

## **PON Governance 2014-2020 - Asse 1, Obiettivo Specifico 1.4, Azione 1.4.1**

**Decreto del Ministero della Giustizia del 05 agosto 2021 e successiva rettifica del 09 agosto 2021 Progetto “Modelli Organizzativi e Innovazione Digitale. Il Nuovo Ufficio per il Processo per l’Efficienza del Sistema-Giustizia” – MOD-UPP - CUP E75F21001650007 – CLP PON\_MDG\_1.4.1\_17**



***Università degli Studi di Napoli “Parthenope”***

### **Linea di intervento 2 - Azione 2.2**

#### **Reporting della sperimentazione**

Il presente reporting della sperimentazione si inserisce nella linea d'intervento 2, ossia, *“Individuazione di modelli per la gestione dei flussi in ingresso e degli arretrati presso gli Uffici Giudiziari”*, nonché nell’azione corrispondente 2.2 della scheda progetto dedicata alla *“Definizione di un modello organizzativo e dei relativi strumenti procedurali e informatici”*.

L'obiettivo è quello di descrivere il processo di sperimentazione delle integrazioni di natura informatica, implementate tramite tecniche innovative di intelligenza artificiale, nonché le metriche e le metodologie utilizzate per valutare la qualità e l’efficacia delle stesse.

## INDICE

<b>1 - Piattaforma sperimentale di AI .....</b>	<b>3</b>
<b>1.1 - Sperimentazione della metodologia cloud native .....</b>	<b>3</b>
<b>1.2 - Modellatore atti.....</b>	<b>6</b>
<b>1.3 - Estrazione automatica di conoscenze dal corpus di sentenze tramite modello di linguaggio BERT .....</b>	<b>6</b>
<b>1.4 - Individuazione dei caratteri di serialità nelle sentenze .....</b>	<b>9</b>
<b>1.5 - Ricerca semantica documentale .....</b>	<b>10</b>
<b>1.6 - Catalogazione dei documenti .....</b>	<b>11</b>
<b>1.7 - Anonimizzazione automatica dei documenti .....</b>	<b>12</b>
<b>1.8 - Digitalizzatore OCR AI.....</b>	<b>13</b>

## **1 - Piattaforma sperimentale di AI**

La fase di sperimentazione ha coinvolto la ricerca e lo studio dello stato dell'arte in tema di gestione ed elaborazione dei dati in ambito giuridico, la disamina delle tecnologie adatte all'integrazione nel complesso ecosistema informatico delle strutture giudiziarie, e un primo collaudo delle funzionalità esposte e in implementazione. Si procede alla discussione dell'approccio tecnologico "*cloud native*", base d'appoggio per l'implementazione e la fruizione dei servizi integrativi proposti, per poi passare al vaglio la sperimentazione delle singole applicazioni individuate, e del contesto informatico, in particolare del ramo di intelligenza artificiale, in cui si inseriscono.

### **1.1 - Sperimentazione della metodologia cloud native**

Come anticipato nei precedenti report, l'implementazione dei prototipi informatici destinati ad un'integrazione nel complesso sistema informatico gestito da Dgisia, è stata ipotizzata usufruire della tecnologia cloud. L'approccio cloud è stato adottato infatti per l'implementazione di un applicativo web accessibile da parte dei funzionari giudiziari, con la quale prevedere interazioni nella più inoltrata fase di prototipazione e integrazione. I funzionari avranno modo di testare le funzionalità individuate nelle precedenti fasi e fornire feedback diretto, semplificando e evitando il laborioso processo di installazione del software su macchine locali. La metodologia cloud, nella sua accezione "*native*", costituisce però anche il *modus operandi* con la quale viene implementato e ingegnerizzato il prodotto informatico. Infatti mentre il termine "*cloud*" fa riferimento alla possibilità di utilizzare risorse hardware attraverso Internet, in questo caso per l'hosting della "piattaforma sperimentale di AI" (**Figura 2**) , il termine "*cloud native*" fa riferimento ad un approccio architetturale allo sviluppo software, progettato per sfruttare appieno le funzionalità di scalabilità, flessibilità e disponibilità. In particolare, un'applicazione "cloud native" è progettata dalle fondamenta per essere eseguita allocando dinamicamente risorse hardware da un pool di risorse virtualmente infinito, scalando in real-time il carico computazionale in base al flusso di richieste in entrata. L'idea è di allontanarsi dal "modello monolitico" attuale, nella quale un server centrale gestisce tutte le richieste in arrivo, rallentando la propria esecuzione in presenza di carichi improvvisi. La sperimentazione ha coinvolto quindi la definizione dei cosiddetti "microservizi", associati ad ogni funzionalità elementare descritta nei seguenti paragrafi. I microservizi sono servizi modulari e indipendenti, ognuno dei quali è progettato per svolgere un'attività specifica e ben definita, indipendente dalle altre. Tale paradigma consente grande flessibilità e scalabilità, poiché il rallentamento o blocco temporaneo di una componente ha conseguenze ben limitate,

