

Università Politecnica delle Marche
Facoltà di Ingegneria
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e
dell'Automazione



Sviluppo di un chatbot personalizzato
per il settore turistico basato su Large
Language Models

Development of a custom chatbot based on Large
Language Models for the tourism sector

Relatrice: Prof.ssa Claudia Diamantini

Correlatore: Prof. Alex Mircoli

Laureanda: Laura Ferretti

Anno Accademico 2022-2023

A mio papà Luigino

Ringraziamenti

Nelle successive righe vorrei ringraziare tutte quelle persone che mi hanno accompagnato in questi tre anni e che hanno contribuito al raggiungimento di questo traguardo.

Prima di tutto vorrei ringraziare mio papà Luigino, colui che mi ha fatto scoprire la bellezza della scienza e della logica, colui che mi ha mostrato cos'è il vero coraggio e la tenacia di fronte ad ogni difficoltà.

Un grazie speciale a mia mamma Annagrazia, mia nonna Silvana, mia sorella Ilaria e mio cognato Edoardo, per il sostegno, la fiducia ma soprattutto per l'amore che non mi hanno fatto mai mancare.

Grazie a Edoardo, Walter, Giansimone, Alessandra, Luca, Valeria e Alessio, per tutte le ore passate insieme, per tutte le chiacchierate, per tutte le risate e per il costante supporto, siete stati la più bella scoperta dell'università.

Grazie a Ester, Annalaura, Beatrice, Sara, Mariachiara, Nina e Margherita per aver portato spensieratezza e vitalità nella mia vita.

Grazie alle ragazze del mio gruppetto per esserci sempre, in ogni tappa fondamentale.

Grazie al professore Alex Mircoli e alla professoressa Claudia Diamantini per avermi accompagnato durante tutta la fase di sviluppo del progetto e poi di stesura della tesi, dimostrando sempre grande disponibilità.

Ed infine, vorrei ringraziare di cuore Luca per la sua costante presenza, per l'amore, per la gentilezza, per la fiducia, per la condivisione, per la poesia e la musica che ha portato nella mia vita.

Contents

Acknowledgment	3
1 Introduzione	11
1.1 Le sfide del settore turistico in Italia	11
1.2 Soluzione Proposta	12
1.3 Struttura della tesi	13
2 Obiettivi del Progetto	15
2.1 Azienda "Hello Vacanze"	15
2.2 Obiettivo del progetto	16
3 Strumenti Tecnologici	17
3.1 Large Language Model GPT-3	17
3.1.1 I transformer	17
3.1.2 Trasferimento di conoscenza	18
3.2 GPT Index	19
3.2.1 Data Ingestion	19
3.2.2 Data Indexing	19
3.2.3 Query Interface	20
4 Realizzazione del Chatbot	21
4.1 Metodologia di progetto	21
4.2 Dataset	23
4.3 Data Preprocessing	24
4.3.1 Estrazione dati di interesse	24
4.3.2 Eliminazione dati sensibili	26
4.3.3 Eliminazione messaggi con particolari tipologie	26
4.3.4 Ristrutturazione dataset	27
4.3.5 Creazione file JSONL	27
4.4 Fine Tuning	28
4.5 Traferimento conoscenza di dominio	30

4.6	Interfaccia grafica	32
5	Valutazione	35
5.1	Valutazione Linguistica	35
5.2	Valutazione finale del LLM	37
5.2.1	Confronto pre e post inserimento knowldge base . . .	39
6	Conclusioni e sviluppi futuri	41
6.1	Sviluppi futuri	41
	Bibliography	43

List of Figures

3.1	Step Query Interface	20
4.1	Diagramma di flusso rappresentante le varie fasi di progetto .	22
4.2	Fasi svolgimento Data Preprocessing	24
4.3	Rappresentazione struttura dataset	25
4.4	Struttura file JSONL	27
4.5	Comando per validazione file	29
4.6	Nome modello personalizzato	29
4.7	Esempio file .txt	30
4.8	Estratto da file index.json	31
4.9	Interfaccia Grafica Chatbot	32

List of Tables

4.1 Campi del dataset 23

Chapter 1

Introduzione

In Italia il settore turistico ricopre un ruolo particolarmente centrale per l'economia del paese. In termini numerici, nel 2023, in Italia, si stimano 442 milioni di turisti, con un aumento del 12,2% rispetto all'anno precedente [1]. Il Prodotto Interno Lordo (PIL) derivante dal settore turistico si stima aggirarsi intorno ai 100 miliardi di euro [2], con un numero di occupati (970.724) pari al 7.3% del totale dei lavoratori [3].

L'imponenza di questo settore, evidenziata dai dati numerici sopra riportati, giustifica l'interesse nella ricerca di nuove tecnologie da utilizzare per il miglioramento dei servizi e delle attività fornite.

Il progetto presentato in questo lavoro di tesi, propone un chatbot implementato facendo uso delle più recenti tecnologie informatiche, come i Large Language Models, al fine di risolvere uno dei problemi tipici del settore turistico: l'impatto delle fluttuazioni del numero di turisti su un tour operator che si occupa di assistenza al cliente.

1.1 Le sfide del settore turistico in Italia

Alcune delle principali caratteristiche del turismo sono l'elevata stagionalità e la difficile prevedibilità del numero di turisti.

Ogni anno l'affluenza delle persone nelle principali mete turistiche varia, a causa di fenomeni sociali, economici e sanitari, rendendo molto complesso fare stime attendibili con largo anticipo. Inoltre, durante un anno solare si registrano picchi di affluenza di turisti in periodi specifici, ad esempio durante le festività natalizie ed i mesi estivi.

Tutto ciò rende particolarmente complesso il dimensionamento del numero dei lavoratori necessari al corretto funzionamento del settore.

Un esempio di azienda che risente fortemente di questa problematica è il tour operator.

Il tour operator è uno dei principali attori del mercato turistico, si occupa dell'interfacciamento con clienti al fine di organizzare vacanze, prenotare strutture ricettive e garantire assistenza durante tutto il periodo del soggiorno. Attualmente molte aziende di questa tipologia forniscono i loro servizi principalmente online, attraverso lo scambio di messaggi in chat.

Le componenti chiave per fornire servizi soddisfacenti sono: la minimizzazione dei tempi di attesa, la disponibilità 24/7 e la competenza del personale.

In base alle caratteristiche esposte sopra, risulta chiaro come il corretto dimensionamento del numero dei dipendenti, addetti al servizio di customer care, sia una sfida importante per aziende di questo tipo.

1.2 Soluzione Proposta

La soluzione proposta per risolvere questa particolare problematica consiste nell'implementazione di un chatbot che possa sostituire il personale addetto all'interazione con il cliente tramite chat.

Un chatbot è un sistema che elabora le domande poste da un essere umano e fornisce risposte attingendo ai dati con cui è stato addestrato o alla knowledge base fornitagli.

Un sistema di questo tipo cambia radicalmente il paradigma classico della conversazione che prevede due esseri umani che si scambiano domande e risposte, introducendo un altro attore, un'intelligenza artificiale, che vada a rimpiazzare uno dei due umani.

Nel caso di un servizio di customer care, un sistema di questo tipo potrebbe potenzialmente sostituire gli agenti dell'agenzia di viaggio, andando a risolvere alla radice il problema del loro dimensionamento, portando enormi vantaggi sia in fattori economici ma anche a livello di miglioramento del servizio offerto.

Infatti, andando ad analizzare le caratteristiche chiave di un servizio di customer care esposte sopra, risulta evidente come un chatbot le possa soddisfare a pieno, apportando anche migliorie rispetto a quanto un essere umano possa offrire.

Le precedenti affermazioni, circa la possibile sostituzione integrale di esseri umani in favore di sistemi chatbot, aprono a dibattiti etici e morali circa l'effettiva capacità di una macchina di sostituire l'operato umano. In questo periodo storico, ancora la tecnologia non sembra essere sufficientemente matura per poter rimpiazzare l'uomo e le sue abilità, ma sicuramente,

nel contesto dell'assistenza al cliente, i chatbot possono adempiere a molti dei compiti più semplici, come ad esempio il fornire informazioni.

