Valutazione della propensione alla mediazione tramite eXplainable AI[1]

Paolo Nesi

Abstract: La mediazione nei processi civili può efficacemente risolvere controversie al di fuori delle procedure giudiziarie, alleviando il carico sui tribunali in caso di successo. L'efficienza nell'identificazione delle dispute è essenziale, poiché un tentativo fallito di mediazione può allungare la durata del processo. La decisione spetta al giudice/tribunale sulla base di numerosi documenti che contengono alcune dichiarazioni significative per la decisione. Questo articolo descrive una soluzione di intelligenza artificiale, AI, per fornire un sistema di supporto alle decisioni in grado di elaborare documenti e (i) produrre suggerimenti affidabili, (ii) produrre motivazioni circostanziate evidenziando le affermazioni che hanno portato al suggerimento, (iii) rispettare la privacy e la sicurezza dei dati. A tal fine sono state utilizzate tecnologie dell'AI tecniche di eXplainable AI (XAI), ottenendo una soluzione che soddisfa gli obiettivi definiti. La soluzione è stata sviluppata nell'ambito del progetto di ricerca Giustizia Agile, finanziato nel PON Governance e Capacità Istituzionale Nazionale Italiana, e validata rispetto a casi reali. La soluzione ha sfruttato il framework Snap4City per la gestione dei dati e della soluzione AI/XAI.

1. Introduzione

Il sistema giudiziario italiano è uno dei più lenti d'Europa. Secondo il rapporto della Commissione Europea per l'Efficienza della Giustizia (CEPEJ)[2] pubblicato nel 2022 e riferito all'anno 2020, la principale causa di inefficienza del sistema giudiziario italiano è l'eccessiva durata dei procedimenti giudiziari, soprattutto di natura civile e commerciale[3]. La misura utilizzata dalla CEPEJ per confrontare la celerità dei sistemi giudiziari nei diversi paesi dell'UE si chiama *disposition time*. Pur essendo progressivamente diminuito dal 2012 al 2018, l'Italia ha registrato il più alto *disposition time* nell'Unione europea per le cause civili di primo grado, con una durata della causa di ben 674 giorni contro una media europea di soli 237.

[1] Ringraziamo per la collaborazione alla realizzazione del progetto e del presente contributo: Enrico Collini, Claudia Raffaelli, Francesco Scandiffio, assegnisti di ricerca del Laboratorio di Sistemi Distribuiti e Tecnologie Internet (DISIT) dell'Università degli Studi di Firenze.

[2] Sistemi giudiziari europei – Rapporto di valutazione CEPEJ – Ciclo di valutazione 2022 (dati 2020) <https://rm.coe.int/cepej-report-2020-22-e-eb/1680a86279>.

[3] Direzione generale di statistica e analisi organizzativa (DG-Stat) <https://webstat.giustizia.it/SitePages/Home.aspx>.

Paolo Nesi, University of Florence, Italy, paolo.nesi@unifi.it, 0000-0003-1044-3107

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Nel caso dei processi civili la mediazione è uno strumento utile a risolvere questo problema. Oltre ai casi per i quali la legge italiana prevede l'obbligo di tentare la mediazione prima dell'inizio del processo, il giudice può invitare le parti a un tentativo di mediazione a sua discrezione[4]. Un tentativo di mediazione concluso con successo alleggerisce il carico dei tribunali perché elimina la necessità di avviare o portare a conclusione un processo; al contrario, a seguito del fallimento del tentativo di mediazione, la causa ritorna in giudizio (solitamente mesi), aggiungendo tempo alla durata del processo. Secondo il rapporto CEPEJ, nel 2020 in Italia ci sono stati 60.110 tentativi di mediazione, di cui solo 15.013 conclusi con un accordo: in tutti gli altri casi i casi sono ritornati in tribunale, e il tempo impiegato nel tentativo di raggiungere un accordo ha contribuito ad allungare i tempi durata del processo contenzioso. Per aumentare l'efficacia della mediazione è necessario valutare con precisione l'attuale propensione delle parti a raggiungere un accordo.

A tal fine, nel contesto del progetto Giustizia Agile ci siamo focalizzati a realizzare un sistema di supporto alle decisioni in grado di fornire informazioni aggiuntive al giudice per determinare se un caso specifico potesse essere ragionevolmente risolto utilizzando lo strumento della mediazione della controversia. Ad esempio, con una classificazione in classi 'non propensione alla mediazione', 'propensione alla mediazione' e 'neutro'. Individuare correttamente la propensione alla mediazione può aiutare ad evitare di gestire la controversia nei complessi e lunghi meccanismi giudiziari, garantendo così una risoluzione della controversia che può essere più rapida riducendo il carico di lavoro del tribunale e dei giudici. Pertanto, ci siamo prefissati di realizzare uno strumento che possa fornire:

- *suggerimenti attendibili* per determinare quando la mediazione può avere successo. Ciò implica fornire al giudice una misura della probabilità delle parti di accettare il processo di mediazione come un modo per trovare un accordo reciproco;
- *motivazioni circostanziate* a supporto del suggerimento fornito. Ciò implica fornire al giudice un motivo, un'evidenza, estratta dai documenti, che le parti coinvolte nella controversia possono essere motivate a mediare, ad esempio fornendo le dichiarazioni e le frasi contenute nel documento ufficiale che porterebbero dedurre tale fatto;
- *un'interfaccia web che possa produrre suggerimenti e motivazioni su richiesta*, a servizio del tribunale e dei giudici coinvolti. La soluzione deve essere sviluppata in modo da integrarsi nel corrente flusso di lavoro e con gli strumenti disponibili e attualmente adottati nei tribunali italiani. La soluzione deve rispettare la privacy dei dati secondo il GDPR Regolamento generale sulla protezione dei dati dell'Unione europea 2016/679[5].

[4] Decreto legislativo n. 28 del 4 marzo 2010, art. 5, Condizioni di procedibilità e rapporti con il processo. Disponibile al sito ufficiale: <https://www.gazzettaufficiale.it/eli/id/2010/03/05/010G0050/sg>.

[5] Regolamento generale sulla protezione dei dati dell'Unione europea 2016/679, GDPR <https://gdpr.eu/what-is-gdpr/> (15-10-2023).

Per questo motivo, la ricerca descritta in questo articolo si è concentrata sull'uso di tecniche di intelligenza artificiale, AI e NLP (elaborazione del linguaggio naturale) per sviluppare un sistema di supporto alle decisioni a servizio del tribunale e dei giudici. L'attività di ricerca si è concentrata sullo sviluppo di una soluzione per elaborare documenti legali, tramite tecniche di intelligenza artificiale e specificamente derivate da BERT (Bidirection Encoder Representations from Transformers) (Devlin et al. 2018), per identificare nell'ampio insieme di documenti e dichiarazioni relativi a controversie se esistono elementi che possono dedurre una propensione alla mediazione delle parti, e perché. A tal fine, l'attività si è concentrata su (i) sviluppare un set di dati da usare in fasi di training, validazione e test set della AI; (ii) definire un modello utilizzando tecniche di *fine tuning* su un modello BERT pre-addestrato per la lingua italiana, per eseguire una classificazione del testo e rilevare la possibilità di mediare o meno; (iii) sviluppare una soluzione di AI spiegabile, eXplainable AI, XAI, mediante Shapley (Lundberg e Lee 2017) come strumento per fornire le motivazioni alla base del modello al punto (ii), affermazione per affermazione; (iv) sviluppare uno strumento integrato di supporto alle decisioni da fornire al tribunale e ai giudici che vogliono analizzare i documenti che hanno a disposizione in modo da ridurre il tempo necessario alla valutazione. Infine, lo stesso strumento di (iv) deve essere in grado di raccogliere suggerimenti e commenti per migliorare ulteriormente il modello del punto (ii) e il set di dati del punto (i) per le prossime versioni della soluzione AI prodotta. Più nello specifico, lo strumento è stato chiamato XAI4MA (Explainable Artificial Intelligence tool for Mediation Agile).

XAI4MA ed il modello di AI proposto sono stati validati da un gruppo di esperti di mediazione, afferenti al Dipartimento di Giurisprudenza dell'Università di Firenze. Per questo sono stati processati svariati atti giudiziari, ed il risultato prodotto è stato confrontato con le loro valutazioni, fornendo tale risultato al sistema stesso. Questo lavoro è stato sviluppato nell'ambito del progetto di ricerca Giustizia Agile, finanziato nel PON Governance e Capacità Istituzionale Nazionale per migliorare l'obiettivo di una migliore organizzazione della macchina giudiziaria. La soluzione proposta ha sfruttato il framework Snap4City per la gestione dei dati e AI/XAI (Garau et al. 2020).

Questo articolo è strutturato come segue, ed in accordo alla Figura 1 in cui è riportato il flusso logico del documento in relazione ai processi sui dati. La Sezione 2 descrive il contesto e alcuni lavori di ricerca correlati. La Sezione 3 delinea i requisiti e gli obiettivi del sistema prodotto. La Sezione 4 discute le tecniche adottate per creare il dataset, tra cui: la raccolta dei dati, la pre-elaborazione, l'etichettatura delle frasi, la normalizzazione del testo; la suddivisione delle frasi, e la suddivisione in blocchi ed infine la preparazione dei data set per l'addestramento, la convalida e il test del modello di AI. La Sezione 5 fornisce un dettaglio sull'architettura di sistema fornendo una descrizione del modello BERT (Devlin et al. 2018) per la classificazione delle frasi.

Le metodologie di XAI fanno riferimento all'approccio Shapley (Lundberg e Lee 2017) e sono descritte nella Sezione 6. La Sezione 7 presenta l'estensione della valutazione di propensione o meno a livello di documento. Nella Sezione

8 viene presentato lo strumento di supporto alle decisioni che potrebbe essere messo a disposizione dei tribunali. Lo strumento è stato denominato XAI4MA e presenta un'interfaccia utente grafica progettata per consentire ai decisori di sfruttare la soluzione AI/XAI in diverse fasi delle procedure, usandolo come un sistema esperto da consultare su richiesta. Lo stesso strumento ha permesso di raccogliere dati aggiuntivi e di eseguire un'ulteriore validazione della soluzione come riportato alla fine della Sezione 8.

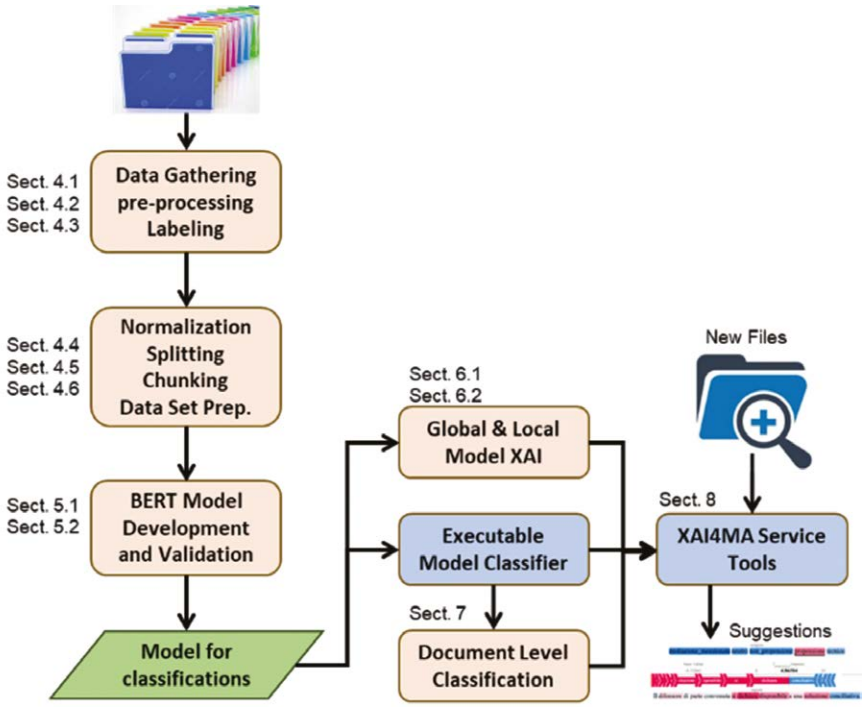


Figura 1 – Flussi di dati e processi nel documento.