non impattando sulle restanti funzionalità del sistema; inoltre qualora una particolare funzionalità, ad esempio “l’anonimizzazione documenti”, sia richiesta contemporaneamente da più uffici, ad essa verranno allocate contestualmente più risorse, in maniera tale che l’utente non percepisca disservizi. La scelta del provider “Google Cloud Platform” e del servizio PaaS “Google Cloud Run” ha rappresentato il punto iniziale della sperimentazione del paradigma “cloud native”. Google Cloud Run è un servizio che permette di eseguire in maniera scalabile container Docker contenenti applicazioni web o moduli computazionali (nel contesto, individuati nelle funzionalità integrative) in un ambiente serverless, in cui l’infrastruttura viene scalata automaticamente in base al carico di lavoro, pagando solo il tempo effettivo di utilizzo delle risorse. In **Figura 1** uno schema illustrativo della suddivisione delle funzionalità integrative in microservizi, implementati come containers Docker.

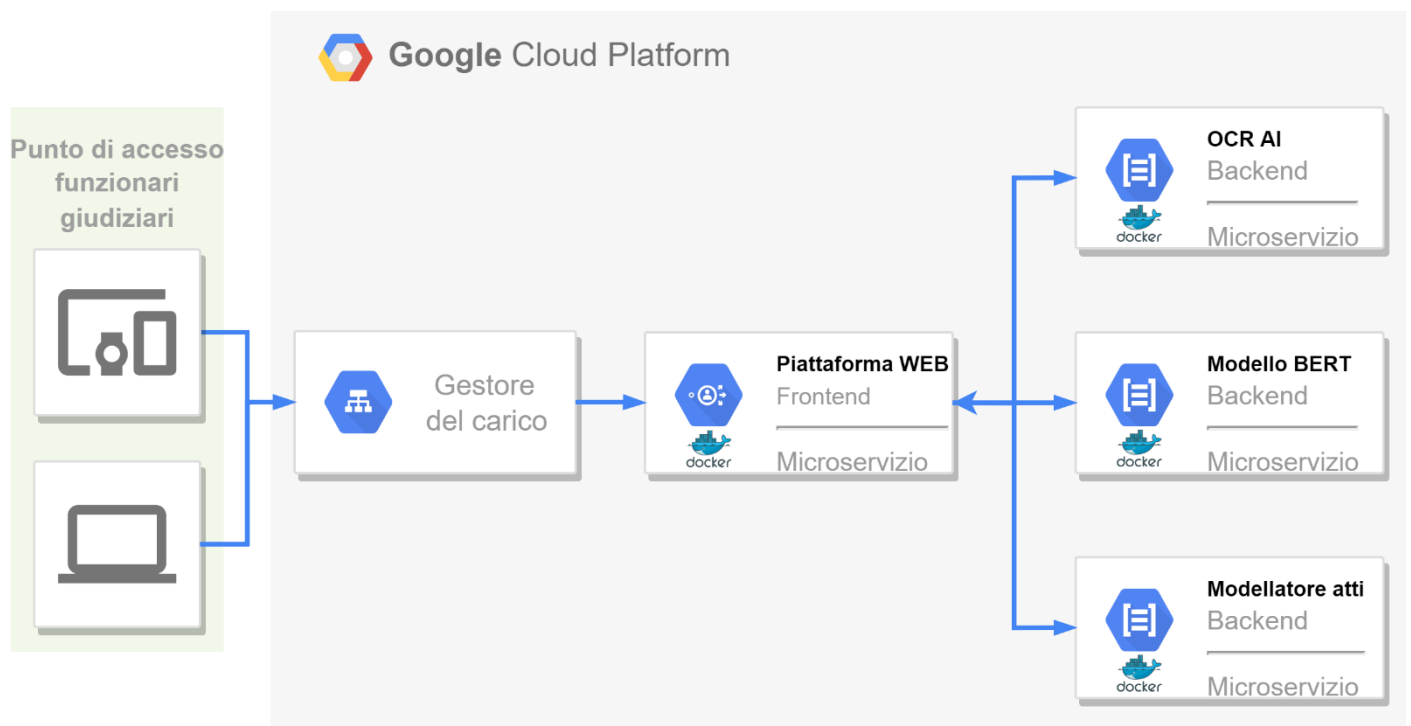


Figura 1: Architettura in microservizi della piattaforma cloud

Il framework di sviluppo che sta interessando la sperimentazione consiste di:

- Linguaggi Python, CSS, HTML e JavaScript
- Librerie BootStrap, Flask, Gunicorn
- Librerie Hugginface e Transformers
- API Google Cloud Platform
- API Only Office



# Piattaforma sperimentale di AI

Per la sperimentazione di funzionalità informatiche integrative basate  
su tecniche di intelligenza artificiale

Ricerca...



