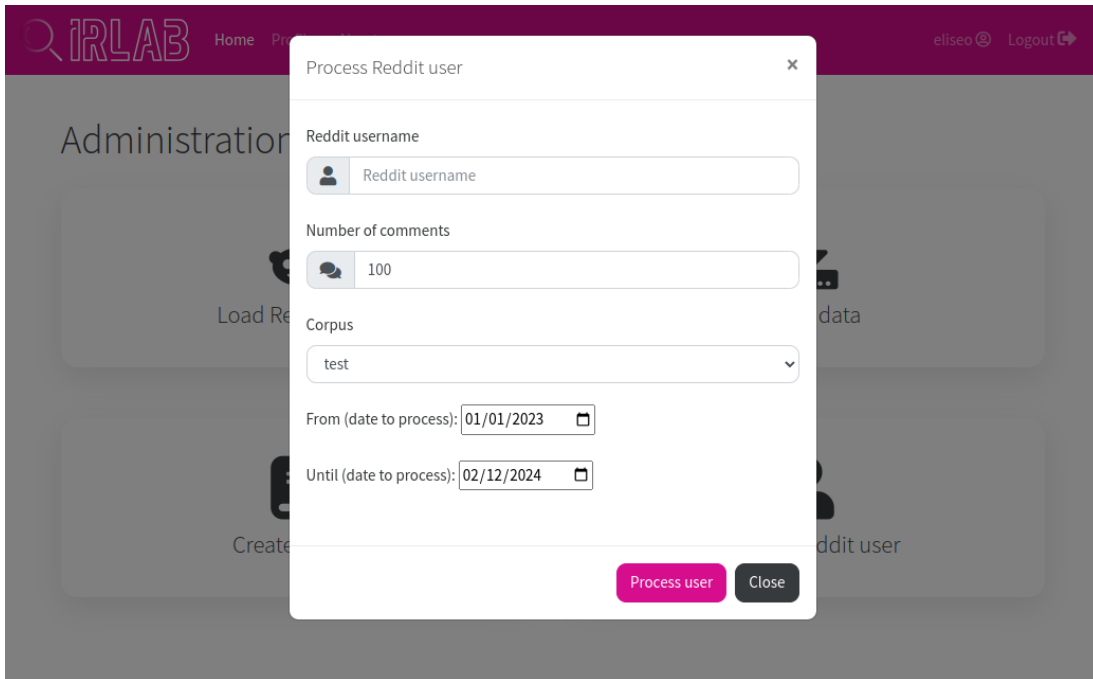


Figura 7.5: Obtendo os comentarios dun perfil (suxeito) concreto de Reddit.

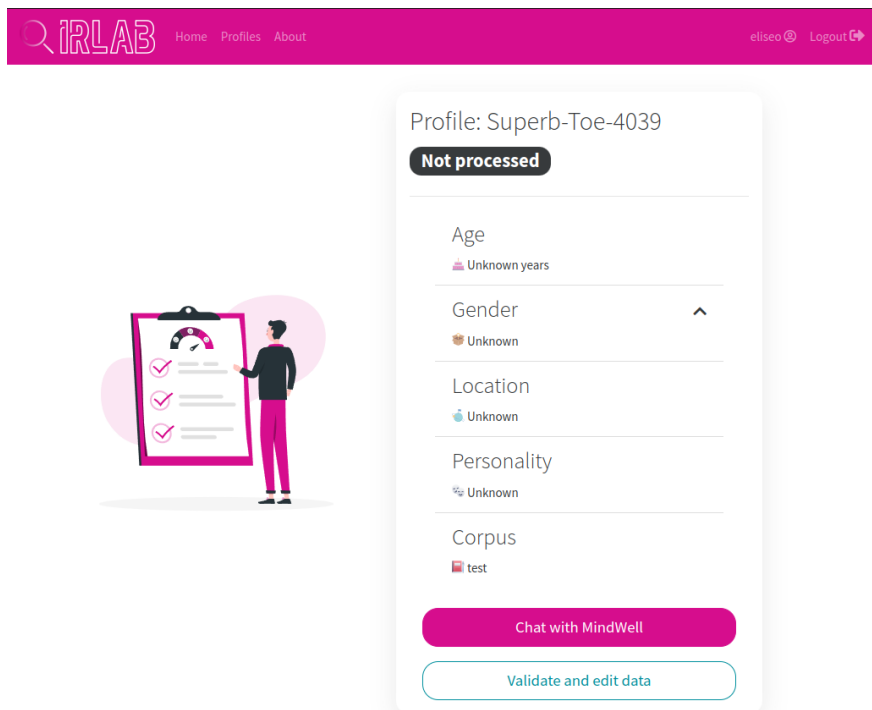


Figura 7.6: Vista da pantalla detalle dun perfil (suxeito).