2. Lavori correlati

La maggior parte delle soluzioni commerciali progettate per assistere gli avvocati si concentrano sull'indicizzazione delle informazioni contenute negli atti e consentono di ricercare informazioni utili utilizzando query in linguaggio naturale. Ne sono esempi Jurimetria[6] e Predictice[7], mediante le quali è possibile

[6] LA LEY. Giurimetria. Estratto il 13 marzo 2023 da <https://jurimetria.laleynext.es/content/QueEs.aspx>.

[7] Previsione <https://predictice.com/fr>.

ottenere una stima dell'importo del risarcimento. Altri sistemi si concentrano sull'analisi dei documenti legali per prevedere la decisione del giudice (Medvedeva, Wieling e Vols 2023). Ad esempio, Katz et al. (2016) hanno proposto un sistema per prevedere le decisioni prese dalla Corte Suprema degli Stati Uniti utilizzando tecniche basate su AI. I casi giudiziari sono stati modellati con un massimo di 240 variabili, la maggior parte delle quali categoriche. Queste soluzioni hanno prodotto accuratezze del 70,2% nella classificazione del caso. Alghazzawi et al. (2022) hanno utilizzato modelli di AI a breve termine (Hochreiter e Schmidhuber 1997) e reti convoluzionali (Gu et al. 2018) (LSTM +CNN) per prevedere i verdetti della Corte Suprema degli Stati Uniti. In questo caso, il modello ha raggiunto una precisione 92,05%. Medveva, Vols e Wieling (2020) hanno sfruttato l'analisi dei big data sulle sentenze della Corte Europea dei Diritti dell'Uomo (CEDU) per prevedere se il caso descrivesse una violazione dei propri diritti o meno. Questo problema di classificazione binaria è stato valutato dagli autori tramite Support Vector Machines (SVM) (Cortes e Vapnik 1995) e ha raggiunto una precisione del 75%. Aletras et al. (2016) sullo stesso dataset hanno raggiunto una precisione del 79% aggiungendo qualche accorgimento statistico sempre utilizzando SVM. Hsun-Ping et al. (2022) si sono concentrati sulla possibilità di successo della mediazione tra le parti. Gli autori hanno proposto un sistema per prevedere il successo delle richieste di mediazione utilizzando informazioni testuali e proprietà del caso come il luogo della controversia, l'ID del mediatore, il numero di partecipanti, ecc. L'obiettivo finale è la riduzione dell'onere a carico del tribunale. I risultati sono stati ottenuti utilizzando un framework basato su LSTM, chiamato LSTMEnsambler in grado di prevedere i risultati della mediazione assemblando più classificatori. I risultati sul set di test hanno raggiunto una precisione del 78,8%. Una delle principali carenze degli attuali approcci al ragionamento giuridico sull'AI è che non sono in grado di fornire una giustificazione del loro ragionamento in termini di concetti giuridici appropriati (Collenette, Atkinson e Bench-Capon 2023). Branting et al. (2021) hanno concentrato i loro sforzi nel proporre un sistema spiegabile di supporto alla previsione delle decisioni legali. La soluzione proposta mirava a evidenziare le porzioni più rilevanti del testo della causa che maggiormente hanno determinato la decisione finale derivata dal modello AI. Abilitare spiegazioni sui modelli di intelligenza artificiale (noti anche come eXplainable Artificial Intelligence, XAI) per il dominio legale consentirebbe ai giudici di massimizzare la possibilità di identificare errori e *bias* all'interno degli algoritmi come riportato in Deeks (2019). Il Regolamento generale sulla protezione dei dati (GDPR) dell'Unione europea contiene disposizioni che richiedono quello che alcuni hanno definito un «diritto a una spiegazione» (Goodman e Flaxman 2017).

Una delle caratteristiche chiave che devono definire ciascuno di questi sistemi è l'assenza di pregiudizi. Nell'apprendimento automatico, un modello è affetto da *bias* quando produce risultati che sono sistematicamente polarizzati a causa di ipotesi errate nel processo decisionale (Van Giffen, Herhausen e Fahse 2022). Sebbene la decisione finale rimanga responsabilità del giudice, è importante che il sistema non produca inavvertitamente suggerimenti distorti

che siano corrotti da elementi discriminatori incorporati nel set di dati di addestramento e/o dal processo di selezione del modello e dei dati/caratteristiche. Geraghty e Woodhams (2015) confrontano diversi strumenti per prevedere la probabilità di recidiva di persone già condannate o sotto processo. È dimostrato che molti degli strumenti sono stati realizzati con dati provenienti da soggetti di sesso maschile, con conseguente minore accuratezza nella previsione relativa alla recidiva delle donne. Un altro esempio di questo aspetto può essere COMPAS (Brennan, Dieterich e Ehret 2009); si tratta di una soluzione commerciale in uso negli Stati Uniti per prevedere il potenziale di recidiva tra gli imputati penali. Tali soluzioni sono state messe all'indice per l'utilizzo di una selezione parziale dei dati che ha prodotto modelli di AI addestrati che tendevano ad essere più sfavorevoli per le persone di determinate etnie. DataJust è stata una soluzione con l'obiettivo di sviluppare un dataset e un sistema per proporre risarcimenti legati sia al danno subito che agli attori coinvolti, essendo il danno considerato più o meno grave a seconda delle caratteristiche della persona che lo subisce. La soluzione è stata disattivata[8] a causa di alcune critiche sull'utilizzo di dati non totalmente anonimi. Il trattamento dei dati sensibili, infatti, è un punto fondamentale da non trascurare durante l'implementazione di qualsiasi tipi di software. Soprattutto quando si tratta di dati provenienti da determinati settori, come quello sanitario, finanziario o giudiziario, che naturalmente fanno riferimento alle persone.

In alcuni casi, le informazioni necessarie per prendere correttamente una decisione sono contenute nei molti documenti di contesto in modo più chiaro che nelle sintesi o nelle dichiarazioni. Quando ciò avviene, l'anonimizzazione del dataset deve consentire di preservare la descrizione del contesto e di salvaguardare la privacy degli utenti (Csányi et al. 2021) Una corretta anonimizzazione dei dati personali comporta la trasformazione dei testi dai quali l'identificativo della persona o l'associazione della vertenza alla persona non può essere ricostruita, né direttamente né indirettamente, né da parte del titolare del trattamento né in collaborazione con alcun altro soggetto, che potrebbe fornire conoscenza addizionale (vedi nota 5). I dati completamente anonimizzati possono essere archiviati e utilizzati senza limitazioni. Inoltre, la pulizia dei dati dovrebbe rimuovere le informazioni che potrebbero introdurre pregiudizi, come lo status sociale, il genere, la nazionalità, l'etnia, ecc., delle persone coinvolte. Nel mondo accademico e commerciale sono stati proposti numerosi strumenti per identificare e rimuovere informazioni private all'interno dei documenti[9]. I sistemi più recenti fanno coincidere la procedura di anonimizzazione sulla base del riconoscimento delle entità o ID private, Named Entity Recognition (Hassan, Domingo Ferrer e Soria-Comas 2018; Licari e Comandè 2021), identificando all'interno dei te-

[8] Datajust <https://acteurspublics.fr/articles/exclusif-le-ministere-de-la-justice-renonce-a-son-algorithme-datajust>.

[9] Super.AI. Documento Redact <https://super.ai/super-redact/document-redact>.

sti particolari entità di interesse come nomi, numeri di telefono, date, indirizzi, e procedendo poi a sostituirle con etichette anonime.

3. Analisi dei requisiti

L'obiettivo principale della ricerca è produrre uno strumento che possa assistere i giudici nella gestione di un procedimento al fine di accelerarne la conclusione nella valutazione dell'esito positivo della mediazione della controversia. Dall'analisi degli obiettivi e delle sfide critiche del progetto Giustizia Agile e in base ai workshop svolti con gli operatori della giustizia, è stato possibile identificare i requisiti che un sistema di supporto decisionale per la mediazione dovrebbe fornire. D'altro canto, prima di fornire una dichiarazione dei requisiti, è necessaria un'ulteriore analisi e descrizione dell'intero contesto e delle procedure.

A. Contesto e procedura

Secondo la legge, alcuni argomenti possono implicare l'obbligo di condurre una procedura di mediazione, in cui le parti, aiutate da un mediatore, cercano di trovare una soluzione che ponga fine alla controversia riducendo il tempo del processo. In altri casi è il giudice che può scegliere di mandare le parti in mediazione. Questo processo di mediazione potrebbe, qualora non si raggiunga un accordo, contribuire al protrarsi nel tempo del caso. Con questi presupposti, affinché la valutazione del giudice sia il più accurata possibile, può essere opportuno pensare all'introduzione di un sistema in grado di stimare la probabilità di successo della mediazione come strumento di supporto alla decisione, come si dovrebbe da un esperto in mediazione, che potrà fornire anche alcune motivazioni a margine del suggerimento/valutazione fornita.

La valutazione della probabilità di successo della mediazione (qui chiamata mediabilità) di un processo può essere analizzata da diverse prospettive, ciascuna delle quali si concentra su un aspetto particolare dell'opportunità di mediazione. In particolare, sono noti i seguenti tre casi principali:

a) *Probabilità di successo del tentativo di mediazione*. Il successo della procedura è influenzato da molti fattori, quali: il contesto della controversia; le personalità e le relazioni interpersonali delle parti in causa nonché i loro interessi personali; la gravità della controversia; l'entità delle risorse disponibili per sostenere il processo di mediazione; e le capacità di negoziazione del mediatore assegnato.

b) *Propensione delle parti in causa a impegnarsi nella mediazione*. Analizza la tendenza delle parti coinvolte di partecipare attivamente ad un tentativo di mediazione per risolvere le loro divergenze. Infatti, uno spirito desideroso di cooperare e di ricercare una soluzione condivisa è un fattore determinante per il buon esito della mediazione.