## Modellatore atti

Costruisci atti di parte e documenti standardizzati in maniera rapida, semplice e guidata.

**Scopri →**



## Individuazione dei caratteri di serialità nelle sentenze

Identifica ed estrapola automaticamente informazioni ricorrenti all'interno dell'archivio giurisprudenziale.

**Scopri →**



## Ricerca semantica documentale

Utilizza tecniche di AI di comprensione del significato per ricercare contenuti con risultati più pertinenti.

**Scopri →**



## Catalogazione dei documenti

Classifica i documenti sulla base del contenuto.

**Scopri →**



## Anonimizzazione automatica documenti

Riconosci e anonimizza automaticamente informazioni sensibili nei documenti con tecniche avanzate di AI.

**Scopri →**



## Digitalizzatore OCR AI


Trasforma una scansione in un documento nativamente digitale con tecniche avanzate di AI.

**Scopri →**

Figura 2: Illustrazione dell'UI della piattaforma web

## 1.2 - Modellatore atti

La funzione modellatore atti della piattaforma sperimentale prevede la possibilità di costruire atti di parte e altri documenti legali tramite una procedura guidata e automatizzata. La sperimentazione di questa funzionalità coinvolge l'uso di API JavaScript della piattaforma “*ONLYOFFICE Docs*”, una suite per ufficio open source che include editor per documenti di testo, fogli di calcolo, presentazioni e moduli compilabili, permettendo di integrare tali editor all'interno della piattaforma web sperimentale proposta tramite il protocollo REST-based WOPI, acronimo di *Web Application Open Platform Interface*. La possibilità di fare uso, nell'applicativo web proposto, delle API di una suite office, permette di implementare funzionalità di compilazione guidata degli atti e redazione di documenti standardizzati. L'idea è infatti di uniformare il processo di generazione dei documenti e delle informazioni all'interno dei vari uffici giudiziari, standardizzando il formato e il contenuto degli atti. Interagendo con la scheda “Modellatore atti” all'interno della piattaforma web, si apre un menu a tendina nella quale è possibile ricercare velocemente il tipo di atto da redigere e accedere ad un form come mostrato in **Figura 3**.



Collegio:  
Collegio A

Nr. RG  
181/2020 Esec.

Giudice:  
Mario Rossi ...

 **Crea documento word**

 **Crea foglio di calcolo**

Figura 3: Esempio di form per la compilazione guidata dei documenti

La funzionalità GUI sopra descritta è stata implementata mediante la sperimentazione con script in linguaggio JavaScript. La **Figura 4** raffigura un esempio di codice per la generazione di un modello “Eccezione giudice di pace.docx” attraverso l'utilizzo dell'API prima descritta.

```
builder.CreateFile("docx");
var oDocument = Api.GetDocument();
var oParagraph, oRun;
oParagraph = oDocument.GetElement(0);
oParagraph = Api.CreateParagraph();
oParagraph.AddText("Collegio");
oDocument.Push(oParagraph);
oParagraph = Api.CreateParagraph();
oParagraph.AddText("Tribunale di Nola");
oDocument.Push(oParagraph);
oParagraph = Api.CreateParagraph();
oRun = Api.CreateRun();
oRun.SetBold(true);
oRun.AddText("Sezione Penale");
oParagraph.AddElement(oRun);
oRun = Api.CreateRun();
oRun.AddText("Il Giudice");
oParagraph.AddElement(oRun);
oRun = Api.CreateRun();
oRun.SetBold(true);
oRun.AddText("PQM");
oParagraph.AddElement(oRun);
oDocument.Push(oParagraph);
oParagraph = Api.CreateParagraph();
oParagraph.AddText("Data");
oParagraph.AddLineBreak();
builder.SaveFile("docx", "Eccezione_Giudice_Di_Pace.docx");
builder.CloseFile();
```