Dunque, l'obiettivo che ci si è posti è quello di realizzare un chatbot che imiti nelle risposte quelle del personale di un'agenzia di viaggio, al fine di garantire agli utenti un'esperienza di interfacciamento con il sistema molto simile a quella che si può sperimentare nel dialogare con un essere umano. In aggiunta, si vuole garantire correttezza e affidabilità nelle risposte.

Nella realizzazione del chatbot si è fatto uso di un dataset, fornitoci dall'azienda Hello Vacanze, contenente vari scambi di messaggi, suddivisi in conversazioni, tra utenti e agenti dell'agenzia di viaggio. L'utilizzo del dataset è stato importante per poter andare a realizzare un modello personalizzato che fosse capace di imitare il comportamento di lavoratori specializzati nel settore dell'assistenza clienti.

Il primo passo, che ha portato alla realizzazione di questo sistema, è consistito nell'effettuare un accurato pre-processing sui dati contenuti nel dataset. In seguito, i dati strutturati in formato JSONL, sono stati usati per effettuare il fine-tuning su un modello base di GPT-3.

Al termine della fase di fine-tuning sono state svolte delle valutazioni sulle risposte fornitoci dal modello personalizzato. Queste valutazioni hanno condotto all'introduzione di una base informativa al fine di consentire al modello di accedere a maggiori informazioni per poter dare risposte più accurate.

Per consentire che il sistema chatbot fosse accessibile agli eventuali utilizzatori è stata realizzata anche un'interfaccia grafica.

L'ultima fase è stata dedicata ad un'ulteriore valutazione del modello finale. Nei capitoli successivi verrà illustrato nello specifico il processo di implementazione del chatbot, indagando più nel dettaglio l'obiettivo che ha motivato la realizzazione del seguente progetto e le motivazioni che hanno condotto a determinate scelte tecnologiche.

1.3 Struttura della tesi

In questa sezione verrà esposta la struttura della tesi, specificando il contenuto di ogni capitolo.

Nel capitolo 2 viene discusso con un maggior livello di dettaglio l'obiettivo del progetto e viene presentata l'azienda "Hello Vacanze", che ha fornito il dataset utilizzato per effettuare il fine-tuning.

Nel capitolo 3 vengono descritti gli strumenti tecnologici utilizzati nella fase di sviluppo del progetto, in particolare si tratterà di GPT_3 e GPT Index (o LlamaIndex).

Nel capitolo 4 viene esposto il flusso di lavoro che ha portato alla realizzazione del chatbot, vengono descritte nel dettaglio tutte le singole fasi, ponendo attenzione alle motivazioni che hanno guidato le scelte progettuali. Nel capitolo 5 vengono esposti i risultati delle valutazioni effettuate in due fasi diverse, al termine del fine tuning e successivamente all'inserimento della base informativa.

Nell'ultimo capitolo, capitolo 6, è presente un breve riassunto del lavoro e vengono esposte alcune considerazioni finali elaborate al termine del progetto. A seguito delle valutazioni conclusive, come risultato delle riflessioni svolte, vengono proposti possibili sviluppi futuri del progetto.

Chapter 2

Obiettivi del Progetto

In questo capitolo verrà presentata l'azienda "Hello Vacanze" che ha fornito il dataset sul quale abbiamo operato, verrà esposto il collocamento della stessa all'interno del settore turistico italiano e si spiegheranno le motivazioni e le necessità che hanno condotto alla scelta di realizzare un chatbot. Inoltre, verranno esposti gli obiettivi del progetto in questione.

2.1 Azienda "Hello Vacanze"

L'azienda "Hello Vacanze" S.r.l. è il primo Tour Operator italiano online specializzato nel segmento Mare Italia, la sede è situata presso Porto San Giorgio (FM) nella regione Marche. [4]

L'azienda opera nel settore turistico italiano fornendo una vasta gamma di servizi, in particolare, si occupa dell'organizzazione di viaggi proponendo ai clienti i prezzi migliori per varie strutture ricettive, come hotel, residence, villaggi, b&B, appartamenti e campeggi. Si occupa, inoltre, di effettuare le prenotazioni presso tali strutture e di fornire assistenza ai clienti. Tutti questi servizi vengono offerti esclusivamente online.

Come detto sopra, l'azienda opera nel settore turistico italiano e per tale motivo si trova ad affrontare tutte le sfide tipiche di tale settore. In particolare, essendo il loro core business quello dell'assistenza al cliente in tutte le fasi di una vacanza, risentono in particolar modo della difficoltà relativa al dimensionamento del numero dei dipendenti in funzione del carico di lavoro. Infatti stimare l'afflusso di persone e, quindi, il numero di dipendenti necessari risulta molto complesso a causa dell'elevata fluttuazione di questo tipo di mercato. Una sovrastima del numero dei dipendenti necessari potrebbe portare all'introduzione di costi non necessari e dall'altra parte, sottostimare

il numero causerebbe una riduzione dell'efficienza del servizio offerto.

2.2 Obiettivo del progetto

Dopo aver individuato le problematiche relative alla gestione di un servizio clienti per il settore turistico, si è fatta un'analisi delle caratteristiche più importanti che dovrebbe avere al fine di poter individuare gli obiettivi del progetto.

Le caratteristiche emerse come fondamentali per garantire un servizio di assistenza clienti di alto livello sono le seguenti:

- Correttezza nelle informazioni fornite
- Rapidità nelle risposte
- Disponibilità 24/7

Il sistema ritenuto potenzialmente in grado di assolvere a questi aspetti è un chatbot personalizzato su un dominio specifico.

L'azienda Hello Vacanze offre il suo servizio esclusivamente online senza nessun tipo di interazione in presenza tra agente e cliente, dunque, l'adozione di un sistema chatbot risulta in linea con la modalità di erogazione del servizio.

Dunque, l'obiettivo del seguente progetto è quello di realizzare un chatbot che possa interagire con i clienti occupandosi di fornire informazioni sulle strutture ricettive e consigliare il posto che maggiormente si adegui alle richieste del cliente.

Entrando maggiormente nel dettaglio, sono stati definiti due sotto-obiettivi che hanno caratterizzato e motivato le scelte di progettazione e degli strumenti tecnologici adottati.

Il primo obiettivo è quello di realizzare un sistema che simuli nello stile e nel tono delle risposte il personale specializzato di un'agenzia viaggi, rendendo l'esperienza di interazione il più simile possibile a quella che si può sperimentare nel dialogare con un altro essere umano. Per fare questo si è individuato come strumento migliore quello di effettuare un fine tuning su un large language model con i dati fornitoci dall'azienda.

Il secondo obiettivo consiste nel fornire al modello una base informativa a cui attingere per rispondere correttamente, in termini di contenuto, alle domande poste dall'utente.

Chapter 3

Strumenti Tecnologici

All'interno di questo capitolo verranno illustrati gli strumenti tecnologici adottati al fine di realizzare il sistema chatbot. Nel dettaglio si parlerà del large language model GPT-3 e del framework GPT index(o LlamaIndex).

3.1 Large Language Model GPT-3

GPT-3 è l'acronimo di Generative Pre-training Transformer 3, un modello generativo basato su transformer, terza generazione di modelli linguistici creati da OpenAI per dialogare con gli esseri umani tramite l'utilizzo di linguaggio naturale.

Questo modello è stato progettato tramite l'uso di una rete neurale di tipo Transformer dalla dimensione di circa 175 miliardi di parametri, con lo scopo di predire la parola successiva a un insieme di frasi già scritte.

3.1.1 I transformer

I transformers sono modelli di apprendimento automatico che utilizzano una struttura chiamata "trasformatore" per processare e comprendere i dati in modo più efficiente rispetto ai metodi tradizionali. Sono stati presentati per la prima volta nel 2017 da Vaswani et al. nella loro pubblicazione "Attention Is All You Need". [5]

"We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely."

Questi modelli forniscono un'ottima alternativa alle architetture di reti neurali ricorrenti (RNN), andando a rivoluzionare le tecniche di elaborazione del linguaggio naturale.