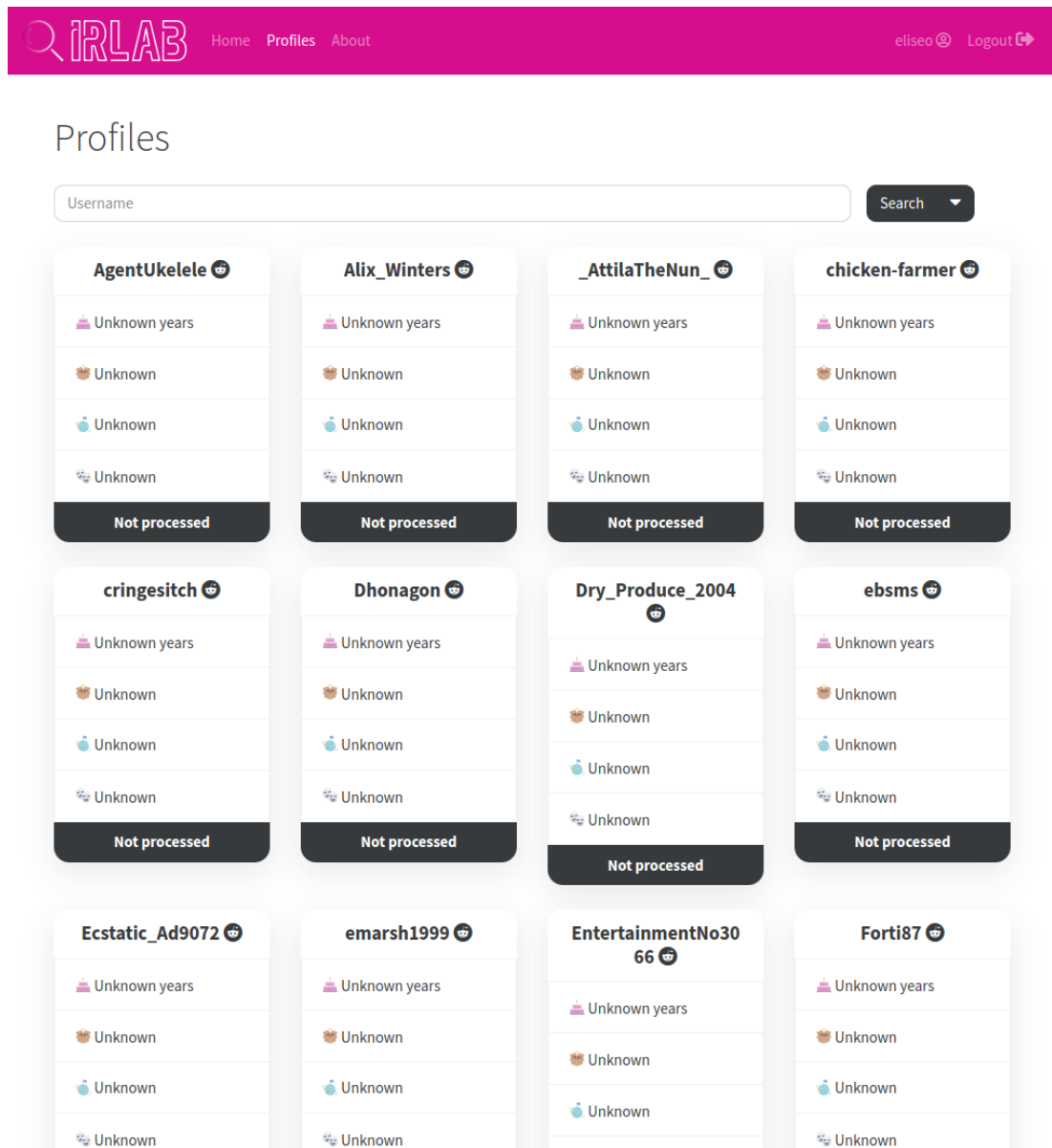


Figura 7.7: Vista da pantalla que amosa os perfís (suxeitos) cos que se está a traballar.

Unha vez chegados a este punto, entra en xogo o *chatbot* creado neste TFM. Escolleuse como exemplo a conversa representada na Figura 7.8 por ilustrar esta moi ben as capacidades do sistema. No primeiro intercambio de mensaxes, MindWell explica as súas habilidades, buscando introducir a perfís con non demasiada experiencia nas funcionalidades básicas agradables do sistema. Así pois, ante a pregunta de “*Que podes facer?*”, o *chatbot* responde que “*Podo analizar os comentarios do suxeito para determinar se presenta algún signo de síntomas depresivos baseándome na lista de síntomas proporcionada. Xustificarei as miñas decisións mediante extractos dos comentarios do suxeito. Se non sei a resposta, direi sinceramente que non o sei..*” Pos-

teriormente, o profesional podería preguntar “*As mensaxes do suxeito revelan algún sintoma de depresión?*”, e o sistema respondería algo similar a “*Tras analizar os comentarios proporcionados, podo identificar os seguintes síntomas de depresión: 1. Pesimismo: Fágoo ás veces, non moi a miúdo, porque necesito coller o tren para chegar ao carreiro de montaña máis próximo. Non creo que iso cambie o meu problema principal, que creo que é a sensación de que non pertenzo a ningún sitio; 3. Perda de pracer: Encantaríame mudarme a un lugar onde puidese atopar a a miña xente, a a miña comunidade, pero non se se existe un lugar así; 4. Culpa: Estou seguro de que hai moitos como eu, inadaptados atrasados. É importante sinalar que os comentarios do suxeito poden non indicar a presenza de todos os síntomas, senón máis ben unha combinación deles. Ademais, é posible que o suxeito estea a experimentar síntomas adicionais que non figuran na lista proporcionada.* Para continuar, ante a pregunta de cando o suxeito fixo un determinado comentario, o *chatbot* ten a capacidade de proporcionar unha ligazón ao *post* no que tivo lugar. O sistema tamén pode proporcionar información sobre a data de publicación do comentario. Por último, no exemplo da Figura 7.8, o *chatbot* ten a capacidade de, neste caso, inferir a idade aproximada do suxeito en base aos seus comentarios.

7.2 Software liberado

- ❖ Código para o *fine-tuning* do modelo:

📄 <https://gitlab.irlab.org/tfm-eliseo/mindwell-finetuning>.

- ❖ LLM resultante do proceso de *fine-tuning*:

📄 <https://huggingface.co/irlab-udc/MindWell>.

- ❖ Código para orquestrar Elasticsearch, Langchain e o LLM:

📄 <https://gitlab.irlab.org/tfm-eliseo/mindwell-api>.

- ❖ Código coa aplicación incrementada:

📄 <https://gitlab.irlab.org/tfm-eliseo/mindwell>.

