Sistemi di supporto alle decisioni basati su tecnologie di Intelligenza Artificiale per il governo dei processi di e-procurement

Pasquale Lops[1], Marco Di Ciano[2], Nicola Lopane[3], Lucia Siciliani[1], Vincenzo Taccardi[1]
[1]Dip. di Informatica - Università degli Studi di Bari Aldo Moro
[2]InnovaPuglia S.p.A.
[3]Regione Puglia
pasquale.lops@uniba.it, m.diciano@innovapuglia.it, n.lopane@regione.puglia.it,
lucia.siciliani@uniba.it, vincenzo.taccardi@uniba.it

Abstract

Gli appalti pubblici rappresentano un potente strumento d'investimento dei fondi pubblici e sono una risorsa strategica per lo sviluppo economico. Risulta quindi fondamentale migliorare l'operatività delle stazioni appaltanti e sviluppare modelli di valutazione che possano agevolare il rilevamento di eventuali anomalie. In questo contributo, presentiamo la nostra ricerca preliminare volta alla creazione di un sistema a supporto delle decisioni e delle attività di monitoraggio dell'intero ciclo degli investimenti e degli appalti (SIAP). Tale sistema prevede l'utilizzo di tecniche di Intelligenza Artificiale basate sull'elaborazione del linguaggio naturale e machine learning, atte a fornire strumenti che consentano di estrapolare informazioni utili a partire da strutture dati sia di tipo strutturato che di tipo non strutturato (testuale).

1 Introduzione

Gli appalti pubblici in generale, e quelli per l'innovazione in particolare, sono uno strumento strategico a disposizione delle politiche di sviluppo economico e rappresentano un potente strumento di investimento di fondi pubblici. Tuttavia, nel campo specifico della trasparenza e del monitoraggio dell'intero ciclo degli investimenti e degli appalti, risulta fondamentale migliorare da un lato il processo di engagement dei RUP[1], delle Stazioni Appaltanti, delle Amministrazioni e degli Enti Aggiudicatori consentendo loro di assolvere a molti degli adempimenti assegnati in maniera più efficace, efficiente e sostenibile e dall'altro, sviluppare schemi di assessment che mettano in correlazione particolari sequenze logico-temporali di fatti e contenuti che possano essere ricondotti a determinati indicatori di anomalia. In questo quadro di riferimento, tecnologie di Intelligenza Artificiale e sistemi di elaborazione automatica del linguaggio naturale incentrati in particolare sulla lingua italiana rappresentano una nuova frontiera per l'interpretazione della semantica, l'estrazione di concetti e la correlazione di testi e documenti. Le attività di ricerca avviate sono pertanto mirate a sviluppare un sistema in grado, tra le altre cose, di interfacciarsi con le basi dati esistenti, predisporre dataset che soddisfino i requisiti ottimali per consentirne una semplice ed efficace fruizione ed analisi, effettuare estrazione automatica di relazioni tra entità testuali, realizzare test di correlazione tra porzioni di testo anche di differente lunghezza (paragrafi vs intero documento), ricevere e gestire interrogazioni (query) e restituire in formato web-based outcome predefiniti (short report, evidence, reference code, etc.).

[1]Responsabile Unico del Procedimento

2 Metodologia

La figura 1 mostra lo schema dell'architettura ad alto livello per la creazione di un sistema in grado di fornire supporto nella gestione del ciclo degli appalti. Il presente schema è stato concepito con la finalità di poter permettere, nelle fasi successive di progetto, la sua specializzazione. L'architettura è suddivisa in quattro diversi moduli:

- Data Collector
- Pre-Processing
- Tender Analyzer
- Service Tools

Il modulo Data Collector si occupa della raccolta dei dati relativi ai bandi e alle gare. Tali dati possono essere estratti da diverse sorgenti di dati: banche dati a livello europeo come quella di TED[2], nazionale come SIMOG/ANAC[3] o anche regionale come nel caso di EmPULIA[4]. Tuttavia, a seconda delle esigenze previste nei casi d'uso, questa ricerca può coinvolgere anche sorgenti assai diverse e non focalizzate sul tendering come Feed RSS. Per questo motivo l'estrazione dei dati è modularizzata e scomposta in diversi Plug-Ins, ognuno dei quali si occupa di recuperare l'informazione da una specifica sorgente. Ciò permette di aggiungere in maniera più semplice nuove sorgenti o modificare quelle già esistenti. Poiché le sorgenti sono per loro natura eterogenee, è necessario includere una componente di Data Integration che si occupi di combinare l'informazione proveniente da ciascuna di esse. Tale componente può verificare che ci sia un corretto overlap tra sorgenti distinte e segnalare eventuali anomalie.

