RIVISTA ITALIANA
DI INFORMATICA E DIRITTO
Periodico internazionale di diritto, giustizia e tecnologie

ISSN 2704-7318 • n. 2/2026 • DOI 10.32091/RIID0287 • articolo sottoposto a peer review • pubblicato in anteprima il 18 set. 2026

Enea Ahmedhodzic – Sara Gagliardi – Andrea Mario Trentini

Aumentare la trasparenza normativa con similarità semantica e Akoma Ntoso

La trasparenza normativa dipende dalla possibilità di ricostruire e verificare i legami tra fonti, oltre che dalla mera disponibilità dei testi. Il recepimento delle direttive Ue è un caso emblematico, perché gli obblighi europei vengono spesso riformulati e distribuiti su più disposizioni nazionali, rendendo la ricostruzione manuale delle corrispondenze onerosa e poco replicabile. Questo lavoro propone una procedura riproducibile (pipeline) per supportare l'analisi comparativa tra direttive Ue e decreti legislativi italiani di attuazione, mantenendo un collegamento esplicito tra risultati computazionali e testo giuridico. Il workflow normalizza le fonti su una rappresentazione strutturata. I decreti italiani sono acquisiti da Normattiva in formato XML Akoma Ntoso; le direttive Ue sono scaricate da EUR-Lex nella versione HTML "OJ" e convertite in una struttura XML coerente per metadati e segmentazione in articoli e paragrafi. Su tali unità si calcolano la similarità semantica (Sentence-BERT, similarità coseno) e un indicatore complementare di overlap lessicale; i candidati vengono ordinati e restituiti tramite output ispezionabili e visualizzazioni (heatmap e grafi) utili alla revisione umana.

Trasparenza – Akoma Ntoso – Similarità semantica – Ricerca multidisciplinare

Enhancing regulatory transparency through semantic similarity and Akoma Ntoso

Regulatory transparency depends on the possibility of reconstructing and verifying the links between sources, rather than on the mere availability of texts. The transposition of EU directives is an emblematic case, because European obligations are often reformulated and distributed across multiple national provisions, making the manual reconstruction of correspondences costly and difficult to replicate. This work proposes a pipeline to support comparative analysis between EU directives and the Italian legislative decrees that implement them, while keeping an explicit connection between computational results and the legal text. The workflow normalizes the sources into a structured representation. Italian decrees are acquired from Normattiva in Akoma Ntoso XML format, while EU directives are downloaded from EUR-Lex in the "OJ" HTML version and converted into a consistent XML structure with coherent metadata and segmentation into articles and paragraphs. Semantic similarity (Sentence-BERT, cosine similarity) and a complementary lexical overlap indicator are calculated on these units; the candidates are then ranked and provided through inspectable outputs and visualizations (heatmaps and graphs) designed for human review.

Transparency – Akoma Ntoso – Semantic similarity – Multidisciplinary research

E. Ahmedhodzic è dottoranda in Informatica presso il Dipartimento di Informatica "Giovanni degli Antoni" dell'Università degli Studi di Milano. S. Gagliardi ha svolto questo lavoro nell'ambito del corso di laurea magistrale in Informatica presso il Dipartimento di Informatica "Giovanni degli Antoni" dell'Università degli Studi di Milano. A.M. Trentini è ricercatore presso il Dipartimento di Informatica "Giovanni degli Antoni" dell'Università degli Studi di Milano

Sommario: 1. Introduzione. – 2. Stato dell'arte e sfide del recepimento automatizzato. – 3. Fonti normative, materiali e criteri di selezione. – 4. Metodologia. – 4.1. Normativa italiana di recepimento. – 4.2. Conversione delle direttive Ue da HTML EUR-Lex a XML pseudo-Akoma Ntoso e criticità. – 4.3. Da Akoma Ntoso a JSONL e calcolo degli embedding. – 4.4. Calcolo della similarità, selezione dei candidati e sintesi dei risultati. – 4.5. Output ispezionabili, visualizzazioni e tracciabilità del confronto. – 5 Caso di studio e risultati. – 5.1. Confronto tra paragrafi, la matrice di similarità come mappa. – 5.2. Sintesi a livello di articolo, heatmap e grafo dei best match. – 5.3. Valutazione quantitativa su confronti incrociati. – 6. Valutazione. – 7. Conclusioni.

1. Introduzione

Quando pensiamo alla pubblica amministrazione e alla sua trasparenza[1], torna spesso alla mente la celebre immagine della "casa di vetro", evocata da Filippo Turati nel 1908 nel dibattito parlamentare sull'agire amministrativo[2]. Oggi, in un contesto segnato dalla digitalizzazione e dall'uso crescente di strumenti automatici, la trasparenza viene spontaneamente associata all'accesso: la pubblicazione di atti, portali che rendono disponibili documenti, archivi aperti da cui scaricare leggi e testi ufficiali. L'accesso, però, è solo il primo passo. La trasparenza diventa effettiva quando le informazioni sono anche rintracciabili, confrontabili e verificabili, almeno nei passaggi essenziali[3].

Nel caso della norma questo punto è particolarmente evidente: i testi possono essere pubblici e disponibili, ma orientarsi tra rinvii, definizioni, eccezioni e formule ricorrenti richiede competenze e tempo. Per "competenze" intendiamo, prima di tutto, quelle del cittadino comune, che non sempre padroneggia il linguaggio giuridico e, spesso, neppure gli strumenti digitali attraverso cui le fonti vengono rese disponibili. Ridurre la distanza tra "testo disponibile" e "testo comprensibile e controllabile" è quindi uno degli aspetti che rende la trasparenza normativa un obiettivo difficile, soprattutto per chi non è giurista e deve comunque capire quali obblighi o diritti derivino da una certa disciplina. Nel diritto penale, questa difficoltà pesa ancora di più: l'ignoranza della legge non è di norma una scusante, a eccezione dei casi di ignoranza inevitabile riconosciuti dalla giurisprudenza costituzionale[4].

Un contesto in cui questa difficoltà emerge è nel recepimento del diritto dell'Unione europea. Le direttive, per loro natura, richiedono misure nazionali di attuazione e lasciano margini di adattamento. In concreto, ciò significa che la relazione tra fonte europea e fonte nazionale non è necessariamente "uno a uno" e raramente è tracciabile in modo automatico. Lo stesso contenuto può essere riformulato, distribuito su più disposizioni, inglobato in norme già esistenti o ricollocato in parti

1. Questa ricerca è stata finanziata dal Ministero dell'Università e della Ricerca (MUR) nell'ambito del Decreto Ministeriale n. 118 del 2 marzo 2023, PNRR Missione 4, Componente 1, Investimento 4.1 ("Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale"). Il soggetto finanziatore non ha avuto alcun ruolo nella progettazione dello studio, nell'esecuzione, nell'analisi o nella redazione dei risultati.

2. Cfr. Atti del Parlamento italiano, Camera dei deputati, sessione 1904-1908, 17 giugno 1908.

3. Lima–Brito–Nascimento et al. 2022.

4. Cfr. Corte cost., 24 marzo 1988, n. 364, sull'art. 5 c.p. e ignoranza inevitabile, in relazione all'art. 27 Cost.

diverse dell'ordinamento interno. Il recepimento può quindi risultare frammentato e disperso su più atti e articoli, rendendo meno immediata la ricostruzione del percorso normativo. Ne consegue che individuare dove "finisce" un obbligo europeo e come viene espresso nel diritto nazionale è spesso un lavoro manuale, lungo e poco replicabile, soprattutto quando non si vuole guardare a un singolo caso, ma a un insieme di direttive o a un intero settore[5].

Il problema, in termini operativi, non è stabilire la correttezza giuridica del recepimento, ma ricostruire in modo tracciabile dove e come gli obblighi della direttiva siano riflessi nelle disposizioni nazionali. Su scala, questo lavoro diventa costoso, poco scalabile e difficilmente riproducibile, rendendo complesso confrontare casi diversi con criteri omogenei. È importante chiarire subito un punto: l'interpretazione giuridica resta centrale e la conoscenza umana rimane il riferimento ultimo. Un sistema informatico non "decide" se un recepimento sia corretto, né può sostituire il contesto che un esperto applica nel valutare un obbligo. Tuttavia, esiste una fascia di attività preliminari e strumentali che può essere efficacemente supportata in modo computazionale, ad esempio individuare rapidamente porzioni di testo plausibilmente corrispondenti, evidenziare dove la somiglianza è alta o bassa, segnalare zone che meritano revisione umana e produrre viste d'insieme utili a non perdersi tra decine o centinaia di paragrafi. La tecnologia non sostituisce il giudizio umano, ma riduce il costo della ricerca e aumenta la verificabilità del percorso analitico[6].

Negli ultimi anni, l'informatica giuridica e le tecniche di Natural Language Processing (NLP) hanno ampliato gli strumenti disponibili per trattare testi normativi, con risultati promettenti ma anche con difficoltà note, tra cui la lunghezza dei documenti, il linguaggio specialistico e la disponibilità limitata di risorse annotate[7]. Un filone particolarmente utile per il confronto tra testi è quello della Semantic Textual Similarity (STS), che mira a stimare con un punteggio continuo la vicinanza semantica tra porzioni di testo[8]. La STS è interessante proprio perché il recepimento raramente è una copia letterale in quanto può esserci corrispondenza anche quando le parole cambiano e la struttura viene riorganizzata.

Operativamente, la STS viene qui implementata tramite embedding a livello di paragrafo calcolati con l'approccio Sentence-BERT (SBERT)[9] usando la libreria sentence-transformers[10] confrontati con similarità coseno. Al tempo stesso, la somiglianza semantica non coincide con equivalenza giuridica e, quindi, il suo impiego deve essere accompagnato da prudenza e da strumenti che rendano i risultati ispezionabili[11].