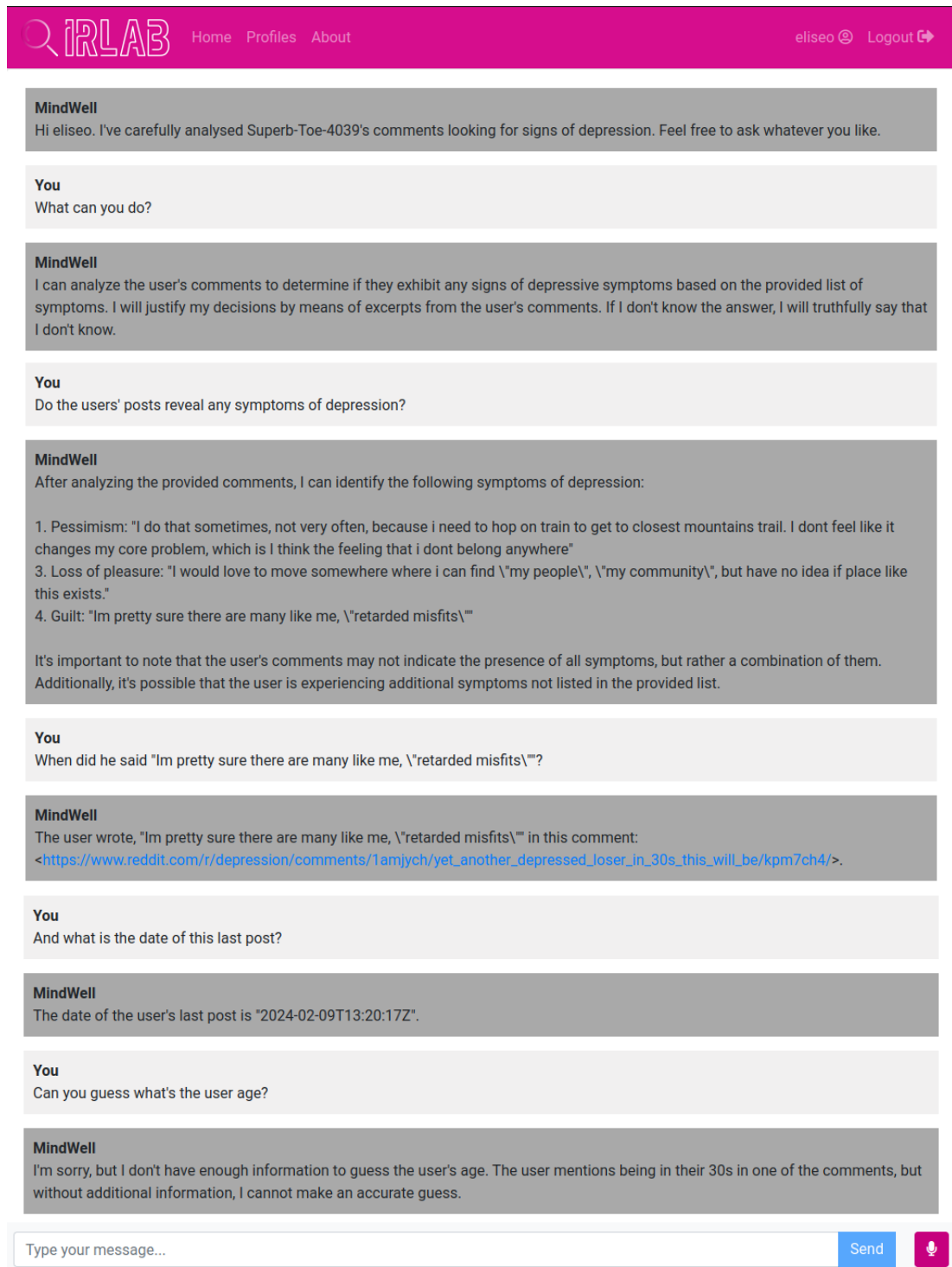


Figura 7.8: Vista do *chat* com MindWell sobre un perfil (sucesso) determinado.

Conclusións, relación co *Master* e traballo futuro

ESTE Capítulo ofrece unha Sección de conclusións sobre o traballo realizado, a 8.1, unha Sección na que se detalla a relación do proxecto co *Master* dentro do que está encadrado, a 8.2, e unha Sección na que se fixan directrices para mellorar e incrementar o proxecto de cara ao futuro, a 8.3.

8.1 Conclusións

Para reflexionar sobre o desenvolvemento do proxecto e poder plasmar as conclusións sobre o mesmo, resulta interesante e útil revisar o éxito ou o fracaso en canto á consecución dos obxectivos fixados. Estes obxectivos poden estar relacionados co resultado do proxecto ou co aspecto formativo que debe ter este traballo, polo que a continuación se detalla a análise dos obxectivos atendendo a esta categorización.

Resultado do proxecto

- Obtívose con éxito un conxunto de datos co que poder levar a cabo o *fine-tuning* do LLM Vicuna-7B-1.5, obtendo como resultado un LLM capaz de detectar e explicar síntomas de depresión en conxuntos masivos de texto.
- Implementouse a lóxica necesaria para manter índices densos onde se representa de xeito vectorial todas as publicacións dos suxeitos analizados xunto coa súa equivalencia a texto e metadatos como a data de publicación. Isto ten dobre valor, xa que proporciona tanto unha parte central do sistema desenvolvido coma un recurso a explotar no futuro.
- Desenvolveuse tanto unha API que permite consumir o modelo de detección e

explicación de síntomas a través dun servizo **REST** coma unha interface *web* moderna para que as persoas usuarias finais poidan interactuar co sistema.

Aspecto formativo

- Principalmente nas fases iniciais, este proxecto supuxo unha revisión da literatura relacionada e o estudo dos fundamentos teóricos e as tecnoloxías escollidas. Todo isto impacta na formación do autor de dúas maneiras: Por unha parte, asimílanse conceptos xa estudados con anterioridade. Complementariamente, o autor acada como resultado do proxecto formación nun eido previamente descoñecido para el.
- O uso dunha metodoloxía, nomeadamente unha versión adaptada de Scrum, implicou responsabilidades a asumir polo autor en tanto que foi el o responsable das tarefas de análise e planificación. Isto permitiulle seguir aprendendo máis sobre este marco metodolóxico e perfeccionar a súa aplicación práctica.
- Como xa se comentou con anterioridade, o traballo consiste nun proxecto de innovación, o que supón unha serie de aprendizaxes. A destacar, este traballo conleva o manexo dunha serie de tecnoloxías de recente aparición, o que en ocasións pode implicar un esforzo extra. Por outra banda, a propia natureza do proxecto promove a apertura a novas ideas e leva asociada a tolerancia aos riscos, os cambios e o posible fracaso. Tamén é un resultado de aprendizaxe a capacidade para realizar lectura académica e científica.