Senza alcun dubbio, il caso A, appare quello più efficace nel determinare se valga la pena intraprendere o meno il percorso della mediazione. Tuttavia, co-

me già descritto, questa opzione è influenzata da fattori difficilmente quantificabili e non ricercabili negli atti giudiziari. Un requisito che il sistema in esame deve soddisfare è che non si basi su dati soggettivi o di difficile interpretazione, ma esclusivamente su documentazione affidabile come i documenti testuali dei tribunali. Nella maggior parte dei casi è molto difficile capire se una controversia si è conclusa con una mediazione oppure è ancora in corso. Una mediazione riuscita non viene generalmente segnalata come risultato in tribunale. Caso B, la valutazione della propensione delle parti alla partecipazione attiva al tentativo di mediazione consentirebbe di quantificare la disponibilità delle parti in causa di impegnarsi nella risoluzione delle controversie. Indicatori della propensione delle parti possono essere individuati all'interno dei testi giudiziari a differenza degli altri due provvedimenti, che necessiterebbero di ulteriori informazioni. *In questo articolo, l'attenzione si è concentrata sulla fornitura di una soluzione per il caso B.*

B. Requisiti identificati

Lo strumento alla base di questo progetto di ricerca ha come obiettivo quello di fornire una valutazione imparziale a partire da atti giudiziari. I dati trattati sono sensibili e comprendono un gran numero di informazioni sensibili relative ai soggetti coinvolti, quali nomi, cognomi, codici fiscali, indirizzi e date. È importante garantire che queste informazioni non influenzino la valutazione effettuata dallo strumento, ad esempio adottando misure per rimuoverle dai testi stessi man mano che vengono utilizzate per sviluppare il sistema. Una procedura simile potrebbe essere applicata anche ai documenti man mano che vengono immessi nel sistema per la loro valutazione. Ciò sarebbe necessario se questi venissero caricati, ad esempio, su un *cloud* per l'archiviazione. La de-identificazione, ovvero la rimozione delle informazioni sensibili dai testi ne consentirebbe la conservazione nel rispetto dei requisiti di privacy.

In sintesi, abbiamo identificato i seguenti requisiti di sistema che la soluzione deve soddisfare. In particolare, lo strumento deve conformarsi al requisito di:

R1. produrre una valutazione complessiva di ciascuna controversia in merito alla propensione delle parti a partecipare attivamente al processo di mediazione. Si prega di notare che ogni controversia è descritta da una serie di documenti. Alcuni di essi possono essere molto significativi ai fini della valutazione altri potrebbero essere semplici attestazioni, come ad esempio: ricezione di documenti, registrazione dell'avvenuto passaggio di alcuni atti legali, ecc;

R2. produrre la valutazione considerando solo documenti testuali giudiziari. Facoltativamente altri documenti o singoli documenti potranno essere valutati in maniera indipendente;

R3. classificare ogni documento di controversia in una delle classi: propensione alla mediazione (M), non propensione alla mediazione (NM), e neutro (N);

R4. fornire una valutazione sulla confidenza della classificazione prodotta di R3: M, NM e N;

R5. fornire una spiegazione per il suggerimento/classificazione fornito di R3, R4, a livello di frase;
R6. garantire che il sistema sia sviluppato senza introdurre pregiudizi/*bias*, ad esempio rimuovendo informazioni sensibili dai documenti utilizzati come base di conoscenza. Supportare l'etica dell'intelligenza artificiale (Brundage et al. 2018; HLEG 2019); e l'etica dei dati (Richterich 2018; Butterworth 2018);
R7. garantire che i documenti immessi nello strumento siano adeguatamente conservati nel rispetto della privacy dei soggetti interessati in base al GDPR;
R8. identificare frasi significative per le classi di interesse M e NM e renderle accessibili;
R9. fornire un'interfaccia semplice in modo che gli utenti senza competenze informatiche avanzate possano ottenere risultati/suggerimenti, che possono essere utilizzati dal giudice o dal team per prendere la decisione finale e rimanere accessibili su carta;
R10. fornire un'interfaccia agli utenti qualificati con la quale sia per loro possibile inserire correzioni/suggerimenti alle valutazioni prodotte dalla soluzione AI, che potranno essere utilizzati dalla soluzione AI stessa per migliorare il modello di classificazione in versioni successive.

In realtà le soluzioni dovrebbero prevedere la propria valutazione sulla base di una serie di documenti caso per caso. È abbastanza frequente che, presi singolarmente, i documenti possano esprimere affermazioni contrastanti circa la propensione delle parti a prendere parte al processo di mediazione, e gran parte di essi possano essere neutrali. Nella stessa controversia composta da X documenti, alcuni di essi potrebbero essere valutati individualmente come appartenenti alla classe M, altri come NM e la maggior parte come N.

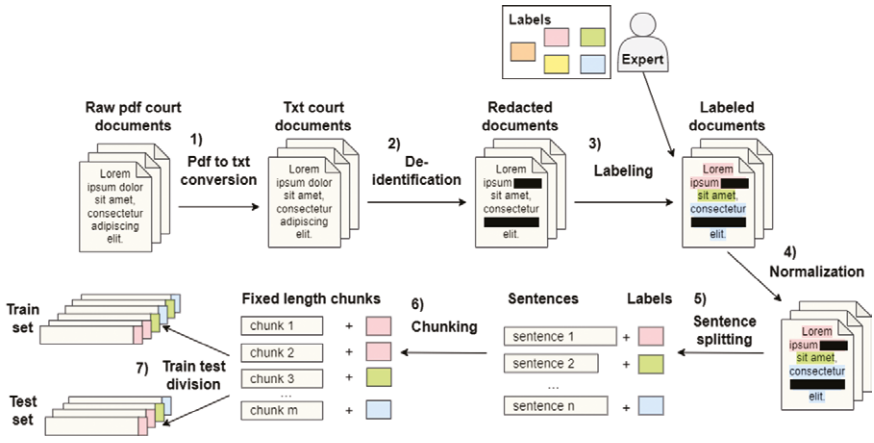


Figura 2 – Procedura di costruzione del dataset.

In genere, il giudice investigherebbe ogni documento alla ricerca di frasi significative che rivelino aspetti propensi alla mediazione o la mancanza di questa.

Un ulteriore requisito del sistema proposto dovrà quindi essere quello di fornire una classificazione complessiva dei documenti nelle classi M, NM e N, filtrando le frasi, all'interno dei testi, evidenziando quelle significative per le due classi di interesse M e NM. Queste informazioni aggiuntive possono essere rilevanti nel processo decisionale, fornendo allo stesso tempo una spiegazione, in termini di frammenti di testo rilevanti, sul motivo per cui il sistema ha prodotto automaticamente tale classificazione. Lo stesso approccio può essere effettuato a livello di singolo documento, che può contenere diverse affermazioni neutre e alcune di esse orientate come NM/M; solo questi ultimi sono significativi per prendere le decisioni.

4. Valutazione ed elaborazione dati

In accordo alle tecniche di machine learning supervisionato è necessario avere o produrre dataset di training, di valutazione e test. Pertanto, questa Sezione è dedicata alla descrizione dei passaggi intrapresi per costruire il dataset che possa essere utilizzato per il training, la validazione e il test della soluzione per la produzione dei suggerimenti su nuovi documenti, si veda R9, R10.

La procedura completa è riportata in Figura 2. Le successive sottosezioni forniscono descrizioni dettagliate dei passaggi. Nello specifico, la Sottosezione 4.A quantifica e descrive i dati nel loro formato grezzo. La Sottosezione 4.B fornisce una descrizione approfondita della procedura di de-identificazione. La Sottosezione 4.C è dedicata alla procedura di etichettatura per ottenere testi commentati. La Sottosezione 4.D descrive la normalizzazione dei dati. La Sottosezione 4.E riporta il processo di suddivisione in blocchi delle frasi per renderle utilizzabili nella fase di apprendimento. La Sottosezione 4.F descrive la suddivisione dei dati in dati di training, validazione e test.

A. Raccolta di file

I dati sono stati resi disponibili attraverso una specifica convenzione tra l'Università degli Studi di Firenze, DISIT Lab del DINFO, e il Tribunale civile di Firenze precedente al progetto Giustizia Agile. Ciò ha consentito l'autorizzazione ad accedere ed elaborare il contenuto dei fascicoli delle cause civili. I dossier sono costituiti da un insieme di documenti in lingua italiana di varia lunghezza e quantità. Si tratta di sentenze, verbali di udienze e documenti redatti dagli avvocati ed inseriti negli atti del processo. Più raramente un fascicolo può contenere anche la trascrizione della mediazione. Il basso numero di trascrizioni della mediazione, tuttavia, non deve essere preso come indicazione del successo o meno della mediazione: spesso le parti comunicano solo verbalmente di aver aggiunto un accordo, senza fornire i documenti prodotti dalla mediazione. Inoltre, è chiaro che ogni processo può seguire un iter più o meno lungo a seconda della complessità, delle materie trattate o degli interessi delle parti. I dati a nostra

disposizione provengono direttamente dal sistema informativo giudiziario denominato SICID (Sistema Informativo Distrettuale sulle Controversie Civili), che archivia gli atti giudiziari, sotto forma di file PDF[10]. Poiché in Italia il passaggio dal formato cartaceo a quello digitale è avvenuto in tempi relativamente recenti, molti dei documenti caricati sull'applicazione non sono nativi digitali, ma frutto di scansioni di documenti cartacei. Per questo motivo è stato deciso di escludere dalla selezione tutti i file PDF non nativi digitali in quanto avrebbe aggravato la fase di pre-elaborazione con il rischio di introdurre imprecisioni nei testi. Il set di dati grezzi selezionato era costituito da 74 fascicoli per un totale di 474 documenti. Come anticipato, i fascicoli contengono un numero variabile di documenti a seconda del numero di udienze svolte e dell'andamento del processo. Nel nostro caso la dimensione di un singolo fascicolo varia da 2 fino a 13 documenti, con un valore medio di 6. Anche i documenti hanno una lunghezza variabile, con un numero di pagine compreso tra 1 e 40. I documenti sono digitali PDF nativi. Questo processo corrisponde al passaggio 1 di Figura 2.

B. De-identificazione dei dati

Il requisito R6 prevedeva la garanzia che il sistema dovesse essere sviluppato secondo l'etica dei dati e l'etica dell'intelligenza artificiale. Informazioni come nomi, cognomi, numeri di previdenza sociale, indirizzi, date di nascita o altri eventi sono particolarmente a rischio di essere interpretate erroneamente. Questo problema viene chiamato di *annotation bias* e si verifica proprio quando il sistema è indotto a creare un'associazione tra etichette e informazioni irrilevanti contenute negli esempi utilizzati come *ground Truth*, come eventuali dati personali. La presenza di queste informazioni all'interno dei testi potrebbe essere associata dal sistema, ad esempio, a determinati reati o ad una probabilità di propensione. Supponiamo che una determinata organizzazione, ad esempio una banca, sia coinvolta in processi in cui tutte le parti sono sempre d'accordo per risolvere la controversia attraverso la procedura di mediazione: lasciare il nome della banca in chiaro porterebbe il sistema a classificare erroneamente tutte le frasi che contengono il nome di quella banca, senza nemmeno tener conto del resto della pena. Sebbene la creazione di metadati di associazione tra una determinata entità e una specifica propensione alla mediazione fornirebbe dati utili per l'analisi statistica, lo scopo di questa ricerca è quello di creare un modello di analisi testuale che sia sufficientemente generico e imparziale da poter essere utilizzato al di fuori dello specifico tribunale.