Figura 4: Sezione di codice per la generazione del docx via API

### 1.3 - Estrazione automatica di conoscenze dal corpus di sentenze tramite modello di linguaggio BERT

BERT, acronimo di "Bidirectional Encoder Representations from Transformers", è un modello di machine learning basato sulla tecnologia dei transformer, utilizzato per la comprensione del linguaggio naturale. Sviluppato da Google e pubblicato nel 2018, rappresenta un importante avanzamento nello sviluppo dei modelli di linguaggio naturale e a differenza dei modelli di linguaggio precedenti, che utilizzavano una sola direzione di lettura per elaborare il testo, BERT utilizza una rappresentazione bidirezionale dei dati di input, ovvero analizza il testo in entrambe le direzioni. In questo modo, il modello può comprendere meglio il contesto dei termini nel testo e fornire risultati più precisi. Il modello è di base addestrato su enormi quantità di dati, inclusi i testi di Wikipedia e libri interi, utilizzando l'algoritmo di apprendimento profondo noto come "apprendimento per trasferimento". Questo significa che il modello viene addestrato a svolgere una determinata attività, come la comprensione del significato di una frase, e poi viene riutilizzato per risolvere problemi simili. Inoltre, BERT ad oggi viene utilizzato per una vasta gamma di applicazioni di linguaggio naturale, come la classificazione del testo, la risposta alle domande, la traduzione automatica, l'analisi dei sentimenti e molto altro ancora. Le sue prestazioni risultano essere superiori rispetto ad altri modelli di linguaggio naturale per cui BERT è diventato uno degli strumenti più utilizzati in tale campo.

Nel contesto specifico la sperimentazione ha riguardato l'utilizzo di tale modello neurale al fine di estrarre delle conoscenze da un corpus costruito ad hoc tramite un processo di web scraping. L'attività di web scraping consiste nell'estrazione automatica di informazioni da pagine web attraverso un software che analizza la struttura delle pagine, individua i tag HTML che contengono le informazioni di interesse, le estrae, le pulisce e le salva in un formato specifico. Il processo è stato automatizzato attraverso l'uso di scripts in linguaggio python per l'estrazione di circa 6000 sentenze prodotte da diverse strutture giudiziarie italiane dalla piattaforma "De Jure", banca dati di giurisprudenza italiana. Si è proceduto poi nell'utilizzare un modello pre-trained, "Italian Legal BERT", che come mostrato in **Figura 5** usa come base d'appoggio il modello "Italian XXL BERT" la cui procedura di training è stata ultimata con 4 epoche di apprendimento su 3.7GB di testi di giurisprudenza italiana, utilizzando l'ottimizzatore AdamW, un learning rate iniziale di  $5e-5$ , una lunghezza di frasi di 512, un batch size di 10 e 8.4 milioni di passi di training su una GPU V100 da 16GB. Terminata la procedura di "*training*" o addestramento, si ottiene un modello neurale del linguaggio con conoscenze approfondite non solo del linguaggio italiano, ma anche della giurisprudenza, che unitamente al corpus raccolto precedentemente descritto, può essere

sfruttato per rispondere a pieno alle attività delineate nella piattaforma sperimentale. I risultati sperimentali mostrano una diminuzione della *perplexity* usando l'adattamento del dominio da 10.98 a 8.98 su un dataset di riferimento e nell'ambito della classificazione un aumento dell'*F1-score* da 0.86 a 0.89.

Infatti, dai documenti del corpus, attraverso questo modello neurale, possono essere estratte rappresentazioni numeriche delle parole in essi presenti, chiamate “*word embeddings*”, che mettono in relazione, in uno spazio algebrico, le parole in base alla loro distanza di