I transformer utilizzano il meccanismo dell'attenzione per elaborare sequenze di input, come le sequenze di parole in un testo. Questo meccanismo permette alla rete neurale di ricavare solamente i dati rilevanti per rispondere ad un determinato quesito, ignorando gli altri. Il tutto al fine di gestire grandi quantità di dati in modo più efficiente e di ottenere risultati più accurati. [6]

Questa capacità di prestare attenzione a parti specifiche dell'input rende i transformers particolarmente adatti all'elaborazione del linguaggio naturale, dove l'informazione rilevante può essere dispersa all'interno di una sequenza di parole. [7]

Inoltre, le reti neurali Transformer sono in grado di "imparare" in modo autonomo, grazie all'utilizzo di algoritmi di machine learning avanzati. Questo significa che possono migliorare nel tempo, adattandosi ai dati che vengono forniti loro e diventando sempre più accurate nel risolvere i problemi a cui sono destinate. [6]

3.1.2 Trasferimento di conoscenza

GPT-3 è stato addestrato su una grande quantità di testo estrapolato dal web, all'incirca 45 terabyte di dati. Per andare a specializzare il seguente modello su un dominio specifico, al fine di migliorare specifiche prestazioni e competenze, esistono varie tecniche, le più usate sono il fine-tuning o il few-shot learning.

Fine-tuning

Per fine-tuning si intende un riaddestramento di un modello che è stato già addestrato, che quindi avrà tutta una serie di logiche già ben definite e comprovate, per la nuova classificazione.

Tecnicamente, esistono due strategie di fine-tuning, o viene modificata la struttura generale del modello oppure viene modificato solamente l'ultimo strato della rete neurale responsabile della classificazione. Modificando solo quest'ultimo strato, i livelli precedenti, ovvero quelli in grado di estrarre le informazioni, rimarranno gli stessi e cambierà solo il modo in cui verranno interpretate quelle informazioni. [8]

Per quanto riguarda il fine-tuning su un modello base di GPT-3, OpenAi mette a disposizione delle API per realizzarlo. Questa chiamata API richiede

che i dati di training siano strutturati in un file JSONL, in cui ogni riga rappresenta una coppia prompt-completion.

Il modello personalizzato prodotto al termine del fine-tuning, viene archiviato sul cloud di OpenAI e non è scaricabile. L'unico modo per accedervi è tramite chiamate API specifiche. Questo aspetto rappresenta una potenziale limitazione dal momento che le chiamate API, necessarie sia per la fase di training che per l'utilizzo, sono a pagamento e il loro costo dipende dal numero di token analizzati.

Few-shot learning

Il few-shot learning è una tecnica di traferimento di conoscenza per i LLM che si può realizzare tramite la prompt-engineering.

Invece di effettuare il fine-tuning ed andare a modificare la struttura della rete neurale, tramite questa tecnica, il prompt viene strutturato in modo da indicare al modello il task che vogliamo che assolva, tramite alcuni esempi, e a specificare caratteristiche di linguaggio e stile che vorremmo che usasse.

3.2 GPT Index

GPT Index (o LlamaIndex) è un framework di dati semplice e flessibile per collegare fonti di dati personalizzate a large language models. Il principio di funzionamento si basa su tre strumenti chiave:

- Data Ingestion
- Data Indexing
- Query Interface

3.2.1 Data Ingestion

La Data Ingestion è quel processo che consente di accedere ed importare i dati da varie sorgenti per andarli poi a connettere con il large language model.

GPT Index consente l'integrazione sia di file **strutturati** come excel o SQL ma anche dati **semi strutturati** proveniente dalle API di Slack e Notion, fino ad arrivare a dati **non strutturati** come PDF, immagini e video.

3.2.2 Data Indexing

Successivamente alla fase di Data Ingestion, i dati importati vengono analizzati ed indicizzati in un formato semplice da interrogare.

Al termine di questa fase viene generato un file index, realizzato calcolando i vector embeddings.

3.2.3 Query Interface

In questa fase, viene interrogata la knowledge base, risultato dei precedenti step, andando a recuperare il context più appropriato in base alla domanda. Una volta individuata viene passato al LLM insieme alla domanda al fine di realizzare una risposta appropriata.

LlamaIndex, inoltre, fornisce moduli componibili che aiutano a costruire e integrare pipeline RAG (Retrieval Augmented Generation) per Q&A (motore di interrogazione), chatbot (motore di chat) o come parte di un agent. Questi blocchi possono essere personalizzati per riflettere le preferenze di classificazione e possono essere composti per ragionare su più basi di conoscenza in modo strutturato. [9]

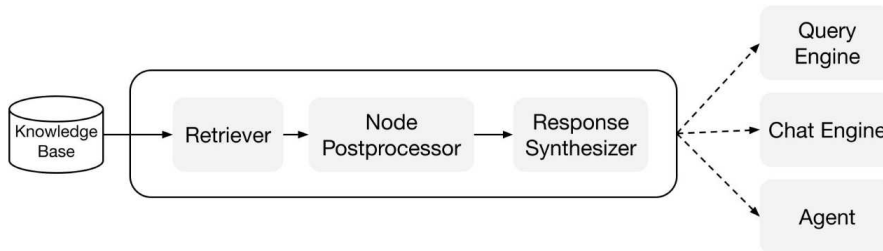


Figure 3.1: Step Query Interface

Chapter 4

Realizzazione del Chatbot

Nel seguente capitolo verrà illustrato il processo di realizzazione del chatbot. Nel dettaglio, saranno esposte tutte le varie fasi di sviluppo ponendo particolare attenzione alle motivazioni che hanno condotto all'adozione di tali soluzioni. Gli aspetti principali che verranno trattati nel seguente capitolo sono: la descrizione delle caratteristiche del dataset utilizzato, la pulizia e la trasformazione dei dati, il fine tuning realizzato sul modello Ada di GPT-3 e l'inserimento della knowledge base. Nella descrizione di questi vari aspetti verrà illustrata anche la motivazione per cui si è scelto di inserire una knowledge base nonostante sia stato precedentemente effettuato il fine tuning. Successivamente verrà brevemente illustrata l'interfaccia grafica realizzata al fine di rendere più user-friendly la fruizione del suddetto sistema.

4.1 Metodologia di progetto

Come detto in precedenza, il processo di realizzazione del chatbot è stato caratterizzato da varie fasi. Nell'immagine che segue è riportato un diagramma di flusso che mira a mettere in risalto gli aspetti cruciali per lo sviluppo di un chatbot personalizzato su un dominio specifico.

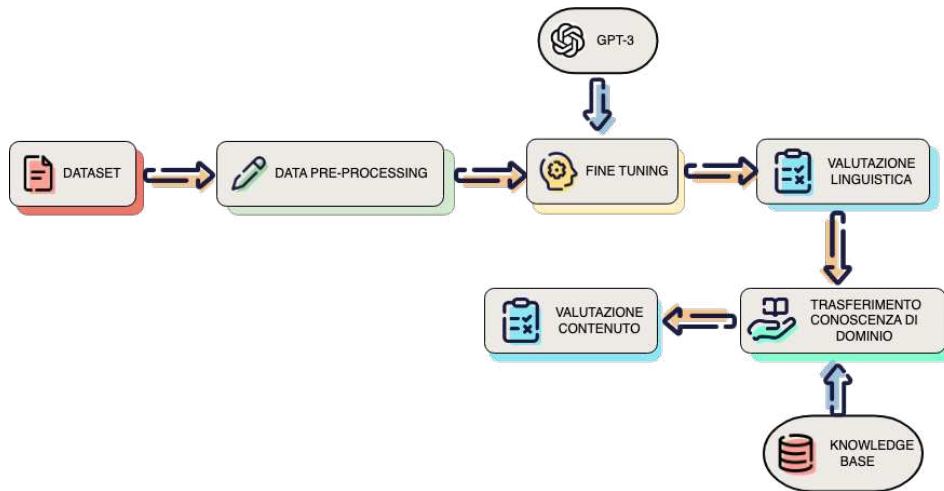


Figure 4.1: Diagramma di flusso rappresentante le varie fasi di progetto

Il diagramma di flusso è stato realizzato al fine di fornire una visione grafica di insieme delle fasi di progetto e del modo in cui le varie attività e i componenti interagiscono tra di loro ad un alto livello di astrazione. Come si può notare dal diagramma, il primo passo è stato quello di effettuare un accurato data pre-processing al fine di eliminare i dati sensibili, di rimuovere tutte quelle informazioni non ritenute utili allo scopo preposto e di trasformare il dataset al fine di renderlo utilizzabile per effettuare il fine-tuning. Successivamente, dal dataset è stato estratto un test set su cui effettuare le valutazioni relative al modello ottenuto come risultato del fine-tuning. Il secondo step è stato quello di effettuare un fine-tuning su un modello di deep learning, al fine di ottenere un modello personalizzato capace di imitare nelle risposte i dipendenti dell'agenzia. Al termine del fine-tuning è seguita una prima fase di valutazione sulle risposte fornite dal modello, alle domande contenute nel test set, rivolta principalmente ad individuare la coerenza e la correttezza grammaticale delle risposte, ciò per andare a verificare se l'obiettivo postosi fosse stato raggiunto. A seguito di questa fase, essendo emersa una carenza a livello contenutistico e avendo individuato aree di miglioramento, si è scelto di inserire una base informativa da cui il modello potesse ricavare le informazioni necessarie per rispondere correttamente alle domande degli utenti. A questo punto è stata svolta una seconda fase di valutazioni, al fine di valutare come l'inserimento di una nuova fonte di dati andasse ad influenzare le risposte del modello precedentemente personalizzato. Al termine di questo processo, è stata creata una semplice interfaccia grafica, con la finalità di rendere il chatbot maggiormente fruibile. Nei suc-

cessivi paragrafi si andranno ad analizzare nel dettaglio le varie fasi esposte brevemente poc'anzi.