La presenza di informazioni personali potrebbe rappresentare un problema, soprattutto alla luce delle attuali normative sulla privacy dei dati. Lavori come quello di Sánchez e Batet (2016) miravano a proporre modelli di privacy per la sanificazione dei documenti. Altri propongono linee guida per il trattamento

[10] SICID (Sistema Informativo Distrettuale sul Contenzioso Civile) <https://www.tribunale.napolinord.giustizia.it/documentazione/D_59023.pdf>.

dei dati personali nei documenti legali per essere conformi al GDPR come in Arajärvi e Holden (2021). Per quanto riguarda questo lavoro, questi aspetti sono coperti dal processo di anonimizzazione specifico del SICID. Inoltre, la valutazione della propensione o meno alla mediazione deve basarsi esclusivamente sugli eventi e sui fatti riportati negli atti, e non sulle caratteristiche delle parti in giudizio. Per scongiurare la formazione di tali presupposti, questa Sezione riporta la procedura intrapresa per rimuovere tali informazioni dai documenti, definendo il processo di de-identificazione. Ciò comporta la rimozione di tutte le informazioni sensibili che potrebbero portare all'identificazione di qualsiasi persona o organizzazione coinvolta nel caso. Ovviamente questo processo deve garantire che i testi rimangano comprensibili e il loro significato intatto, cioè semanticamente comprensibili. La rimozione dei dati identificativi è una parte importante del processo preparatorio per: (i) evitare potenziali distorsioni della base di conoscenze del sistema, (ii) migliorare il grado di privacy delle persone coinvolte, i cui dati sensibili meritano una particolare cura nel trattamento.

Questo processo di rimozione corrisponde al passaggio 2 di Figura 2. Questo processo non deve essere confuso con una fase di anonimizzazione poiché in questo caso occorre preservare il significato delle frasi. Per chiarire questo fatto, dobbiamo descrivere che esistono diversi modi per rimuove le identificazioni:

1. la purificazione o approccio brutale è l'approccio più semplice, che prevede la sostituzione di tutti i dati personali con un'unica etichetta, come OMISSIS (Licari e Comandè 2021). Questo metodo garantisce la totale de-identificazione; perde però del tutto l'informazione associata alla categoria a cui appartengono i dati rimossi.
2. In alternativa, l'utilizzo di etichette relative alle entità, come #PERSON o #ORGANIZATION, consente di anonimizzare e mantenere pressoché invariato il contesto della frase.
3. Come ulteriore estensione, potrebbe essere possibile assegnare delle etichette numerate con cui mantenere una distinzione anonima tra le istanze di un dato Ente, come #PERSONA1 per Mario Rossi e #PERSON2 per Luigi Verdi, e per tutte le altre identità coinvolte i documenti e per tutti i documenti con le stesse etichette.

Per la costruzione del dataset per l'analisi del testo tramite machine learning è molto importante eseguire una de-identificazione per evitare di includere dati personali, e allo stesso tempo preservarne il significato per identificare gli elementi rilevanti per la mediazione. Pertanto, è imperativo conservare le informazioni che consentono di distinguere tra entità come litigante, giudice e avvocato, preferendo così modalità di rimozione dei dati che preservino il contesto. Per questo motivo è escluso il caso 1 di forza bruta. L'aggiunta di numeri distintivi come suffissi alle etichette aumenta notevolmente la complessità di un processo di anonimizzazione automatizzata, soprattutto considerando che un file è composto da più documenti in cui deve essere mantenuta la coerenza dell'identificatore. Inoltre, ogni numero rappresenta una variazione ad un'etichetta che altrimenti sarebbe identica per tutti gli esempi/documenti,

portando ad un conseguente aumento delle caratteristiche che il modello deve apprendere. Tali caratteristiche non costituiscono una differenza rilevante ai fini dell'analisi del testo, essendo quei numeri esclusivamente utili per una migliore comprensione del contesto a favore di un essere umano. Per questi motivi anche la strategia illustrata nel caso 3 è esclusa. Queste osservazioni ci hanno portato a considerare l'uso di etichette legate al contesto (caso 2) senza numeri è stata la scelta ideale.

Secondo il caso 2, è stato sviluppato uno strumento di sostituzione supervisionata guidato dai metadati definiti in SICID (queste sono le condizioni tipiche in cui i documenti legali sono anonimizzati in Italia) direttamente nello strumento SICID fornito per ciascun tribunale in Italia. Per ogni documento una serie di etichette viene mappata sul segnaposto. Ad esempio, l'entità 'e-mail' può riferirsi a più soggetti presenti nei testi, per questo abbiamo predisposto una serie di etichette 'e-mail', una per ogni possibile soggetto: attore, convenuto, terzo, giudice, attore avvocato, difensore dell'imputato, difensore di parte, testimone, ctp/ctu (consulenti tecnici), notaio, amministratore, banca, società generica, ecc. Si applica a tutti i tipi di soggetti, come nomi e cognomi, data di nascita, luogo di nascita, codice fiscale e indirizzi di residenza; e a soggetti non strettamente legati ad un argomento, quali altri luoghi, date o codici (SSN, partita IVA, codici fiscali, codici Camera di Commercio, ecc.).

C. Etichettatura degli elementi del data set

Questa Sezione descrive la procedura di annotazione utilizzata per creare un insieme di dati supervisionati per l'addestramento, la validazione ed il test del modello (passaggio 3 della Figura 2). Per l'attività di etichettatura, è stato utilizzato Doccano[11]. Ci riferiamo da ora in poi come frase, per le porzioni di testo che vanno da un periodo all'altro. Con questo termine si intendono però anche altri casi, se ritenuto opportuno dall'esperto annotatore, ad esempio una sottostringa, che pur appartenendo ad un punto più ampio, richiede un'etichetta diversa rispetto al punto restante, o anche blocchi di testo composti da più punti contigui periodi e appartenenti alla stessa classe, e che quindi per comodità vengono etichettati insieme come un'unica frase. Le etichette adottate per annotare manualmente le frasi sono le seguenti:

- *Propensione alla mediazione (M)*: sentenze dalle quali emerge chiaramente la volontà delle parti alla partecipazione attiva al procedimento di mediazione.
- *Non propensione alla mediazione (NM)*: sentenze in cui emerge la riluttanza di almeno una delle parti a non essere coinvolta nel processo di mediazione.
- *Tecnico coinvolto*: sentenze in cui vengono menzionati i periti tecnici, quali i consulenti tecnici d'ufficio (CTU) ed i consulenti tecnici di parte (CTP). Ciò implica che una delle parti o il giudice li abbiano coinvolti.

[11] Doccano, annotazione testuale per esseri umani <https://doccano.herokuapp.com> (15-10-2023).

- *Mediazione menzionata*: sentenze in cui si rinvia al procedimento di mediazione, ed escludendo quelle in cui emerge l'orientamento negativo o positivo delle parti.
- *Neutro*: identifica tutte le altre frasi non classificate nei casi precedenti.

L'uscita del processo di etichettatura è un file per ogni dossier. Ciascuno di questi file contiene tanti segmenti quanti sono i documenti annotati nel file del caso. Ogni segmento ha le seguenti chiavi: id, testo, etichetta. Dove testo è il contenuto dell'intero documento ed etichetta è un elenco di annotazioni. Ogni annotazione è composta da un indice iniziale e finale rispetto al testo del documento e da un'etichetta. Si tenga presente che, in base alla distribuzione delle etichette assegnate alle frasi del documento, si è riscontrato un gran numero di documenti neutri, e quindi non significativi ai fini della produzione della valutazione, del suggerimento. Le porzioni di testo rilevanti ai fini dell'analisi della propensione alla mediazione sono di piccole dimensioni rispetto alla dimensione complessiva dei documenti.

D. Normalizzazione del testo

Questa Sezione descrive il passaggio 4 della Figura 2 riguardante la normalizzazione del testo. Questo passo è necessario perché il gergo utilizzato in ambito giuridico è ricco di abbreviazioni e riferimenti legali che devono essere normalizzati per evitare comportamenti incoerenti durante il successivo processo di segmentazione della frase in elementi semplici comprensibili, detti token, e anche per evitare una polarizzazione (*bias*) provocata dalle forme gergali adottate. Inoltre, le abbreviazioni forniscono anche punti/interpunzioni che possono essere interpretate erroneamente come punti di fine frase. Esistono anche diverse abbreviazioni (per stile o per errori di battitura) che in realtà si riferiscono ad un'unica versione estesa. Ad esempio, le sigle s.r.l s.r.l. srl. Srl fanno tutte riferimento alla versione estesa di 'SRL, Società a Responsabilità Limitata'. Per normalizzare le abbreviazioni è stata utilizzata della conoscenza aggiuntiva codificata con tutte le versioni e la rispettiva alternativa estesa.

L'analisi dei dati ha rivelato anche l'abbondanza di informazioni specifiche non rilevanti per l'analisi di interesse, come date, orari e altri dati numerici. Tali dati potrebbero introdurre distorsioni nella rete poiché sono contenuti anche in frasi rilevanti per decidere. Anche se contenuti in frasi neutre, la loro specificità si traduce in una ridotta somiglianza tra le frasi. Per questo motivo è stato deciso di sostituire tali occorrenze con dei segnaposto come DATE, TIME, CODE. Per quanto riguarda date e orari, abbiamo notato anche l'assenza di uno standard nella formattazione, con punti, due punti e spazi utilizzati in modo intercambiabile come separatori. Inoltre, se il numero di ore o giorni a cifra singola, lo standard non veniva soddisfatto aggiungendo uno 0 come prefisso, con conseguente modifica dei modelli. Per quanto riguarda i codici e gli importi, si è scelto di sostituire qualsiasi sequenza numerica, eventualmente intervallata da virgole, spazi e punti, che non sia già stata identificata dagli schemi di date e orari.

La fase di normalizzazione deve mantenere i simboli necessari per la separazione logica in punti, come virgole e due punti, ma deve rimuovere altri considerati non essenziali per la comprensione del testo, come virgolette e parentesi. Per normalizzare questa condizione, sono stati separati i testi in frasi, separando le frasi con ogni punto e punto e virgola. Sono quindi stati rimossi gli altri segni di punteggiatura e spazi non necessari. Inoltre, sono state ridotte e unificate più parole con un singolo alias o variante, nonché separare le parole dai segni di punteggiatura. Questo processo presenta l'ulteriore vantaggio di ridurre il numero di token in cui verrà convertito il testo. Minore è la dimensione dello spazio dei token, più facile si rivelerà l'apprendimento del modello AI basato su BERT.

E. Suddivisione e suddivisione delle frasi

Secondo il processo di normalizzazione descritto nella Sezione precedente, la fase di separazione delle frasi come passo 5 della Figura 2 viene eseguita dividendo i blocchi di testo in corrispondenza di ciascun punto e punto e virgola. Il risultato di questa fase di pre-elaborazione è un elenco di esempi costituiti da coppie del tipo <frase, etichetta di addestramento>. L'architettura BERT opera con token di input e quindi il testo è stato ulteriormente suddiviso in tokens. Con il termine Chunking ci si riferisce alla fase di pre-processing 6 della Figura 2, durante la quale il testo viene trasformato in blocchi composti da token per l'addestramento del modello di machine learning. La dimensione del blocco, o dei blocchi, deve essere al massimo di 512 token, che è un limite imposto dall'architettura BERT come meglio descritto nella Sezione V. Questo processo trasforma in serie di token le frasi in modo indipendente, in modo che ogni blocco contenga solo token derivati da una singola frase. Per definire il numero massimo di token in un blocco, è stata analizzata la distribuzione del numero di token per frasi dell'intero dataset. Analizzando questa distribuzione, il limite di 128 token ha permesso di trovare un compromesso tra la dimensione effettiva del blocco e l'occupazione del blocco.

