Federico Andrea Galatolo, Gabriele Martino, Mario G.C.A. Cimino, and Chiara Ombretta Tommasi

SERICA Digital Library: ricerca di testi antichi attraverso Neural NLP

1 Introduzione

I sistemi di Information Retrieval (IR) sono diventati una componente essenziale della moderna gestione delle informazioni. Questi sistemi sono progettati per recuperare informazioni rilevanti da grandi collezioni di documenti, in risposta alle richieste degli utenti. In particolare, gli approcci DIR (Dense IR), che si basano sulla tecnologia *deep learning*, sono sempre più utilizzati in vari ambiti per trovare informazioni rilevanti a partire da dati ampi ed eterogenei in modo più rapido ed efficiente rispetto agli IR basati sulla tradizionale elaborazione del linguaggio naturale (NLP). I recenti progressi nel campo del NLP hanno portato allo sviluppo di modelli di trasformazione pre-addestrati, come BERT e LaBSE, che hanno mostrato prestazioni impressionanti. Questi modelli sono addestrati su *corpora* ampi, perché possano apprendere ricche rappresentazioni delle lingue, e sono adatti a una varietà di compiti in ambito NLP.

In questo senso, le biblioteche digitali storiche pongono sfide maggiori. I testi antichi, medievali o della prima età moderna sono scritti in lingue non più in uso, presentano un'ortografia e una grammatica in evoluzione nei vari periodi, e quindi sono difficili da elaborare con i sistemi NLP tradizionali. Per superare queste sfide, questo lavoro presenta un sistema DIR per una biblioteca latina che utilizza modelli di *machine learning* detti "transformer" pre-addestrati. Il metodo adoperato richiede di calcolare gli embeddings (rappresentazioni numeriche delle parole) utilizzando i modelli disponibili, come Latin BERT e LaBSE, per ogni frase nei documenti e per le interrogazioni. Per recuperare le frasi più simili per una determinata interrogazione, viene utilizzata una distanza come il coseno. Il presente lavoro illustra lo sviluppo del sistema proposto, che comprende una sequenza di pre-elaborazione dei dati, un motore di elaborazione delle interrogazioni, e un'interfaccia di ricerca per gli utenti. Per valutare le prestazioni del sistema, una metrica quantitativa basata sulla classifica dei documenti per interrogazioni è stata fornita da una persona esperta in latino e confrontata con i risultati ottenuti dal sistema proposto. I risultati sperimentali mostrano che il modello basato su LaBSE supera quello basato su Latin BERT per lo scopo di recuperare informazioni pertinenti dalla biblioteca latina. Per favorire ulteriori svi-

Open Access. © 2024 the author(s), published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783111308289-015

luppi, i dati e il codice sorgente sono stati resi pubblici [Galatolo and Martino (2023)].

Il presente articolo è strutturato come segue. La Sezione 2 tratta lo stato dell'arte. Il metodo è discusso nella Sezione 3. Gli studi sperimentali sono descritti nella Sezione 4. Infine, la Sezione 5 trae le conclusioni.

2 Stato dell'arte

Con l'avvento dell'apprendimento automatico, i modelli di IR si sono evoluti dai metodi classici alle funzioni di classificazione basate sull'apprendimento. Uno dei fattori critici per la progettazione di modelli IR efficaci è il modo in cui apprendere le rappresentazioni del testo e modellare una corrispondente rilevanza. Con i recenti progressi dei modelli di linguaggio pre-addestrati (LLM), come BERT e GPT, è possibile apprendere efficacemente le rappresentazioni dense delle interrogazioni e dei testi nello spazio cosiddetto "latente" e costruire una funzione di corrispondenza semantica per la modellazione della rilevanza. Questo approccio è noto come *dense retrieval*, in quanto impiega vettori o embeddings densi per rappresentare i testi [Zhao et al. (2022)]. BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al. (2019)] è un modello di rappresentazione linguistica all'avanguardia che ha ottenuto ottimi risultati su una serie di compiti di elaborazione del linguaggio naturale. Si tratta di un modello di deep learning che utilizza un'architettura a "transformer" per elaborare sequenze di testo e generare rappresentazioni di alta qualità. L'innovazione chiave di BERT è l'uso dell'elaborazione bidirezionale, che gli consente di catturare le informazioni contestuali sia in avanti che indietro su una determinata parola. Ciò si ottiene dividendo il testo in ingresso in segmenti di lunghezza fissa e poi elaborando ogni segmento in entrambe le direzioni, da sinistra a destra e da destra a sinistra. In questo modo, BERT è in grado di acquisire informazioni sul contesto in cui appare una parola, comprese le parole che la precedono e la seguono. Oltre all'elaborazione bidirezionale, BERT utilizza anche diverse altre tecniche per migliorare le sue prestazioni. Queste includono l'uso delle seguenti caratteristiche:

(i) l'auto-attenzione multi-testa, che consente al modello di concentrarsi selettivamente su diverse parti del testo in ingresso;
(ii) un obiettivo di modellazione linguistica mascherata, che incoraggia il modello a prevedere le parole mancanti in base al contesto in cui appaiono;
(iii) un compito di previsione della frase successiva, che incoraggia il modello a comprendere le relazioni tra le diverse frasi di un documento.

Recentemente, David Bamman e Patrick Burns [Bamman and Burns (2020)] hanno presentato Latin BERT, un modello linguistico contestuale per la lingua latina, addestrato su un ampio corpus di 642,7 milioni di parole provenienti da varie fonti che vanno dall'epoca classica al XXI secolo. Gli autori hanno dimostrato le capacità di questo modello specifico per la lingua latina attraverso diversi casi di studio, tra cui il suo utilizzo per l'etichettatura "part-of-speech", dove Latin BERT ha raggiunto un nuovo stato dell'arte per tre dataset *Universal Dependency Latin*. Il modello viene utilizzato anche per la previsione del testo mancante, comprese le emendazioni critiche, e supera i word embeddings statici per la disambiguazione del senso delle parole. Inoltre, lo studio dimostra che Latin BERT può essere utilizzato per la ricerca semanticamente informata, interrogando i vicini contestuali.

Anche LaBSE è un modello di embedding di frase multilingue basato sull'architettura BERT [Feng et al. (2022)]. Gli autori hanno studiato sistematicamente i metodi per l'apprendimento di embeddings di frasi multilingue, combinando i migliori metodi per l'apprendimento di rappresentazioni monolingue e multilingue, tra cui il Masked Language Modeling (MLM), il Translation Language Modeling (TLM), il Dual Encoder Translation Ranking e l'Additive Margin Softmax. Gli autori hanno dimostrato che l'introduzione di un modello linguistico multilingue pre-addestrato riduce drasticamente la quantità di dati di addestramento paralleli necessari per ottenere buone prestazioni. La combinazione dei migliori metodi ha prodotto un modello che raggiunge l'83,7% di accuratezza nel recupero di testi in più di 112 lingue sul set di dati Tatoeba, contro il 65,5% di accuratezza ottenuto dai precedenti modelli allo stato dell'arte, ottenendo allo stesso tempo prestazioni competitive in benchmark di transfer learning monolingue. Gli autori hanno inoltre dimostrato l'efficacia del modello LaBSE estraendo dati paralleli dal repository CommonCrawl e utilizzandoli per addestrare modelli di traduzione automatica neurale (Neural Machine Translation, NMT) competitivi per inglese-cinese e inglese-tedesco.

Un lavoro recente nella comprensione del linguaggio, che sfrutta le caratteristiche contestualizzate, è Semantics-aware BERT (SemBERT) [Zhang et al. (2020)]. SemBERT incorpora la semantica contestuale esplicita dall'etichettatura dei ruoli semantici pre-addestrata, migliorando le capacità di rappresentazione linguistica di BERT. SemBERT è in grado di assorbire la semantica contestuale senza sostanziali modifiche specifiche al compito, con un design più potente e semplice rispetto a BERT. Ha ottenuto nuovi risultati relativamente allo stato dell'arte in diversi compiti di comprensione automatica e di inferenza del linguaggio naturale. Per l'IR basata sul latino, Piroska Lendvai e Claudia Wick hanno perfezionato Latin BERT per la *Word Sense Disambiguation* sul *Thesaurus linguae Latinae* (*TLL*) [Lendvai and Wick (2022)]. Questo lavoro propone di utilizzare Latin BERT per creare un nuovo set di dati basato su un sottoinsieme di rappresentazioni del *Thesaurus linguae Latinae*. I risultati dello studio hanno mostrato che le rappre-