4.2 Dataset

Il dataset di cui ci si è serviti al fine di realizzare il fine tuning è stato fornito dall'agenzia turistica Hello Vacanze, della quale si è già parlato nel capitolo 2. Il dataset fornitoci riporta al suo interno vari scambi di messaggi, avvenuti tra agenti ed utenti, suddivisi per conversazione. Nel dettaglio, il dataset è costituito da 5 file con estensione .csv, ognuno contenente le conversazioni di un'intera settimana. Ogni file è costituito da circa 10000 righe per un totale di 56000 righe. I suddetti file, in forma tabulare, presentano 19 campi, ogni riga rappresenta e racchiude varie informazioni circa il singolo messaggio inviato, l'attore responsabile dell'invio del messaggio e la conversazione all'interno della quale è stato inviato il messaggio. Nella tabella 4.1 sono riportati i campi che appaiono nei file che costituiscono il dataset con una brevissima descrizione.

Campo	Descrizione
conversation_id	ID della conversazione
conversation_url	URL della conversazione
interaction_raw_id	ID dell'interazione
message_id	ID del messaggio
message_type	Tipologia di messaggio
message_parts	Testo del messaggio
created_time	Data e ora di invio
message_source	Origine del messaggio
actor_id	ID dell'attore
actor_type	Tipo di attore (user/agent)
actor_sub_entity	Tipo specifico di attore
actor_email	Email dell'attore
actor_phone	Numero di telefono dell'attore
actor_first_name	Nome dell'attore
actor_last_name	Cognome dell'attore
reference_id	ID di riferimento
channel_id	ID del canale
channel_name	Nome del canale
detailed_message_type	Tipo di messaggio specifico

Table 4.1: Campi del dataset

Dopo un'attenta analisi del contenuto dei messaggi si è notato che sono principalmente relativi a richieste di prenotazione di strutture ricettive, emissione di preventivi e assistenza al cliente durante il soggiorno. Non sono presenti informazioni sulle strutture di cui trattano. Quest'ultimo aspetto ha influenzato molto le successive scelte di sviluppo del chatbot, in quanto, non fornendo il dataset informazioni esaustive si è reso necessario andare a creare una nuova base informativa 4.5 al fine di rendere il chatbot capace di rispondere in maniera corretta alle domande inerenti le varie strutture.

4.3 Data Preprocessing

Una volta individuato il dataset di partenza, una fase fondamentale per il processo di realizzazione di un chatbot personalizzato su un dominio specifico è stata la realizzazione del fine tuning di un Large Language Model base. Per massimizzare l'efficacia del fine tuning, però, è risultato essenziale effettuare una selezione, una pulizia e una trasformazione dei dati a disposizione. Infatti, è dimostrato come i risultati finali ottenuti al termine del processo di addestramento dipendono in gran parte dalla qualità del dataset a disposizione e dall'accuratezza con cui viene svolta la fase di preprocessing dei dati.

Nel diagramma in figura 4.2 si possono visualizzare le varie fasi di manipolazione dei dati che hanno portato alla realizzazione del file JSONL che è stato successivamente utilizzato per effettuare il fine tuning.

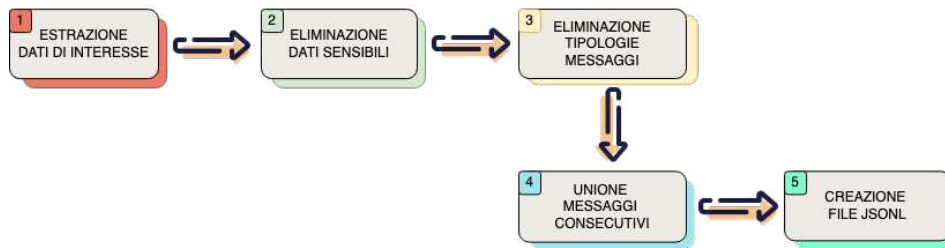


Figure 4.2: Fasi svolgimento Data Preprocessing

Tutte queste fasi sono state realizzate mediante l'utilizzo del linguaggio di programmazione Python, usufruendo di particolari librerie per la manipolazione dei dati che il linguaggio mette a disposizione come Pandas e Re.

4.3.1 Estrazione dati di interesse

La prima azione di preprocessing svolta sul dataset è stata un'attenta analisi del contenuto e della struttura dei dati presenti, al fine di individuare i dati

utili allo scopo del progetto, le criticità che il dataset presenta e le possibili soluzioni, il tutto per poter garantire un fine tuning ottimale. In questa fase di analisi è emerso come molti campi presenti nel dataset contenevano informazioni non utili per l'obiettivo postosi, per cui, sono stati eliminati e conservati solamente quei campi necessari ad individuare il messaggio inviato e il suo contenuto, la conversazione all'interno della quale era stato inviato il messaggio, il tipo di mittente e la data di creazione. La struttura della tabella ottenuta al termine di questa fase è riportata in figura 4.3. I dati riportati sono a puro scopo esemplificativo, non sono dati reali.

conversation_id	message_id	message_parts	created_time	actor_id	actor_type	detailed_message_type
00001	000002	"Buongiorno"	2022-02-11	00003	user	NORMAL
00001	000004	"Buongiorno, come posso aiutarla?"	2022-02-11	00005	agent	NORMAL

Figure 4.3: Rappresentazione struttura dataset

I campi ritenuti necessari e, dunque, estratti dal dataset originale, sono solamente 7, di seguito è riportata una breve descrizione di ognuno:

- **conversation_id:** ID identificativo della conversazione in cui è inserito il messaggio
- **message_id:** ID identificativo del messaggio
- **message_parts:** Testo completo del messaggio
- **created_time:** Data e ora di invio del messaggio
- **actor_id:** ID identificativo del mittente del messaggio
- **actor_type:** Tipologia del mittente, in particolare questo campo può assumere solo due valori "user" o "agent"
- **detailed_message_type:** Tipologia dettagliata del messaggio inviato, può assumere i seguenti valori:
 - NORMAL

- STATUS_CHANGE_REOPENED
- STATUS_CHANGE_ASSIGNED
- PRIVATE_NOTE
- OFFLINE_MESSAGE
- STATUS_CHANGE_RESOLVED
- AWAY_MESSAGE
- FRESHDESK_TICKET_CREATE_MESSAGE

4.3.2 Eliminazione dati sensibili

Una volta individuati i campi del dataset contenenti informazioni utili per lo sviluppo del chatbot, si è ritenuto necessario eliminare tutti i dati sensibili presenti all'interno del testo dei messaggi per evitare che il modello potesse riproporli durante l'utilizzo. In particolare, sono stati individuati nel dataset tutti quei messaggi contenenti al loro interno **numeri di telefono** o **indirizzi email**.

Successivamente sono state svolte delle valutazioni per capire se convenisse eliminare l'intera conversazione o il singolo messaggio contenente il dato sensibile. La scelta di eliminare l'intera conversazione sarebbe stata maggiormente corretta in quanto, l'eliminazione di un solo messaggio, potrebbe portare a difficoltà di comprensione della chat ma alla fine la scelta è ricaduta sull'eliminazione del solo messaggio. Questa decisione è stata presa in quanto, andando ad analizzare il dataset, si è riscontrato che nel 20% delle conversazioni compariva un indirizzo email e nel 25% delle conversazione un numero di telefono, dunque, eliminare l'intera chat avrebbe portato a rilevanti perdite di informazioni che sono state valutate come una perdita troppo cospicua.

