Report progetto Giustizia Agile, Univ Firenze, DINFO dept.

periodo 01/09/2022 – 31/12/2022
versione: 0.1
data: 08-01-2023

Ref. di unità: Prof. Paolo Nesi

Il progetto Giustizia Agile ha come obiettivo l'abbattimento dell'arretrato e riduzione della durata media dei processi giudiziari. In particolare, la ricerca affrontata mira alla realizzazione di un prototipo di un sistema di supporto alle decisioni del giudice in materia di propensione alla mediazione. Un altro tema su cui si è concentrata la ricerca è quello relativo all'anonimizzazione automatica dei fascicoli, obiettivo su cui sono stati spesi alcuni mesi di lavoro interrotto dalla prelazione del ministero.

Sono riportati in seguito i report dei gruppi di S. Marinai, P. Nesi e P. Cappanera

Attività Gruppo Simone Marinai

Nel primo periodo di ricerca ci siamo concentrati sull'analisi di strumenti da utilizzare per l'estrazione di dati da sentenze che abbiamo supposto appartenere a due tipi di domini: scannerizzazioni e documenti digitali. Le scannerizzazioni sono immagini in vari formati, mentre per documenti digitali intendiamo testi in formato PDF ottenuti da programmi di elaborazione testo (e.g. MSWORD)
Avendo a che fare con due tipi di file, sono state seguite due strade diverse per recuperare il testo:

- Immagini (scannerizzazioni): utilizzando un software di riconoscimento di testo (OCR)
- PDF (documenti digitali): utilizzando una pipeline in grado di estrarre il testo

Entrambi i percorsi convergono nella generazione di file in formato txt, in modo da poter facilmente lavorare con il testo in fasi successive di elaborazione.

Fatta questa premessa, un altro passo fondamentale per l'inizio dello sviluppo è stato quello di disporre di dati verosimili, ovvero, sentenze pubbliche. Per questo motivo abbiamo sviluppato un semplice script per scaricare sentenze (in formato PDF) dal sito della Corte di Cassazione (https://www.italgiure.giustizia.it/sncass/).
Dal momento che i documenti scaricati sono in formato PDF, per i nostri scopi abbiamo convertito ciascuna pagina dei documenti in formato PNG, in modo da poter sperimentare con i due approcci elencati sopra: OCR ed estrazione di testo.
Entrando più nel merito degli strumenti utilizzati, dopo un'analisi di possibili alterntive abbiamo optato per i seguenti:

- Software OCR: tesseract (https://github.com/tesseract-ocr/tesseract)
- Documenti di testo: PDFMiner

In particolare, per tesseract abbiamo utilizzato il wrapper Python pytesseract (https://pypi.org/project/pytesseract/); per quanto riguarda i documenti di testo, abbiamo utilizzato PDFMiner (https://pypi.org/project/pdfminer/).

E' importante giustificare il motivo della necessità di lavorare con questi due formati diversi: per quanto evidenziato negli incontri con i colleghi di Giurisprudenza, i processi nel processo Civile hanno un certo grado di informatizzazione, per cui si lavora con documenti digitali, mentre tutt'altra situazione esiste nel processo Penale, dove i magistrati, gli Avvocati, i PM e tutti gli altri organi corollari lavorano su carta.

Avendo a disposizione questi strumenti siamo in grado di convertire sentenze in formato scannerizzato o digitale in documenti testuali, dai quali estrarre dati da utilizzare come input a modelli di NLP per scopi di clusterizzazione, classificazione ed indicizzazione delle sentenze. Tuttavia, per problemi di privacy, abbiamo

dovuto fare un passaggio ulteriore: l'anonimizzazione dei dati. Questa operazione è stata implementata utilizzando due modelli di NER (Named Entity Recognition), atti al riconoscimento e alla classificazione di entità nel testo, nel nostro caso nelle sentenze. Le entità da anonimizzare sono organizzazioni, nomi, cariche, luoghi, date, valute. In questo modo, nei passi successivi (task NLP) è più agevole utilizzare tali dati.
Il processo di anonimizzazione è stato eseguito utilizzando due modelli pre addestrati; per il momento, infatti, non siamo stati in grado di eseguire un addestramento specifico sui dati di interesse, non avendo dati a sufficienza. In particolare, abbiamo usato presidio di Microsoft (https://microsoft.github.io/presidio/) e NerIta, (https://huggingface.co/bullmount/it_nerIta_trf). Per caricare questi due modelli ci siamo serviti di Spacy (https://spacy.io/).