In questo quadro, il presente lavoro propone una pipeline riproducibile per l'analisi comparativa tra direttive Ue e norme nazionali italiane di attuazione, con un obiettivo concreto: aumentare la tracciabilità tra fonti e rendere più leggibile un processo normativo complesso, collegando direttamente informatica e trasparenza.

Il sistema produce report e visualizzazioni pensati per accompagnare la revisione umana dei candidati. La scelta metodologica di base è semplice: partire dalla struttura del documento e lavorare su paragrafi come unità di analisi. Per questo, un elemento abilitante è l'uso di uno standard aperto e strutturato per rappresentare il testo giuridico, e il riferimento è Akoma Ntoso[12] (AKN nel seguito).

Il workflow è progettato come un processo end-to-end in quattro passaggi.

(i) Le direttive Ue vengono convertite dall'HTML "OJ" di EUR-Lex in una struttura pseudo-AKN, con una segmentazione coerente in articoli e paragrafi e con metadati essenziali rappresentati mediante gli elementi FRBR di Akoma

5. Nanda–Siragusa–Di Caro–Boella 2019.
6. Ariai–Mackenzie–Demartini 2025.
7. *Ibidem*.
8. Agirre–Cer–Diab–Gonzalez-Agirre 2012.
9. Sentence-BERT (SBERT), sbert.net.
10. Libreria sentence-transformers: *https://pypi.org/project/sentence-transformers*.
11. Agirre–Cer–Diab–Gonzalez-Agirre 2012.
12. Akoma Ntoso: *https://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html*.

Ntoso. Tali metadati sono utilizzati nel workflow solo come riferimenti identificativi, per identificare il documento processato e mantenere il collegamento con la fonte di origine.

(ii) I paragrafi vengono estratti e normalizzati in un formato lineare "analysis-ready" (JSONL): una riga corrisponde a un paragrafo, con campi omogenei per Ue e Italia, così che la similarità sia calcolata su unità testuali confrontabili e non risenta delle differenze di struttura tra le fonti.

(iii) Su questi segmenti si calcolano gli embedding e si producono le corrispondenze candidate tramite ranking per similarità.

(iv) Infine, la pipeline genera output ispezionabili che rendono verificabile il percorso di confronto tra direttiva e recepimento. L'output non è un "giudizio" sul recepimento, ma un insieme ordinato di candidati di corrispondenza che evidenzia i match più probabili e, soprattutto, le aree ambigue che richiedono revisione umana.

Questa impostazione consente una lettura operativa, perché orienta l'analisi verso i punti di maggiore similarità e verso le zone incerte, e una lettura diagnostica, perché la dispersione delle corrispondenze o la difficoltà stessa di trovarle può essere interpretata come segnale misurabile di complessità e scarsa tracciabilità del processo di attuazione.

Per mantenere netta la distinzione tra indicatore computazionale e valutazione giuridica, includiamo una valutazione manuale su un campione annotato e riportiamo misure ordinali di concordanza tra predizione del sistema e valutazione umana (kappa quadratica) e di errore medio (MAE) sulla scala ordinale. Inoltre, prima dell'embedding, vengono eseguiti controlli di qualità sui segmenti (identificativi mancanti, paragrafi vuoti) e una pulizia mirata per rimuovere parti ripetitive o non informative (boilerplate, intestazioni ricorrenti, formule standard).

2. Stato dell'arte e sfide del recepimento automatizzato

Negli ultimi anni l'analisi computazionale del diritto si è consolidata come linea di ricerca che mira a supportare l'esplorazione e il confronto di grandi collezioni di testi giuridici, difficili da gestire esclusivamente con procedure manuali. La letteratura recente evidenzia opportunità e limiti. I modelli linguistici e misure semantiche possono migliorare la ricerca e l'organizzazione dell'informazione, ma nel dominio legale incontrano ostacoli strutturali, tra cui linguaggio specialistico, documenti lunghi e forte dipendenza dal contesto. Questi limiti diventano più visibili quando l'obiettivo non è il retrieval generico, ma la ricostruzione di relazioni tra fonti diverse, come nel recepimento.

Un filone rilevante in questo quadro è la Semantic Textual Similarity, che stima con un punteggio continuo la vicinanza semantica tra porzioni di testo anche in assenza di sovrapposizione lessicale[13]. Con l'introduzione dei transformer[14], e poi con la diffusione dei modelli pre-addestrati basati su questa architettura, è diventato comune rappresentare frasi e paragrafi tramite embedding confrontabili; in molti scenari di ricerca e allineamento testuale, questo approccio consente un matching più robusto rispetto alle sole parole chiave. In particolare, modelli come Sentence-BERT producono embedding a livello di frase o paragrafo, consentendo di ordinare le possibili corrispondenze per similarità, così da guidare l'ispezione manuale[15].

Nel dominio giuridico la similarità va interpretata come un supporto e non come equivalenza in quanto le definizioni, eccezioni e condizioni procedurali possono introdurre differenze giuridicamente rilevanti anche in testi con punteggi alti. L'analisi basata su "best match", intesa come punteggio più alto di similarità tra due elementi, può ignorare il contesto normativo più ampio e non catturare l'equivalenza logica, oltre a soffrire quando il recepimento è disperso in più punti del testo nazionale. Per questo, oltre al ranking, serve che il risultato resti controllabile sul testo, con evidenze che rendano chiaro da dove arriva la similarità.

Nel caso del recepimento, la difficoltà è infatti anche strutturale. Gli obblighi possono essere riformulati, frammentati e distribuiti su più disposizioni, oppure integrati in norme nazionali preesistenti. Questo rende molto difficile ottenere un allineamento lineare articolo per articolo. Di

13. Agirre–Cer–Diab–Gonzalez-Agirre 2012.

14. Vaswani–Shazeer–Parmar et al. 2017.

15. Reimers–Gurevych 2019.

conseguenza, anche quando la similarità individua candidati plausibili, i risultati devono restare ispezionabili, così che l'analista possa verificare dove la corrispondenza è Forte, dove è Moderata e dove Debole.

Per rendere questo tipo di analisi praticabile, la rappresentazione strutturata del testo normativo è un prerequisito. AKN è uno standard di riferimento per rappresentare documenti normativi e giuridici in XML, separando struttura, contenuto e metadati e favorendo interoperabilità e riuso. Uno standard comune consente di rappresentare testi normativi secondo una struttura condivisa (ad esempio, articoli e paragrafi), facilitando segmentazioni coerenti e confronti più comparabili tra fonti eterogenee[16].

Nella pratica, la disponibilità di strutture omogenee non è simmetrica tra le fonti: i testi italiani sono acquisibili da Normattiva in AKN, mentre per le direttive Ue è necessario un passaggio preliminare di normalizzazione, discusso nella metodologia.

In letteratura si parla di National Implementing Measures (NIM), ossia gli atti nazionali che gli Stati membri notificano come misure di attuazione di una direttiva. Sul tema NIM e tracciabilità, i problemi ricorrenti sono due: trovare automaticamente buoni candidati e valutarli in modo comparabile, su dati confrontabili. Nanda et al.[17] affrontano il primo punto: usano similarità testuale e semantica per riconoscere le misure nazionali e confrontano approcci supervisionati e non supervisionati.

Nella pratica, senza dataset condivisi è difficile confrontare i risultati tra studi diversi perché cambiano casi, fonti e criteri. Ferrod et al.[18] mettono a disposizione un dataset per il pairing tra direttive Ue e misure nazionali. Impostano il compito come semantic search e ranking e riportano baseline che vanno da metodi lessicali a modelli basati su Sentence Transformers.

Nei testi normativi lunghi, il punteggio è utile solo se resta collegato alle porzioni testuali che lo producono.

Rispetto a questi lavori, il contributo qui non è proporre una nuova metrica di similarità, ma rendere il problema affrontabile nel concreto con un processo riproducibile e verificabile.

Il nostro obiettivo è lavorare su articoli e paragrafi, portando fonti diverse su una rappresentazione comune. La metodologia descrive come questa normalizzazione venga realizzata tramite pseudo-AKN, AKN e JSONL, così da calcolare la similarità su segmenti confrontabili.

Inoltre, AKN non è trattato come una garanzia automatica di qualità. Prima dell'embedding eseguiamo controlli e filtri sugli XML, ad esempio identificativi incoerenti, paragrafi vuoti e boilerplate, così da rendere più chiaro quali segmenti entrano effettivamente nel confronto.

3. Fonti normative, materiali e criteri di selezione

In questo studio è stato costruito un corpus di dieci coppie, mettendo insieme le direttive Ue e i relativi decreti legislativi italiani di recepimento. Le direttive Ue sono acquisite da EUR-Lex in formato HTML. I recepimenti italiani sono acquisiti da Normattiva (normattiva.it) in formato XML AKN. I testi italiani sono disponibili da Normattiva in XML AKN; per le direttive Ue, invece, non avendo un AKN pronto all'uso, abbiamo normalizzato gli HTML di EUR-Lex in una struttura pseudo-AKN. Usiamo il termine "pseudo" per indicare che l'obiettivo non è produrre un XML formalmente validato secondo lo standard ufficiale in ogni dettaglio, ma ottenere una rappresentazione coerente rispetto alle esigenze analitiche di segmentazione e tracciabilità interna.

La selezione del corpus segue un criterio di eterogeneità: abbiamo incluso ambiti diversi, dall'accessibilità al whistleblowing, dagli open data ai giocattoli, fino a strumenti di misura e attrezzature a pressione. Le fonti sono esclusivamente istituzionali e abbiamo verificato la relazione tra direttiva e testo nazionale tramite riferimenti espliciti nell'atto. La Tabella 1 riassume le coppie utilizzate nell'analisi. La nomenclatura it_xxx ed eu_xxx è adottata anche nelle figure delle sezioni 5 e 6.