sentazioni BERT contestualizzate e messe a punto sui dati *ThLL* hanno prestazioni migliori rispetto agli embeddings statici utilizzati in un classificatore LSTM bidirezionale sullo stesso dataset. Inoltre, i modelli BERT per ogni singolo template hanno ottenuto prestazioni più elevate e più robuste rispetto ai risultati precedenti, basati su dati provenienti da un dizionario bilingue latino. Più recentemente, Zhengbao Jiang et al. hanno proposto X-FACTR: a Multilingual Factual Knowledge Retrieval from Pretrained Language Models [Jiang et al. (2020)]. Gli autori hanno proposto un benchmark di interrogazioni cloze-style per 23 lingue tipologicamente diverse per valutare il recupero della conoscenza fattuale nei LLM (Large Language Model). Lo studio ha ampliato i metodi di sondaggio da entità a parola singola a entità a più parole, e ha sviluppato diversi algoritmi di decodifica per generare previsioni a più token. I risultati dello studio hanno fornito indicazioni sulle prestazioni degli attuali LLM in lingue con più o meno risorse disponibili. I ricercatori hanno inoltre proposto un metodo basato sul code-switching per migliorare la capacità dei LLM multilingue di accedere alla conoscenza, che è stato verificato essere efficace in diverse lingue di riferimento. I dati e il codice di riferimento di questo articolo sono stati resi pubblici per facilitare ulteriori ricerche in questo settore.

3 Nuovi paradigmi di ricerca nella SERICA Digital Library

3.1 I livelli di rappresentazione del testo nella SERICA Digital Library

La SERICA Digital Library è stata progettata per ospitare diversi paradigmi di elaborazione del testo, corrispondenti a diverse metodologie di analisi testuale. La Figura 1 mostra i sei livelli di rappresentazione del testo. In sostanza, il testo viene archiviato in formati incrementali per poter essere elaborato in modo indipendente con vari metodi.

La struttura multilivello è composta come di seguito. Il primo livello è quello del *testo semplice*, ossia il contenuto puro senza alcun'altra informazione aggiuntiva. Il secondo livello, del *testo normalizzato*, si ottiene dopo aver convertito in una singola forma canonica ogni termine che si manifesta in diverse varianti storiche. Il *testo strutturato* corrisponde invece al terzo livello, in cui vengono aggiunte informazioni sulla struttura testuale, quali riferimenti, sezioni, pagine, ecc. Il quarto livello è il *testo annotato* semanticamente, in cui elementi testuali vengono associati a concetti, quali luoghi, persone, eventi ecc., di interesse dello studioso. Poi si passa al livello del