8.2 Relación co *Master*

O traballo levado a cabo durante a realización deste *Traballo de Fin de Master* (TFM) garda relación tanto con diferentes asignaturas do Mestrado Universitario en Enxeñaría Informática (8.2.1) como con múltiples das competencias que garante o título (ver 8.2.2).

8.2.1 Relación con asignaturas

Das materias da titulación¹ coas que está estreitamente relacionada este proxecto, destácanse as seguintes:

614502002 - Dirección de Proxectos

A materia de Dirección de Proxectos presenta un marco práctico para o desenvolvemento profesional na dirección de proxectos de propósito xeral, o cal está estreitamente relacionado co desempeño de labores de manexo do proxecto asumidas polo alumno

¹https://guiadocente.udc.es/guia_docent/index.php?centre=614&ensenyament=614502&consulta=assignatures

durante o desenvolvemento do seu TFM. A xustificación radica na necesidade e a capacidade para poder reaccionar ante os problemas que tiveron lugar e os axustes que foron necesarios.

614502007 - Deseño de Sistemas de Información

A materia revisa e consolida conceptos relacionados co deseño *software*, incluíndo patróns de deseño e arquitectura, deseño orientado a compoñentes, calidade no deseño ou accesibilidade. Todos estes aspectos son aplicados en maior ou menor medida ao longo deste TFM para desenvolver unha aplicación construída con compoñentes (*templates*) baixo un patrón *Model-View-Template* (MVT) sumado ao cumprimento dos principios SOLID.

614502006 - Análise de Sistemas de Información

No marco desta asignatura introdúcense modelos de análise de requisitos que permiten identificar necesidades de cliente e casos de uso, especificar os requisitos, analízalos, documentalos e validalos. Todos estes conceptos resultan imprescindibles para a fase de análise do proxecto, onde é necesario comprender como debe ser a solución a desenvolver, que situacións se poden dar e cales son os casos de uso esperados polo cliente.

614502004 - Informática como Servizo

De entre a variedade de contidos tratados ao longo da materia, é especialmente útil para este TFM o manexo de Docker, tecnoloxía fundamental do empaquetado e despregamento da aplicación. Así pois, ser coñecedor das boas prácticas e ter un contexto previo coa elaboración de configuracións resultou de gran axuda.

614502010 - Recuperación da Información e Web Semántica

A materia introduce modelos, técnicas e algoritmos de recuperación de información que están directamente conectados coa solución desenvolvida en tanto que son a base funcional do sistema. A recolección de información en *Internet*, o almacenamento da mesma en índices, a busca dentro destes e os *embeddings* son conceptos estudados no marco da asignatura que foron despois utilizados todos eles neste Proxecto de Fin de Master.

8.2.2 Relación con competencias do título

Das competencias que garante a titulación² coas que está estreitamente relacionada este TFM, destácanse 3 de cada tipo (A, B e C) ordenadas por código identificador:

² https://guiadocente.udc.es/guia_docent/index.php?centre=614&ensenyament=614502&consulta=competencies

AP3 - Capacidade para a dirección de proxectos de investigación, desenvolvemento e innovación en empresas e centros tecnolóxicos, con garantía da seguridade para as persoas e bens, a calidade final dos produtos e a súa homologación

O alumno adopta durante o transcurso do TFM labores de dirección de proxecto, considerándose este un proxecto de innovación ao integrar e explotar algunhas tecnoloxías coma os LLMs ou LangChain (ver Capítulo 3) en pleno proceso de crecemento e auxe.

AP4 - Capacidade para modelar, deseñar, definir a arquitectura, implantar, xestionar, operar, administrar e manter aplicacións, redes, sistemas, servizos e contidos informáticos

O alumno é responsable da solución durante todo o seu ciclo de vida completo, é dicir, desde a fase de modelado ata o mantemento pasando polo deseño, a implementación e a implantación.

AP8 - Capacidade para analizar as necesidades de información que se presentan nun contorno e levar a cabo en todas as súas etapas o proceso de construción dun sistema de información

En sucesivas conversas cos seus directores (que xogan o rol de *Product Owners*), o alumno debe ser quen de levar a cabo unha elicitación básica dos requisitos da aplicación, identificando tamén os casos de uso e proponendo solución para os mesmos.

BP4 - Capacidade para organizar e planificar

O alumno é responsable de organizar o seu propio traballo e planificar as tarefas e a súa descomposición, de xeito que os obxectivos establecidos poidan ser completados de maneira exitosa.

BP6 - Toma de decisións

Ante as eventualidades que poidan xurdir durante o transcurso do proxecto, o alumno debe tomar a responsabilidade de decidir como atallalas e corrixir os problemas e as desviacións ocasionadas. Tamén se adquire esta competencia durante o proceso de selección de tecnoloxías e ferramentas.

BP15 - Capacidade para a dirección xeral, dirección técnica e dirección de proxectos de investigación, desenvolvemento e innovación, en empresas e centros tecnolóxicos, no ámbito da Enxeñaría Informática

As responsabilidades de dirección anteriormente comentadas están encadradas dentro dun proxecto de innovación levado a cabo nun centro tecnolóxico.

CP2 - Dominar a expresión e a comprensión de forma oral e escrita dun idioma estranxeiro

O proxecto, tanto a nivel de interface como a nivel de código, desenvólvese en inglés. Así mesmo, toda a documentación requirida, como a literatura consultada, coma os conxuntos de datos están escritos neste mesmo idioma.

CP3 - Utilizar as ferramentas básicas das tecnoloxías da información e as comunicacións (TIC) necesarias para o exercicio da súa profesión e para a aprendizaxe ao longo da súa vida

Todas as tecnoloxías e ferramentas descritas no Capítulo 3 son de máxima vixencia e fundamental manexo na actualidade para exercer a profesión.