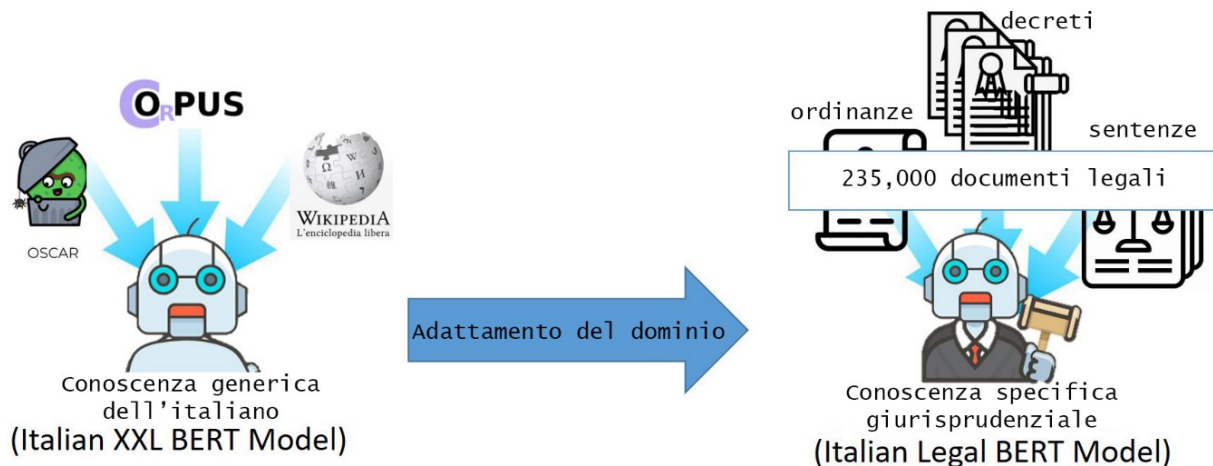


Figura 5: Processo di adattamento del modello neurale per il particolare contesto informativo

significato. Termini vicini nello spazio avranno quindi un'analogia somiglianza e relazione semantica (**Figura 6**). Tale caratteristica può essere estesa a frasi, paragrafi e interi documenti, emulando il processo cognitivo umano di raffronto documentale e sfruttandolo a proprio vantaggio per le funzionalità successivamente individuate.

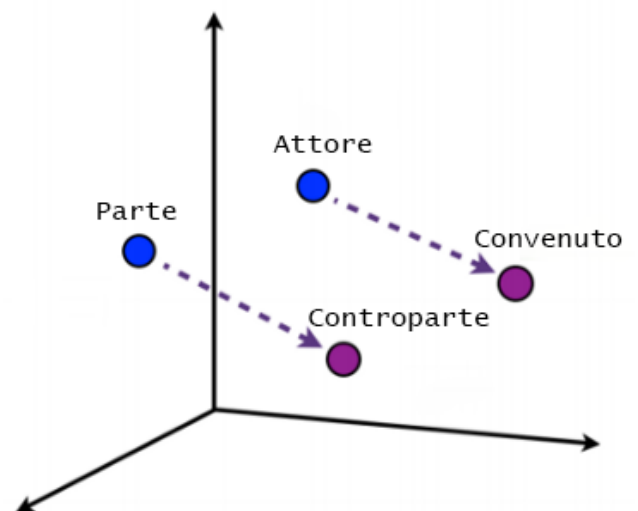
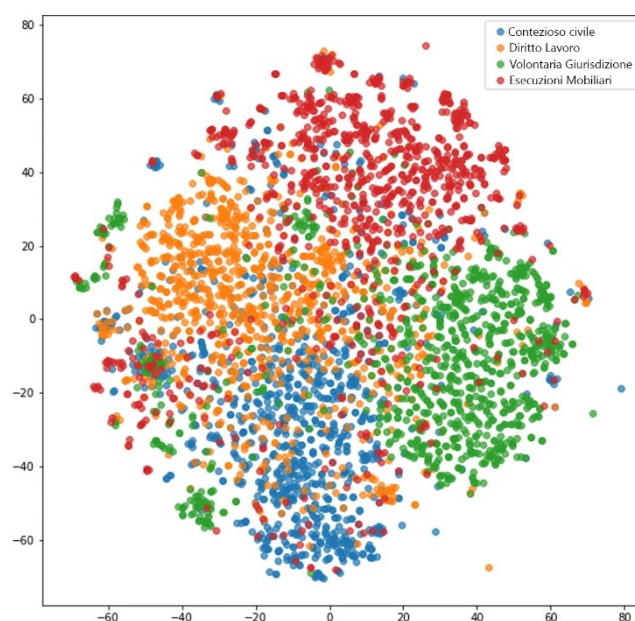


Figura 6: La distanza spaziale tra "Attore" e "Convenuto" è la medesima che intercorre tra "Parte" e "Controparte"; tale distanza numerica trova un corrispettivo nell'interpretazione umana della distanza semantica tra queste coppie di termini