4. Metodologia

Questa sezione descrive la metodologia applicata per confrontare direttive Ue e decreti legislativi

16. Palmirani–Vitali 2011.
17. Nanda–Siragusa–Di Caro–Boella 2019.
18. Ferrod–Bondarenko–Audrito–Siragusa 2023.

Ambito	Direttiva UE	Decreto legislativo italiano (recepimento)
Accessibilità PA	Direttiva (UE) 2016/2102 (ue_accessibilita)	D.Lgs. 10 agosto 2018, n. 106 (it_accessibilita)
Whistleblowing	Direttiva (UE) 2019/1937 (ue_whistleblowing)	D.Lgs. 10 marzo 2023, n. 24 (it_whistleblowing)
Open data	Direttiva (UE) 2019/1024 (ue_opendata)	D.Lgs. 8 novembre 2021, n. 200 (it_opendata)
Accesso a informazione ambientale	Direttiva 2003/4/CE (ue_accesso_ambientale)	D.Lgs. 19 agosto 2005, n. 195 (it_accesso_ambientale)
Attrezzature a pressione	Direttiva (UE) 2014/68 (ue_attrezzature_pressione)	D.Lgs. 25 febbraio 2000, n. 93 - testo consolidato al 19 luglio 2016 (it_attrezzature_pressione)
Strumenti di misura	Direttiva (UE) 2014/32 (ue_strumenti)	D.Lgs. 2 febbraio 2007, n. 22 - testo consolidato al 26 maggio 2016 (it_strumenti)
Imbarcazioni da diporto	Direttiva (UE) 2013/53 (ue_imbarcazioni)	D.Lgs. 11 gennaio 2016, n. 5 (it_imbarcazioni)
Sicurezza dei giocattoli	Direttiva 2009/48/CE (ue_giocattoli)	D.Lgs. 11 aprile 2011, n. 54 (it_giocattoli)
Condizioni di lavoro trasparenti e prevedibili	Direttiva (UE) 2019/1152 (ue_lavoro)	D.Lgs. 27 giugno 2022, n. 104 (it_lavoro)
Ambiente: plastica monouso	Direttiva (UE) 2019/904 (ue_plastica_monouso)	D.Lgs. 8 novembre 2021, n. 196 (it_plastica_monouso)

Tab. 1 — *Coppie direttiva–decreto legislativo analizzate*

italiani di recepimento a livello di paragrafo, con aggregazione dei risultati anche a livello di articolo e, in forma riassuntiva, anche a livello di legge considerando esclusivamente gli articoli. Questa scelta introduce un limite operativo: il sistema non ricostruisce la corrispondenza a livello di "intera legge" su tutto il testo, ma attraverso l'aggregazione delle unità strutturate disponibili. Il punteggio attribuito al confronto è interamente computazionale ed è semplicemente un indicatore di vicinanza semantica, non un giudizio di equivalenza giuridica. L'intera pipeline è sviluppata in Python e organizza le fasi di acquisizione, normalizzazione, segmentazione e confronto in moduli eseguibili in modo riproducibile, a parità di input e parametri.

Il lavoro sfrutta le potenzialità dello standard AKN in quanto consente di rappresentare documenti giuridici preservando struttura e metadati, rendendo possibile collegare ogni risultato analitico a un punto identificabile del testo[19].

Il confronto avviene a livello di paragrafo, ed è essenziale sapere a quale articolo appartiene ogni segmento e da quale punto del testo proviene. Questa informazione strutturale permette di riorganizzare e aggregare i risultati (paragrafo, articolo, legge) mantenendo la tracciabilità.

4.1. Normativa italiana di recepimento

Normattiva rende disponibili gli atti in XML AKN, quindi la struttura e gli identificativi sono in gran parte già presenti e riutilizzabili. Tuttavia, anche in questo caso la disponibilità di uno standard non elimina i problemi di qualità. Nel corpus emergono identificativi mancanti, segmenti vuoti o non informativi e porzioni ripetitive che, se non gestite, influenzano direttamente gli embedding e quindi i risultati di similarità. Si sono analizzati tutti i recepimenti scaricati da Normattiva per verificare che ogni unità sia pronta per un'analisi semantica a livello di paragrafo.

I formati sono risultati tutti conformi allo schema di validazione di AKN tramite un validatore online (*The Akoma Ntoso Schema Validator*), ma, nonostante ciò, hanno creato problemi quando si

19. Palmirani–Vitali 2011.

vuole estrarre un singolo paragrafo: in alcuni casi, tag <p> contengono stringhe non informative (ad esempio parentesi, separatori, marcatori editoriali, oppure token come "AGGIORNAMENTO"), che introdurrebbero rumore nel calcolo di similarità.

Dunque, prima della fase di embedding, è stato introdotto un passaggio preliminare di controllo qualità e pulizia, con tre obiettivi concreti:

(i) Verificare la presenza degli eId nei paragrafi, poiché sono necessari per mantenere un collegamento stabile tra risultati e testo. Quando l'eId è assente, il segmento resta comunque analizzabile, ma perde una parte della tracciabilità. Nel nostro lavoro l'eId mancante comporta riduzione di informazioni, il paragrafo non viene numerato ma mantenuto come segmento "paragrafo" riconducibile al solo articolo di appartenenza.
(ii) Escludere i paragrafi troppo corti, al di sotto dei 20 caratteri, perché nella pratica corrispondono spesso a contenuti non informativi citati precedentemente e aumenterebbero la produzione di corrispondenze spurie.
(iii) Contabilizzare le criticità più comuni (paragrafi vuoti, heading-only, boilerplate e contenuti ripetitivi), perché incidono direttamente su ciò che entra nel confronto semantico e quindi sulla riproducibilità dei risultati.

4.2. Conversione delle direttive Ue da HTML EUR-Lex a XML pseudo-Akoma Ntoso e criticità

Le direttive Ue sono acquisite da EUR-Lex in formato HTML (versione "OJ"). Accanto all'HTML va considerato anche FORMEX, formato XML adottato dall'Ufficio delle pubblicazioni dell'Unione europea per rappresentare in modo strutturato le pubblicazioni ufficiali dell'Ue[20]. FORMEX rende più esplicite componenti come metadati, preambolo, considerando, articoli e paragrafi, e potrebbe quindi offrire una base più solida per una conversione verso AKN. In questo lavoro si è scelto tuttavia di partire dall'HTML OJ in quanto il formato era direttamente accessibile da EUR-Lex, sufficientemente regolare nel corpus considerato e adeguato all'obiettivo principale, cioè ricostruire una struttura pseudo-AKN utile alla segmentazione e al confronto semantico. La soluzione metodologicamente più coerente sarebbe la disponibilità di testi Ue già forniti in Akoma Ntoso. A oggi, tuttavia, AKN4EU[21] si configura come formato in sviluppo per lo scambio strutturato dei documenti giuridici nel processo decisionale dell'Ue. L'uso di FORMEX richiederebbe quindi un modulo specifico di acquisizione e normalizzazione, che potrà essere valutato come estensione futura.

Per rendere l'HTML utilizzabile nel confronto paragrafo–paragrafo, è stato introdotto un passaggio di normalizzazione verso XML pseudo-AKN.

La conversione inizia con l'estrazione dei metadati principali (tipo di atto, numero, data, titolo). La data viene normalizzata in formato ISO (YYYY-MM-DD) a partire dai formati ricorrenti (testuale o numerico) e, in caso di mancata estrazione, viene applicato un valore di fallback. Su questi elementi viene poi costruito un blocco FRBR (Work/Expression/Manifestation) con un URI base derivato da tipo di atto, data e numero. Nel contesto di questa pipeline, gli elementi FRBR sono impiegati come metadati identificativi, con la funzione di mantenere stabile il riferimento interno al documento convertito e di ricondurre i segmenti analizzati alla fonte di partenza.

Successivamente il testo viene riorganizzato distinguendo porzioni con funzioni diverse. Il convertitore analizza i paragrafi che precedono il primo articolo e, tramite pattern ricorrenti, individua la formula di emanazione (ad es. "IL PARLAMENTO EUROPEO", "LA COMMISSIONE EUROPEA"), le citazioni introdotte da "visto/vista/visti", l'avvio e la sequenza dei "considerando" (ad es. "considerando quanto segue") e la formula di adozione (ad es. "HANNO ADOTTATO LA PRESENTE").

Questa separazione è utile per il confronto semantico perché evita di mescolare preambolo e corpo normativo, che hanno registri e funzioni diverse. In caso contrario, il sistema rischia di produrre similarità "di contorno" poco informative rispetto agli obblighi e alle disposizioni che si vogliono allineare.

Il corpo normativo viene ricostruito individuando gli articoli e, per ciascuno, estrapolando: numero dell'articolo, titolo (heading) e paragrafi.

20. Ufficio delle pubblicazioni dell'Unione europea, FORMEX, EU Vocabularies, in op.europa.eu.
21. Ufficio delle pubblicazioni dell'Unione europea, *AKN4EU*, EU Vocabularies, in op.europa.eu.

Ogni paragrafo riceve un identificativo interno (`eId`) che consente di mantenere la tracciabilità tra segmento e posizione nel testo.

La segmentazione dei paragrafi segue una logica progressiva. Quando la struttura della pagina lo consente, il convertitore riconosce i blocchi già separati e genera un paragrafo per blocco. Quando sono presenti elenchi, recupera e conserva i marker (ad esempio "a)", "b)", ecc.) così da non perdere la struttura logica dell'articolazione. Se il paragrafo inizia con una numerazione interna ("1.", "2)"), la numerazione viene separata dal testo descrittivo e memorizzata come elemento distinto. Infine, vengono esclusi i segmenti privi di contenuto normativo (ad esempio titoli isolati o "Articolo X" senza testo), perché non informativi per l'analisi semantica.