4.3.3 Eliminazione messaggi con particolari tipologie

Il passo successivo in questa fase è stato quello di eliminare tutti i messaggi che contenevano informazioni non ritenute utili allo scopo del fine tuning. Sono stati eliminati tutti i messaggi di sistema che segnalavano l'inizio e la fine di una conversazioni, l'apertura di un ticket verso il reparto tecnico e le note private degli utenti. Questi messaggi nel dataset vengono identificati da particolari tipologie, pertanto il processo di pulizia si è svolto individuando ed eliminando tutti i messaggi che presentavano nel campo `detailed_message_type` una delle seguenti stringhe di testo:

STATUS_CHANGE_REOPENED, STATUS_CHANGE_ASSIGNED, STATUS_CHANGE_RESOLVED, PRIVATE_NOTE, OFFLINE_MESSAGE,

FRESHDESK_TICKET_CREATE_MESSAGE e AWAY_MESSAGE.

4.3.4 Ristrutturazione dataset

Il terzo step è stato quello di ristrutturare il dataset ed, in particolare, di unire tutti i messaggi consecutivi provenienti da uno stesso utente. Come detto in precedenza, il dataset è costituito da una serie di messaggi, ogni riga corrisponde ad un particolare messaggio e per ogni messaggio abbiamo varie informazioni. Gli scambi di messaggi tramite chat sono spesso informali e, durante l'attenta analisi del dataset, si è riscontrato come fenomeno frequente l'invio di più messaggi consecutivi da parte degli attori. La necessità che si aveva nella ristrutturazione del dataset era quella di ottenere ed individuare chiaramente la suddivisione tra le domande dell'utente e le risposte dell'agente, al fine di poterle usare per effettuare il fine tuning, quindi, è risultato necessario unire tutti quei messaggi consecutivi proveniente dallo stesso attore. Questo procedimento è stato realizzato ordinando, come prima cosa, il dataset per `conversation_id` e `created_time`, in modo da avere una struttura dati ordinata con tutti i messaggi raggruppati per conversazioni e ordinati per data di invio. Il tutto per poter individuare facilmente i messaggi consecutivi.

Svolta questa azione di ordinamento, si è proceduto con l'unione del testo dei messaggi consecutivi inviati dallo stesso attore. Per semplicità, si è scelto di unire il testo aggiungendo il separatore ", " tra un messaggio e l'altro.

4.3.5 Creazione file JSONL

L'ultima fase è stata dedicata alla creazione di un file JSONL necessario per effettuare il fine tuning. Il file è stato realizzato seguendo la struttura richiesta da OpenAI per poter utilizzare il tool messo da loro a disposizione per effettuare il fine tuning su uno dei modelli base di GPT-3. La struttura del file, che viene richiesta per effettuare il fine tuning, è quella mostrata in figura:

```
1 {"prompt": "<prompt text>", "completion": "<ideal generated text>"}
2 {"prompt": "<prompt text>", "completion": "<ideal generated text>"}
3 {"prompt": "<prompt text>", "completion": "<ideal generated text>"}

```

Figure 4.4: Struttura file JSONL

Prima di realizzare il nuovo file JSONL, però, si è reso necessario eseguire altri passaggi di pulizia del dataset. In particolare, sono state eliminate tutte le conversazioni iniziate dall'agente e le chat contenenti un numero dispari di righe.

Questo ulteriore processo di pulizia si è reso necessario per poter realizzare il file JSONL con la struttura richiesta. Infatti, come si può notare in figura 4.4, ogni riga deve essere costituita da una coppia chiave-valore, le due chiavi devono essere "prompt" e "completion", in cui il valore corrispondente alla chiave "completion" deve essere il testo che vorremmo che il nostro modello generi al seguito di un determinato prompt.

Essendo il nostro scopo quello di realizzare un modello che risponda in maniera simile ad un agente, in particolare per quanto riguarda lo stile, il tono e il formato, dunque, risulta chiaro come il prompt debba coincidere con la domanda dell'utente e la completion con la risposta dell'agente. La necessità di eliminare le conversazioni con numero di righe dispari è sorta per evitare di avere dei prompt senza una relativa completion mentre la necessità di eliminare le rare conversazioni iniziate dall'agente è emersa per garantire la struttura in cui la domanda venga dall'utente e la risposta dall'agente.

Il risultato ottenuto dalla fase di data preprocessing è, dunque, un file JSONL contenente le informazioni estrapolate dal nostro dataset, disponibile per effettuare il fine tuning di un modello base di GPT-3. Questo file è stato successivamente suddiviso in due file, uno usato come training set, contenente all'incirca 12000 coppie prompt-completion e il secondo come test set, contenente all'incirca 200 coppie prompt-completion. La motivazione per cui il test set presenta pochi elementi è dovuta al fatto che si è scelto di valutare manualmente le risposte del nostro sistema personalizzato. Nella sezione dedicata alla valutazione del modello ottenuto in seguito al fine tuning verranno chiariti gli aspetti critici della valutazione.

4.4 Fine Tuning

Il passo successivo è stato quello di effettuare il fine tuning. La scelta del modello su cui realizzare il fine tuning è ricaduta sul modello base **ADA** di GPT-3.

Tale modello presenta varie caratteristiche riscontrate ottime per il nostro caso specifico. Come prima cosa, OpenAi per la realizzazione del fine tuning, mette a disposizione una Api e un tool che oltre ad effettuare il processo di addestramento, forniscono anche una valutazione sulla struttura dei dati da utilizzare; ciò riduce gli errori causati da errate formattazioni dei dati, velocizzando e migliorando il processo di fine tuning. Inoltre, le prestazioni

del modello, anche in assenza di fine tuning, risultano essere piuttosto buone, caratteristica molto importante nel caso in cui i dati a disposizione non siano in quantità elevate.

L'ultimo punto che ha segnato la scelta del modello di GPT-3 da utilizzare è stata la valutazione del rapporto qualità-costo. OpenAI, infatti, mette a disposizione vari modelli di GPT-3 su cui fare il fine tuning, tra cui ADA, BABBAGE, CURIE e DAVINCI. Ogni modello è caratterizzato da differenti performance e da un costo diverso per effettuare il fine tuning. Il modello ADA che è stato scelto presenta il minor prezzo per token, offrendo comunque prestazioni adeguate agli obiettivi del progetto.

Dopo aver individuato il modello base è stato utilizzato il tool messo a disposizione da OpenAI per valutare il file preparato durante il processo di data preprocessing. Questo strumento, oltre a fornire una valutazione, fornisce anche suggerimenti per migliorare la struttura del file. In figura è riportato il comando utilizzato per la validazione del file JSONL.

```
openai tools fine_tunes.prepare_data -f <LOCAL_FILE>
```

Figure 4.5: Comando per validazione file

In seguito è iniziata la vera fase di fine tuning al termine della quale è stato generato il modello personalizzato sul nostro dataset specifico.

```
Uploaded model: ada:ft-personal-2023-07-11-18-47-52
```

Figure 4.6: Nome modello personalizzato

4.5 Traferimento conoscenza di dominio

Il passo successivo è stato quello di fornire al modello dei dati circa alcune strutture ricettive per poter risolvere le problematiche dovute a mancanza di informazione.