F. Suddivisione Formazione-Convalida-Test

L'elevata mole delle frasi neutre, unita alla bassa quantità di esempi per le classi interessate, ci ha portato a decidere di ridurre la dimensione dei set di validazione e di test per favorire un training set più bilanciato. Questa scelta è stata motivata anche per il fatto che si è previsto un secondo livello di validazione del modello. Come descritto nella Sezione 8, è stata eseguita una seconda convalida utilizzando nuovi dati quando gli esperti legali hanno utilizzato la soluzione e hanno fornito anche i loro commenti e valutazioni utilizzando il software sviluppato. Come descritto nella Sezione 4.C, il dataset è stato etichettato dagli esperti a livello di casistica di singola frase. Come riportato nella Tabella 1, la maggior parte delle frasi sono state etichettate come neutre come previsto fin dall'inizio. Considerando un fascicolo generico, documenti come le citazioni in giudizio delle parti non contengono – per loro natura – alcuna informazione utile circa la pro-

pensione alla mediazione. Anche quando un documento contiene informazioni rilevanti, queste sono sempre limitate a un numero relativamente piccolo di frasi dell'intera lunghezza del documento. Poiché il numero di esempi neutrali costituisce oltre il 90% del set di dati, l'utilizzo dell'intero set di dati per la formazione risulterebbe in un modello sbilanciato, fornendo risultati soddisfacenti solo per la classe neutra. Al contrario, il nostro obiettivo è produrre un modello in grado di classificare correttamente le affermazioni di propensione e non propensione. Pertanto, è stato prodotto un training set ri-bilanciato, mantenendo 1800 frasi neutre sulle 14115 totalmente disponibili scelte casualmente, mentre tutte le altre classi sono rimaste invariate. Il numero di esempi neutri da conservare, K, è stato determinato come segue. Partendo dal principio di bilanciamento secondo cui il numero di esempi appartenenti alla classe K più popolosa dovrebbe essere pari a 10 volte il numero di esempi della classe meno frequente, poniamo K = 600 come limite per il numero di esempi. È tuttavia fondamentale considerare che l'elevato numero di frasi neutre non solo è rappresentativo della reale composizione dei documenti, ma contiene anche un'elevata variabilità tipica del linguaggio naturale. Il dataset finale è stato composto da 1800 esempi della classe neutra, equivalenti al 12,75% del numero totale di neutrali originariamente disponibili e 30 volte il numero di esempi nella classe più piccola. I dettagli sul dataset ribilanciato risultante da questa riduzione sono riportati nella Tabella 1. Il dataset complessivo è stato suddiviso in dataset di training, validazione e test in base a percentuali 80-10-10 (risultando così nel passaggio 6 della Figura 2).

5. Sviluppo e Validazione del modello di AI

Questa Sezione contiene i dettagli sul training del modello di AI per classificare le frasi nelle 5 classi sopra menzionate. Le classificazioni delle frasi hanno permesso di avere una granularità di classificazione simile a quella che si otterrebbe dall'analisi logica del linguaggio naturale (Sottosezione 5.B). Considerando che un file è composto da più documenti con contenuti molto diversi, è stato utile implementare un meccanismo per mappare i risultati in una classificazione a livello di documento (Sezione 6). Le classificazioni delle frasi sono state valutate anche utilizzando le tecniche Shap di XAI (vedere Sezione 6) (Lundberg e Lee 2017).

A.Sviluppo del modello

I modelli NLP tradizionali sono spesso addestrati utilizzando set di dati etichettati su attività specifiche, richiedendo una quantità rilevante di dati etichettati per ciascuna attività e un gran numero di risorse. La tecnica di AI dei Transformer (Vaswani et al. 2017) usa un'architettura encoder-decoder con un meccanismo di Attention per catturare le dipendenze tra le diverse parti della sequenza di input. Ponendo l'attenzione su parti rilevanti dell'input, i Transformer possono modellare efficacemente le dipendenze a lungo raggio e acquisire informazioni contestuali in modo più efficace rispetto ai modelli precedenti, rendendoli particolarmente utili per la classificazione del testo. Il modello Transformer di Devlin

et al., nel 2018, introduce l'approccio BERT che sfrutta il contesto bidirezionale per acquisire una comprensione più completa del testo, a differenza dei modelli precedenti che si basavano prevalentemente sulla modellazione linguistica unidirezionale. L'innovazione di BERT risiede nel suo approccio di pre-training e fine-tuning. In Figura 3 è riportato uno schema dell'architettura BERT. Durante la fase di pre-training, il modello è stato addestrato su corpora su larga scala utilizzando un obiettivo di modellazione del linguaggio di tipo *masked* (MLM) (Vaswani et al. 2017). In questo processo, una certa percentuale di token di input viene mascherata in modo casuale e il modello ha il compito di prevedere questi token mascherati in base al contesto circostante. Addestrandosi su grandi quantità di dati di testo, BERT apprende rappresentazioni ricche che catturano caratteristiche semantiche e sintattiche profonde. In sostanza, ciò consente al modello di apprendere rappresentazioni/pattern/modelli dal linguaggio naturale in ingresso tramite esempi. Queste rappresentazioni permettono di modellare una lingua in modo generale, apprendendo le relazioni semantiche tra parole e strutture specifiche di una determinata lingua. La fase di pre-training è seguita da un perfezionamento su compiti specifici come l'analisi del sentiment o il riconoscimento delle entità nominate, solo per citarne alcuni.

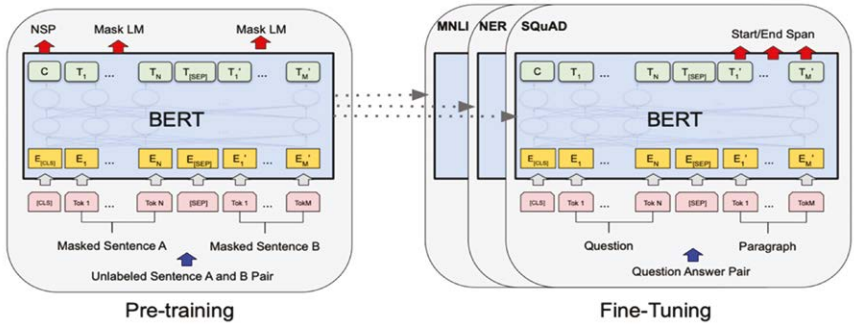


Figura 3 – Procedura tipica generale di pre-addestramento e messa a punto del BERT.

Per il nostro scopo, è stato adottato un approccio di perfezionamento incentrato sulla creazione di un modello per la classificazione del testo (vedere Figura 4). A tal fine, come modello pre-addestrato è stato utilizzato il modello italiano BERT XXL Cased (Chung et al. 2021) (che tine conto delle maiuscole), che è stato addestrato su testi italiani provenienti dalle raccolte di dati di Wikipedia[12], OPUS[13] e del progetto OSCAR[14]. Il corpus formativo finale era di

[12] Wikipedia <https://en.wikipedia.org/wiki/Main_Page>.

[13] OPUS è una raccolta crescente di testi tradotti dal web <https://opus.nlpl.eu> (15-10-2023).

[14] Il progetto OSCAR (Open Super-large Crawled Aggregated coRpus) è un progetto open source che mira a fornire risorse e set di dati multilingue basati sul web per applicazioni di machine learning (ML) e intelligenza artificiale (AI) <https://oscar-project.org> (15-10-2023).

81GByte comprendente oltre 13 miliardi di token. È stata utilizzata la versione Cased anziché quella Uncased poiché la prima non rimuove le lettere maiuscole e gli accenti. Nella lingua italiana le maiuscole si usano solo all'inizio delle frasi e per i nomi propri e, data l'unicità del contesto giuridico, tutti i sostantivi sono stati sostituiti con etichette anonime (si tenga presente che in alcuni cognomi italiani sono presenti anche sostantivi generici o aggettivi, ad esempio: Rossi, Bellini; la maiuscola può aiutare). Tali etichette sono semplicemente parole maiuscole molto comuni alle quali è anteposto un simbolo cancelletto, ad esempio #GIUDGE e #AVVOCATO. Anche saper riconoscere le parole accentate può essere fondamentale per comprendere una frase; basti pensare, ad esempio, che solo un accento distingue la parola 'e', cioè congiunzione finale, da 'è' terza persona singolare del verbo essere.

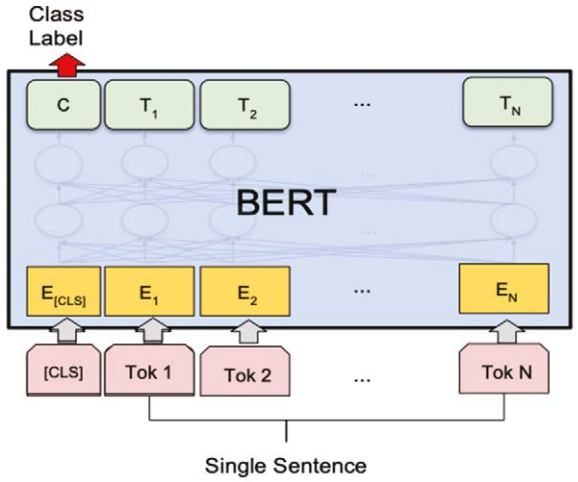


Figura 4 – Ottimizzazione del BERT sulla classificazione di singole frasi.

Gli iperparametri del modello italiano BERT XXL Cased pre-addestrato sono: 12 attention-heads, 768 come dimensione degli strati nascosti, 12 strati nascosti, 512 come numero massimo di token, un vocabolario di 31102 possibili token. BERT utilizza uno strumento di trasformazione in token per preparare l'input testuale, detto tokenizzatore. Il tokenizzatore divide le istruzioni in modo da avere una parola per token o in parti di parole in cui una parola viene suddivisa in più token. Un'altra limitazione del modello pre-addestrato adottato impone una dimensione fissa per gli ingressi, e quindi questi devono essere regolarizzati introducendo frasi di riempimento e/o troncamento. Nel modello pre-addestrato il numero di token è stato fissato a 128, per le motivazioni riportate nella Sezione 4.E. Una iperparametrizzazione (procedura per la scelta dei parametri migliori) per massimizzare il valore della metrica F1 (che risulta essere un indice di qualità del risultato) sulla base della validazione è stata inclusa nel processo di fine-tuning. L'intervallo degli iperparametri è riportato nella

Tabella 2. I migliori risultati sono stati ottenuti con un learning-rate di 3,15E-05, un decay factor di 7,85E-03 e una dimensione del batch di 16; ottenendo un valore di F1 di 0,944, dove il massimo è 1.

Tabella 2 – Gamma di iperparametri per la messa a punto del modello BERT.

Iperparametro	*Cerca dominio*
Early Stopping	15
Learning Rate	da 1e-6 a 1e-3
Weight Decay	da 0,005 a 0,01
Batch Size	[2,4,8,16]

Nella Figura 5, l'andamento di F1, della precisione e della capacità di richiamo (recall) nella classificazione sono riportati in funzione dell'epoca, iterazioni del processo. Secondo il grafico il miglior valore in F1 è risultato essere all'undicesima epoca.