Quando non sono disponibili blocchi numerati, la conversione utilizza un fallback: raccoglie i `<p>` con classi "normali" dell'articolo e li concatena in un unico paragrafo. Questa strategia evita la perdita di contenuto, ma produce unità più grandi e quindi meno comparabili con la granularità tipica dei paragrafi italiani.

Questa procedura ha limiti che incidono sulla qualità del confronto e vanno dichiarati. L'estrazione dipende dalla struttura `OJ` e dalle classi HTML, quindi varianti di markup possono ridurre la precisione. Anche la separazione di citazioni e "considerando" si basa su pattern linguistici (ad esempio "visto…", "considerando…"), e può risultare incompleta quando cambiano formulazioni o impaginazione.

Per queste ragioni il risultato viene definito pseudo-AKN. L'obiettivo non è garantire la piena conformità a uno standard AKN "ufficiale" per ogni HTML, ma ottenere una struttura abbastanza stabile da supportare segmentazione e tracciabilità interna. La conversione è pensata per rendere osservabile la similarità tra paragrafi in modo controllabile e trasparente, con il compromesso che in alcuni casi la fedeltà strutturale possa essere parziale.

4.3. Da Akoma Ntoso a JSONL e calcolo degli embedding

Dopo la normalizzazione delle fonti in una struttura XML omogenea, il passo successivo è rendere il corpus adatto alle fasi NLP senza perdere la tracciabilità verso il testo giuridico. Per questo la pipeline introduce un formato intermedio lineare (JSONL), in cui ogni riga corrisponde a una singola unità di analisi e conserva i metadati necessari per risalire al punto preciso del documento. In questo modo separiamo la gestione della gerarchia normativa (articoli, paragrafi, identificatori) dal calcolo di rappresentazioni vettoriali e similarità.

4.3.1. Estrazione e segmentazione in JSONL

La pipeline legge i documenti AKN ed estrae, quando disponibili, i riferimenti FRBR (FRBRuri e FRBRthis). Questi riferimenti sono conservati come metadati identificativi del documento processato e non intervengono nel calcolo degli embedding né nella misura di similarità. Il testo viene dunque elaborato a partire dagli articoli e dai paragrafi con l'obiettivo di ricostruire unità che siano allo stesso tempo confrontabili e richiamabili.

Nella pratica, i file reali presentano piccole variazioni e inconsistenze. Ad esempio, quando un paragrafo non possiede un identificativo, esso viene ricostruito con una regola stabile basata su articolo e posizione, così da non perdere la possibilità di "puntare" al segmento nei report successivi.

Un altro caso ricorrente riguarda articoli privi di paragrafi espliciti. Invece di scartarli, il nostro workflow li tratta come un unico segmento, evitando la perdita di contenuto e mantenendo comunque un collegamento all'articolo di appartenenza.

Prima di salvare i segmenti in JSONL viene svolta una pulizia mirata con la normalizzazione degli spazi, decodifica di entità HTML e correzione di caratteri degradati, in modo da evitare che problemi di encoding si trasformino in rumore nei punteggi di similarità. In parallelo, vengono esclusi i segmenti non informativi, in particolare paragrafi "solo titolo" e stringhe che contengono soltanto separatori, punteggiatura o numerazioni isolate (ad esempio numerazioni senza testo).

4.3.2. Embedding dei segmenti e gestione della lunghezza

A ciascun segmento JSONL viene associato un embedding calcolato tramite Sentence-BERT. Il modello adottato di default è *sentence-transformers/paraphrase-multilingual-mpnet-base-v2*[22], scelto come baseline per il confronto UE–IT perché è

22. Sentence-transformers/paraphrase-multilingual-mpnet-base-v2: *https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2*.

multilingue e quindi adatto anche alla lingua italiana; inoltre, essendo un bi-encoder, consente di pre-calcolare e riutilizzare gli embedding dei paragrafi e di stimare la distanza semantica tramite similarità coseno[23], con un costo computazionale contenuto.

I Transformer hanno un limite di lunghezza in input; per questo, i segmenti che superano la soglia massima vengono suddivisi in chunk di 512 token con stride 128. Ogni chunk viene rappresentato tramite un embedding separato e l'embedding finale del segmento è ottenuto tramite media pesata sulla lunghezza dei chunk. Questa scelta consente di evitare la troncatura dei segmenti più lunghi e di non concentrare il calcolo della similarità solo sulla parte iniziale del testo. L'overlap tra chunk mantiene una continuità minima tra porzioni adiacenti dello stesso paragrafo, mentre la media pesata permette di ottenere un'unica rappresentazione del paragrafo. La segmentazione a lunghezza fissa, in questa fase sperimentale, è stata adottata per mantenere espliciti e controllabili i parametri della procedura. Strategie alternative, più sensibili ai confini linguistici o strutturali del testo, potranno essere valutate in applicazioni future, purché mantengano il collegamento con la struttura AKN. Per favorire la replicabilità degli esperimenti e contenere i costi computazionali, gli embedding possono essere memorizzati in cache.

L'intera pipeline è sviluppata in Python 3.12 e potrà essere rilasciata con licenza libera.

4.4. Calcolo della similarità, selezione dei candidati e sintesi dei risultati

Una volta calcolati gli embedding per ciascun segmento (paragrafo), la pipeline costruisce una matrice di similarità tra i segmenti Ue e quelli italiani usando la similarità coseno, dove per ogni paragrafo della direttiva viene calcolata la vicinanza semantica con ogni paragrafo del recepimento. Questo passaggio permette di osservare l'intero spettro delle corrispondenze, da quelle più forti a quelle più deboli, e di individuare i candidati da ispezionare.

4.4.1. Matrice di similarità e metriche descrittive

Per ogni coppia direttiva–recepimento, la pipeline produce una matrice M in cui ogni cella $M_{i,j}$ rappresenta la similarità tra il paragrafo i del testo Ue e il paragrafo j del testo italiano. A partire da M vengono calcolate alcune statistiche descrittive (media, deviazione standard, quantili), utili come sintesi del comportamento della coppia e come supporto alle scelte di visualizzazione.

Queste statistiche aiutano anche a confrontare coppie diverse in termini di distribuzione dei punteggi e a mettere in evidenza, nelle viste sintetiche e nei grafi, solo le relazioni più forti.

4.4.2. Selezione dei candidati e best-match

L'obiettivo è ottenere una traccia ispezionabile di corrispondenze plausibili. Per ogni paragrafo Ue la pipeline produce una lista ordinata dei paragrafi italiani più vicini (top-k) e individua il best-match. In questo modo chi revisiona può partire dai candidati con similarità più alta e verificare rapidamente se la corrispondenza è davvero pertinente. In parallelo al punteggio di similarità coseno, la pipeline calcola e salva anche un indicatore di sovrapposizione lessicale ("overlap") per le coppie candidate, riportandolo nei risultati.

Oltre al best-match, vengono salvate informazioni utili a leggere e contestualizzare il risultato, come l'identificativo del segmento e il riferimento all'articolo. Così un punteggio alto non resta un numero astratto, ma rimanda subito al punto del testo che lo ha generato.

4.4.3. Aggregazione a livello di articolo e sintesi a livello di legge

Anche se il calcolo avviene a livello di paragrafo, i risultati vengono poi riorganizzati per articolo per ottenere viste più leggibili. L'aggregazione permette di capire se un articolo della direttiva trova corrispondenze concentrate in uno specifico articolo italiano oppure se risulta "distribuito" su più articoli.

A livello di legge, la sintesi viene ulteriormente riassunta considerando l'insieme degli articoli. In questo modo si ottiene un indicatore globale del grado di allineamento tra i due testi, utile come vista d'insieme prima di scendere nei dettagli.

4.4.4. Fasce di similarità e copertura del testo

Per rendere confrontabili le coppie anche in forma sintetica, la pipeline assegna i best-match dei paragrafi a tre fasce di intensità. Queste fasce non vanno lette come accuratezza, ma come "fasce di similarità" che descrivono quanta parte del testo

23. Reimers–Gurevych 2019.

Ue trova nel recepimento un candidato con vicinanza alta o molto alta.

Il riepilogo percentuale per coppia, calcolato sul numero di paragrafi Ue, rende visibili differenze tra domini e tra atti. Alcune coppie presentano una quota maggiore di best-match nelle fasce alte, mentre altre restano concentrate nella fascia debole. Questo può indicare una riformulazione più marcata, una frammentazione su più disposizioni o un disallineamento strutturale tra i testi.

4.5. Output ispezionabili, visualizzazioni e tracciabilità del confronto

Dopo il calcolo delle similarità, la pipeline produce output pensati per l'ispezione e la verifica. L'idea è mantenere un collegamento diretto tra punteggi e testo giuridico, così che le corrispondenze proposte possano essere controllate e discusse.

Le visualizzazioni aiutano a leggere il confronto su testi lunghi. La heatmap mostra la matrice di similarità e mette in evidenza pattern di concentrazione o dispersione. I grafi mostrano i collegamenti sopra soglia e permettono di vedere rapidamente dove emergono relazioni forti. La selezione delle soglie è gestita in modo esplicito e ripetibile, con diverse modalità (percentili, clustering, statistiche della matrice), così da poter motivare cosa viene evidenziato e mantenere accesso al dato completo quando serve.