[2]https://ted.europa.eu/
[3]https://simog.anticorruzione.it/
[4]http://www.empulia.it/

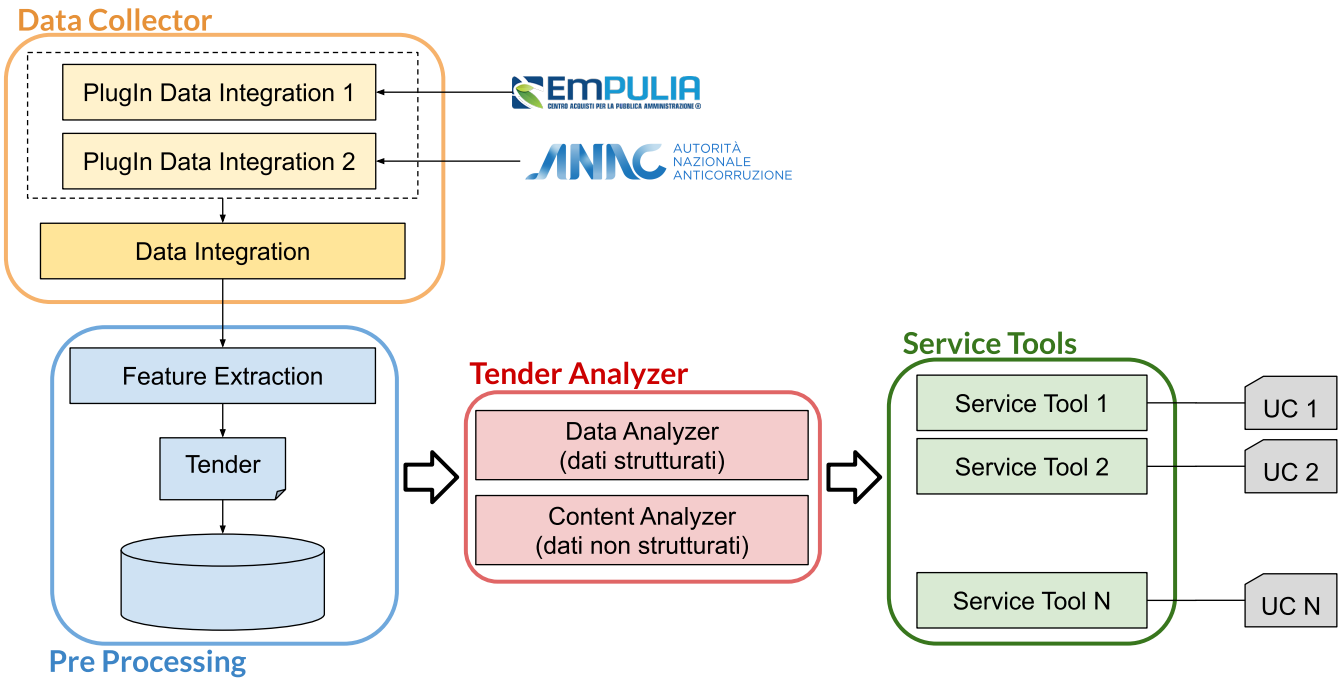

Figura 1: Architettura ad alto livello del sistema SIAP.

Il modulo di Pre-Processing si occupa di sublimare le informazioni estratte tramite il modulo di Data Collection in entità sulle quali possano essere efficacemente effettuati i successivi passi di analisi. Per questo motivo, è necessario stabilire quali sono le caratteristiche di interesse per poter rappresentare in maniera idonea un bando (o a qualsiasi altra forma di entità rilevante per il problema in esame) e successivamente memorizzarle mediante l'ausilio di database o di particolari indici.