Attività Gruppo DISIT, Paolo Nesi

Le attività del periodo si sono principalmente concentrate su:

- Studio dello stato dell'arte dei sistemi di anonimizzazione
- Studio dello stato dell'arte dei sistemi di NLP per la giustizia
- Studio dello stato dell'arte dei modelli predittivi in ambito giustizia
- Sviluppo di modelli BERT per la stima della propensione alla mediabilità

Anonimizzazione
L'anonimizzazione è il processo di rimozione di tutti i dati che permetterebbero l'identificazione di una persona. Anonimizzare i fascicoli di un tribunale è un'azione necessaria qualora si volesse procedere alla condivisione tra tribunali oppure per la creazione di una banca dati pubblica accessibile per la consultazione da parte di tutti i cittadini. I punti fondamentali del sistema di anonimizzazione sviluppato sono i seguenti:

1. Deve rimuovere tutti i dati che permetterebbero il riconoscimento del cittadino: non solo gli identificatori diretti - ad esempio nome e cognome, indirizzi - ma anche identificatori indiretti come il codice di fascicolo, le date delle udienze, gli importi di risarcimento.
2. Il processo di rimozione dei dati non deve comunque intaccare la leggibilità e facilità di comprensione dei documenti.
3. Il sistema deve anonimizzare automaticamente nuovi documenti basandosi esclusivamente sul loro contenuto.

Per poter anonimizzare il sistema deve innanzitutto essere in grado di comprendere il testo e individuare automaticamente quali sono gli elementi da anonimizzare: per raggiungere questo obiettivo abbiamo utilizzato BERT, una architettura per il Natural Language Processing con cui poter eseguire anche task di tipo Named Entity Recognition (NER).
BERT ci ha permesso di partire da un modello NLP pre-addestrato per la comprensione del testo italiano sul quale noi abbiamo effettuato ulteriori addestramenti *(fase di fine-tuning)* finalizzati al riconoscimento degli elementi da anonimizzare. Per poter eseguire il fine-tuning di un task NER è necessario definire un insieme di etichette di anonimizzazione e addestrare il sistema con una serie di documenti di esempio anonimizzati con tali etichette: a partire da alcuni dati utilizzati da un precedente progetto DISIT in ambito giuridico – Giustizia Semplice – abbiamo definito un set di etichette di anonimizzazione e anonimizzato manualmente alcuni fascicoli di processi civili. La realizzazione di questo sistema ha seguito i seguenti passi:

- Scelta della architettura NLP da utilizzare e del tipo di task con cui individuare il testo da anonimizzare: abbiamo scelto BERT ed il task Named Entity Recognition.
- Scelta del modello BERT in grado di comprendere testi in lingua italiana: abbiamo scelto dbmdz/bert-base-italian-uncased
- Definizione delle etichette di anonimizzazione.
- Estrazione automatica dei testi in formato txt da pdf non anonimi.
- Etichettatura manuale dei txt estratti dai pdf.
- Sviluppo di una pipeline di pre-processing per rendere compatibili alcuni files di Giustizia Semplice semi-anonimizzati per l'addestramento con BERT.