Per ogni coppia direttiva–recepimento, gli output vengono salvati in una struttura di cartelle coerente. Il report HTML raccoglie metriche e visualizzazioni e guida l'ispezione. In parallelo, tabelle CSV rendono la matrice consultabile come elenco di candidati, includendo best-match e top-k per i paragrafi Ue. Ogni match resta collegato a riferimenti che consentono di risalire al paragrafo e all'articolo, rendendo la verifica più rapida e meno ambigua.

Su più coppie, la pipeline produce riepiloghi aggregati e supporta una valutazione su campione annotato. Inoltre, salva parametri e artefatti intermedi (testi normalizzati, segmentazione, embedding) così da poter rieseguire l'analisi e confrontare in modo trasparente l'effetto di cambiamenti su modello, filtri o soglie.

5. Caso di studio e risultati

Questa sezione mostra, su alcune coppie del corpus, come si distribuiscono i punteggi e quali pattern emergono nelle visualizzazioni.

Per rendere il caso di studio comprensibile abbiamo separato le coppie in due gruppi. Il gruppo dei positivi attesi comprende le coppie in cui il decreto italiano è effettivamente l'atto di attuazione della direttiva considerata, dove ci aspettiamo di trovare corrispondenze sistematiche tra parti dei due testi. I negativi attesi sono invece coppie costruite apposta tra atti che non sono in relazione di recepimento, utili come riferimento per capire quanto spesso compaiano similarità spurie anche quando la relazione normativa non esiste.

Operativamente, i positivi attesi coincidono con gli abbinamenti "corretti" direttiva–decreto definiti nella sezione 3. I negativi attesi si ottengono invece confrontando ogni direttiva con tutti gli altri decreti del dataset che non sono il suo recepimento, cioè usando gli abbinamenti "incrociati" generati dal confronto esaustivo tra direttive e decreti, escluso il pairing corretto.

All'interno dei positivi attesi, per dare un lessico comune ai risultati, usiamo due soglie ricavate in modo semplice dai punteggi dei positivi. I punteggi ricavati sono valori di similarità coseno calcolati tra gli embedding dei paragrafi. In particolare, per ciascun paragrafo della direttiva la pipeline costruisce la matrice paragrafo–paragrafo e registra il best match, cioè il massimo valore di similarità coseno rispetto a tutti i paragrafi del decreto. È su questa distribuzione di best match dei positivi attesi che calcoliamo la *mediana* pari a circa 0.84 e il *75 percentile* pari a circa 0.92.

Le soglie vengono calcolate esclusivamente sui positivi attesi e poi applicate anche ai negativi attesi, così da usare una scala unica quando si confrontano i due gruppi.

Nella seguente sezione useremo sempre le stesse soglie numeriche: *Debole* sotto 0.84, *Moderato* tra 0.84 e 0.92, *Forte* da 0.92 in su. Le fasce sono un supporto alla lettura, e Moderato indica la fascia di revisione in cui la similarità segnala vicinanza ma l'allineamento può dipendere dal contesto. Le soglie permettono di stabilire le priorità di controllo e accompagnano il passaggio dalle figure al testo nei report ispezionabili.

Nella Figura 1 riportiamo una vista d'insieme che mette a confronto la distribuzione dei punteggi nei positivi attesi e nei negativi attesi. La figura va letta come confronto tra due distribuzioni: sull'asse orizzontale ci sono i valori di similarità coseno e sull'asse verticale la densità, così da vedere se e

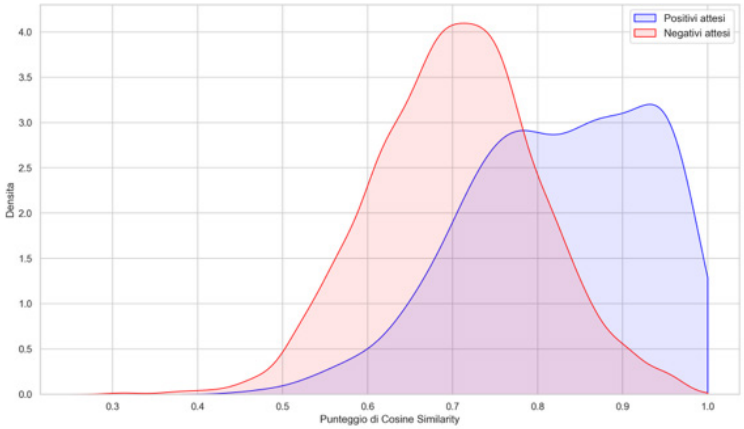

Fig. 1 — *Distribuzione dei punteggi di similarità coseno nei positivi attesi e nei negativi attesi*

quanto i positivi si spostano verso punteggi più alti rispetto ai negativi. La figura serve a inquadrare il comportamento generale del sistema e a ricordare che la similarità, da sola, va sempre letta insieme al testo. In generale, le coppie negative possono raggiungere in pochi casi valori molto alti, ma tendono a concentrarsi su punteggi mediamente più bassi. Le coppie positive, invece, mostrano un picco più marcato verso valori elevati, intorno a 0.95, e risultano complessivamente più spostate verso destra. La distribuzione riassume l'insieme dei cento confronti ottenuti dal confronto incrociato tra tutte le direttive e tutti i decreti del dataset, distinguendo tra gli abbinamenti corretti, che costituiscono i positivi attesi, e gli abbinamenti incrociati, usati come negativi attesi.

5.1. Confronto tra paragrafi, la matrice di similarità come mappa

Per ogni coppia direttiva–recepimento la pipeline calcola la similarità coseno tra tutti i paragrafi UE e tutti i paragrafi IT, producendo una matrice che viene resa leggibile tramite heatmap.

La matrice viene usata come una mappa per mostrare dove si concentrano le corrispondenze e dove emergono dispersioni dovute a riformulazioni, spostamenti o frammentazione del recepimento.

Nella figura 2, relativa alla Direttiva 2009/48/CE sulla sicurezza dei giocattoli e al relativo recepimento italiano D.Lgs. 11 aprile 2011, n. 54, la heatmap mostra una banda diagonale visibile per buona parte della matrice, indice di un andamento complessivamente parallelo tra l'ordine dei contenuti Ue e quello dei contenuti nel decreto. Allo stesso tempo la banda non è uniforme e presenta interruzioni e zone più "granulari", con addensamenti in aree laterali. Questa combinazione è tipica dei casi in cui molti paragrafi mantengono un allineamento sequenziale, ma alcuni passaggi vengono spezzati in più unità, accorpati o ricollocati, generando corrispondenze multiple e quindi isole di similarità fuori diagonale.

Questo tipo di lettura è utile perché consente una prima diagnosi strutturale prima ancora di aprire i testi. Se la banda resta compatta e vicina alla diagonale, il recepimento tende a seguire una trasposizione più lineare. Quando invece la banda si interrompe e compaiono isole più scure distanti dalla diagonale, diventa probabile che il contenuto UE sia stato redistribuito su più articoli o paragrafi del decreto, oppure che alcune parti siano state riformulate e agganciate ad altre sezioni normative.

La Figura 2 mostra uno screenshot della versione HTML interattiva della heatmap. Passando il puntatore sulle celle è possibile visualizzare l'incrocio esatto tra i due paragrafi e il relativo valore di similarità coseno. Questo supporto è centrale nel caso di studio, perché permette di verificare rapidamente, senza dover leggere per intero i testi normativi, se le zone più scure corrispondono a una relazione plausibile oppure a una somiglianza solo superficiale.

5.2. Sintesi a livello di articolo, heatmap e grafo dei best match

Il livello di paragrafo è quello computazionale, ma spesso la lettura giuridica parte dagli articoli. Per questo, la pipeline costruisce viste articolo–articolo che riassumono i segnali osservati nei paragrafi sottostanti.

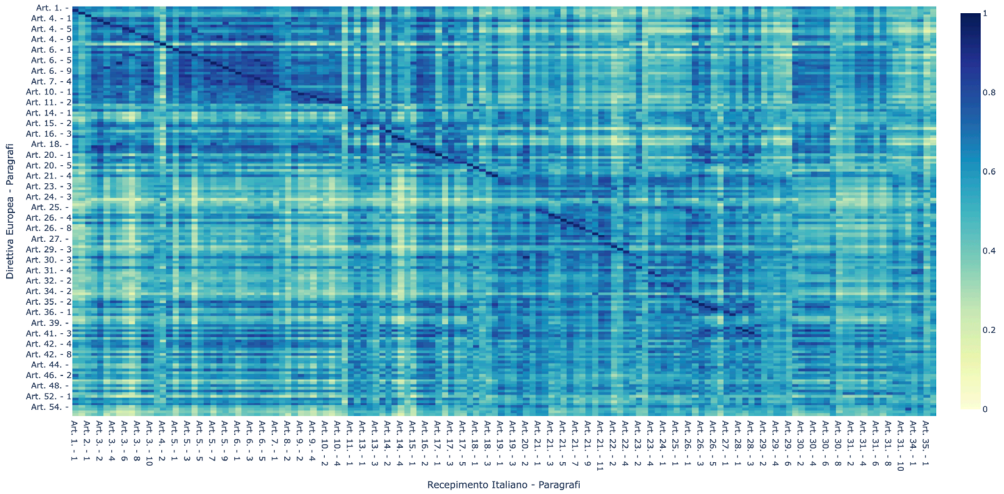

Fig. 2 — *Matrice paragrafo–paragrafo in versione HTML interattiva per Direttiva 2009/48/CE e D.Lgs. 11 aprile 2011, n. 54*