Il Tender Analyzer ha lo scopo di effettuare le analisi a partire dalle informazioni estratte dalle diverse sorgenti. Data la natura delle informazioni che possono essere associate ad un bando, questo modulo è suddiviso in due diverse componenti:

- il *Data Analyzer* si occupa di analizzare l'informazione strutturata associata ai bandi: codici (come CUP[5], CIG[6], etc), date, importi, etc.;
- il *Content Analyzer*, invece, effettua l'analisi dell'informazione non strutturata associata ai bandi. Esempi possono essere rappresentati dagli eventuali allegati inerenti al bando (determine, capitolati, etc.). Il Content analyzer effettua l'analisi testuale utilizzando tecniche di NLP.

L'ultimo modulo previsto dall'architettura è rappresentato da una serie di Service Tools. Date le informazioni e le analisi svolte all'interno del modulo precedente, con i service tools vengono effettuate le operazioni specifiche per la realizzazione di particolari casi d'uso. Pertanto, ogni singolo Service Tool sarà collegato a un insieme ben definito di Use Cases.

[5]Codice Unico di Progetto
[6]Codice Identificativo di Gara

3 Applicazioni

Come anticipato nella sezione 2, il framework proposto lavora sia su dati strutturati che su dati non strutturati con l'obiettivo di sfruttare interamente l'informazione associata al bando e catturarne tutti gli aspetti.

3.1 Analisi dei Dati strutturati

L'applicazione sui dati strutturati prevede di ricavare degli indicatori in grado di rilevare anomalie o condizioni di difformità rispetto agli standard normativi e regolamentari nelle attività di procurement [Tóth *et al.*, 2014]. Questi possono essere calcolati sulla base dei dati contenuti all'interno dei database di appalti pubblici disponibili:

- *Valore relativo del bando*: rappresenta il rapporto tra l'offerta vincitrice e il prezzo stimato del bando
- *Variazione delle offerte*: la varianza e cattiva distribuzione delle offerte
- *Differenza tra la prima e la seconda offerta*: data l'importanza fondamentale del primo e del secondo miglior offerente per l'esito di una procedura di gara
- *Struttura di mercato concentrata*: uno dei principali risultati delle offerte di natura potenzialmente collusiva è che la struttura del mercato diventa concentrata su pochi soggetti
- *Struttura di mercato statica*: indica che vi è una varianza molto bassa tra le quote di mercato
- *Vittorie cicliche*: se vi sono particolari pattern che si ripetono ciclicamente (ad esempio, date due aziende A

e B, il pattern delle vincite del bando risulta essere A-B-A-B e così via) questo potrebbe essere indice di difformità regolamentare, e di potenziale comportamento fraudolento

- *Mancanza di offerte*: la mancanza di offerte da parte di un'azienda precedentemente attiva in un dato mercato può indicare la presenza di potenziali condizioni di irregolarità
- *Offerenti superflui*: rappresenta uno dei modi più semplici per simulare situazioni di normale concorrenza mentre in realtà le offerte potrebbero essere concordate
- *Prevalenza di domande erronee*: la concorrenza nei mercati degli appalti può essere simulata da concorrenti che presentano offerte deliberatamente errate
- *Prevalenza dei consorzi*: le offerte congiunte riducono il numero effettivo di parti in competizione, il che può diminuire l'effettiva pressione competitiva.
- *Prevalenza del subappalto*: questo è un modo conveniente per condividere i profitti tra le parti collusive e può anche servire come strumento di garanzia contro eventuali sconfitte nell'assegnazione dell'appalto

I nuovi indicatori così ottenuti vengono utilizzati per arricchire il dataset originale. Questa operazione di *features engineering* può agevolare l'applicazione dei metodi di machine learning, con la finalità di rilevare appalti sospetti la cui assegnazione sia il risultato di eventuali accordi collusivi tra le imprese partecipanti al bando o afferenti a quel mercato.

