Sistemi di supporto alle decisioni basati su tecnologie semantiche per la gestione delle politiche di innovazione

Marco Di Ciano[1], Pasquale Lops[2], Marianna Cavone[1], and Giovanni Semeraro[2]

[1] InnovaPuglia S.p.A., m.diciano@innova.puglia.it,m.cavone@innova.puglia.it
[2] Dip. di Informatica - Università degli Studi di Bari Aldo Moro, nome.cognome@uniba.it

Abstract

In questo contributo presentiamo i risultati preliminari relativi al progetto INTERREG-MED TALIA, il cui obiettivo principale è la realizzazione di sistemi di supporto alle decisioni basati su tecniche di Intelligenza Artificiale, che possano aiutare i policy maker in una corretta ed efficace gestione delle politiche di innovazione. L'utilizzo di tecniche di Natural Language Processing e di semantica distribuzionale consente di estrarre ed organizzare conoscenza rinveniente dall'analisi massiva di documenti testuali prodotti dai numerosi progetti finanziati. Tale conoscenza si rivela preziosa per la gestione delle politiche di innovazione e per la capitalizzazione dei risultati dei progetti di maggior successo.

1 Introduction

La proposta di regolamento del Parlamento Europeo e del Consiglio che istituisce il programma Europa Digitale per il periodo 2021-2027, identifica tra gli obiettivi di attuazione tecnica, il rafforzamento delle capacità di base dell'Intelligenza Artificiale (IA) in Europa, rendendole accessibili a tutte le imprese e le Pubbliche Amministrazioni (PA), puntando sulle esperienze di sperimentazione dell'IA esistenti negli Stati membri. Ugualmente il Piano d'Azione dell'UE per l'eGovernment 2016-2020 si fonda sulla visione in base alla quale entro il 2020 le amministrazioni e le istituzioni pubbliche nell'Unione Europea dovrebbero essere aperte, efficienti e inclusive e fornire servizi pubblici digitali end-to-end personalizzati e intuitivi a tutti i cittadini e a tutte le imprese nell'UE.

Il ricorso ad approcci innovativi permette di progettare e fornire servizi migliori, in linea con le esigenze e le richieste di cittadini e imprese. Le PA possono sfruttare le opportunità offerte dal nuovo ambiente digitale per interagire più facilmente tra di loro e con le parti interessate. Anche la precedente pubblicazione della Commissione Europea per i servizi pubblici[1] afferma che l'evoluzione della società richiede alle PA di affrontare molte nuove sfide, questo anche perché le pressioni economiche e i vincoli di bilancio impongono ai governi investimenti sempre più mirati, facendo leva su strategie di ricerca ed innovazione per lo sviluppo socio-economico locale.

Un'enorme quantità di informazioni provenienti da progetti finanziati a livello regionale (POR, FESR, FSE, ...), nazionale (PON, ...) o di cooperazione territoriale sovranazionale (Interreg MED, Horizon, ...), o azioni di sostegno specifiche all'innovazione e allo sviluppo socio-economico dei territori (PMI, Ricercatori, Università, Centri di ricerca etc.) sono archiviate in decine di terabyte di documenti digitali. La progettazione e il monitoraggio delle strategie di innovazione, l'analisi comparativa delle evidenze con i risultati esistenti o i risultati già raccolti derivanti dalle azioni politiche implementate, richiedono una conoscenza permanente ed aggiornata dei territori e delle capacità locali. Alcune di queste evidenze, quali contenuti di ricerca, tecnologie abilitanti, relazioni e network tra gli stakeholder, potrebbero essere rintracciate ed evidenziate a partire da studi appropriati e analisi dettagliate dei risultati dei progetti documentati e archiviati in database digitali contenti prevalentemente informazioni non strutturate (testi). La comprensione delle peculiarità territoriali e delle ricadute degli investimenti sostenuti, può quindi beneficiare dell'IA facendo leva su algoritmi intelligenti di identificazione ed estrazione di concetti, di calcolo della similarità semantica tra documenti, etc., da applicarsi a specifiche collezioni di documenti digitali. Natural Language Processing (NLP), tecnologie semantiche, tecniche di apprendimento automatico e sistemi di recommendation sono alcune delle metodologie offerte dall'IA e dalla Linguistica Computazionale in grado di supportare le interazioni tra esseri umani e dati digitali, consentendo di governare al meglio il processo in corso di Digital Data Deluge[2].

In questo articolo presentiamo i primi risultati conseguiti nel progetto *TALIA "Territorial Appropriation of Leading-edge Innovation Actions"*, finanziato a valere sul bando INTERREG-MED, Priority Axis 1: Promoting Mediterranean innovation capacities to develop smart and sustainable growth, Programme specific objective 1.1 To increase transnational activity of innovative clusters and networks of key sectors of the MED area, il cui obiettivo principale è quello

[1] https://ec.europa.eu/digital-single-market/en/news/vision-public-services

[2] https://www.ibm.com/thought-leadership/institute-business-value/report/cognitivemarketingsales

di sviluppare un sistema di *supporto alle decisioni* in grado di supportare gli stakeholder, ognuno in base a specifici obiettivi. La comunità del programma MED ha una crescente consapevolezza circa la necessità di andare oltre l'*analisi isolata* e a se stante dei singoli progetti, esplorando il loro potenziale di *scalabilità* in maniera integrata, efficace e coerente. Per scalabilità, si intende la capacità dei risultati dei progetti di raggiungere un numero maggiore di beneficiari nel tempo e nello spazio, garantendo in tal modo un maggiore impatto sulle politiche tematiche e pratiche. Questo approccio consente di impostare approcci nuovi ed innovativi alla *capitalizzazione* dei risultati dei progetti pilota di successo a livello di stati membri, regioni o macro regioni.