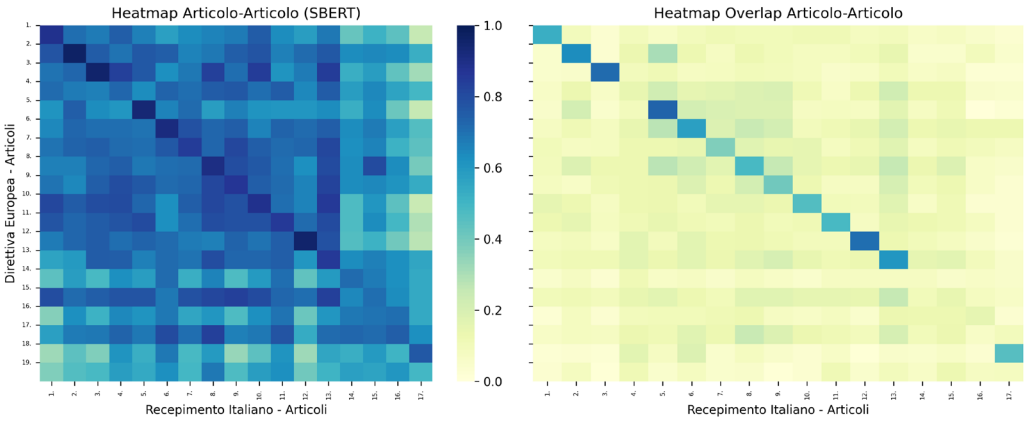

Fig. 3 — *Heatmap articolo–articolo per la coppia Direttiva 2019/904/UE e D.Lgs. 8 novembre 2021, n. 196. A sinistra la similarità semantica basata su* SBERT*, a destra l'overlap lessicale come indicatore di sovrapposizione terminologica*

La vista articolo–articolo serve a capire se la corrispondenza è concentrata o dispersa. Se un articolo UE trova i punteggi più alti quasi sempre nello stesso articolo italiano, il recepimento appare più "agganciato" a un punto preciso, se invece i valori alti sono distribuiti, è plausibile una trasposizione frammentata su più articoli e quindi meno immediata da ricostruire.

La prima vista è la Figura 3, una heatmap articolo–articolo. Ogni cella rappresenta la relazione tra un articolo UE e un articolo IT. In questo modo si osserva se un articolo europeo tende ad appoggiarsi su un singolo articolo del recepimento oppure se il contenuto viene distribuito su più articoli. Nel caso studio sulla plastica monouso, relativo alla Direttiva 2019/904/UE e al D.Lgs. 8 novembre 2021, n. 196, la heatmap semantica basata su SBERT è affiancata da una heatmap di overlap lessicale. L'overlap è un indicatore di sovrapposizione terminologica tra i due articoli e aiuta a interpretare il valore di similarità in modo più concreto.

Le due mappe vanno lette insieme: un valore alto in SBERT con overlap basso può indicare riformulazione o parafrasi, mentre valori alti in entrambi sono più spesso compatibili con riprese testuali dirette o con un lessico molto condiviso.

Per rendere più intuitiva la differenza tra similarità e overlap, riportiamo un esempio breve in cui direttiva e decreto riprendono quasi la stessa formulazione, e quindi entrambi i valori risultano alti.

– UE, Direttiva 2019/904/UE, art. 2, par. 1. *La presente direttiva si applica ai prodotti di plastica monouso elencati nell'allegato, ai prodotti di*

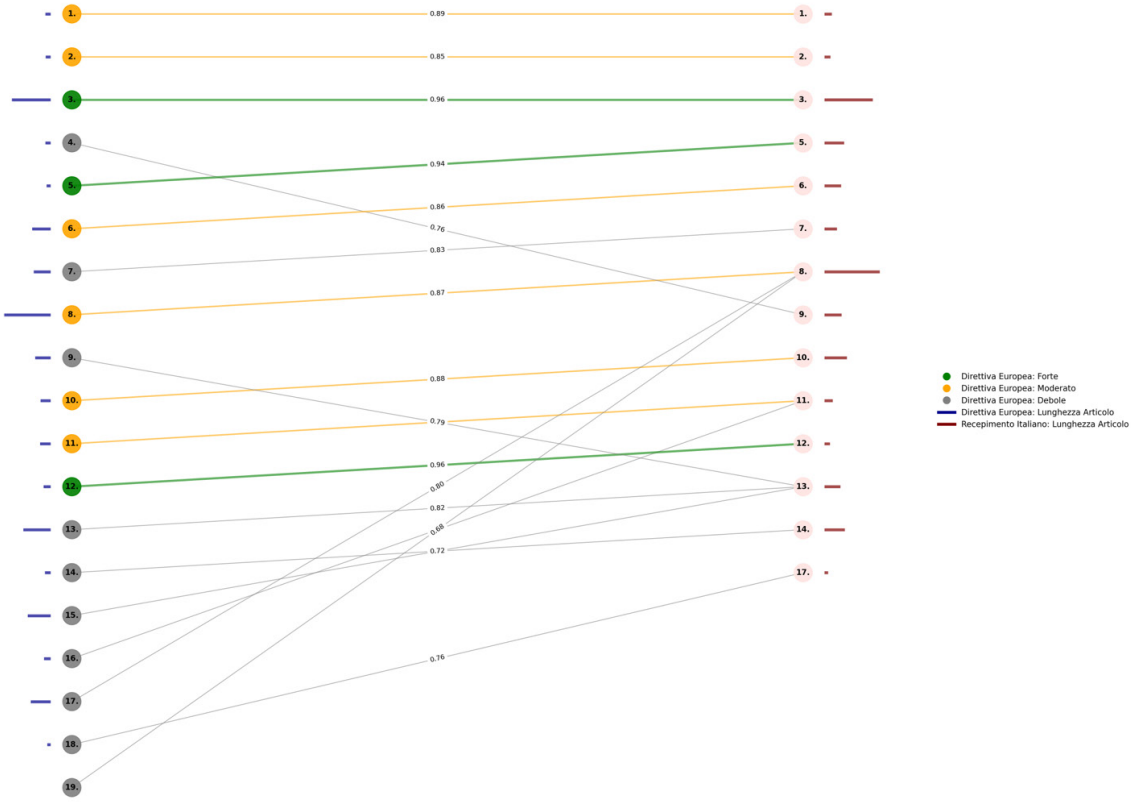

Fig. 4 — *Grafo dei best match articolo–articolo per la coppia Direttiva 2019/904/UE e D.Lgs. 8 novembre 2021, n. 196. Ogni arco collega un articolo UE al miglior candidato nel recepimento e la visualizzazione include le lunghezze in token per contestualizzare i punteggi*

plastica oxo degradabile e agli attrezzi da pesca contenenti plastica.

– IT, D.Lgs. 8 novembre 2021, n. 196, art. 2, par. 1. *Il presente decreto si applica ai prodotti in plastica monouso, di cui all'Allegato, ai prodotti in plastica oxo degradabile, nonché agli attrezzi da pesca contenenti plastica*.

In questo caso la similarità coseno è 0.97 e l'overlap lessicale è 0.63.

La seconda vista è la Figura 4, un grafo dei best match articolo–articolo che, per mantenere leggibile la figura, mostra solo i collegamenti selezionati sopra soglia. Il grafo collega ogni articolo UE al suo miglior candidato nel recepimento e integra anche informazioni diagnostiche, tra cui le lunghezze in token rappresentate come barre accanto ai nodi. Il grafo si legge seguendo gli archi in uscita dagli articoli UE: ogni arco indica il miglior candidato italiano, e la selezione sopra soglia serve a far emergere i collegamenti più informativi. Questo aiuta a contestualizzare i punteggi, perché un collegamento alto su un testo molto breve non va interpretato automaticamente come evidenza forte senza un controllo sul contenuto.

Le due visualizzazioni sono complementari. La Figura 3 offre una vista d'insieme su copertura e dispersione tra articoli, mentre la Figura 4 mette in evidenza la struttura del mapping e rende immediato vedere dove il best match è univoco e dove invece più articoli UE convergono sugli stessi articoli italiani. In pratica, la heatmap aiuta a vedere "dove si concentra" l'allineamento, il grafo aiuta a vedere "come si distribuisce" il best match.

5.3. Valutazione quantitativa su confronti incrociati

Le Figure 5-7 riportano, per ciascuna coppia direttiva–decreto, la percentuale di articoli UE il cui best match articolo–articolo ricade nelle tre fasce Debole, Moderato e Forte. Ogni heatmap visualizza una sola fascia, quindi le tre mappe vanno lette insieme come tre viste della stessa distribuzione percentuale. Per ogni coppia direttiva–decreto, le percentuali nelle tre fasce sommano a 100 perché ogni articolo UE viene conteggiato una sola volta, in base alla fascia

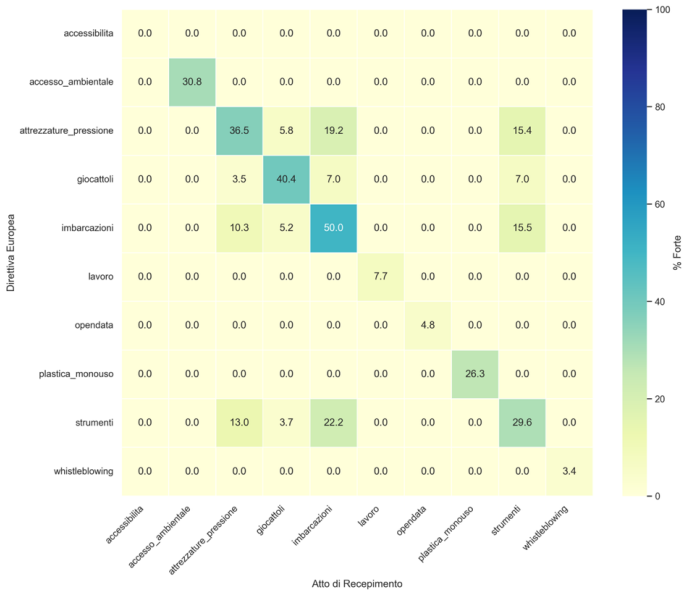

Fig. 5 — *Heatmap della percentuale di articoli UE in fascia Forte per ciascuna coppia direttiva–decreto*

