INAPP - Open access e machine learning

Descrizione sintetica dell'iniziativa

La situazione di partenza è quella di un archivio istituzionale **oa.inapp.org** che rende disponibili in formato aperto le **risorse informative e documentali** (lavori scientifici e documentazione tecnica) pubblicate come risultato delle proprie attività di ricerca. Questi oggetti informativi - che impattano sulla **società civile** e che sono utili ad aumentare la **consapevolezza** generale su temi come l'inclusione e le politiche attive del lavoro - sono redatti con **linguaggi specifici** conosciuti solo all'interno della comunità di riferimento mentre sono ricercati usando termini di **linguaggio naturale**; questo ***mismatch* informativo** viene gestito da questa iniziativa proposta applicando una soluzione di **AI** che avvicina le risorse agli utenti con risposte precise e veloci. L'iniziativa si propone pertanto di avvicinare gli utenti a **fonti di qualità realizzate dalla ricerca pubblica** e di potenziare le loro **competenze digitali** nella **consultazione** di **informazioni** e **documenti in rete**. Per arrivare a questo risultato si **amplifica l'effetto normalizzatore** che già svolge il **thesaurus** multidisciplinare, usato in fase di catalogazione semantica delle risorse informative e documentali, proponendo agli utenti una soluzione di **intelligenza artificiale** che ha lo scopo di **facilitare le ricerche** interpretando, traducendo e correggendo le parole del linguaggio naturale utilizzate come chiave di ricerca innescando così un ciclo continuo di affinamento nella ricerca. Dal momento che la **rete neurale utilizzata dalla soluzione AI lavora** tanto **sui metadati** quanto **sul testo dei documenti** pubblicati indipendentemente dal formato (pdf, jpeg, tiff, png) è possibile accompagnare gli **utenti generici** nel reperire le pubblicazioni/informazioni più attinenti alle parole chiave utilizzate per effettuare la ricerca sia fattuali che semantiche. Al tempo stesso la soluzione **AI** proposta **arricchisce di metadati** la catalogazione stessa, sia descrittiva che semantica, creando una ulteriore **base dati formata dall'aggregazione delle informazioni usate in fase ricerca**, base dati che offre così agli **utenti esperti**, bibliotecari/documentalisti o professionisti della PA, un panorama significativo dei temi di maggior interesse e delle parole del linguaggio naturale maggiormente utilizzate per ricercarli.

GESTIRE IL MISMATCH INFORMATIVO TRA UTENTI E SPECIALISTI GRAZIE A SOLUZIONI AI APPLICATE AL LINGUAGGIO

INTERPRETARE, TRADURRE E CORREGGERE I TERMINI DEL LINGUAGGIO NATURALE PER ARRIVARE AI CONCETTI

Motivazione/i dell'iniziativa

La **forza della condivisione** e della disponibilità di contenuti qualificati in modalità aperta è emersa significativamente in risposta alla crisi pandemica ed INAPP con la propria iniziativa intende rendere il proprio **operato** a livello tecnico-scientifico più **trasparente**, **partecipativo** e **inclusivo.** L'iniziativa intrapresa di applicazione di una soluzione di **AI** (artificial intelligence) su risorse già disponibili in modalità open agevola, nell'attuale contesto caratterizzato da un gran numero di risorse informative virtualmente disponibili, l'accesso all'informazione **prodotta dalle istituzioni pubbliche** che è la migliore **fonte** cui il cittadino possa attingere per qualità e autorevolezza e quindi l'accesso ai documenti di studio e di indirizzo degli enti di ricerca pubblici in base al loro specifico ambito e raggio di azione deve essere favorito e garantito grazie al ricorso ad **archivi elettronici pubblici** organizzati ed indicizzati per favorirne l'individuazione da parte di tutti, stakeholders e cittadini. La soluzione AI aiuta anche il sistema documentale ad arricchirsi di dati statistici utili al miglioramento del sistema stesso. L'ambizione è di mettere così la **ricerca pubblica in ascolto** rispetto alle esigenze reali di ricerca dei propri utenti in modo da poter consolidare il rapporto di fiducia tra istituzioni e società civile.