I dati delle strutture turistiche sono stati raccolti da internet e salvati in svariati file di testo con estensione .txt. Nel complesso è stata realizzata una base informativa costituita da 10 diversi file, ognuno contenente informazioni su una diversa struttura. Le strutture di cui sono state raccolte informazioni si dividono in B&b, hotel, residence e camping, tutti collocati nelle regioni Marche, Abruzzo e Molise. E' stato scelto di utilizzare dei semplici file di testo invece che dei file più strutturati per poter indagare l'efficacia del passaggio di un tipo di informazione rappresentata in modo molto semplice, che permetta, eventualmente, a chiunque fosse interessato ad utilizzare un sistema di questo tipo di personalizzare agevolmente la base informativa a cui attinge il modello. Anche la scelta di passare solo 10 file contenenti informazioni su solo 10 strutture nasce dallo scopo preposto che era solamente quello di testare l'effettiva efficacia di una soluzione di questo tipo. Di seguito è riportato, a scopo esemplificativo, uno dei file .txt riportante informazioni sulla struttura ricettiva EGO Hotel situata nella città di Ancona.

```
EGO HOTEL
Situato in una splendida cornice fuori Ancona nella regione Marche, l'Ego Hotel offre il
WiFi gratuito in tutte le aree, una piacevole vista sul Mare Adriatico e sul porto della
città, e un centro benessere con sauna, vasca idromassaggio, bagno turco, fontana di
ghiaccio e doccia sensoriale. Le camere dispongono di aria condizionata, TV a schermo
piatto e bagno privato con asciugacapelli e set di cortesia. Al mattino vi attende una
colazione all'americana a buffet. In loco è presente un ristorante di cucina italiana
contemporanea. L'Ego Hotel dista 3 km dalla stazione ferroviaria di Ancona Centrale e 5
minuti in auto dal porto turistico Marina Dorica di Ancona. Per ulteriori info consultare
il sito web: https://www.egohotelancona.it/
Quali sono i servizi benessere disponibili presso l'Ego Hotel? L'Ego Hotel offre un centro
benessere completo con diverse strutture per il relax e il benessere. Tra i servizi
disponibili ci sono la sauna, la vasca idromassaggio, il bagno turco, la fontana di
ghiaccio e la doccia sensoriale. Gli ospiti possono godere di momenti di puro relax
durante il loro soggiorno.
Quali servizi sono inclusi nella colazione all'americana a buffet dell'Ego Hotel? La
colazione all'americana a buffet servita presso l'Ego Hotel offre una varietà di deliziosi
piatti, tra cui scelta di bevande calde, assortimento di pane e paste, selezione di
cereali, frutta fresca e succhi di frutta. Gli ospiti possono gustare una colazione
completa per iniziare bene la giornata.
Quali luoghi sono facilmente raggiungibili? L'Ego Hotel dista circa 3 km dalla stazione
ferroviaria di Ancona Centrale e circa 5 minuti di auto dal porto turistico Marina Dorica
di Ancona. Gli ospiti possono raggiungere facilmente la struttura utilizzando i mezzi
pubblici o taxi, garantendo un comodo arrivo e partenza dalla struttura.
```

Figure 4.7: Esempio file .txt

Tecnicamente per rendere possibile il passaggio della base informativa al LLM, senza effettuare il fine tuning, si sono utilizzate due librerie: Langchain

e GPTIndex. L'utilizzo di questi strumenti rende possibile il passaggio di dati non strutturati attraverso un'interfaccia molto semplice. Infatti, sarà proprio il framework a codificare i dati in un file strutturato "index.json" che viene interrogato per estrapolare il contenuto appropriato alla domanda da fornire come risposta. Nella figura 4.8 si può vedere parte del file index.json che è stato generato a partire dal file .txt precedentemente mostrato in figura 4.7.

```
{
  "index_struct_id": "dfd5a4af-b855-415c-b420-db83145a18bd",
  "docstore": {
    "docs": {
      "dfd5a4af-b855-415c-b420-db83145a18bd": {
        "text": null,
        "doc_id": "dfd5a4af-b855-415c-b420-db83145a18bd",
        "embedding": null,
        "extra_info": null,
        "nodes_dict": {
          "6265018860198485245": {
            "text": "EGO HOTEL\nSituato in una splendida cornice fuori Ancona nella regione Marche, l'Ego Hotel offre il WiFi gratuito in tutte le aree, una piacevole vista sul Mare Adriatico e sul porto della citt\u00e0, e un centro benessere con sauna, vasca idromassaggio, bagno turco, fontana di ghiaccio e doccia sensoriale. Le camere dispongono di aria condizionata, TV a schermo piatto e bagno privato con asciugacapelli e set di cortesia. Al mattino vi attende una colazione all'americana a buffet. In loco \u00e8 presente un ristorante di cucina italiana contemporanea. L'Ego Hotel dista 3 km dalla stazione ferroviaria di Ancona Centrale e 5 minuti in auto dal porto turistico Marina Dorica di Ancona. Per ulteriori info consultare il sito web: https://www.egohotelancona.it/\nQuali sono i servizi benessere disponibili presso l'Ego Hotel? L'Ego Hotel offre un centro benessere completo con diverse strutture per il relax e il benessere. Tra i servizi disponibili ci sono la sauna, la vasca idromassaggio, il bagno turco, la fontana di ghiaccio e la doccia sensoriale. Gli ospiti possono godere di momenti di puro relax durante il loro soggiorno.\nQuali servizi sono inclusi nella colazione all'americana a buffet dell'Ego Hotel? La colazione all'americana a buffet servita presso l'Ego Hotel offre una variet\u00e0 di deliziosi piatti, tra cui scelta di bevande calde, assortimento di pane e paste, selezione di cereali, frutta fresca e succhi di frutta. Gli ospiti possono gustare una colazione completa per iniziare bene la giornata.\nQuali luoghi sono facilmente raggiungibili? L'Ego Hotel dista circa 3 km dalla stazione ferroviaria di",
            "doc_id": "94d69784-83b7-4c24-8c82-d2eb9670526c",
            "embedding": null,
            "extra_info": null,
            "index": 0,
            "child_indices": [],
            "ref_doc_id": "eea88adf-ae70-4561-907b-b66532511884",
            "node_info": {
              "start": 0,
              "end": 1611
            }
          },
          "5994297118294231740": {
            "text": "Hotel dista circa 3 km dalla stazione ferroviaria di Ancona Centrale e circa 5 minuti di auto dal porto turistico Marina Dorica di Ancona. Gli ospiti possono raggiungere facilmente la struttura utilizzando i mezzi pubblici o taxi, garantendo un comodo arrivo e partenza dalla struttura.",
            "doc_id": "b988d256-c4f4-4fbb-95ba-cf00254bc484",
            "embedding": null,
            "extra_info": null,
            "index": 1,
            "child_indices": [],
            "ref_doc_id": "eea88adf-ae70-4561-907b-b66532511884",
            "node_info": {
              "start": 1559,
              "end": 1845
            }
          }
        },
        "id_map": {
          "2359665a-8d80-4065-91c1-c2c8aeba6dc0": 6265018860198485245,
          "06f07925-7f36-485c-8ac3-41089e8cefb1": 5994297118294231740,
          "embeddings_dict": {
            "2359665a-8d80-4065-91c1-c2c8aeba6dc0": [
              -0.0076854475773870945, -0.001702244859188795,
              0.014193550683557987, -0.024776576086878777, -0.010857738554477692, 0.0020276501309126616,
              -0.02246112935245037, -0.015972649678587914, -0.02353382110595703, 0.009170210920274258,
              0.012630297802388668, 0.011917349882423878, -0.010759626515209675, 0.0018772114999592304,
              -0.015226997435092926, -0.008156386204063892, 0.004460830241441727, 0.012407910078763962,
              -0.010785790160298347, 0.0025002232287079096, 0.00879084412008524, 0.025522226467728615,
              -0.00012172032438684255, -0.01189772691577673, -0.015070018358528614, -0.012525644153356552,
              -0.0018461426952853799, -0.019766317680478096, 0.010975473560392857, -0.03300288928747177,
              0.009451465681195259, -0.028622569516301155, -0.04039601981639862, 0.006835143081843853,
              0.011053962633013725, -0.009078639559447765, -0.005353650078177452, -0.0008396760676987469,
              0.026568755507469177, 0.013748775236308575, 0.01697993464767933, -0.009464547038078308,
              0.009418761357665062, -0.015174671076238155, -0.0026882714591920376, 0.015946486964821815,
              -0.008077896200120449, 0.0018199793994426727, -0.020891336724162102, 0.005893266759812832,
              0.0039735399186611176, 0.00997473020106554, -0.0050756661221385, -0.010138249956071377,
              -0.0013956446200609207, 0.014533672481775284, -0.02412249520421028, 0.0023350678384304047,
            ]
          }
        }
      }
    }
  }
}
```

Figure 4.8: Estratto da file index.json

Il LLM risponde ad uno specifico prompt (domanda dell'utente) andando ad interrogare il file index.json. Per garantire uno scambio di messaggi più agevole e l'esperienza della conversazione soddisfacente, si è scelto di passare come prompt, oltre alla domanda corrente, tutte le domande e le risposte fornite in precedenza. Questo sistema di costruzione del prompt fa sì che il modello abbia sempre presenti tutti i precedenti scambi di messaggi così da poter rispondere alle successive domande tenendo conto delle informazioni fornite durante tutta la conversazione.