CP8 - Valorar a importancia que ten a investigación, a innovación e o desenvolvemento tecnolóxico no avance socioeconómico e cultural da sociedade

Froito da súa inmersión dentro dun proxecto de innovación no marco dun centro tecnolóxico, o alumno comprende e reforza a súa concepción sobre a importancia do I+D+i para o global da sociedade e o seu progreso.

8.3 Traballo futuro

As posibilidades e ideas de cara á iteración, ampliación e mellora deste proxecto son multitude. Por unha parte, sería interesante tratar de adaptar os modelos resultantes a outros dominios de aplicación e incluso idiomas, sendo este un punto crítico para o seu rendemento. Neste sentido, tratar outras redes sociais como Twitter, Facebook ou Instagram é o paso inmediato. Así pois, a capacidade de xeneralizar é aínda a día de hoxe un desafío. Por outra banda, sería desaxable poder dar soporte tamén a outros manuais de psicoloxía ademais de *Beck Depression Inventory, Second Version* (BDI-II), como poder ser o caso de *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5). Levando máis alá esta idea, incluso se podería pensar en estender o traballo a outros trastornos relacionados/comórbidos á depresión, nomeadamente as alteracións na alimentación, a ansiedade, a bipolaridade, etc. Moitos dos síntomas son comúns entre estas enfermidades e a depresión, sendo tamén similar o diagnóstico na práctica clínica. Para finalizar, destacar tamén que existe interese en desenvolver un marco metodolóxico que mellore a avaliación experta, proporcionando unha análise da aplicación máis profunda e técnica.

Compromiso ético

Este último Capítulo da memoria serve como peche da mesma, e nel as Seccións 9.1, 9.2 e 9.3 reflexionan a supervisión do traballo por parte dun perfil clínico experto, o cumprimento dos acordos de uso dos conxuntos de datos e as emisións de CO₂ relacionadas co proxecto, respectivamente.

9.1 Supervisión dun perfil clínico experto

Calquera desenvolvemento deste tipo debe estar supeditado á validación das persoas usuarias finais, que neste caso son profesionais da psicoloxía [62]. É importante ter sempre en mente que estes traballos buscan ser un apoio, e nunca algo substitutorio. É por isto que o seu *feedback* resulta especialmente enriquecedor, sobretudo tendo en conta que o obxectivo é que a ferramenta resultante sexa útil no seu traballo diario. A este efecto, dispónse dun profesional da psicoloxía a través do laboratorio de investigación no que se desenvolve este proxecto, xa que se contan con recursos deste tipo no marco de proxectos coma o PLEC2021-007662 (Big-eRisk: Early prediction of personal risks on massive data)¹.

9.2 Cumprimento dos acordos de uso dos conxuntos de datos

Os datos utilizados neste traballo obtivéronse de fontes de acceso público, respectando o *status* de exención en virtude do título 45 CFR §46.104². O uso de BDI-Sen e PsySym levouse a cabo en total conformidade coas súas respectivas políticas de uso. Para manter a privacidade, aplícanse medidas para garantir que non se poida identificar ningunha información persoal e que todos os usuarios permanezan anónimos. Os datos utilizados proceden de Reddit, e

¹<https://www.aei.gob.es/convocatorias/buscador-convocatorias/proyectos-colaboracion-publico-privada-2021>

²<https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104>

axústanse estritamente a todos os termos especificados por esta plataforma. É importante destacar que o sistema proposto pretende complementar o traballo dos profesionais sanitarios, non substituílos. O desenvolvemento deste tipo de tecnoloxías require un enfoque prudente, poñendo énfase no seu uso ético e mantendo o respecto pola privacidade e a autonomía dos pacientes.

9.3 Emisións de CO₂ relacionadas co proxecto

Este traballo realizouse nunha infraestrutura privada cunha eficiencia de carbono estimada de 0,432 Kg CO₂/KWh (media anual da OCDE de 2014). Os recursos utilizados incluíron 20 horas de cómputo nun dispositivo RTX 6000 Ada Generation (cun TDP de 300 W). As emisións totais estímase en 2,59 Kg CO₂. Para poñer a cifra en perspectiva, isto equivale a conducir un coche promedio durante 10 quilómetros. Estas cifras determináronse coa axuda da calculadora MachineLearning Impact Calculator [63].

Apéndices

Licenzas *software*

<i>Software</i>	Licenza
Vicuna 7B 1.5	Llama2
Python	PSF 2.2
Transformers	Apache License 2.0
DeepSpeed	Apache License 2.0
FastChat	Apache License 2.0
FastAPI	MIT License
<i>Elasticsearch</i>	<i>Server Side Public License</i>
LangChain	MIT License
Django	3-Clause BSD
VSCodim	MIT License
Pip	MIT License
Docker	Apache License 2.0
Taiga	Mozilla Public License
GitLab	MIT License

Táboa A.1: Relación entre os distintos *softwares* empregados e as súas licenzas.

Questionario BDI-II

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully. And then pick out the one statement in each group that best describes the way you have been feeling during the past two weeks, including today. Circle the number beside the statement you have picked. If several statements in the group seem to apply equally well, circle the highest number for that group. Be sure that you do not choose more than one statement for any group, including Item 16 (Changes in Sleeping Pattern) or Item 18 (Changes in Appetite).

1. Sadness

- 0. I do not feel sad.
- 1. I feel sad much of the time.
- 2. I am sad all the time.
- 3. I am so sad or unhappy that I can't stand it.

2. Pessimism

- 0. I am not discouraged about my future.
- 1. I feel more discouraged about my future than I used to.
- 2. I do not expect things to work out for me.
- 3. I feel my future is hopeless and will only get worse.

3. Past Failure

- 0. I do not feel like a failure.
- 1. I have failed more than I should have.
- 2. As I look back, I see a lot of failures.