L'INFORMAZIONE DI FONTE PUBBLICA È UN PATRIMONIO CONOSCITIVO CHE RAPPRESENTA UNO STRUMENTO UTILE NEI CONFRONTI DELLE *FAKE NEWS*

LA RICERCA PUBBLICA ASCOLTA LE REALI ESIGENZE DI RICERCA PER CONSOLIDARE IL RAPPORTO DI FIDUCIA TRA ISTITUZIONI E SOCIETÀ CIVILE

Obiettivi e Target, specificare anche i target, se fissati nella pianificazione

Obiettivo del progetto è "applicare" aspetti legati al **machine learning** all'ambiente open access di INAPP **per facilitare l'accesso a tutti i cittadini a risorse, informazioni esperienze e conoscenze prodotte dall'Istituto**. Ciò permetterà di rinforzare il patrimonio messo a disposizione da INAPP attraverso forme open, riusabili e personalizzabili, ampliandone le prospettive. I target pianificati sono 1. cittadini/società civile, intesi come utenti che beneficiano della ricerca ed in prospettiva anche come valutatori della risposta ottenuta nonché come risorsa per eventuali sviluppi e/o criticità 2. altri enti di ricerca che ospitino soluzioni di pubblicazioni open access 3. professionisti dell'informazione 4. sviluppatori di soluzioni AI.

- ✓ **CITTADINI/SOCIETÀ CIVILE**
- ✓ **ENTI DI RICERCA**
- ✓ **PROFESSIONISTI DELL'INFORMAZIONE**

Principali Attività realizzate

A partire dalle risorse presenti è stato realizzato uno **studio di fattibilità** per l'applicazione di soluzione **AI** all'open archive. L'open archive INAPP è una realtà consolidata ed è utilizzato come ambiente verticale interrogato dal sito web per popolare delle sezioni specifiche (pubblicazioni/eventi) ed i suoi descrittori semantici (derivati dal thesaurus multidisciplinare) sono utilizzati anche per indicizzare i contenuti del sito stesso. **Per lo studio si è proceduto incrociando le risorse documentali disponibili in open archive con i parametri delle ricerche degli utenti sia provenienti dal monitoraggio degli *insights* del sito istituzionale sia presenti nel motore di ricerca interno all'open archive stesso**. Questo incrocio ha permesso di **identificare lo scarto tra i contenuti reali e utenti potenziali**. È stata successivamente simulata la ricerca di un *utente 0* con e senza la soluzione di **AI** per misurarne il funzionamento e per verificare come si adeguassero ed arricchissero i risultati della ricerca in presenza/assenza della soluzione stessa. Si è quindi realizzato un **modello prototipale** che lavora su sommarizzazione estrattiva, estrazione di parole chiave, ricerca di entità nominate (persone, luoghi, date). **Sommarizzare** un testo significa estrarre le informazioni ed i concetti più importanti da un documento, offrendo una panoramica chiara, rapida e concisa dei contenuti presenti all'interno di esso. **Sull'estrazione lavora poi una soluzione di intelligenza artificiale che effettua operazioni di *embedding* restituendo una lista di frasi riordinata per importanza**. Approccio molto simile viene poi utilizzato per l'**estrazione delle parole chiave** con l'unica differenza che si lavora in termini di singole parole, restituendo la lista di parole più importanti rispetto al tema del testo in esame (**l'obiettivo è l'applicazione di funzioni di tipo semantico-lessicali che consentano di raggruppare i documenti e le risorse in base ai concetti chiave**). Per l'**individuazione delle "entità nominate"** o NER (*Named-entity recognition*) si è invece proceduto in modo diverso ossia attraverso il *training* di una rete neurale alla quale è stato fornito un insieme di testi nei quali ogni parola è stata "taggata" consentendo così alla rete di "imparare" a riconoscere nomi di persone, luoghi ed altre entità. L'algoritmo utilizzato si chiama BERT, acronimo inglese di *Bidirectional Encoder Representations from Transformers.* BERT consiste in un sistema per l'elaborazione del linguaggio naturale, il Natural Language Processing (NLP), che si basa sul concetto, appunto, di rete

STUDIO DI FATTIBILITÀ

↓

SCARTO TRA CONTENUTI REALI E UTENTI POTENZIALI

SIMULAZIONE

PROTOTIPO

neurale ossia un modello informatico che imita il funzionamento neuronale umano ed il modo in cui i singoli neuroni si mettono in relazione tra loro. Bert è stato originariamente pre-allenato sull'intera Wikipedia inglese e sul Brown Corpus ed è stato perfezionato proprio per lo svolgimento di attività di *natural language processing*. Dalla prototipizzazione effettuata si è confermato l'approccio seguito durante lo studio di fattibilità perché è non solo replicabile ma anche scalabile e potrebbe offrire alla PA maggiore interazione tra enti e, quindi, permettere al cittadino di ricercare informazioni a prescindere dal singolo soggetto pubblico o argomento trattato. A partire dalla prototipizzazione INAPP **realizzerà la soluzione AI applicata al proprio OA**.

A quanto ammonta la spesa sostenuta per realizzare l'iniziativa?

200.000,00 euro

Quali sono le fonti di finanziamento utilizzate?

Fondi nazionali e comunitari

Quali sono gli stakeholder coinvolti?