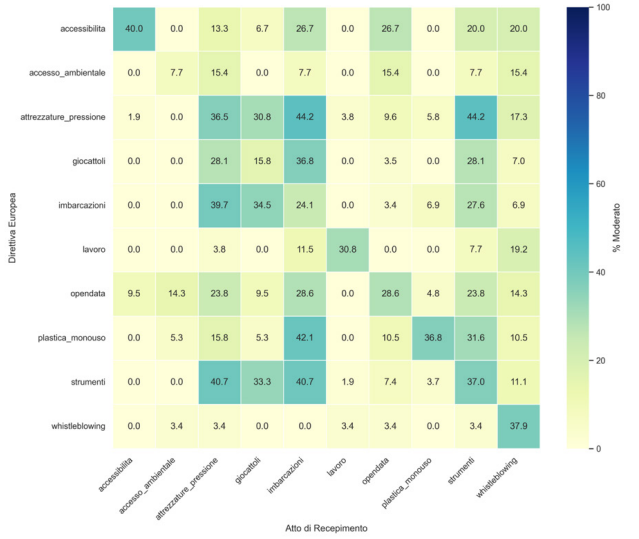

Fig. 6 — *Heatmap della percentuale di articoli UE in fascia Moderato per ciascuna coppia direttiva–decreto*

in cui cade il suo best match nel decreto. Il colore delle heatmap indica quanta parte degli articoli UE, in quella coppia, cade nella fascia considerata.

L'obiettivo di questa valutazione è osservare se la ripartizione tra le tre fasce cambia in modo sistematico quando la coppia è corretta rispetto ai confronti incrociati. La sintesi è netta. Nei positivi attesi la quota media di collegamenti Forte è 22.95%, quella Moderato è 29.53% e Debole è 47.52%. Nei negativi attesi, invece, Forte scende in media a 1.42%, Moderato a 11.15% e Debole sale a 87.43%. Osservata sulle heatmap, questa differenza si riflette in un pattern coerente.

La Figura 5 evidenzia che i valori di Forte tendono a concentrarsi nei pairing corretti e restano per lo più nulli o marginali nei pairing incrociati. In lettura, celle più scure significano che una quota più alta di articoli UE trova un best match sopra 0.92 nel decreto considerato.

La Figura 6 mostra che la fascia Moderato è più frequente nei pairing corretti e intercetta casi plausibili ma meno allineati sul piano testuale,

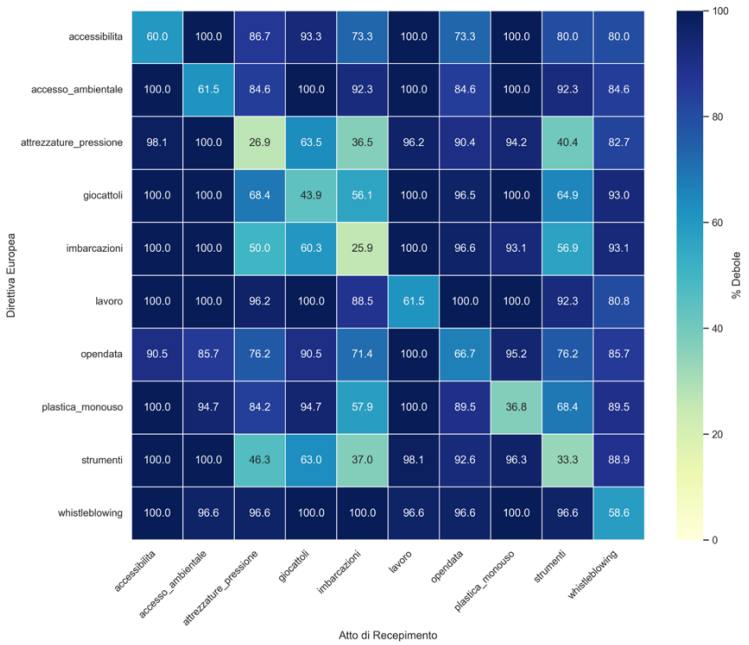

Fig. 7 — *Heatmap della percentuale di articoli UE in fascia Debole per ciascuna coppia direttiva–decreto*

per esempio per riorganizzazione del contenuto, accorpamenti o rinvii. Qui il colore indica quanta parte degli articoli UE ricade tra 0.84 e 0.92, cioè nella zona in cui il segnale è utile per orientare ma richiede controllo sul testo.

La Figura 7 conferma che nei pairing incrociati prevale la fascia Debole, che costituisce il comportamento atteso quando non esiste una relazione di recepimento. Celle più scure in Debole indicano che, per quella coppia, la maggior parte degli articoli UE non trova candidati con similarità sopra 0.84 nel decreto.

Nel complesso, le tre heatmap forniscono una lettura quantitativa compatta che aiuta a distinguere pairing corretti e incrociati e, soprattutto, a isolare la fascia Moderato come area in cui l'allineamento è plausibile ma richiede una verifica sul testo.

6. Valutazione

Per verificare quanto le etichette assegnate automaticamente dalla pipeline siano coerenti con una lettura umana, abbiamo annotato manualmente un campione di 60 coppie di paragrafi, associando a ciascuna coppia una classe tra Debole, Moderato e Forte (valutazione umana) e confrontandola con la classe prodotta dal sistema. Il campione è stato selezionato in modo stratificato a partire dalle fasce operative generate dalla pipeline, con 20 coppie per ciascuna fascia. Durante la lettura umana alcune coppie, soprattutto in fascia Moderato, sono state riclassificate perché la valutazione dipendeva dal contesto, ad esempio rinvii, riorganizzazioni e formulazioni standard. Nell'annotazione manuale, Forte è riservato ai casi in cui il contenuto normativo appare sostanzialmente allineato, Moderato ai casi ambigui che richiedono controllo contestuale, Debole ai casi in cui la somiglianza è prevalentemente superficiale o generica.

La Figura 8 sintetizza come cambiano le etichette quando si passa dalla predizione alla valutazione umana. Le barre sull'asse orizzontale rappresentano le tre classi predette dal sistema e, per ciascuna, mostrano la ripartizione delle classi assegnate dall'annotatore. Nel campione considerato, la classe Predetto Forte rimane stabile, mentre la maggiore variabilità si concentra su Predetto Moderato, che raccoglie molti casi borderline e viene talvolta rivalutato verso Debole o verso Forte. Questo comportamento è atteso perché la fascia intermedia è quella in cui è più facile che un punteggio alto rifletta somiglianze parziali o formulazioni standard, senza una corrispondenza normativa davvero "pulita".

La Tabella 2 riporta precision, recall e F1-score, metriche standard nella valutazione di compiti

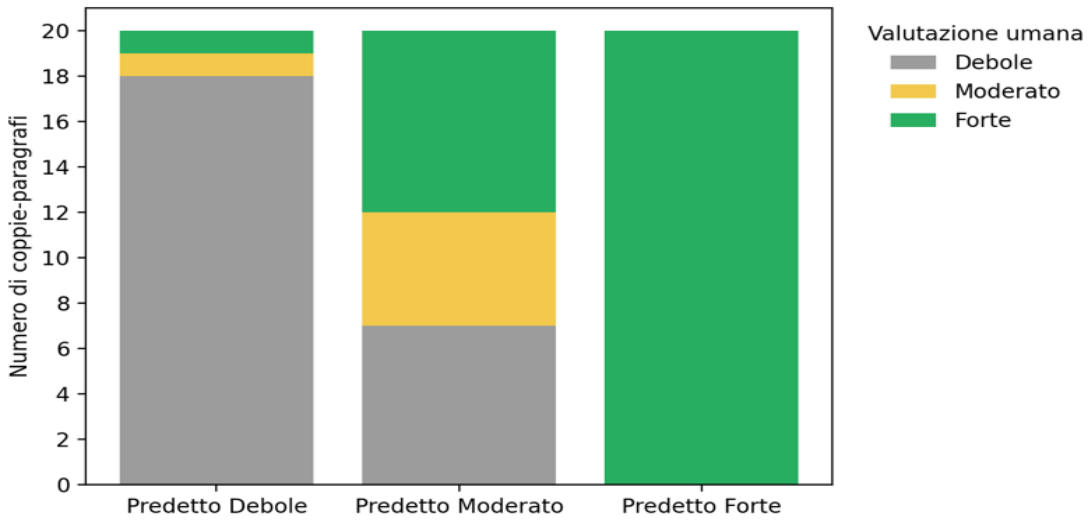

Fig. 8 — *Riassegnazione delle classi nella valutazione umana per ciascuna classe predetta (campione di 60 coppie paragrafo–paragrafo)*

di classificazione e retrieval[24]. Qui, per ciascuna classe, la precisione misura la quota di predizioni del sistema su quella classe che risultano corrette secondo la valutazione umana, mentre il recall misura la quota di casi di quella classe (secondo la valutazione umana) che il sistema riesce a individuare; l'F1-score è una misura di sintesi che combina precisione e recall (media armonica), e quindi scende quando una delle due è bassa. L'accuracy complessiva è pari a 0.72 su 60 coppie. La classe Forte mostra una precisione alta, mentre la classe Moderato risulta la più delicata, con recall elevato ma precisione bassa, segnale che molti casi predetti come Moderato vengono giudicati umanamente in modo diverso.

Questo comportamento non è solo un limite, è anche un'indicazione utile sul ruolo della fascia intermedia. Moderato è pensato come un contenitore di casi borderline e quindi è atteso che, a parità di soglia, una parte venga rivalutata verso Debole o verso Forte quando si legge il testo. In altre parole, la pipeline "sbaglia" soprattutto dove, anche per un umano, l'interpretazione dipende dal contesto, ed è proprio lì che serve più supporto ispezionabile.