- Implementazione delle metriche di valutazione della rete per task NER:
    - Precision, Recall, F-measure valutate a livello di token
    - Matrice di confusione sul riconoscimento delle entità da anonimizzare.
- Valutazione della rete

In data 20 ottobre 2022 si è tenuto un incontro con i vari gruppi di ricerca dell'Università di Firenze coinvolti nel progetto durante il quale abbiamo presentato il nostro sistema di anonimizzazione, confrontandone le caratteristiche rispetto ad altri metodi di anonimizzazione utilizzati in Italia.
In seguito alla riunione il set di etichette utilizzato è stato revisionato ed approvato da due costituzionalisti che ne hanno verificato la completezza e grado di conservazione della semantica.
Nonostante il dataset di addestramento fosse ridotto nel numero di documenti di addestramento utilizzati, il sistema sviluppato ha prodotto risultati molto buoni sia in precision che in recall; un ampliamento del dataset rappresenterebbe sicuramente un miglioramento in quanto aiuterebbe la rete ad apprendere in modo più accurato tutte le classi di entità, soprattutto quelle per cui erano presenti pochissimi esempi, aumentando ancora di più la bontà dei risultati.
A tal fine stiamo lavorando alla modifica di un software pubblico di etichettatura – Doccano – che potrebbe essere dato in dotazione al personale del tribunale. Questo potrebbe essere impiegato per la produzione di nuovi documenti giudiziari etichettati, da utilizzare in fase di train.

Le competenze acquisite in materia di reti Transformers, BERT ed anonimizzazione sono state utilizzate per tenere due seminari durante i corsi di Knowledge Engineering e Big Data Architectures.

Propensione alla mediazione

Il giudice e i suoi collaboratori stabiliscono, in seguito alla lettura delle carte prodotte dalle prime udienze, se per il caso in esame può essere possibile una risoluzione tramite mediazione. I tempi necessari per la conclusione di un tentativo di mediazione sono però molto lunghi: nel caso in cui la mediazione fallisca, la causa ritorna molti mesi dopo a dover essere ridiscussa in tribunale. Tutti i mesi trascorsi nel tentativo di mediazione concorrono alla durata del processo: all'aumentare del numero di tentativi di mediazione falliti, la durata media dei processi aumenta. Potrebbe essere elaborato un sistema di supporto alle decisioni in grado di suggerire al giudice la probabilità che la causa in esame possa risolversi con una mediazione, evidenziando anche all'interno dei documenti quali sono gli elementi che influenzano il risultato prodotto, indicandone anche il peso.

Per la produzione di questo sistema viene utilizzata l'architettura BERT basata su modello pre-addestrato in lingua italiana. Sul tema della mediazione sono in sviluppo due sistemi in parallelo relativi ai seguenti obiettivi:

1. Stabilire la propensione alla mediazione. Abbiamo prodotto un dataset in cui sono applicate delle etichette indicanti la propensione alla mediazione (o meno), a livello di fascicolo. Ad un intero fascicolo viene dunque associata una stessa label. Il modello di partenza viene addestrato con il dataset così costruito, ovvero, viene applicato un fine-tuning con come obiettivo la realizzazione di un task detto di text classification. Poiché l'architettura BERT produce sempre predizioni a livello di blocchi da 512 tokens ciascuno, mentre i fascicoli sono composti da testi ben più lunghi, le predizioni dei blocchi devono essere post-processate per poter stabilire una sola etichetta a livello del fascicolo che contiene i blocchi etichettati. Per ogni etichetta viene anche indicato uno score di confidenza e a partire dagli score dei vari blocchi calcoliamo la percentuale che indica la propensione alla mediazione.
2. Comprendere quando l'argomento del fascicolo riguarda materie per le quali deve obbligatoriamente essere effettuato un tentativo di mediazione prima dell'avvio del processo e di conseguenza verificare se tale tentativo sia stato davvero esperito. Per questo obiettivo è stato necessario definire un set di etichette diverso dal precedente, più ampio e a granularità fine, con le quali intendiamo etichettare i documenti a livello di blocco.