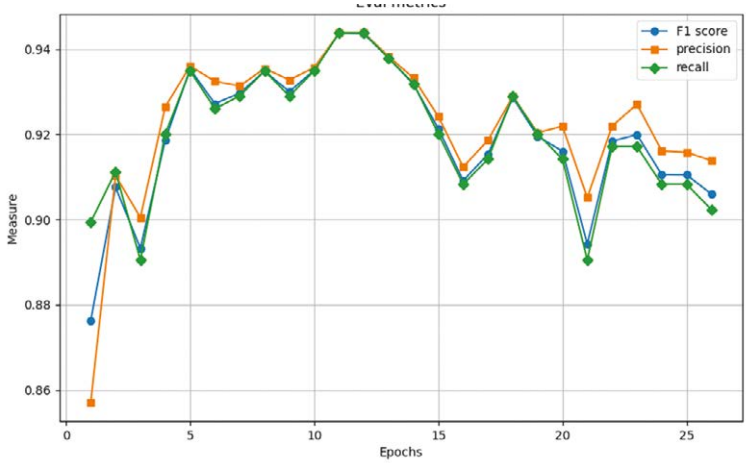


Figura 5 – Andamento di F1, precisione e recall in funzione del numero delle epoche in fase di allenamento/validazione.

B. Sviluppo del modello

Tutto ciò è possibile sulla base dei risultati forniti dal modello di classificazione a livello di frase. Nello specifico, per ogni frase del documento, sono stati utilizzati i punteggi probabilistici in ciascuna classe prodotti dalla classificazione del modello. Come primo passo, per ogni frase, vengono sommati i punteggi delle classi non caratterizzanti: in altre parole, i punteggi delle classi 'neutro', 'tecnico' e 'mediazione menzionata' vengono tutti sommati in 'neutro_sum'.

Questo raggruppamento è giustificato dal fatto che il contenuto delle frasi classificate come 'tecnico' e 'mediazione menzionata', pur essendo più rilevante di un generico neutro, non è sufficientemente distintivo per essere utilizzato durante questa analisi automatica.

Il modello perfezionato presentato nella Sezione 5.A è stato valutato sul set di test (vedere Tabella 1). Il modello ha classificato correttamente 321 delle 328 frasi di test. Le metriche di precisione, recall e F1 calcolate a livello di singola classe sono riportate nella Tabella 3. Considerando come valori di classe 'Negativo', 'Neutro', 'Positivo', richiamando come risultato le definizioni di Vero Positivo (TP) dove il modello identifica correttamente la classe, Falso Positivo (FP) come risultato in cui il modello identifica erroneamente la classe considerata, e Falso Negativo (FN) come risultato in cui il modello identifica erroneamente la classe non considerata, sono state calcolate le seguenti metriche per ogni classe:

Tabella 3 – Risultati della valutazione a livello di classe sul set di test.

classi	Precisione	Richiamare	F1
Propensione alla mediazione	1.000	0,923	0,956
Non propensione alla mediazione	0,500	1.000	0,667
Neutro	0,959	0,989	0,974
Tecnico coinvolto	1.000	0,915	0,956
La mediazione menzionata	0,944	0,864	0,903
Globale	0,958	0,950	0,952

È stata stimata una valutazione globale utilizzando il punteggio medio ponderato calcolato prendendo la media di tutti i punteggi per classe, considerando il supporto di ciascuna classe (vedere l'ultima riga della Tabella 3). Il supporto di una classe è definito come il numero di occorrenze vere per la classe. Il 'peso' è la proporzione di ciascun supporto di ciascuna classe rispetto alla somma di tutti i supporti. Con la media ponderata, l'uscita tiene conto della distribuzione dei casi per ciascuna classe ponderata in base al numero di istanze di una determinata classe. Ciò è particolarmente utile nei casi in cui i dati di apprendimento sono sbilanciati tra le categorie come nel nostro caso. Inoltre, la metrica di Accuratezza è stata stimata come rapporto tra il numero totale di classificati correttamente rispetto al numero di sentenze, e in questo caso l'Accuratezza è risultata essere del 94,9%.

Dai risultati, si può osservare che il modello produce ottimi risultati F1-score. Tenendo presente che l'obiettivo del sistema è quello di assistere il giudice nell'individuazione degli elementi di propensione e di non propensione, si ritiene che queste ultime due classi siano quelle più significative su cui focalizzare l'attenzione della valutazione. In particolare, è estremamente importante che il modello faciliti e acceleri l'identificazione di quelle poche frasi rilevanti (per quelle classificazioni) che sono tipicamente contenute in documenti lunghi e prevalentemente poco rilevanti. Per questo motivo è importante che il modello

presenti una recall elevata per queste due classi, come infatti accade con 0,923 per la propensione alla mediazione e 1,0 per la non propensione alla mediazione.

Sulla base dei risultati previsti è stata creata una matrice di confusione sulle 5 classi considerate e sono state calcolate metriche classiche come precisione, recall e F1 che aggregano tutte le classi (vedere Figura 6). Secondo la matrice di confusione è possibile osservare che per la classe non incline alla mediazione vengono commessi 6 errori, di cui 4 comportano un'errata classificazione verso la classe di mediazione citata. Si possono derivare le seguenti due considerazioni:

- il numero di errori commessi è relativamente basso rispetto al numero totale di frasi analizzate. Ciò significa che se un utente avesse voluto identificare tutte le frasi che esprimevano un sentimento di riluttanza, avrebbe potuto identificarle tutte analizzando solo 12 frasi (le 6 identificate correttamente con recall 1.0 e le 6 errate) invece di analizzarne 338 (il numero totale di frasi nel set di dati);
- gli errori commessi riguardano principalmente la classe di propensione alla mediazione, i cui esempi sono in realtà simili a quelli della classe di non propensione.

Mentioned mediation	51	4	4	0	0
Neutral	1	186	1	0	0
Not propensity to Mediate	0	0	6	0	0
Propensity to Mediate	0	0	1	12	0
Technician involved	2	4	0	0	65
	Mentioned mediation	Neutral	Not propensity to Mediate	Propensity to Mediate	Technician involved

Figura 6 – Matrice di confusione calcolata sul set di test. I valori effettivi sono quelli riportati sull'asse X.

6. Spiegabilità dei risultati, XAI

Uno strumento di supporto alle decisioni deve anche motivare i suggerimenti/valutazioni calcolate, spiegando così all'utente i risultati delle previsioni prodotte dal sistema identificato come R7 (dovrebbe identificare frasi significative per le classi di interesse M e NM, R8). A tal fine, abbiamo adattato una tecnica XAI per identificare e fornire evidenze nel testo che hanno contribuito alla produzione della classificazione. Secondo l'approccio BERT, le caratteristiche adottate sono le parole nel testo tramite Shapely (SHAP) (Lundberg e Lee 2017), sia

nella versione di spiegazioni locali che nella versione globale. La forma globale mira a identificare le parole più influenti nel dare una classificazione secondo il modello, mentre la forma locale ad ottenere spiegazioni sulle affermazioni/testi inviati in inferenza e sul perché di ciascuna classificazione a livello frase.

A. Spiegabilità globale

La spiegazione globale mira a rappresentare l'impatto complessivo di alcune caratteristiche sul modello di classificazione sviluppato in termini di importanza rispetto alla classe determinata da parole specifiche nelle frasi considerate. Per questa analisi è stato considerato il test set riportato nella Sezione 4.F. La libreria XAI SHAP assegna un valore SHAP che rappresenta una quantificazione del contributo alla classificazione della sentenza in una delle classi considerate. Questo valore può essere positivo (la parola ha contribuito positivamente alla classificazione della classe determinata) oppure negativo. Di conseguenza, per ciascuna classe, sono stati calcolati i valori SHAP relativi alle parole contribuite alla classificazione. I risultati sono riportati nella Figura 7 in cui le prime 10 parole più rilevanti per classe sono riportate in una nuvola di parole. La nuvola di parole è caratterizzata da una rappresentazione delle parole in termini di colore e dimensione. Più grande è la parola, più importante è il valore SHAP associato. Il colore rappresenta la classe associata. I risultati riportati in Figura 7 sono utili per comprendere il funzionamento del modello, ad esempio la presenza determinante delle parole chiave CTP/CTU nella classificazione di un'affermazione riferita al consulente tecnico, che nella maggior parte dei casi impone la decisione ad un ulteriore esame tecnico analisi della controversia.

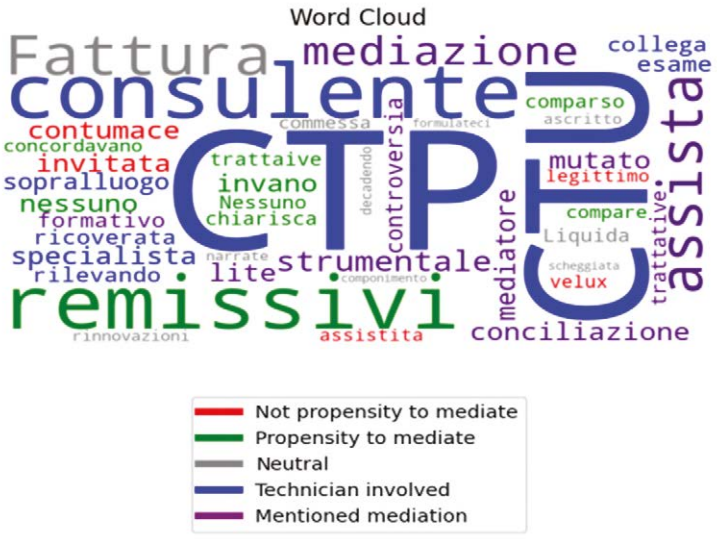


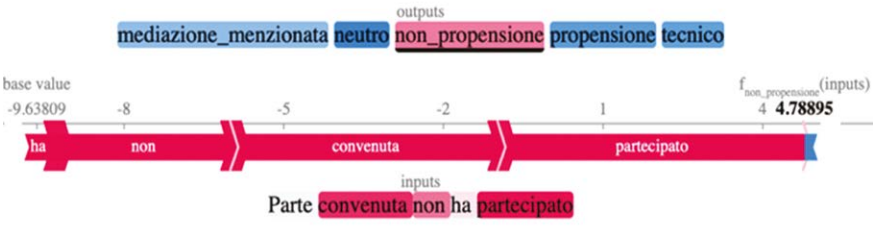
Figura 7 – Feature cloud, in lingua italiana.

B. Spiegabilità locale

Oltre all'interpretazione globale del modello prodotto rispetto all'intero test set, una valutazione specifica può essere effettuata nel momento in cui si adotta il modello per classificare la singola frase. In questo caso, le parole più rilevanti identificate da valori SHAP più alti sono quelle che maggiormente hanno influenzato la decisione/classificazione suggerita, positivamente o negativamente. Questo è l'approccio locale di SHAP, ogni frase viene analizzata in modo indipendente e ogni parola nella frase corrisponde a una caratteristica. Ciò fornisce la prova della conformità con R5.

Dal set di test, sono stati selezionati 2 esempi chiave riguardanti frasi classificate come 'propensione alla mediazione' e 'non propensione alla mediazione' per approfondire il funzionamento del modello sviluppato. I grafici generati e riportati in Figura 8 sono utili per: (i) valutare i meccanismi del modello, (ii) comunicare ai decisori quali sono le parole chiave che hanno contribuito maggiormente alla classificazione prodotta dal modello AI, in ciascuna specifica classe (e in particolare nella classe identificata come più probabile individuata dal modello/classificatore). Questi grafici sono proposti all'utente del sistema pe ogni frase analizzata.