Al termine dell'implementazione e integrazione di questo strumento è seguita

una fase di test per valutare l'effettiva funzionalità del sistema creato.

4.6 Interfaccia grafica

Nell'ultima fase del processo di realizzazione del chatbot ci si è concentrati sull'implementazione di una semplice interfaccia grafica, allo scopo di rendere il sistema facilmente e piacevolmente fruibile da un eventuale utente. L'interfaccia grafica è stata realizzata mediante l'uso della libreria Gradio, la quale consente di creare semplicemente dell'interfacce web che rendono possibile la fruizione del modello di machine learning in modo agevole senza doversi interfacciare con il terminale.

In figura 4.9 si può vedere il risultato finale dell'implementazione dell'interfaccia.

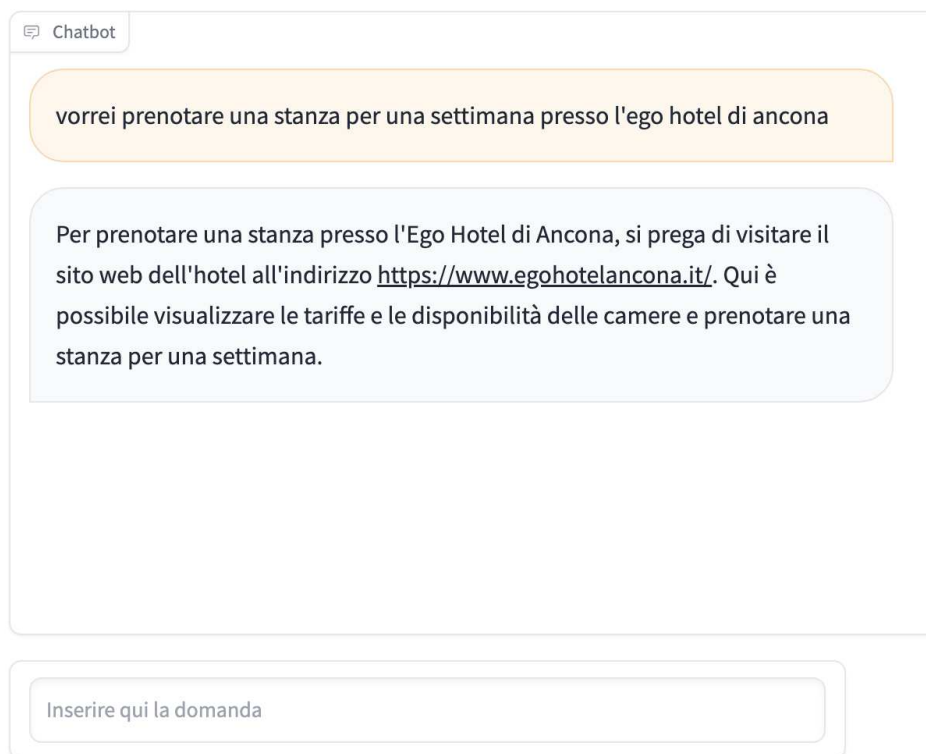


Figure 4.9: Interfaccia Grafica Chatbot

Come si evince dalla foto è un'interfaccia molto semplice ma che evidenzia bene la separazione tra domanda e risposta, inoltre, vuole imitare lo stile

grafico di un attuale sistema di chat, il tutto per rendere il sistema più semplice e familiare per un eventuale utente.

Chapter 5

Valutazione

In questo capitolo sono riportate le valutazioni effettuate sui risultati prodotti dal modello al termine della fase di fine tuning e a seguito dell'inserimento della knowledge base. Nel fare le due valutazioni ci si è soffermati su aspetti diversi, in particolare, con la prima si valutato lo stile e la correttezza grammaticale delle risposte, con la seconda la correttezza del contenuto delle risposte.

5.1 Valutazione Linguistica

La prima fase di valutazione è iniziata con l'effettuazione di alcune ricerche sui possibili strumenti a disposizione per valutare un chatbot ed è risultato che la migliore soluzione fosse quella di effettuare una valutazione manuale. La motivazione principale di questa scelta è che, in questa prima fase, si voleva valutare il tono e lo stile delle risposte fornite dal modello per verificare se fosse simile a quello degli esempi contenuti nel file usato durante la fase di fine tuning. La scelta di questo tipo di valutazione, però, ha portato alla necessità di selezionare un test set piuttosto ristretto per poter valutare in breve tempo i risultati del modello. Il test set usato è costituito da 200 domande estratte casualmente dal file originale. Di seguito sono riportate, in una prima parte, le domande e le risposte passate come esempio al LLM in fase di fine tuning e, a seguire, le risposte fornite dal modello personalizzato ad alcune domande estratte dal test set.

*"prompt": "Salve preventivo pensione completa dal 2307 al 3007 grazie ",
"completion": " Mi indica gentilmente il numero dei componenti del suo nu-*

cleo familiare con la relativa età dei bambini ?”

”prompt”: *”Grazie ”,*
”completion”: *” di nulla si figuri :)”*

”prompt”: *”Buon Giorno chiedevo un preventivo per la settimana di ferragosto o quella prima siamo 2 adulti e un bambino di 12 anni vorrei sapere le procedure per la presentazione grazie mille.”,*
”completion”: *” Buongiorno; :) cliccando sul link potrà controllare la disponibilità ed effettuare velocemente, e in automatico, il preventivo:”*

Di seguito possiamo vedere alcuni esempi delle risposte **fornite dal modello** in risposta ad alcuni prompt estratti dal test set.

”prompt”: *”Vorrei avere un preventivo pensione completa”,*
”completion”: *”Mi indica gentilmente il numero dei componenti del suo nucleo familiare con la relativa età dei bambini e il periodo di soggiorno che le interessa?”*

”prompt”: *”Grazie”,*
”completion”: *”di nulla ;)”*

”prompt”: *”Salve volevo sapere se ci sono disponibilità... 14 /28 agosto 2 adulti +2 bambini Marche ”,*
”completion”: *”Buongiorno; :) cliccando sul link potrà controllare la disponibilità ed effettuare velocemente, e in automatico, il preventivo”*

In seguito ad una prima analisi si può notare come il modello risponda effettivamente imitando lo stile e il tono delle risposte fornite da un agente.

Nel dettaglio, possiamo vedere un uso delle emoticon del tutto simile a quello presente nel training set ma anche la scelta dei termini e la formulazione delle frasi è coerente con quanto mostrato al modello nella fase di fine tuning.

Si sono riscontrate, però, anche delle criticità. Ad esempio, nell'ultima risposta il link che viene fornito per effettuare il preventivo è di una struttura situata in Abruzzo, non coerentemente con la richiesta di prenotare in una strutture delle Marche. Questo errore consegue dal fatto che il dataset non presenta all'interno delle informazioni sulle strutture a cui il LLM potesse attingere.

In generale, su un totale di 200 coppie domanda-risposta analizzate, risultano 88 risposte corrette grammaticalmente e appropriate in relazione alla domanda, come nel caso del primo e secondo esempio sopra riportato; 73 risposte risultano solamente corrette grammaticalmente ma presentano inesattezze nel contenuto in relazione alla domanda, come nell'ultimo esempio; Infine, 5 risposte risultano appropriate a livello di contenuto ma presentano alcuni errori grammaticali. Le restanti 34 risposte presentano o gravi errori nella struttura grammaticale della frase o gravi incoerenze rispetto alla domanda posta.

In conclusione, si sono ritenute accettabili l'83% delle risposte fornite dal modello, questo perchè l'obiettivo postosi in questa fase era quello di ottenere un modello che rispondesse con stile simile a quanto passato come esempio e con una buona correttezza grammaticale. Gli errori a livello di contenuto presenti nelle 73 risposte sono del tutto in linea con quanto ci si aspettava, non avendo il modello una effettiva base informativa da cui attingere.