Attività Gruppo Paola Cappanera

L'obiettivo è stato quello di analizzare lo stato dell'arte della letteratura sull'uso dell'ottimizzazione per l'efficientamento dei sistemi giudiziari confrontando input, output e metodologie utilizzate nei lavori analizzati. Abbiamo visto come lo stato dell'arte del tema proposto, almeno in termini di valutazione del grado di efficienza, presenti un certo grado di omogeneità nelle considerazioni effettuati dai vari autori interessati. È risultata evidente la preferenza nei confronti di determinate variabili caratteristiche dei processi interni dei sistemi giudiziari, quali numero di giudici e casi risolti. Queste sono decisamente le più influenti, in quanto stiamo parlando del principale artefice del processo e della più rilevante forma di output presente.
Al contempo, per diverse motivazioni, questa relazione input-output è ritenuta tanto essenziale quanto incompleta data la varianza di informazioni e caratteristiche delle diverse tipologie di cause sostenibili. Al momento, si può affermare che la letteratura odierna presenti alcune lacune e che queste trattino aspetti altrettanto importanti. Il numero di articoli che trattano l'aspetto economico del tema è infatti molto esiguo, solo uno sul sistema svedese ha preso in considerazione l'idea di utilizzare dei costi come input. Un altro possibile argomento di studio e dibattito che sarebbe molto interessante da valutare è relativo all'ottimizzazione dell'utilizzo delle aule da parte dei giudici. Questi sono i due punti principali su cui aspettarsi degli sviluppi futuri nella letteratura.
Se gli studi riguardanti la semplice analisi della prestazione dei tribunali hanno espresso concetti chiari e semplici da ipotizzare, quelli che hanno approfondito l'impatto di alcuni specifici fattori hanno fornito alcune delucidazioni che portano con sé un certo grado di sorpresa.
Il reciproco rapporto tra giudici, carico di lavoro e sentenze presenta infatti più complicazioni di quanto fosse possibile immaginare a priori. I dati mostrano come l'aumento dei giudici non porti gli effetti desiderati e che l'incremento dei casi sopravvenuti provochi un miglioramento della performance dei magistrati. Questa conclusione potrebbe essere tuttavia parzialmente fuorviante, dato che anche un ammontare eccessivo del carico di lavoro presente potrebbe inibire le capacità dei soggetti addetti al suo smaltimento di mantenere degli standard elevati, giungendo così all'ingorgo che si forma tra cause in entrata e in uscita e provocando liste di attesa e ritardi.
Per quanto riguarda la possibilità di ottenere benefici da una diversa organizzazione del proprio lavoro da parte dei giudici, risulta che a parità di casi sopravvenuti, lavorando contemporaneamente su meno processi se ne conclude un numero maggiore, impiegando in media meno tempo a partire dalla data di iscrizione a ruolo. Questa modalità sarebbe facilmente applicabile da parte dei giudici se questi sfruttassero a pieno il calendario del processo, che è peraltro previsto dall'anno 2009 per i processi civili. Tramite questo strumento il magistrato ha la possibilità di definire le date di tutte le udienze sin dall'inizio, soddisfacendo così l'esigenza di avere tempi certi. Al contrario, ciò che accade nella maggior parte dei casi è che venga fissata un'udienza alla volta, facendo sì che le cause che necessitano di un numero maggiore di esse finiscano alla fine della coda. Se la calendarizzazione fosse attuata, le cause si svolgerebbero in sequenza secondo l'ordine di iscrizione, producendo così un sistema FIFO (*First In First Out*). Chiedendosi quale siano i motivi che portino i giudici ad ignorare il sistema di calendarizzazione, si ottengono risposte chiare e facili da comprendere.

- Necessità di maggiore impegno: organizzare un alto numero di udienze ha sicuramente le sue complicazioni, per questo tale strumento va perfettamente d'accordo con un'organizzazione sequenziale del lavoro.
- L'estinzione di un procedimento richiede la liberazione degli slot occupati dalle restanti udienze, che dovranno così essere riassegnati; questo compito non è richiesto dal giudice "seriale" che non usufruisce del calendario, mentre nel caso contrario si possono avere grosse complicazioni.