(i) Propensione alla mediazione.



(ii) Non propensione alla mediazione.

Figura 8 – Diagrammi SHAP locali relativi a due esempi di spiegabilità. Esempio in alto (i) per 'propensione alla mediazione' e in basso per (ii) 'non propensione alla mediazione'.

All'interno di ogni frase, le parole in rosso sono associate positivamente all'etichetta scelta dal modello, e in blu quelle che hanno contribuito alla classificazione in un'altra classe. Si può osservare l'intensità del colore con cui vengono evidenziate le caratteristiche: maggiore intensità corrisponde a maggiore rilevanza, cioè a maggiore peso (dimensione della barra freccia sottostante) con cui la parola influenza il risultato del modello per la frase in esame. Per il caso (i), l'affermazione analizzata è stata «Il difensore di parte convenuta si dichiara disponibile a una soluzione conciliativa». Lo XAI ha evidenziato come veramente importante ai fini della classificazione della 'propensione alla mediazione' la dicitura «si dichiara disponibile a una soluzione» che indica la propensione alla mediazione. Nel caso (ii), la frase «parte convenuta non ha partecipato» in cui la visualizzazione esplicativa giustificava la non propensione alla mediazione, disinteresse per la procedura. Questa funzionalità XAI presenta due vantaggi principali, uno per la parte di sviluppo del progetto per comprendere il funzionamento del modello AI e il secondo per l'utente finale per comprendere meglio la decisione supportata. Questo strumento XAI è stato integrato nel prototipo del sistema di supporto alle decisioni sviluppato i cui dettagli sono riportati in una sezione successiva.

7. Valutazione del suggerimento a livello di documento

La descrizione riportata nella sezione precedente si è concentrata sulla classificazione delle singole dichiarazioni tra quelle che figurano nei diversi documenti relativi ad un caso di contenzioso. Tra i requisiti individuati, la capacità di classificare i documenti del caso contenzioso (R3) può essere molto utile per guidare il tribunale/i giudici verso i documenti rilevanti alla determina. La classificazione a livello di documento deve essere secondo le classi: propensione alla mediazione (M), non propensione alla mediazione (NM), e neutrale (N). Inoltre, abbiamo mirato a fornire un punteggio di confidenza sulla classificazione prodotta (R4).

Come sopra accennato, la presenza nel documento di affermazioni classificate come 'propensione alla mediazione' e 'non propensione alla mediazione' è sporadica. Poiché lo scopo della classificazione dei documenti è identificare quei documenti che possono aiutare a prendere la decisione, le classi 'Neutro', 'Tecnico coinvolto' e 'Mediazione menzionata' sono considerate come 'somma_neutra'. Questo raggruppamento è giustificato dal fatto che le frasi classificate come 'Tecnico coinvolto' e 'Mediazione menzionata' sono di fatto affermazioni non orientate e quindi Neutre. Pertanto, se almeno una frase è classificata come propensione o non propensione, consideriamo solo quelle frasi per classificare il documento. Utilizziamo quindi i punteggi di queste frasi per calcolare la media ponderata nelle tre categorie: NM, M e N. A seconda dei risultati dei punteggi per ciascuna classe/raggruppamento, distinguiamo tra i seguenti casi:

- *nessuna frase del documento è stata classificata né come propensione né come non propensione*. Pertanto, tutte le frasi del documento vengono classificate come 'Neutro', 'Tecnico coinvolto' o "'Mediazione menzionata', e la classi-

ficazione del documento può essere stimata sulla base della media ponderata di queste classi.
- *Almeno una frase è classificata come propensione o non propensione*, le frasi verranno considerate per classificare il documento orientate alla propensione alla mediazione (M), non alla propensione alla mediazione (NM), ovvero 'somma neutra', e la classificazione e confidenza del documento sarà stimata sulla base della media ponderata dei corrispondenti dichiarazioni in una delle 3 classi.
- *Il risultato ottenuto è per entrambi i casi un insieme di tre medie ponderate*, che vengono trasformate in percentuali e rappresentano la confidenza della classificazione del documento.

8. Strumento XAI4MA per il supporto alle decisioni

In questa sezione vengono presentati i sistemi di supporto alle decisioni consegnati al Tribunale di Firenze e denominati XAI4MA (Explainable Artificial Intelligence tool for Mediation Agile). Come descritto nella Figura 1, lo strumento sviluppato sfrutta: (i) il modello BERT per la classificazione a livello di sentenze i risultati dell'approccio XAI in Shapely per fornire supporto in tribunale e ai giudici per prendere una decisione sull'invio o meno di una controversia alla mediazione. Secondo le regole imposte dal GDPR per tutelare la privacy dei dati dei cittadini europei, viene eseguita un'operazione di anonimizzazione estesa e dettagliata sui dati sensibili contenuti negli atti giudiziari prima di poterli trattare. Infatti, né il servizio di analisi né il processo di apprendimento del modello necessitano di dati personali delle parti in causa. XAI4MA accetta i documenti anonimizzati provenienti da SICID che in Italia è uno strumento in dotazione a tutti i tribunali. Questa scelta ci ha permesso di soddisfare R6 e di garantire tempi rapidi di elaborazione dei documenti.

L'architettura di XAI4MA è riportata nella Figura 9. Questa, permette di elaborare i dati anonimizzati provenienti da SICID, così come qualsiasi altro tipo di documento che il tribunale o il giudice desideri valutare. Attraverso XAI-4MA, gli utenti possono:
- caricare documenti testuali anonimizzati da analizzare, tipicamente quelli prodotti dallo strumento SICID del Ministero. Si tratta semplicemente della possibilità di valutare documenti separati, vedere R2.
- Caricare documenti non anonimizzati che vengono anonimizzati secondo lo standard SICID del governo italiano, vedi R6. I documenti ricevuti e i risultati prodotti sono gestiti nella piattaforma Snap4City che è conforme al GDPR (Badii et al. 2020), (R7).
- Ricevere risultati con suggerimenti di XAI sulle classificazioni a livello di frase.
- Visualizzare e stampare su carta i risultati delle classificazioni sia a livello di frase che a livello di documento. Da utilizzare durante la discussione in tribunale con i giudici (R9, R10).
- Fornire una classificazione a livello di documento, per aiutare i giudici a identificare i documenti e le dichiarazioni più rilevanti che possono essere

utilizzate per prendere una decisione sull'invio o meno della controversia alla mediazione (R9, R10).
- Facoltativamente, gli utenti possono valutare i risultati forniti, e quindi validare i risultati del modello, confermando o proponendo una modifica alle classificazioni fornite a livello di sentenza.
- Organizzare i documenti in fascicoli, in modo simile al sistema SICID, vedere R1.

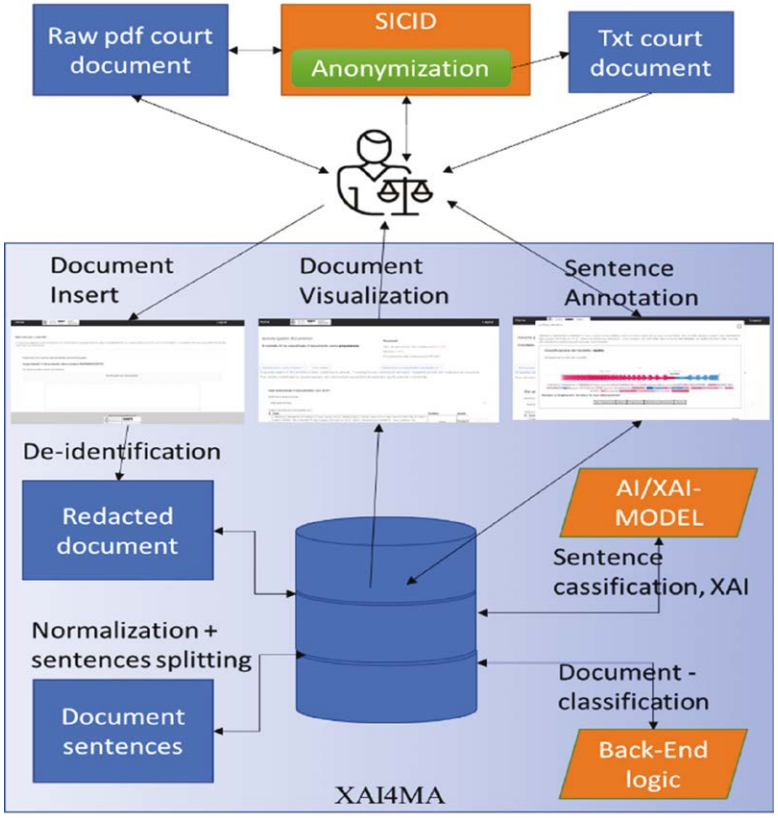


Figura 9 – Architettura XAI4MA, basata sull'infrastruttura Snap4City (Badii et al. 2020; Garau et al. 2020).

L'architettura XAI4MA è un'applicazione multiutente in cui è possibile utilizzare il front-end del sistema per ricevere documenti/testi. In XAI4MA, i documenti e le sentenze elaborate sono archiviati in un database. L'elaborazione da parte del modello AI/XAI genera le classificazioni e le spiegazioni a livello di frase e una logica back-end di elaborazione gestisce la classificazione finale del modello come riportato nella Figura 10.

In particolare, il processo si completa utilizzando un'interfaccia di monitoraggio per accedere al documento con i risultati della classificazione nella parte in alto a destra dell'interfaccia riportati tramite le percentuali delle classi, e la spiegabilità del modello AI con la relativa classe nel riquadro al centro dell'interfaccia utente della Figura 10.

XAI4MA è in grado di raccogliere i documenti forniti da elaborare, salvare la valutazione proposta e raccogliere eventuali correzioni e giudici degli esperti utilizzando lo strumento. Gli strumenti XAI4MA sono stati utilizzati con successo da diversi esperti per eseguire un'ulteriore convalida. A tal fine sono stati valutati 25 nuovi documenti per un totale di 6.060 nuove sentenze (di cui oltre 5.600 classificate neutre dagli esperti). In questa ulteriore validazione, la precisione ponderata globale è stata di 0,99, la recall di 0,97 e un F1-score di 0,98. L'accuratezza ottenuta su questi nuovi dati è del 97%, andando oltre le aspettative della Tabella 3.