La criticità principale di quanto sopra proposto è l'assenza di dataset che registrino per un dato appalto la concomitanza di un'inchiesta dell'autorità giudiziaria che abbia accertato la presenza di accordi collusivi tra i partecipanti. Ciò dato, uno dei principali campi di analisi è l'elaborazione di modelli *non supervisionati*, ad esempio clustering o anomaly detection. Inoltre qualora si intendano associare eventuali implicazioni registrate dalle autorità giudiziarie preposte e disponibili in dataset esterni è possibile far ricorso all'elaborazione di modelli di apprendimento *supervisionato*, creando dataset annotati incrociando informazioni ottenute con tecniche di *data mining* su fonti esterne non strutturate (sentenze dell'autorità giudiziaria, Feed RSS, etc) ed i dati sugli appalti invece disponibili (ANAC, TED, etc).

3.2 Analisi di Dati non strutturati

L'attività di analisi di questi dati viene eseguita per mezzo di strumenti automatici capaci di rivelare informazioni tacite tramite l'esame delle strutture grammaticali e semantiche presenti nel testo. Le tecniche usate per raggiungere tali scopi seguono approcci basati sul Natural Language Processing (NLP) [Jurasky e Martin, 2000] e sull'analisi semantica (Semantic analysis).

I risultati che è possibile ottenere mediante l'applicazione di tali tecniche sono molteplici. Il più immediato è la creazione di un motore di ricerca in grado di ricevere interrogazioni e restituire, sulla base di queste ultime, i documenti più rilevanti non solo sulla base di misure di co-occorrenza tra i termini che compaiono nella richiesta (query) e quelli presenti all'interno dei documenti, ma anche sfruttando la similarità semantica.

Inoltre è possibile utilizzare tecniche di Natural Language Understanding e Generation per la creazione di outcome predefiniti a partire dai testi disponibili, come ad esempio riassunti in grado di condensare, con la granularità desiderata, le informazioni presenti all'interno di testi caratterizzati da dimensioni considerevoli [Rossiello *et al.*, 2017]. Questo permette di snellire l'accesso alle informazioni sugli appalti e, di conseguenza, di agevolare la realizzazione di processi aziendali innovativi.

L'analisi dei documenti relativi al bando può permettere il riconoscimento e l'estrazione di informazioni estremamente importanti le quali non sono sempre incluse tra i metadati disponibili sulle varie piattaforme. Un esempio è rappresentato dal codice CPV[7] che identifica la tipologia dell'oggetto del bando ed è usato per la classificazione degli appalti pubblici. Questo codice può essere dunque utilizzato per addestrare un classificatore in grado di assegnare una categoria ad una gara sulla base del contenuto dei documenti ad esso relativi.

Oltre a ciò, sfruttando tecniche di Open Information Extraction [Cassotti *et al.*, 2021], [Siciliani *et al.*, 2021], è possibile effettuare l'estrazione automatica di relazioni tra entità testuali contenute all'interno del testo associato ad un bando. Questo tipo di analisi si rivela di fondamentale importanza per la risoluzione di diversi scenari. Ad esempio, può permettere di incrociare l'informazione estratta direttamente dal bando con quella già presente in forma strutturata per l'individuazione di eventuali anomalie ed utilizzare entrambe le sorgenti dati per effettuare analisi preliminari di mercato che consentano di tener traccia dell'andamento di determinati settori o stazioni appaltanti. Disponendo di una quantità sufficiente di dati, queste informazioni possono essere utilizzate per tracciare dei profili che permettono di identificare in maniera più puntuale situazioni di anomalia che possono essere ulteriormente indagate da parte della stazione appaltante o delle autorità competenti.

4 Quadro sintetico delle informazioni

Considerato il rilevante apparato informativo disponibile, si vuole sintetizzare questa informazione per renderla disponibile al servizio della stazione appaltante. Data la modularità del sistema e la capacità di elaborare dati di diversa natura l'obiettivo è quindi quello di contribuire alla definizione di un "Passaporto" per le imprese o operatori economici che consenta alla stazione appaltante e al RUP di accedere a detto documento per una presa visione in tempo reale delle informazioni a disposizione.