3. I feel I am a total failure as a person.

4. Loss of Pleasure

0. I get as much pleasure as I ever did from the things I enjoy.

1. I don't enjoy things as much as I used to.

2. I get very little pleasure from the things I used to enjoy.

3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings

0. I don't feel particularly guilty.

1. I feel guilty over many things I have done or should have done.

2. I feel quite guilty most of the time.

3. I feel guilty all of the time.

6. Punishment Feelings

0. I don't feel I am being punished.

1. I feel I may be punished.

2. I expect to be punished.

3. I feel I am being punished.

7. Self-Dislike

0. I feel the same about myself as ever.

1. I have lost confidence in myself.

2. I am disappointed in myself.

3. I dislike myself.

8. Self-Criticalness

0. I don't criticize or blame myself more than usual.

1. I am more critical of myself than I used to be.

2. I criticize myself for all of my faults.

3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes

0. I don't have any thoughts of killing myself.

1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying

0. I don't cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can't.

11. Agitation

0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated, it's hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest

0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It's hard to get interested in anything.

13. Indecisiveness

0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness

0. I do not feel I am worthless.
1. I don't consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to others.
3. I feel utterly worthless.

15. Loss of Energy

- 0. I have as much energy as ever.
- 1. I have less energy than I used to have.
- 2. I don't have enough energy to do very much.
- 3. I don't have enough energy to do anything.

16. Changes in Sleeping Pattern

- 0. I have not experienced any change in my sleeping.
- 1a. I sleep somewhat more than usual.
- 1b. I sleep somewhat less than usual.
- 2a. I sleep a lot more than usual.
- 2b. I sleep a lot less than usual.
- 3a. I sleep most of the day.
- 3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability

- 0. I am not more irritable than usual.
- 1. I am more irritable than usual.
- 2. I am much more irritable than usual.
- 3. I am irritable all the time.

18. Changes in Appetite

- 0. I have not experienced any change in my appetite.
- 1a. My appetite is somewhat less than usual.
- 1b. My appetite is somewhat greater than usual.
- 2a. My appetite is much less than before.
- 2b. My appetite is much greater than usual.
- 3a. I have no appetite at all.
- 3b. I crave food all the time.

19. Concentration Difficulty

- 0. I can concentrate as well as ever.

1. I can't concentrate as well as usual.
2. It's hard to keep my mind on anything for very long.
3. I find I can't concentrate on anything.

20. Tiredness or Fatigue

0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex

0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely.

Total Score: _____

THE PSYCHOLOGICAL CORPORATION

Harcourt Brace & Company

Copyright 1996, by Aaron T. Beck. All rights reserved.

Equivalencia entre DSM-5 e BDI-II

DSM-5	BDI-II
1. Depressed mood	1. Sadness 10. Crying 17. Irritability
2. Diminished interest or pleasure	4. Loss of pleasure 12. Loss of interest 21. Loss of interest in sex
3. Decrease or increase in appetite	18. Changes in appetite
4. Insomnia or hypersomnia	16. Changes in sleeping pattern
5. Psychomotor agitation or retardation	11. Agitation
6. Fatigue or loss of energy	15. Loss of energy 20. Tiredness or fatigue
7. Feeling of worthlessness	2. Pessimism 3. Past failure 5. Guilty feelings 6. Punishment feelings 7. Self-dislike 8. Self-criticalness 14. Worthlessness
8. Diminished ability to think	13. Indecisiveness 19. Concentration difficulty
9. Recurrent thoughts of death, suicidal ideation	9. Suicidal thoughts or wishes

Táboa C.1: Equivalencia entre DSM-5 e BDI-II.

Equivalencia entre *PsySym* e BDI-II

<i>PsySym</i>	BDI-II
1. Depressed mood	1. Sadness 10. Crying 4. Loss of pleasure
2. Sleep disturbance	16. Changes in sleeping patter
3. Anger or irritability	17. Irritability
4. Decreased energy, tiredness or fatigue	15. Loss of energy 20. Tiredness or fatigue
5. Genitourinary symptoms	21. Loss of interest in sex
6. Hyperactivity or agitation	11. Agitation
7. Inattention	19. Concentration difficulty
8. Indecisiveness	13. Indecisiveness
9. Suicidal ideas	9. Suicidal thoughts or wishes
10. Worthlessness and guilty	14. Worthlessness 5. Guilty feelings 7. Self-dislike 3. Past failure 6. Punishment feelings 8. Self-criticalness
11. Loss of interest or motivation	12. Loss of interest
12. Pessimism	2. Pessimism
13. Poor memory	19. Concentration difficulty
14. Weight and appetite change	18. Changes in appetite

Táboa D.1: Equivalencia entre *PsySym* e BDI-II.

Relación de Acrónimos

- AJAX** *Asynchronous JavaScript and XML.* 16
- API** *Application Programming Interface.* 3, 17, 34, 35, 38, 46, 53, 55, 60, 71
- ASR** *Automatic Speech Recognition.* 11, 16
- BDI-II** *Beck Depression Inventory, Second Version.* VII, 2, 4, 36, 37, 75, 89, 91
- CSS** *Cascading Style Sheets.* 16
- CV** *Computer Vision.* 16
- DSM-5** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.* VII, 4, 36, 37, 75, 89
- GPL** *General Public License.* 18, 96
- GPU** *Graphics Processing Unit.* V, 29, 30, 41
- GUI** *Graphical User Interface.* 46, 60
- HTML** *HyperText Markup Language.* 16, 33
- IA** *Inteligencia Artificial.* V, 10–12, 48
- IR** *Information Retrieval.* 1, 36
- JSON** *JavaScript Object Notation.* 17, 18
- LIME** *Local Interpretable Model-Agnostic Explanations.* 12
- LLM** *Large Language Model.* 2–4, 7–13, 17, 18, 35, 40, 43, 44, 46, 48, 50, 51, 53, 55, 57, 60, 61, 68, 71, 74