Il progetto TALIA mira dunque a costruire e sviluppare la comunità *Social&Creative* del programma InterregMED a partire dall'orchestrazione dei risultati dei singoli progetti modulari in ambito culturale, industria creativa e innovazione sociale (dati aperti, imprenditoria sociale, innovazione del settore pubblico, etc.), promuovendo la *capitalizzazione* di risultati e attività, garantendo in tal modo la *cross-fertilization* tra progetti differenti e tra i loro principali stakeholder.

Per raggiungere questi obiettivi il progetto si basa sulle funzionalità del *semantic framework* (sezione 2), componente trasversale il cui scopo principale è la definizione di un *modello semantico*, ovvero una rappresentazione strutturata di concetti e relazioni estratti dall'analisi profonda dei risultati prodotti da singoli progetti, al fine di definire una struttura tematica coerente in grado di abilitare servizi complessi di policy making.

2 Semantic Framework

Il semantic framework consente di estrarre conoscenza e informazioni strategiche da tutta la documentazione tecnica prodotta dai singoli progetti finanziati dal programma MED, al fine di sostenere politiche di innovazione e la pianificazione strategica in ambito di Ricerca e Sviluppo. Il sistema analizza documenti contenenti le descrizioni dei progetti già finanziati, i deliverables tecnici, i risultati finali e svariate fonti di informazioni testuali. A causa della grande quantità di informazioni da analizzare, il semantic framework fornisce agli utenti un accesso efficace ed intelligente ad una enorme mole di informazioni. Le principali funzionalità fornite dal sistema sono le seguenti:

- *Indexing*: il sistema fornisce servizi per l'indicizzazione dei documenti testuali e per la loro organizzazione in collezioni.
- *Ricerca semantica*: oltre alla ricerca classica, basata sul match esatto tra i termini nelle query e nei documenti, lo strumento offre funzionalità di ricerca semantica, in grado di ritrovare documenti in base alla loro correlazione semantica con i termini presenti nella query.
- *Visualizzazione*: il sistema fornisce diversi strumenti grafici per la visualizzazione dei concetti principali contenuti nei documenti e delle correlazioni semantiche esistenti tra documenti o tra concetti.

La prima versione del sistema implementa una pipeline di NLP in grado di processare documenti testuali attraverso operazioni di tokenizzazione, part-of-speech tagging, lemmatizzazione, stemming, chunking e phrase extraction. Il semantic framework si basa su tecniche di *semantica distribuzionale*, ovvero su metodi in grado di costruire *spazi geometrici di concetti* noti come WORDSPACE. Questi approcci si basano sulla cosiddetta *ipotesi distribuzionale*, che afferma che *"le parole che occorrono negli stessi contesti linguistici tendono ad avere significati simili"* [Harris, 1968], ovvero che *il significato delle parole è legato al contesto in cui queste vengono utilizzate*. Quindi, come gli esseri umani deducono il significato di una parola analizzando le altre parole con cui i termini compaiono, così gli algoritmi estraggono le informazioni sul significato di una parola analizzandone il loro uso in collezioni molto vaste di documenti testuali.

Nel progetto TALIA il semantic framework adotta un particolare modello di semantica distribuzionale basato sul Random Inxdexing [Sahlgren, 2005], che rappresenta le parole come vettori in uno spazio geometrico, e dove parole simili hanno vettori simili che sono vicini nello spazio. I servizi di rappresentazione semantica forniti dal semantic framework sono utilizzati per supportare differenti stakeholder, come ad esempio i policy maker, implementando casi d'uso specifici. Ad esempio, chi ha in carico l'analisi dei documenti di progetto, può utilizzare il servizio di *automatic summarization* del testo, che consente di riassumere documenti molto lunghi e complessi in poche frasi [Rossiello *et al.*, 2017], preservando il significato e il contenuto chiave del testo originale. Un altro servizio avanzato possibile grazie al semantic framework è la scoperta di *connessioni latenti* tra concetti presenti nei documenti testuali, vale a dire relazioni non esplicitamente citate nel testo, ma comunque presenti. Un altro aspetto fondamentale è legato all'estendibilità del modello semantico, che prevede l'arricchimento dello spazio dei concetti attraverso la definizione di nuovi concetti. I nuovi concetti possono essere definiti attraverso la combinazione di concetti già esistenti e possono rispecchiare il punto di vista e le esigenze di analisi di singoli stakeholder. La definizione di nuovi concetti si ottiene attraverso la combinazione delle rappresentazioni vettoriali di concetti già presenti nel modello.

In definitiva, attraverso l'uso di tecniche di IA, TALIA mira a governare il processo di policy making attraverso l'utilizzo della conoscenza prodotta dagli investimenti pregressi.

Riferimenti bibliografici

[Harris, 1968] Z. S. Harris. *Mathematical Structures of Language*. Interscience, New York,, 1968.

[Rossiello *et al.*, 2017] G. Rossiello, P. Basile, e G. Semeraro. Centroid-based text summarization through compositionality of word embeddings. In *Proc. of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, 2017.

[Sahlgren, 2005] M. Sahlgren. An Introduction to Random Indexing. In *Proc. of the Methods and Applications of Semantic Indexing, Workshop at the 7th Int. Conf. on Terminology and Know. Eng., TKE*, 2005.