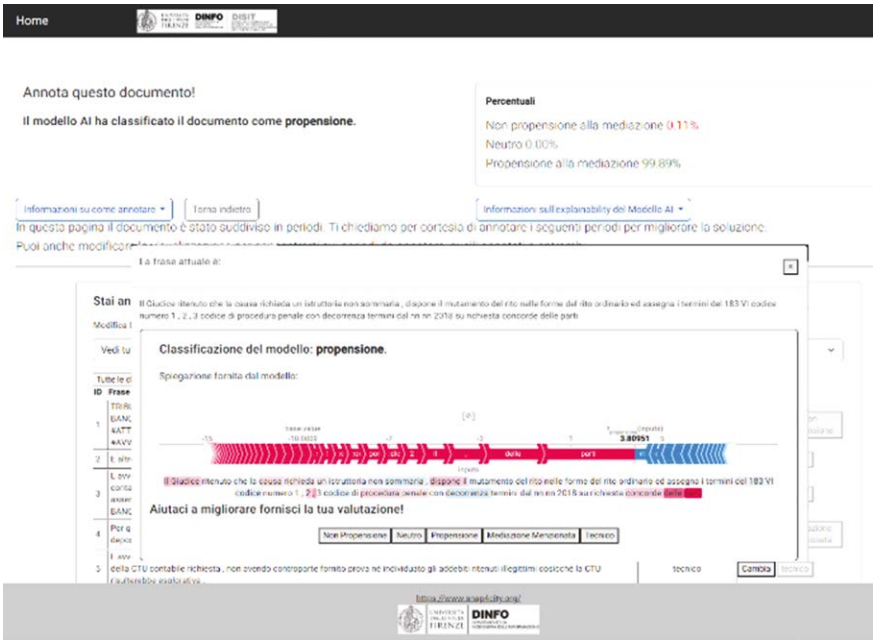


Figura 10 – Pagina del documento principale XAI4MA, con risultati XAI.

9. Conclusioni

Il *disposition time* del sistema giudiziario italiano è uno tra i più alti in Europa. Nei processi civili, l'adozione di un processo di mediazione offre la possibilità di una risoluzione rapida della disputa, consentendo alle parti coinvolte di concludere amichevolmente le controversie al di fuori delle formalità delle

procedure giudiziarie. Nel caso in cui la mediazione non venga raggiunta, questo comporta un aumento nel tempo complessivo per la risoluzione del caso specifico. La decisione se mandare in mediazione o meno un determinato caso, spetta ai giudici/tribunali dopo aver esaminato un'ampia documentazione, spesso composta da centinaia di pagine e materiale giuridico vario, sulla base di poche sporadiche dichiarazioni. Per affrontare questa sfida, il presente studio introduce una soluzione basata su algoritmi di intelligenza artificiale, sotto forma di un innovativo sistema di supporto alle decisioni noto come XAI4MA (Explainable Artificial Intelligence tool for Mediation Agile). Questo strumento non solo facilita la valutazione delle prospettive di mediazione con una precisione del 97% a livello di frase e, cosa ancora più significativa, attraverso l'utilizzo di XAI, chiarisce le clausole e i segmenti specifici all'interno dei documenti che hanno influenzato in modo significativo il processo decisionale. Il sistema proposto non solo potrebbe aiutare i giudici nel processo decisionale finale, ma tiene conto anche del 'diritto a una spiegazione' richiesto dal GDPR per quanto riguarda i risultati da soluzioni di AI. Il sistema si prende cura anche del processo di de-identificazione dei dati verso la generalizzazione dell'applicabilità della soluzione sviluppata e dei requisiti riguardanti gli aspetti legati alla privacy sempre in conformità al GDPR[15].

Riferimenti bibliografici

Aletras, N., Tsarapatsanis D., Preoțiuc-Pietro D., e V. Lampos. 2016. "Predire le decisioni giudiziarie della Corte europea dei diritti dell'uomo: una prospettiva di elaborazione del linguaggio naturale." *PeerJ Informatica* 2: e93. https://doi.org/10.7717/peerj-cs.93

Alghazzawi, D., Bamasag O., Albeshri A., Sana I., Ullah H., e M. Z. Asghar. 2022. "Previsione efficiente delle sentenze dei tribunali utilizzando un modello di rete neurale LSTM+ CNN con un set di funzionalità ottimale." *Matematica* 10, 5: 683.

Arajärvi, N., e L. Holden. 2021. *Linee guida conformi al GDPR per il trattamento dei dati personali nei documenti legali*.

Badii, C., Bellini P., Difino A., e P. Nesi. 2020. *Piattaforma IoT per smart city che rispetta gli aspetti di privacy e sicurezza del GDPR*. Accesso IEEE, 8, 23601-23623.

Branting, L. K., Pfeifer C., Brown B., Ferro L., Aberdeen J., Weiss B., e B. Liao. 2021. "Previsione giuridica scalabile e spiegabile." *Intelligenza artificiale e diritto* 29: 213-38.

[15] Gli autori desiderano ringraziare Giustizia Agile, il progetto nazionale Giustizia Agile e i partner (progetto Giustizia Agile del PON Governance e Capacità Istituzionale <https://www.unitus.it/it/unitus/mappatura-della-ricerca/articolo/giustizia-agile>). Il progetto è stato finanziato dal Ministero della Giustizia italiano, con una collaborazione tra università e tribunali per l'ASSE I, Obiettivo Specifico 1.4, Azione 1.4.1 del Programma Operativo Nazionale (PON) Governance e Capacità Istituzionale 2014-2020, con il finanziamento di scopo di produrre una migliore organizzazione della macchina giudiziaria. Un sentito ringraziamento alla prof.ssa Paola Lucarelli per averci guidato e introdotto nel mondo della mediazione, e ai tanti esperti che hanno contribuito allo sviluppo dei dati annotati, e per le valutazioni generali della soluzione. Snap4City (<https://www.snap4city.org>) è una tecnologia aperta e una ricerca del DISIT Lab, Università di Firenze, Italia.

Brennan, T., Dieterich W., e B. Ehret. 2009. "Valutazione della validità predittiva del sistema di valutazione dei rischi e dei bisogni Compas." *Giustizia penale e comportamento* 36, 1: 21-40. https://doi.org/10.1177/0093854808326545

Brundage, M., Avin S., Clark J., Toner H., Eckersley P., Garfinkel B., e D. Amodei. 2018. *L'uso dannoso dell'intelligenza artificiale: previsione, prevenzione e mitigazione*. arXiv prestampa. arXiv:1802.07228.

Butterworth, M. 2018. "L'ICO e l'intelligenza artificiale: il ruolo dell'equità nel quadro del GDPR." *Revisione di diritto informatico e sicurezza* 34, 2: 257-68.

Chung, Y. A., Zhang Y., Han W., Chiu C. C., Qin J., Pang, R., e Y. Wu, Y. 2021. *W2v-bert: Combinazione di apprendimento contrastivo e modellazione linguistica mascherata per la pre-formazione vocale autosuperata. Nel 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 244-50. IEEE <https://huggingface.co/dbmdz/bert-base-italian-xxl-cased>.

Collenette, J., Atkinson K., e T. Bench-Capon. 2023. "Strumenti di intelligenza artificiale spiegabili per il ragionamento legale sui casi: uno studio sulla Corte europea dei diritti dell'uomo." *Intelligenza artificiale* 317: 10386. https://doi.org/10.1016/j.artint.2023.103861

Cortes, C., e V. Vapnik. 1995. "Reti di vettori di supporto." *Apprendimento automatico* 20, 3: 273-97.

Csányi, G. M., Nagy D., Vági R., Vadasz J. P., e T. Orosz. "Sfide e problemi aperti dell'anonimizzazione dei documenti legali." *Simmetria* 13: 1490. https://doi.org/10.3390/sym13081490

Deeks, A. 2019. "La richiesta giudiziaria di un'intelligenza artificiale spiegabile." *Columbia Law Review* 119, 7: 1829-50.

Devlin, J., Chang M.-W., Lee K. e K. Toutanova. 2018. *BERT: Pre-formazione di trasformatori bidirezionali profondi per la comprensione del linguaggio* (arxiv:1810.04805).

Garau, C., Nesi P., Paoli I., Paolucci M., e P. Zamperlin. 2020. "Una piattaforma big data per città intelligenti e sostenibili: casi di studio sul monitoraggio ambientale in Europa." In *Conferenza internazionale sulla scienza computazionale e le sue applicazioni*, 393-406. Cham: Springer International Publishing.

Geraghty, Kate Anya, e Jessica Woodhams. 2015. "La validità predittiva degli strumenti di valutazione del rischio per le donne delinquenti: una revisione sistematica." *Aggressività e comportamento violento* 21: 25-38. https://doi.org/10.1016/j.avb.2015.01.002.

Goodman, B., e S. Flaxman. 2017. "Regolamenti dell'Unione europea sul processo decisionale algoritmico e un 'diritto alla spiegazione'." *AI Magazine* 38, 3: 50-7. https://doi.org/10.1609/aimag.v38i3.2741

Gu, J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., e T. Chen. 2018. "Recenti progressi nelle reti neurali convoluzionali." *Riconoscimento di modelli* 77: 354-77.

Hassan, F., Domingo-Ferrer J., e J. Soria-Comas. 2018. *Anonimizzazione dei dati non strutturati tramite riconoscimento di entità nominate. Decisioni di modellazione per l'intelligenza artificiale.*

HLEG, A. 2019. *Linee guida etiche per un'intelligenza artificiale affidabile*. Gruppo di esperti ad alto livello sull'intelligenza artificiale, 8, della Commissione europea.

Hochreiter, S., e J. Schmidhuber. 1997. "Memoria lunga a breve termine." *Neural Computation* 9, 8: 1735-80.

Hsun-Ping, Hsieh, Jiawei Jiang, Tzu-Hsin Yang, Renfen Hu e Cheng-Lin Wu. 2022. "Predire il successo delle richieste di mediazione utilizzando le proprietà del caso e

le informazioni testuali per ridurre l'onere a carico del tribunale." *Cifra. Governatore: Ris. Pratica* 2, 4, articolo 30: 1-18. https://doi.org/10.1145/3469233

Katz, Daniel, Bommarito II Michael J., e Josh Blackman. 2016. "Un approccio generale per prevedere il comportamento della Corte Suprema degli Stati Uniti." *PLOS ONE* 12.10.1371/journal.pone.0174698.

Licari, D., e G. Comandè. 2021. "ITALIAN-LEGAL-BERT: un modello linguistico trasformatore pre-addestrato per il diritto italiano." Atti del Knowledge Management for Law Workshop (KM4LAW), 26 settembre 2022, Bolzano, Italia.

Lundberg, S. M., e S.-I. Lee. 2017. "Un approccio unificato per interpretare le previsioni dei modelli". *Progressi nei sistemi di elaborazione delle informazioni neurali 30*.

Medvedeva, M., Vols M., e M. Wieling. 2020. "Utilizzo del machine learning per prevedere le decisioni della Corte Europea dei Diritti dell'Uomo." *Intelligenza artificiale e diritto* 28: 237-66.

Medvedeva, M., Wieling M., e M. Vols. 2023. "Ripensare il campo della previsione automatica delle decisioni giudiziarie." *Intelligenza artificiale e diritto* 31, 1: 195-212.

Richterich, A. 2018. *L'agenda dei big data: etica dei dati e studi critici sui dati*, 154. Stampa dell'Università di Westminster.

Sánchez, D., e M. Batet. 2016. "C-sanitized: un modello di privacy per la redazione e la sanificazione dei documenti." *Giornale dell'Associazione per la scienza e la tecnologia dell'informazione* 67, 1: 148-63.

Van Giffen, Benjamin, Herhausen Dennis, e Tobias Fahse. 2022. "Superare le trappole e i pericoli degli algoritmi: una classificazione dei pregiudizi e dei metodi di mitigazione dell'apprendimento automatico." *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2022.01.076

Vaswani, A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., e I. Polosukhin. 2017. "L'attenzione è tutto ciò di cui hai bisogno." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett, 30: 5998-6008. Curran Associati, Inc.