Ad esempio, simulando le verifiche che normalmente vengono effettuate dal RUP e/o autodichiarate dall'operatore economico, l'applicativo si connette ai database disponibili (Registro imprese, Agenzia delle Entrate, DURC online, etc), estrae ed elabora le informazioni rivelanti a ciascuna verifica e restituisce le stesse in forma sintetica. A queste si aggiungono eventuali risultanze dallo storico degli appalti a cui l'azienda ha partecipato (appalti vinti, statistiche sulle gare, la localizzazione degli interventi, KPIs, etc) al fine di ot-

[7] Common Procurement Vocabulary

tenere un contenuto informativo esaustivo ed automatizzato dell'operatore oggetto d'interesse.

5 Conclusioni e sviluppi futuri

Il progetto di ricerca fin qui esposto si propone di investigare le possibilità offerte dalle tecnologie nel campo dell'intelligenza artificiale per offrire agli attori impegnati nel processo degli appalti pubblici una serie di strumenti utili ad agevolarne il lavoro sia nella fase di engagement che di assessment.

Il framework proposto è in grado di operare sia su dati strutturati che dati testuali con un'architettura modulare è scomposta in diversi Plug-Ins, ognuno dei quali si occupa di recuperare l'informazione da una specifica sorgente. Ciò permette di aggiungere in maniera più semplice nuove sorgenti o modificare quelle già esistenti. Allo stesso modo il modulo Tender Analyzer conserva la natura modulare al fine di poter sviluppare applicativi specifici ad ogni fonte di dati che si vuole utilizzare.

Questa capacità di aggiornamento ed espansione lascia aperta la possibilità di sviluppi ed aggiunte futuri, qualora la stazione appaltante proponga nuove esigenze e richieda funzionalità aggiuntive sia su dataset già utilizzati nonché su sorgenti di nuovo interesse.Le politiche di digitalizzazione sempre più incisive a livello centrale e periferico e la disponibilità crescente di dati in formato digitale aprono a scenari di utilizzo ed integrazione di strumenti per l'analisi e l'elaborazione assistita dall'Intelligenza Artificiale. Il prototipo qui proposto quindi si pone all'avanguardia di questo scenario, esplorandone le possibilità ed allo stesso tempo rimanendo predisposto a futuri sviluppi e miglioramenti.

Riferimenti bibliografici

[Cassotti *et al.*, 2021] Pierluigi Cassotti, Lucia Siciliani, Pierpaolo Basile, Marco de Gemmis, e Pasquale Lops. Extracting relations from italian wikipedia using unsupervised information extraction. In Vito Walter Anelli, Tommaso Di Noia, Nicola Ferro, e Fedelucio Narducci, editors, *Proceedings of the 11th Italian Information Retrieval Workshop 2021, Bari, Italy, September 13-15, 2021*, volume 2947 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[Jurasky e Martin, 2000] Daniel Jurasky e James H Martin. Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey*, 2000.

[Rossiello *et al.*, 2017] Gaetano Rossiello, Pierpaolo Basile, e Giovanni Semeraro. Centroid-based text summarization through compositionality of word embeddings. In George Giannakopoulos, Elena Lloret, John M. Conroy, Josef Steinberger, Marina Litvak, Peter A. Rankel, e Benoît Favre, editors, *Proceedings of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres, MultiLing@EACL 2017, Valencia, Spain, April 3, 2017*, pages 12–21. Association for Computational Linguistics, 2017.

[Siciliani *et al.*, 2021] Lucia Siciliani, Pierluigi Cassotti, Pierpaolo Basile, Marco de Gemmis, Pasquale Lops, e Giovanni Semeraro. Extracting relations from italian wikipedia using self-training. In Elisabetta Fersini, Marco Passarotti, e Viviana Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[Tóth *et al.*, 2014] Bence Tóth, Mihály Fazekas, Ágnes Czibik, e István János Tóth. Toolkit for detecting collusive bidding in public procurement. with examples from hungary. *Report number: CRC-WP/2014:02*, 2014.