#### 1.4 - Individuazione dei caratteri di serialità nelle sentenze

L'utilizzo del modello "BERT" prima descritto permette di giungere a rappresentazioni vettoriali consistenti dei documenti legali. Ottenute tali rappresentazioni, esse possono essere analizzate tramite "algoritmi di clustering" per l'individuazione dei caratteri di serialità nei documenti. La procedura sperimentata implica l'utilizzo di tali algoritmi allo scopo di suddividere l'insieme di documenti in gruppi o cluster omogenei, in maniera tale che i documenti all'interno di ogni cluster siano simili tra loro mentre quelli in cluster differenti siano dissimili. In particolare, usando un clustering di tipo gerarchico, si inizia con la creazione di un singolo cluster che contiene tutti i documenti nella collezione. Successivamente, il cluster viene diviso in due o più sottoclusters, sulla base della somiglianza di contenuto tra gli oggetti al loro interno. Questo processo viene ripetuto iterativamente, creando una gerarchia che può essere visualizzata in forma di dendrogramma. Infine, generati i clusters, per identificare gli argomenti ricorrenti all'interno della collezione di documenti, si possono visualizzare i documenti all'interno di ciascun cluster e osservare la presenza di agglomerazioni più o meno evidenti, segnalanti la ricorrenza di argomenti frequenti. Come si evince dalla **Figura 7**, i documenti possono agglomerarsi in gruppi circoscritti e ispezionabili, che suggeriscono la presenza di concetti ricorrenti nei documenti, in questo caso di argomenti afferenti alla stessa materia. L'attività di visualizzazione descritta può essere interpretata quindi come supporto a operazioni di analisi massive di dati come la massimazione delle sentenze, obiettivo della sperimentazione con i dati provenienti da De Jure.



*Figura 7: Visualizzazione 2D del corpus documentale. "Clusters" di documenti che hanno per oggetto la stessa tematica, corrispondono ad agglomerati di punti vicini*

L'analisi con clustering può essere utilizzata per identificare non solo gli argomenti ricorrenti, ma anche le tendenze e le relazioni tra gli argomenti all'interno della collezione di documenti. Questa tecnica è particolarmente utile per analizzare grandi collezioni di documenti, nel caso in esame il corpus estratto da De Jure, in cui la categorizzazione manuale dei documenti può essere troppo laboriosa e costosa in termini di ore uomo.

### **1.5 - Ricerca semantica documentale**

La ricerca semantica basata su AI è un tipo di ricerca che utilizza l'intelligenza artificiale per comprendere il significato implicito dei termini di ricerca e delle frasi dell'utente interrogante. Questo approccio si basa sulla comprensione profonda dei documenti nella collezione in cui cercare (basata sul modello BERT descritto precedentemente) e dell'intenzione dell'utente che esegue la ricerca, piuttosto che sulla semplice corrispondenza di parole chiave o di espressioni esatte. L'obiettivo è quello di fornire i funzionari giudiziari di uno strumento di ricerca che generi risultati più precisi e pertinenti, avvalendosi di tecniche di machine learning e di algoritmi di elaborazione del linguaggio naturale (NLP) per trovare corrispondenze anche latenti nei dati. Nel caso specifico si è proceduto nella sperimentazione del modello "Sentence BERT" (SBERT), un modello di embedding che utilizza e codifica rappresentazioni numeriche non al livello di parola (come nel modello BERT), ma al livello di frase o macrosequenze di testo, producendo per esse un vettore codificante l'associato contenuto semantico. La metrica di "somiglianza coseno" è stata poi utilizzata per testare la somiglianza semantica tra blocchi di testo, cercando di verificare se alla somiglianza numerica corrispondesse una somiglianza concettuale. L'idea è di far corrispondere ad ogni documento presente nella banca dati della piattaforma web, un associato vettore multidimensionale. All'utente è data poi la possibilità di inserire una frase query per la ricerca. Anche quest'ultima viene trasformata in un vettore e comparata con la collezione di vettori per il recupero degli  $N$  vettori più simili. Tali vettori restituiti sono quindi ripresentati all'utente in formato documentale per la visione e corrispondono ai documenti direttamente o indirettamente più pertinenti con la query fornita. Un esempio del processo di valutazione delle performance per questo modulo di ricerca è rappresentato dalla matrice in **Figura 8**. La matrice confronta le somiglianze reciproche di tre segmenti di testo dalla forma sintattica completamente diversa. Ciò nonostante il primo segmento e il secondo hanno, per un esperto in materia, un chiaro nesso logico-semantico, non individuabile dal testo, ma latente (Legge 1 dicembre 1970, n. 898. "Disciplina dei casi di scioglimento del matrimonio"), e catturato correttamente dall'intelligenza artificiale.