Dato che le classi sono ordinate, affianchiamo alle metriche standard anche due misure ordinali. La kappa di Cohen pesata quadraticamente pesa gli errori in base alla loro distanza, penalizzando di più uno scarto tra estremi rispetto a uno scarto tra classi adiacenti. Nel nostro caso la kappa quadratica è 0.79, indicando un accordo sostanziale oltre quanto ci si aspetterebbe per caso, con un peso coerente con la scala Debole, Moderato e Forte. In altre parole, la kappa quadratica misura la concordanza tra classe predetta e classe umana correggendo per l'accordo casuale e penalizzando maggiormente gli scarti "lontani" (Debole vs Forte) rispetto a quelli "adiacenti" (Debole vs Moderato). Il MAE ordinale misura l'errore medio in "passi" lungo la scala, per esempio uno scambio tra Forte e Moderato vale 1. Il valore osservato, 0.30, suggerisce che gli errori, quando presenti, restano quasi sempre entro un solo livello. Nel nostro caso, il MAE ordinale si interpreta come distanza media assoluta tra classe umana e classe predetta su una codifica Debole=0, Moderato=1, Forte=2, quindi 0.30 indica scarti medi ben inferiori a un livello.

Un esempio aiuta a capire perché, soprattutto nella fascia Moderato, sia utile affiancare al punteggio semantico un controllo testuale. Nella coppia negativa attesa ue_attrezzature_pressione vs it_strumenti il sistema produce una similarità coseno elevata (0.92) e un overlap lessicale non trascurabile (0.55), pur trattandosi di atti non in relazione di recepimento, e la valutazione umana etichetta la coppia come Debole. Questo accade perché i due paragrafi condividono una struttura molto standard, tipica dei testi tecnici (documentazione, procedura di valutazione della conformità, dichiarazione UE, marcatura), e differiscono

24. Manning–Raghavan–Schütze 2008.

Classe	Precision	Recall	F1-score	Support
Debole	0.90	0.72	0.80	25
Moderato	0.25	0.83	0.38	6
Forte	1	0.69	0.82	29
Accuracy			0.72	60
Macro avg	0.72	0.75	0.67	60
Weighted avg	0.88	0.72	0.77	60

Tab. 2 — *Metriche di classificazione sul campione annotato (60 coppie paragrafo–paragrafo)*

soprattutto nel dominio specifico, "attrezzature a pressione" contro "strumento di misura".

- UE, art. 6, par. 2. *I fabbricanti preparano la documentazione tecnica … ed eseguono o fanno eseguire … la procedura di valutazione della conformità … per le attrezzature a pressione … i fabbricanti redigono una dichiarazione di conformità UE e appongono la marcatura CE.*
- IT, art. 4-*bis*, par. 2. *I fabbricanti preparano la documentazione tecnica … ed eseguono o fanno eseguire … la procedura di valutazione della conformità … Qualora la conformità di uno strumento di misura … sia stata dimostrata … i fabbricanti redigono una dichiarazione di conformità UE e appongono la marcatura CE e la marcatura metrologica supplementare.*

Questi casi mostrano che la similarità cattura anche convergenze dovute a formulazioni ricorrenti dei testi tecnici. Per questo l'interpretazione resta legata alla verifica sul contenuto, che consente di distinguere una vicinanza linguistica generica da una corrispondenza normativa sostanziale.

7. Conclusioni

In questo lavoro abbiamo cercato di rendere più "leggibile" un passaggio che, nella pratica, resta spesso opaco. Quando una direttiva viene recepita, capire dove finiscono gli obblighi europei e in che forma riemergono nel diritto nazionale richiede tempo, competenze e molta lettura. La pipeline che presentiamo non nasce per dire se un recepimento sia giusto o sbagliato, ma per aiutare a ricostruire il percorso in modo controllabile, facendo vedere subito dove vale la pena guardare e permettendo di tornare sempre al testo.

La scelta che regge tutto il workflow è lavorare su unità normative riconoscibili, cioè articoli e paragrafi, e qui entra in gioco la potenzialità di Akoma Ntoso. AKN non è solo un formato "comodo" ma è ciò che permette di tenere insieme struttura, metadati e identificativi, e quindi di collegare un risultato numerico a un punto preciso del documento.

Per i testi italiani, la disponibilità di AKN tramite Normattiva facilita l'estrazione delle unità normative. Per le direttive Ue, invece, abbiamo scelto di normalizzare l'HTML di EUR-Lex in una struttura pseudo-AKN, in quanto sufficiente a rendere confrontabili articoli e paragrafi.

Nel caso di studio, la differenza tra pairing corretti e pairing incrociati emerge già nelle viste sintetiche a livello di articolo. Nelle coppie in cui il decreto è davvero l'atto di recepimento, una quota non trascurabile di articoli UE trova un best match nelle fasce Moderato e Forte. Nei confronti incrociati, invece, quasi tutti gli articoli UE ricadono in Debole e la fascia Forte diventa rara. Questa separazione è coerente con i valori medi riportati nella sezione 5.3, dove nei positivi attesi la quota di Forte è nettamente più alta rispetto ai negativi attesi.

La fascia Moderato resta quella che richiede più attenzione perché è il punto in cui i segnali possono essere ambigui. Da un lato intercetta casi reali di recepimento in cui il contenuto è stato riorganizzato, spezzato o riscritto; dall'altro può includere somiglianze "di stile", dovute a formule ricorrenti e linguaggio tecnico standardizzato, che

alzano la similarità anche quando la relazione normativa non è quella attesa. Per questo la pipeline è pensata per accompagnare la lettura: non basta sapere qual è il best match, serve poterlo verificare rapidamente nel testo e capire se il punteggio è sostenuto da contenuto effettivamente corrispondente oppure da espressioni generiche. Il corpus utilizzato rappresenta una prima base di lavoro. Le dieci coppie direttiva–recepimento permettono di testare la pipeline su ambiti diversi, ma non consentono ancora di generalizzare i risultati al recepimento del diritto europeo nell'ordinamento italiano. Una valutazione più solida richiederà un dataset più ampio e annotazioni indipendenti da parte di più esperti.

Gli sviluppi futuri che vediamo sono principalmente tre.

Il primo riguarda l'interazione: spostare il sistema da report statici a un ambiente in cui chi analizza possa filtrare i candidati, regolare soglie e passare dalle viste di insieme alle evidenze testuali senza perdere il collegamento con la struttura AKN.

Il secondo è rendere esplicito che la similarità semantica non è l'unico segnale utile. Oltre al punteggio SBERT, vogliamo integrare fattori che, nella revisione manuale, entrano naturalmente nella decisione. Per esempio, la lunghezza dei paragrafi e la loro asimmetria, l'overlap lessicale, la presenza di keyword e dei loro sinonimi o termini correlati. L'idea è arrivare a un profilo di evidenze per ciascuna coppia, così che un match "alto" possa essere letto insieme ai motivi che lo sostengono.

Il terzo passo, più sperimentale, è confrontare la pipeline con l'uso di LLM nei casi ambigui. Non per sostituire l'interpretazione giuridica, ma per verificare se un modello generativo possa aiutare a leggere insieme questi segnali e proporre una motivazione del match, basata su più fattori e ancorata alle evidenze disponibili nel testo. Questo passaggio richiederebbe comunque una valutazione autonoma in quanto le spiegazioni generate da un LLM dovrebbero restare controllabili e verificabili sul testo. In definitiva, l'idea è semplice. Se la trasparenza normativa passa anche dalla possibilità di ricostruire e controllare i collegamenti tra fonti, allora servono strumenti che non si limitino a produrre punteggi, ma che aiutino a seguirne le tracce. In questo senso, AKN permette di ancorare quelle tracce al documento, e la pipeline prova a trasformare un confronto difficile da scalare in un percorso più verificabile.

Riferimenti bibliografici

E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre (2012), *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*, in "SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Proceedings", Association for Computational Linguistics, 2012

F. Ariai, J. Mackenzie, G. Demartini (2025), *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges*, in "ACM Computing Surveys", 2025

Commissione europea (2023), *Better Regulation Toolbox 2023. Chapter 4 -Compliance, implementation and preparing proposals*, 2023

R. Ferrod, D.A. Bondarenko, D. Audrito, G. Siragusa (2023), *Pairing EU directives and their national implementing measures: A dataset for semantic search*, in "Computer Law & Security Review", vol. 51, 2023

T. Lima, K. Brito, A. Nascimento, G. Valença, F. Pedrosa (2022), *Using Natural Language Processing to Improve Transparency by Enhancing the Understanding of Legal Decisions*, in "Proceedings of EGOV-CeDEM-ePart 2022", CEUR Workshop Proceedings, vol. 3399, 2022

C.D. Manning, P. Raghavan, H. Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press, 2008

R. Nanda, G. Siragusa, L. Di Caro, G. Boella (2019), *Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives*, in "Artificial Intelligence and Law", vol. 27, 2019

M. Palmirani, R. Sperberg, G. Vergottini, F. Vitali (2018), *Akoma Ntoso Version 1.0 Part 1: XML Vocabulary*, OASIS Standard, 2018

M. Palmirani, F. Vitali (2011), *Akoma-Ntoso for Legal Documents*, in G. Sartor, M. Palmirani, E. Francesconi, M.A. Biasiotti (eds.), "Legislative XML for the Semantic Web", Springer, 2011

N. Reimers, I. Gurevych (2019), *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", Association for Computational Linguistics, 2019

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin (2017), *Attention Is All You Need*, in "Advances in Neural Information Processing Systems (NeurIPS)", 2017