Gli **"oggetti" informativi** dell'open archive INAPP **attivano delle interconnessioni con altri ambienti informativi strutturati**. Tra queste connessioni sono significative quelle con gli stakeholder che, a vario titolo, possono beneficiare del progetto e delle risorse messe a disposizione, sia in quanto a loro volta produttori di risorse informative, sia perché potenziali beneficiari dell'iniziativa. I principali stakeholder sono: 1. Cittadini/società civile (non solo in quanto utenti che beneficiano della ricerca ma anche in qualità di valutatori rispetto alla risposta ottenuta e risorsa per eventuali sviluppi e/o criticità) 2. Amministrazioni pubbliche (Istituzioni, Regioni, Enti/Organizzazioni internazionali, EPR) 3. Professionisti dell'informazione e sviluppatori soluzioni AI.

- ✓ **CITTADINI/SOCIETÀ CIVILE**
- ✓ **PA**
- ✓ **PROFESSIONISTI DELL'INFORMAZIONE**

Descrivere le metodologie o le soluzioni tecnologiche definite

L'obiettivo principale dell'attività tecnica è stato quello di **estendere** il dominio **ricerca** oltre che ai **metadati** anche alle **informazioni** scientifiche **contenute** nei file allegati alle schede delle risorse dell'open archive. Sulla parte metadati si è

intervenuti utilizzando il thesaurus multidisciplinare di INAPP ed algoritmi di conversione di stringhe e sinonimi al fine di migliorare la ricerca effettuata sull'Open Archive superando anche i limiti tecnologici del software stesso. Per quanto riguarda i file allegati alle schede descrittive delle risorse contenute nell'open archive (item) si è invece tenuta presente la diversa tipologia di documenti trattati: inizialmente si è proceduto individuandone la tipologia e con i relativi strumenti di riferimento quali OCR e analisi ed estrazione del testo si è creata una base dati "raw" con le informazioni necessarie per la parte di ricerca semantica. L'output di questa prima fase viene elaborato tramite strumenti di analisi statistica del testo e Machine /deep Learning. In prima battuta vengono estratti i blocchi di testo; successivamente con strumenti di analisi statistiche del testo si scompongono i blocchi in frasi al fine di analizzarle singolarmente tramite il rapporto caratteri/punteggiatura in modo da escludere le frasi con un contenuto minimo di informazioni che avrebbero solamente aumentato il grado di entropia della rete Neurale. A questo punto BERT analizza il documento destrutturato e tramite un'operazione di *embedding* e successiva clusterizzazione assegna una priorità alle singole frasi. In funzione del rapporto di sommarizzazione scelto dall'utente il sistema ne restituisce le relative parti salienti ossia la sommarizzazione estrattiva. Analogamente è stato effettuato per l'estrazione delle KEYWORDS. Anche in questo caso si è effettuata un'analisi statistica dei caratteri e/o simboli contenuti nella frase ma in aggiunta si sono dovute eliminare le STOPWORDS, ossia quelle parole prive di un significato specifico, ad esempio articoli e pronomi. Un meccanismo del tutto simile è stato utilizzato per estrarre le frasi più rilevanti del documento che permette di estrarre le keywords. Un approccio diverso è stato usato per il NER. In questo caso la rete neurale è stata addestrata con un dataset contenente generiche frasi "taggando" ogni parola, specificandone la "classe" di appartenenza e riuscendo a far apprendere alla rete neurale come riconoscere le "entità nominate" quali ad esempio persone o entità geopolitiche. Per questo task è stato necessario fornire in ingresso il testo "raw", quindi non processato statisticamente, a differenza di come è avvenuto nei casi precedenti.

Descrivere le iniziative di promozione realizzate

La promozione del progetto avviene in duplice forma: una continuativa, l'altra in occasione di eventi e momenti di incontro più specifici:
- **oa.inapp.org** è oggetto di costante *auto-promozione* in quanto i suoi contenuti sono utilizzati in maniera sinergica al sito web istituzionale: vengono cioè "catturati" e veicolati in occasione in tutti i momenti di promozione delle attività dell'INAPP (eventi, comunicati stampa, prodotti divulgativi, manifestazioni pubbliche)
- **oa.inapp.org** è pubblico dal 2009, nel corso del tempo è stato oggetto di numerosi momenti di promozione e divulgazione su target diversi utilizzando momenti di incontro (seminari, gruppi di lavoro) con i ricercatori, altre istituzioni di ricerca sia con un pubblico più generico in occasione di fiere e manifestazioni pubbliche, finalizzati a incoraggiare la conoscenza e l'utilizzo delle risorse in esso contenute ad un pubblico più vasto possibile
- attraverso la **partecipazione a gruppi di lavoro promossi da consorzi e associazioni pubbliche, finalizzate alla diffusione e allo sviluppo dell'open data e della ricerca semantica automatizzata**
- la fase prototipale è stata sperimentata su un nucleo di utenti interni con competenze specifiche legate alla diffusione di contenuti tecnico-scientifici.