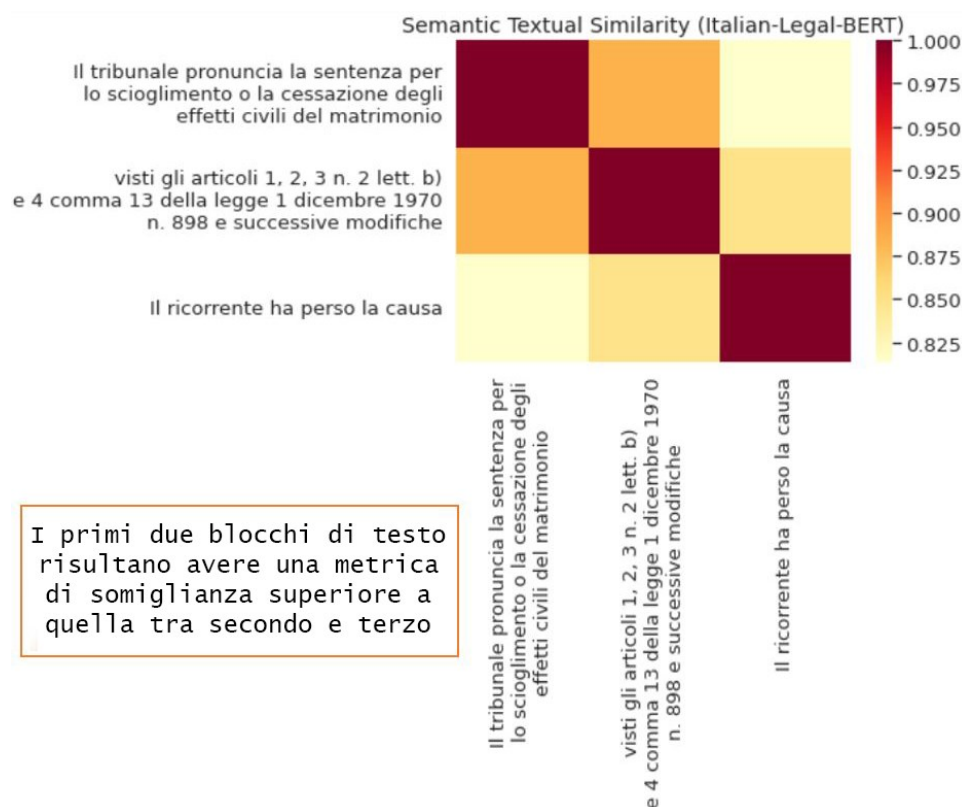


Figura 8: Matrice delle somiglianze coseno tra blocchi di testo. La somiglianza semantica è catturata correttamente dalla somiglianza numerica coseno

## 1.6 - Catalogazione dei documenti

Il modello BERT può essere anche impiegato per automatizzare il processo di indicizzazione e categorizzazione dei documenti, semplificando così la successiva ricerca e l'accesso alle informazioni contenute negli stessi. Il processo sperimentale, in maniera del tutto simile a quanto descritto nei precedenti paragrafi, mira a generare rappresentazioni vettoriali dei documenti che tengano conto delle informazioni semantiche o di forma al loro interno. Ottenuto un cospicuo numero di documenti, nel contesto di riferimento individuati dalle circa 6000 sentenze estrapolate dalla piattaforma De Jure, e trasformati nella loro rappresentazione vettoriale, si procede a definire il criterio di classificazione, per esempio la data di pubblicazione, l'autore o l'area del diritto trattata. Segue quindi una fase di annotazione, ovvero di associazione di ogni documento alla corrispondente categoria sulla base del criterio scelto, di una porzione di questi documenti. Viene infine eseguita una fase di “fine-tuning”, cioè di addestramento del modello neurale su questo nuovo set di dati annotati, affinché il modello impari, date le coppie “documento-annotazione”, come classificare autonomamente nuovi documenti non annotati, senza input umano. Il processo ha riscontro empirico positivo dalla letteratura odierna, la quale suggerisce che il modello del linguaggio BERT può essere

“educato” con successo a svolgere compiti altamente specifici (in questo caso la catalogazione dei documenti sulla base di un criterio prescelto).

### 1.7 - Anonimizzazione automatica dei documenti

Il processo di anonimizzazione dei documenti in ambito giurisprudenziale consiste nel rimuovere o sostituire le informazioni sensibili che potrebbero identificare le persone coinvolte in una determinata causa legale e citate in un documento. Ad esempio, potrebbe essere necessario rimuovere nomi, cognomi, date di nascita, indirizzi e altre informazioni personali dal testo dei documenti. La funzionalità è presente ad oggi, in chiave “deterministica” (vale a dire senza utilizzo di tecniche di AI), nell’applicativo *Consolle del Magistrato*. Tuttavia, le imprecisioni compiute dal relativo algoritmo sono a volte non trascurabili. L'utilizzo di tecniche di AI e quindi del modello BERT, specificamente adattato al task della name entity recognition (NER), possono essere determinanti nel processo di perfezionamento dei risultati. La NER è un ramo dell’AI che mira ad identificare le entità presenti in un testo, come nomi, luoghi, organizzazioni, date e così via. Per anonimizzare i documenti, si può quindi utilizzare un modello BERT preaddestrato per la NER, come ad esempio quello fornito dalle librerie spaCy e HuggingFace, per identificare le entità sensibili presenti nel testo, previa poi l’utilizzo di espressioni regolari per rimuovere o sostituire queste entità con informazioni generiche o placeholder. La **Figura 9** mostra un esperimento di esecuzione del modello BERT-NER su di segmento di sentenza estratta dal corpus De Jure. Le entità riconoscibili dal modello preaddestrato utilizzato sono: PER (persone), LAW (leggi), GPE (entità geografiche o politiche), FAC (indirizzi), DATE (date), CARDINAL (numeri) e ORG (organizzazioni). Notare come queste siano state riconosciute accuratamente, anche in presenza di abbreviazioni. Il processo può ulteriormente essere perfezionato mediante tecniche di fine-tuning analoghe a quella descritta precedentemente.