LoRA *Low-Rank Adaptation of Large Language Models*. 3, 39, 40

LSTM *Long Short-Term Memory*. 8

MIT Massachusetts Institute of Technology. 17–19, 96

ML *Machine Learning*. 1

MVT *Model-View-Template*. 73

NLP *Natural Language Processing*. 1, 8, 16, 36

NLU *Natural Language Understanding*. 11

OMS Organización Mundial da Saúde. 1

OSI Open Source Initiative. 7

PyPI Python Package Index. 19

RAG *Retrieval Augmented Generation*. V, 7, 13, 48

RAM Random Access Memory. 29

REST *Representational State Transfer*. 17, 18, 35, 53, 55, 72

RNN *Recurrent Neural Network*. 8

SCM Source Code Management. 20

SHAP *Shapley Additive exPlanations*. 12

SOTA *State-of-the-Art*. 4, 9, 46, 49, 60

SSPL Server Side Public License. 18

TFM Trabajo de Fin de *Master*. V, 1, 2, 7, 10, 23, 27, 33, 34, 63, 67, 72–74

TTS *Text to Speech*. 11

Glosario

branching No contexto do control de versións, replicación de obxectos para que se poida traballar sobre estes de forma separada e en paralelo. 21

chatbot *software* de intelixencia artificial deseñado para interactuar a través de conversacións automáticas. Utiliza algoritmos e procesamento da linguaxe natural para comprender preguntas e proporcionar respostas relevantes. O seu propósito abarca desde a asistencia en diferentes contextos ata o mero entretemento. 1–3, 10, 17, 18, 33, 44, 45, 63, 67, 68

deep learning Subconxunto do *machine learning* que implica redes neuronais con múltiples capas. Aproveita grandes conxuntos de datos para aprender automaticamente e extraer patróns complexos, o que permite aos modelos tomar decisións ou realizar predicións complexas.. 8

fine-tuning Técnica de *machine learning* na que un modelo *pre-trained* se axusta mediante a actualización duns poucos parámetros para unha tarefa ou conxunto de datos específicos de cara a mellorar o seu rendemento, aproveitando así os coñecementos adquiridos no seu adestramento inicial nun conxunto de datos máis amplo. V, VII, 2, 9, 11, 29, 39–41, 43, 44, 46, 49, 50, 60, 61, 68, 71, 96

kernel Parte fundamental e básica de calquera sistema operativo. 20

machine learning Campo da intelixencia artificial que se centra no desenvolvemento de algoritmos e modelos que procuran aprender patróns a partir de datos e facer predicións ou tomar decisións sen programación explícita. 95, 96

open source *software* ou proxectos cuxo código fonte está dispoñible libremente, permitindo aos usuarios velo, modificalo e redistribuílo. O open-source fomenta a colaboración,

a transparencia e a participación da comunidade, e promove un modelo de desenvolvemento descentralizado. As licenzas comúns, como a [GPL](#) ou a [MIT](#), rexen as condicións de uso, garantindo unha apertura e flexibilidade constantes. [1](#), [2](#), [7](#), [8](#), [17–19](#), [53](#)

pre-trained Modelo de *machine learning* que se adestrou previamente nun gran conxunto de datos para unha tarefa xeral. Este modelo, que aprendeu características e patróns valiosos, pode especializarse ou adaptarse a tarefas específicas mediante técnicas de *fine-tuning*, o que aforra tempo e recursos en comparación co adestramento desde cero. [2](#), [95](#)

software Sistema formal dun sistema informático, que comprende o conxunto das compoñentes lóxicas necesarias para a realización de tarefas específicas, tipicamente en contraposición ás componentes físicas. [VII](#), [1](#), [4](#), [7](#), [8](#), [19](#), [20](#), [23](#), [24](#), [29–31](#), [33](#), [63](#), [81](#), [95](#)

token Unidade de texto que un modelo procesa durante o seu adestramento ou en inferencia cando xenera respostas. Os *tokens* poden ser tan cortos coma un carácter ou tan longos coma unha palabra, dependendo da granularidade elexida para a representación do texto. [VII](#), [37](#), [40](#)

Bibliografía

- [1] T. N. S. of Mental Health and Wellbeing, “Mental health: prevalence and impact,” Australian Institute of Health and Welfare, Tech. Rep., 2022.
- [2] W. H. Organization *et al.*, “Depression and other common mental disorders: global health estimates,” World Health Organization, Tech. Rep., 2017.
- [3] A. Picardi, I. Lega, L. Tarsitani, M. Caredda, G. Matteucci, M. P. Zerella, R. Miglio, A. Gigantesco, M. Cerbo, A. Gaddini, F. Spandonaro, M. Biondi, and SET-DEP Group, “A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care,” *J Affect Disord*, vol. 198, pp. 96–101, Mar. 2016.
- [4] A. Gulliver, K. M. Griffiths, and H. Christensen, “Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review,” *BMC Psychiatry*, vol. 10, p. 113, Dec. 2010.
- [5] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Proceedings of the international AAAI conference on web and social media*, vol. 7, no. 1, 2013, pp. 128–137.
- [6] F. Sadeque, D. Xu, and S. Bethard, “Measuring the latency of depression detection in social media,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ser. WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 495–503.
- [7] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2968–2978.
- [8] F. CACHEDA, D. Fernandez, F. J. Novoa, and V. Carneiro, “Early detection of depression: social network analysis and random forest techniques,” *Journal of medical Internet research*, vol. 21, no. 6, p. e12554, 2019.

-
- [9] A. Beheshti, V. Moraveji-Hashemi, S. Yakhchi, H. R. Motahari-Nezhad, S. M. Ghafari, and J. Yang, “Personality2vec: Enabling the analysis of behavioral disorders in social networks,” in *Proceedings of the 13th International Conference on Web Search and Data Mining*, ser. WSDM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 825–828.
- [10] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, “It’s just a matter of time: Detecting depression with time-enriched multimodal transformers,” in *European Conference on Information Retrieval*. Springer, 2023, pp. 200–215.
- [11] M. Aragon, A. P. Lopez Monroy, L. Gonzalez, D. E. Losada, and M. Montes, “DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 15 305–15 318.
- [12] A. Callahan and K. Inckle, “Cybertherapy or psychobabble? a mixed methods study of online emotional support,” *British Journal of Guidance & Counselling*, vol. 40, no. 3, pp. 261–278, 2012.
- [13] E. A. Rissola, D. E. Losada, and F. Crestani, “A survey of computational methods for online mental state assessment on social media,” *ACM Transactions on Computing for Healthcare*, vol. 2, no. 2, pp. 1–31, 2021.
- [14] C. G. Walsh, B. Chaudhry, P. Dua, K. W. Goodman, B. Kaplan, R. Kavuluru, A. Solomnides, and V. Subbian, “Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence,” *JAMIA Open*, vol. 3, no. 1, pp. 9–15, 01 2020.
- [15] T. U. Hauser, V. Skvortsova, M. De Choudhury, and N. Koutsouleris, “The promise of a model-based psychiatry: building computational models of mental ill health,” *The Lancet Digital Health*, vol. 4, no. 11, pp. e816–e828, 2022.
- [16] A. Pérez, J. Parapar, and Álvaro Barreiro, “Automatic depression score estimation with word embedding models,” *Artificial Intelligence in Medicine*, vol. 132, p. 102380, 2022.
- [17] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, “Improving the generalizability of depression detection by leveraging clinical questionnaires,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8446–8459.