Quali indicatori di output e di outcome sono/sono stati utilizzati per la misurazione dei risultati?

L'iniziativa è in corso e prevede degli indicatori per la misurazione dei risultati di apertura delle pubblicazioni scientifiche in open access che si effettuerà con un sistema ibrido di indicatori legati a metriche tradizionali (rilevamento delle visite on line alla risorsa, geolocalizzazione relativa alla località (città) dalla quale si accede all'open archive, numero dei download) ed indicatori innovativi volti a cogliere le conversazioni on line relative alle risorse stesse (*sentiment analysis*).

METRICHE TRADIZIONALI

CONVERSAZIONI ON LINE

Quali sono i principali risultati conseguiti?

L'open data, fin qui raggiunto, è un **patrimonio** importante per la **società civile** e per la **comunità di riferimento** che ha permesso di rimuovere le barriere verso l'**accesso**, la **comprensione** dei dati, intesi nel sono più estensivo.

AVVICINARE IL CITTADINO ALLA DOCUMENTAZIONE PUBBLICA

Quali sono gli elementi di innovatività introdotti dall'iniziativa che ne hanno determinato il successo?

L'iniziativa è in corso, tra gli elementi di innovatività da segnalare:
- utilizzo di software **open source** che consente 1. disponibilità del codice sorgente 2. indipendenza rispetto al *vendor* 3. partecipazione a community di sviluppo per migliorare i moduli operativi 4. *uses cases* testati
- applicazione di **soluzione AI alla ricerca semantica**
- utilizzo di **algoritmi di Natural Language Processing** per agevolare tanto il lavoro dell'operatore in fase di pubblicazione quanto il cittadino anche meno esperto nella comprensione dei contenuti
- utilizzo di **licenze CC** (Creative Common)
- impiego di **standard** e protocolli internazionali che garantiscono un facile e trasparente reperimento della produzione scientifica presente nell'open archive.

- ✓ **OPEN SOURCE**
- ✓ **NATURAL LANGUAGE PROCESSING**
- ✓ **STANDARD E LICENZE**

Indicare quali sviluppi futuri sono previsti per l'iniziativa

L'iniziativa proposta oltre a rendere nell'immediato l'utente in grado di acquisire informazioni aggiuntive rispetto alla risorsa cercata, permette di guardare a prospettive future di sviluppo.
Nello specifico si prevede di:

1. **Ricercare interlocutori** negli altri enti di ricerca (**EPR**) che possano adottare - perché già in possesso di open archive - soluzioni ed approcci analoghi. L'approccio seguito da questa iniziativa è non solo replicabile ma anche scalabile e potrebbe offrire alla PA maggiore interazione tra enti e, quindi, permettere al cittadino di ricercare informazioni a prescindere dal singolo soggetto pubblico o argomento trattato. **La frammentarietà e la difficoltà con la quale le**

- ✓ **INTEROPERABILITÀ SEMANTICA**
- ✓ **LINKED OPEN DATA**
- ✓ **COMUNITÀ DI PRATICA**

informazioni sono reperibili in rete costituiscono un vero e proprio freno alla diffusione di conoscenza e l'architettura proposta dall'INAPP ha l'ambizione di risolvere, anche parzialmente, tali ostacoli.

2. **Investire** sull'**interoperabilità semantica** che sfrutta i metadati per arrivare alla struttura ontologica di riferimento rispetto alle risorse informative. La soluzione consente infatti di **intercettare collegamenti ad altre ontologie/soggettari/vocabolari controllati** che descrivono le medesime risorse.
3. Operare a livello di **collegamento** tra i **dati (linked data)** per **costruire una rete di risorse collegate** con dataset esterni favorendo una crescita esponenziale del patrimonio e delle informazioni iniziali. Per esempio, con la soluzione tecnica di **AI** presentata si estraggono parole chiave ma anche ad esempio persone (chi) e luoghi (dove). È quindi possibile arricchire la descrizione delle persone (chi) che vengono nominate (quale ruolo assume una persona nel ciclo della risorsa documentale estratta?) ed ottenere informazioni (dove) che possono essere utili per arricchire la descrizione dei luoghi estratti dalle risorse, per esempio per geolocalizzazione gli elementi spaziali.
4. **Promuovere e/o individuare community** che condividano l'interesse all'apertura del patrimonio pubblico, la sua disponibilità e possibilità di utilizzo, finalizzate al confronto reciproco, alla formazione e ad un arricchimento di competenze e idee.