IL TRIBUNALE DI ANCONA SEZIONE PRIMA CIVILE Composta dai magistrati: Dott. RASCONI Valentina PER - Giudice - ha pronunciato la seguente: SENTENZA nella causa iscritta al n. 970/2012 R.G.A.C. LAW , avente ad oggetto: RISARCIMENTO DANNI promossa da: G.L. ORG elettivamente domiciliata in Ancona GPE , viale della Vittoria n. 32 FAC (studio avv. Campanati P. ORG ) unitamente all'avv. Pizzarulli Roberta PER , la quale la rappresenta e difende in virtù di procura a margine dell'atto di citazione; - attrice - contro Funivia Seceda S.p. ORG A, in persona del legale rappresentante pro tempore, con sede in Or GPE . GPE , via (omissis...) d'Anna FAC , 2 CARDINAL ; - convenuta contumace - sulle CONCLUSIONI: precisate dalla sola parte attrice all'udienza del 18.11.2014 DATE , da intendersi di seguito integralmente trascritte. SVOLGIMENTO DEL PROCESSO E MOTIVI DELLA DECISIONE Risulta non contestato ed è stato comunque ulteriormente comprovato dalla prova orale resa dal teste escusso all'udienza del 14.03.2014 DATE che la sig.ra G.L. PER in data (omissis DATE ...), mentre si trovava lungo

Figura 9: Esempio di riconoscimento delle entità in una sentenza pubblicata su De Jure

## 1.8 - Digitalizzatore OCR AI

La fase di digitalizzazione è un aspetto rilevante e delicato nella gestione documentale dei tribunali, poiché permette l'analisi informatica del contenuto dei documenti nonché la successiva adozione di tutte le tecniche precedentemente descritte. Il tool informatico di riferimento per la conversione da scansione cartacea a documento nativamente digitale va sotto il nome di OCR (Optical Character Recognition). A tal riguardo, le strutture della macroarea, in risposta ai questionari di rilevazione, hanno espresso la necessità di disporre di funzionalità OCR più accurate, che non risentano cioè di imprecisione nella fase di conversione in digitale. La sperimentazione ha quindi coinvolto l'adozione di tool OCR basati su modelli neurali di AI, i quali offrono diversi vantaggi rispetto ai tradizionali OCR basati su regole e modelli statistici. Primo fra tutti la maggiore precisione e affidabilità nella lettura e riconoscimento di caratteri, grazie alla capacità dei modelli neurali di apprendere le caratteristiche più importanti del testo dalle immagini. Inoltre, grazie all'utilizzo di modelli di deep learning come le reti neurali convoluzionali (CNN) o i transformer, l'OCR può restituire risultati soddisfacenti anche su documenti cartacei logorati o occlusi da deformazioni che ne invalidano la corretta lettura. Vista la natura cloud della piattaforma web e l'utilizzo del provider Google Cloud Platform, si è proceduto a sperimentare il tool fornito dalla stessa piattaforma Google Cloud, denominato "Document AI", che espone API per l'utilizzo di modelli di OCR basati su reti neurali. La **Figura 10** mostra l'implementazione python della chiamata API all'OCR Google.

```
def riconosci_testo(path):  
    """Riconosci testo in un documento scansionato"""  
    from google.cloud import vision  
    import io  
    client = vision.ImageAnnotatorClient()  
  
    with io.open(path, 'rb') as image_file:  
        content = image_file.read()  
    image = vision.Image(content=content)  
    response = client.text_detection(image=image)  
    texts = response.text_annotations  
    print('Testo digitalizzato:')  
  
    for text in texts:  
        print('\n{}'.format(text.description))  
  
        vertices = ([ '{}{}'.format(vertex.x, vertex.y)  
                    for vertex in text.bounding_poly.vertices])  
  
        print(': {}'.format(''.join(vertices)))
```

*Figura 10: Blocco di codice richiamante le API di OCR AI di Google Cloud Platform*