- [18] Z. Zhang, S. Chen, M. Wu, and K. Zhu, “Symptom identification for interpretable detection of multiple mental disorders on social media,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9970–9985.
- [19] Z. Zhang, S. Chen, M. Wu, and K. Q. Zhu, “Psychiatric scale guided risky post screening for early detection of depression,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 5220–5226, aI for Good.
- [20] A. Pérez, N. Warikoo, K. Wang, J. Parapar, and I. Gurevych, “Semantic similarity models for depression severity estimation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16 104–16 118.
- [21] A. T. Beck, R. A. Steer, and G. Brown, “Beck depression inventory–II,” 1996.
- [22] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” March 2023.
- [23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esio-bu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023.
- [24] OpenAI, “Introducing chatgpt,” November 2022.
- [25] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Proc. Interspeech 2010*, 2010, pp. 1045–1048.
- [26] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.
- [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.

-
- [28] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [30] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*, ser. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.
- [31] O. Vinyals and Q. Le, “A neural conversational model,” 2015, cite arxiv:1506.05869Comment: ICML Deep Learning Workshop 2015.
- [32] R. Guidotti and S. Ruggieri, “On the stability of interpretable models,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.
- [33] M. Tulio Ribeiro, S. Singh, and C. Guestrin, “Model-Agnostic Interpretability of Machine Learning,” *arXiv e-prints*, p. arXiv:1606.05386, Jun. 2016.
- [34] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.
- [35] M. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, J. DeNero, M. Finlayson, and S. Reddy, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101.
- [36] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [37] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, “How to explain individual classification decisions,” *J. Mach. Learn. Res.*, vol. 11, p. 1803–1831, aug 2010.
- [38] B. Ustun and C. Rudin, “Supersparse linear integer models for optimized medical scoring systems,” *Machine Learning*, vol. 102, no. 3, pp. 349–391, Mar 2016.

- [39] H. Lakkaraju, S. H. Bach, and J. Leskovec, “Interpretable decision sets: A joint framework for description and prediction,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1675–1684.
- [40] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.
- [41] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, “Explainable prediction of medical codes from clinical text,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1101–1111.
- [42] H. Song, J. You, J.-W. Chung, and J. C. Park, “Feature attention network: Interpretable depression detection from social media,” in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics, 1–3 Dec. 2018.
- [43] H. Amini and L. Kosseim, “Towards explainability in using deep learning for the detection of anorexia in social media,” in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, H. Horacek, and P. Cimiano, Eds. Cham: Springer International Publishing, 2020, pp. 225–235.
- [44] S. Wiegrefe and Y. Pinter, “Attention is not not explanation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20.
- [45] M. Tutek and J. Šnajder, “Toward practical usage of the attention mechanism as a tool for interpretability,” *IEEE Access*, vol. 10, pp. 47 011–47 030, 2022.
- [46] S. Jain and B. C. Wallace, “Attention is not Explanation,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3543–3556.
- [47] S. Serrano and N. A. Smith, “Is attention interpretable?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951.

- [48] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, “e-snli: Natural language inference with natural language explanations,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.
- [49] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.
- [50] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, mar 2023.
- [51] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [52] “index | tiobe - the software quality company,” 2023, consultado o 2024-02-18. [En línea]. Disponible en: <https://www.tiobe.com/tiobe-index/>
- [53] “Why Use Python for AI and Machine Learning? — globaldev.tech,” 2023, consultado o 2024-02-18. [En línea]. Disponible en: <https://globaldev.tech/blog/python-for-ai-and-machine-learning>
- [54] “Stack overflow developer survey 2023,” 2023, consultado o 2024-02-18. [En línea]. Disponible en: <https://survey.stackoverflow.co/2023/#section-most-popular-technologies-integrated-development-environment>
- [55] J. Sutherland and K. Schwaber, “Scrum guide,” consultado o 2024-02-18. [En línea]. Disponible en: <https://scrumguides.org/scrum-guide.html>
- [56] “Salary guide 2023 spain - technology,” consultado o 2024-02-18. [En línea]. Disponible en: https://www.michaelpage.es/sites/michaelpage.es/files/protected-documents/2023-05/ER%20Tecnolog%C3%ADa%202023_EN.pdf
- [57] G. Jackson-Koku, “Beck depression inventory,” *Occupational Medicine*, vol. 66, no. 2, pp. 174–175, 2016.

- [58] F. Crestani, D. E. Losada, and J. Parapar, *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*. Springer Nature, 2022, vol. 1018.
- [59] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, “erisk 2023: Depression, pathological gambling, and eating disorder challenges,” in *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer, 2023, pp. 585–592.
- [60] A. Perez, J. Parapar, A. Barreiro, and S. Lopez-Larrosa, “Bdi-sen: A sentence dataset for clinical symptoms of depression,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023*.
- [61] C. C. Nuckols and C. C. Nuckols, “The diagnostic and statistical manual of mental disorders,(dsm-5),” *Philadelphia: American Psychiatric Association*, 2013.
- [62] E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim, “Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023.
- [63] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, “Quantifying the carbon emissions of machine learning,” *arXiv preprint arXiv:1910.09700*, 2019.