5.2 Valutazione finale del LLM

Dopo aver predisposto il sistema per l'inserimento di una base informativa ed aver proceduto con l'inserimento della stessa è seguita una seconda fase di valutazione. In questo caso la valutazione si è concentrata principalmente sull'analisi del contenuto delle risposte più che del formato, al fine di andare ad individuare se l'apprendimento delle informazioni passate avvenisse nella maniera corretta. Anche in questo caso, come già per la prima fase di test, la valutazione è stata effettuata manualmente. Il numero di prompt a cui è stato sottoposto il modello sono stati ancora più limitati, circa 50 domande. Questo perchè, oltre alla difficoltà di valutare manualmente le completion, si sono aggiunte problematiche a livello economico in quanto interrogare il sistema con l'inserimento della knowledge base e con la particolare struttura del prompt scelta risultava molto più costoso. Di seguito riportiamo alcuni esempi di risposte fornite dal modello dopo aver inserito la knowledge base.

Prompt: Buongiorno, vorrei avere delle informazioni sul camping Don Diego

Completion: Buongiorno :), il Camping Village Don Diego offre ampie piazzole immerse nel verde, completamente ombreggiate e delimitate da siepi di separazione, adatte per camper e tende. Inoltre, è stata recentemente realizzata un'area piazzole con erba sempreverde per una scelta ancora più confortevole. Il campeggio offre anche un gruppo di servizi igienici, che sono curati e puliti come tradizione. All'interno, sono presenti docce calde gratuite, lavatrici ed asciugatrici a gettone per il comfort degli ospiti. Inoltre, il campeggio offre connessione Wi-Fi illimitata inclusa nella tariffa della piazzola. Il campeggio offre anche un programma completo di animazione che include attività come il mini-club, feste serali, baby dance e acquagym.

Prompt: dove è situato il villaggio Lido D'abruzzo?

Completion: Il villaggio Lido d'Abruzzo è situato nella città di Roseto degli Abruzzi nella regione Abruzzo.

Prompt: dove è situato il villaggio Lido D'abruzzo? Il villaggio Lido d'Abruzzo è situato nella città di Roseto degli Abruzzi nella regione Abruzzo. **Che servizi offre?**

Completion: Il Villaggio offre ombrelloni, sdraio, lettini, docce calde, chiosco per bevande e gelati, un servizio di salvataggio, tre piscine dell'acquapark, un parco giochi attrezzato, una zona Baby Club riservata, videogiochi, solarium ed idromassaggio con fungo, e piatti della buona tavola dalla genuinità insuperabile.

Prompt: vorrei prenotare una stanza per una settimana presso l'ego hotel di ancona.

Completion: Per prenotare una stanza presso l'Ego Hotel di Ancona, si prega di visitare il sito web dell'hotel all'indirizzo <https://www.egohotelancona.it/>. Qui è possibile visualizzare le tariffe e le disponibilità delle camere e prenotare una stanza per una settimana.

Dagli esempi qui riportati si può notare come le informazioni vengono riprese e trasmesse correttamente all'utente. La risposta che il modello fornisce

ad un determinato prompt risulta chiara e precisa. Per quanto riguarda il secondo esempio è interessante evidenziare che la seconda domanda inserita dall'utente in realtà fosse semplicemente "*Che servizi offre?*". La presenza nel prompt dei precedenti messaggi è tale in virtù della strutturazione del prompt implementata, ciò al fine di fornire al modello tutte le informazioni contenute nei messaggi precedenti. Qui possiamo anche notare l'utilità di questa scelta, l'utente infatti, nonostante non abbia specificato nella domanda la struttura di cui conoscere i servizi riceve ugualmente una risposta corretta. Questo perchè il modello ricava dai messaggi precedenti l'argomento di cui si sta parlando.

In conclusione, su un totale di 50 prompt passati, 43 risposte (l'86%) risultano corrette nel contenuto e grammaticalmente, ossia, a seguito del prompt passato la risposta fornita risulta corretta ed esaustiva. Le altre 7 risposte presentano delle imprecisioni sul contenuto anche se continua a risultare buona la correttezza grammaticale.

Nota: i prompt passati indagavano sulle informazioni facenti parte della knowledge base fornita al modello.

5.2.1 Confronto pre e post inserimento knowledge base

Al termine dell'implementazione del chatbot, risulta interessante focalizzare l'attenzione sul confronto tra le due versioni del chatbot, quella precedente e quella successiva all'inserimento della knowledge base.

La principale differenza tra le risposte fornite dai due sistemi è chiaramente relativa al contenuto. Al termine del fine tuning il LLM non aveva una base informativa da cui potesse attingere informazioni e l'aspetto più rilevante risultava essere l'apprendimento della forma e dello stile di risposta. I risultati, come già esposto in precedenza, risultavano essere ottimi sotto questo punto di vista.

L'inserimento della knowledge base ha sicuramente portato a delle risposte chiare e dettagliate a livello di contenuto ma si è leggermente persa quella forte personalizzazione nel tono delle risposte che invece era presente al termine della sola fase di fine tuning.

Chapter 6

Conclusioni e sviluppi futuri

L'obiettivo principale del progetto era creare un chatbot personalizzato su un dominio specifico che rispondesse alle domande degli utenti in maniera simile a dei veri agenti di un'agenzia di viaggio. Gli obiettivi che ci eravamo posti al fine di realizzare questo sistema erano, in una prima fase, di creare un modello personalizzato con i dati a disposizione e successivamente di fornire una knowledge base che il modello potesse interrogare per rispondere adeguatamente alle domande dell'utente.

Il progetto è stato sviluppato servendosi di alcuni strumenti tecnologici come le API e i tool messi a disposizione da OpenAi per effettuare il fine tuning sul modello base ADA di GPT-3. Inoltre, ci si è serviti del framework LlamaIndex (o GPT Index) per consentire il passaggio di dati non strutturati al fine di creare la knowledge base che il modello interroga per fornire risposte accurate alle domande poste dall'utente.

Quest'ultimo aspetto è di grande valore, in quanto GPT Index permette di passare i dati in quasi ogni formato disponibile e ciò da la possibilità all'utente di passare dati al modello senza la necessità di strutturarli, rendendo l'interazione molto più semplice.

La fase di valutazione del modello, anche se limitata a causa di motivi di budget e dell'onerosità della valutazione manuale, ha prodotto ottimi risultati, coerenti con quelli che erano gli obiettivi posti in fase iniziale.

In ogni caso, in seguito alle valutazioni effettuate si può affermare che, gli obiettivi posti in fase progettuale, sono stati soddisfatti adeguatamente.

6.1 Sviluppi futuri

Gli sviluppi futuri che possono essere previsti per il seguente progetto sono vari.

Prima di tutto, potrebbe essere estesa la fase di valutazione al fine di effettuare un'analisi più approfondita sui risultati prodotti dal modello per individuare le possibili aree di miglioramento.

Le tecnologie utilizzate, inoltre, sono in costante evoluzione e dunque potrebbe essere interessante utilizzare LLM con migliori prestazioni per effettuare il fine tuning, come ad esempio gpt-3.5-turbo.

Un altro aspetto interessante potrebbe essere quello di migliorare l'interfaccia grafica al fine di fornire un'esperienza utente migliore.

Ed infine, una visione di lungo periodo potrebbe prevedere la realizzazione di un sito web che permetta agli utenti di realizzare un proprio chatbot personalizzato da integrare alla propria attività imprenditoriale, passando semplicemente i dati che si vuole fornire al modello in qualsiasi formato a disposizione.

Bibliography

- [1] Francesca Di Feo. “Turismo da record in Italia nel 2023: le previsioni per l’estate”. In: *partitaiva.it* (2023).
- [2] “Nel 2023 il pil del turismo a 100 miliardi di euro”. In: *ANSA.it* (2023).
- [3] Mariarita Persichetti. “Lavorare nel turismo in Italia: la situazione attuale”. In: *BuoneNotizie.it* (2023).
- [4] URL: <https://www.hellovacanze.com/it/>.
- [5] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].
- [6] *Reti neurali Transformer*. URL: <https://www.intelligenzaartificialeitalia.net/>.
- [7] *Cosa sono i transformers e come vengono utilizzati nell’elaborazione del linguaggio naturale*. URL: <https://smartstrategy.eu/>.
- [8] *Cos’è il fine-tuning nel Machine Learning*. URL: <https://pulplearning.altervista.org/>.
- [9] URL: <https://gpt-index.readthedocs.io/en/stable/>.