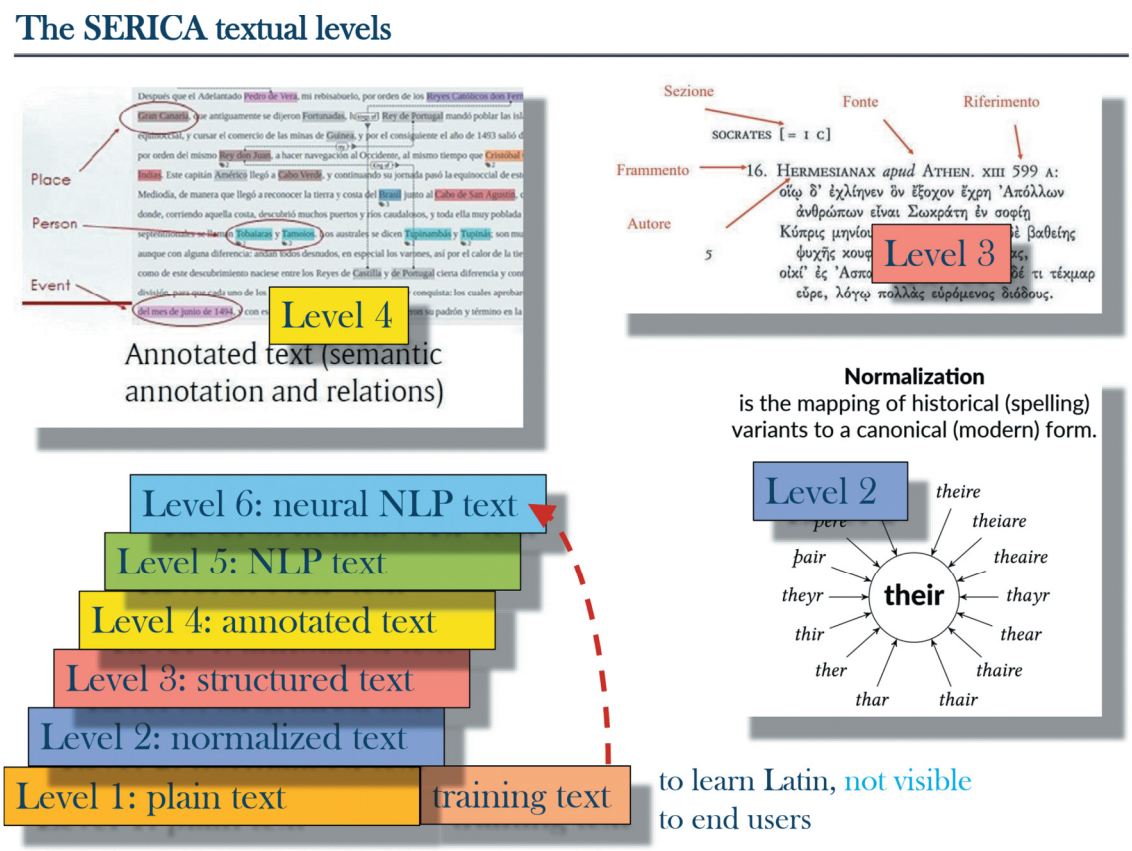

Figura 1: I livelli della SERICA Digital Library.

testo elaborato tramite NLP, in cui si possono produrre analisi elaborate del testo, ad esempio per studiare l'evoluzione del linguaggio, oppure per studiare il sentimento presente in un racconto, attraverso algoritmi finalizzati allo scopo. L'ultimo livello è quello del *testo elaborato tramite neural NLP*, ossia metodi di rappresentazione basati sulle reti neurali, oggetto del presente lavoro di ricerca. Si tratta di metodi che adoperano un corpus di testi di riferimento (*training text*) per modellare la struttura del linguaggio tramite un paradigma computazionale, e adoperano tale modello per esplorare una digital library, ad esempio ai fini della ricerca.

3.2 Il processo di *word embedding* e i modelli multilingua

Più precisamente, si dice *word embedding* il processo di conversione dei testi in numeri, con l'obiettivo di ottenere numeri simili per parole dal significato simile. La conversione numerica avviene secondo un principio puramente distribuzionale: secondo la *ipotesi distribuzionale*, parole che occorrono in contesti simili tendono ad avere significati simili. Pertanto, viene contrapposta la somiglianza basata sulla semantica con la somiglianza basata su come le parole sono distri-

buite. Il *machine learning* può generare rappresentazioni numeriche dove parole con significato simile hanno una rappresentazione numerica simile. Ad esempio, dato un gruppo di parole, si genera una rappresentazione numerica multidimensionale in cui i termini maschili sono più vicini tra loro, e analogamente quelli femminili, rispetto a termini di diverso genere. Oppure termini che rappresentano verbi sono vicini tra loro se hanno lo stesso tempo. Vi sono infine anche *modelli di embedding multilingue*, in cui parole che in diverse lingue corrispondono allo stesso concetto sono vicine tra loro. La Figura 2 mostra le diverse lingue disponibili nel modello LaBSE e il numero di frasi per lingua.

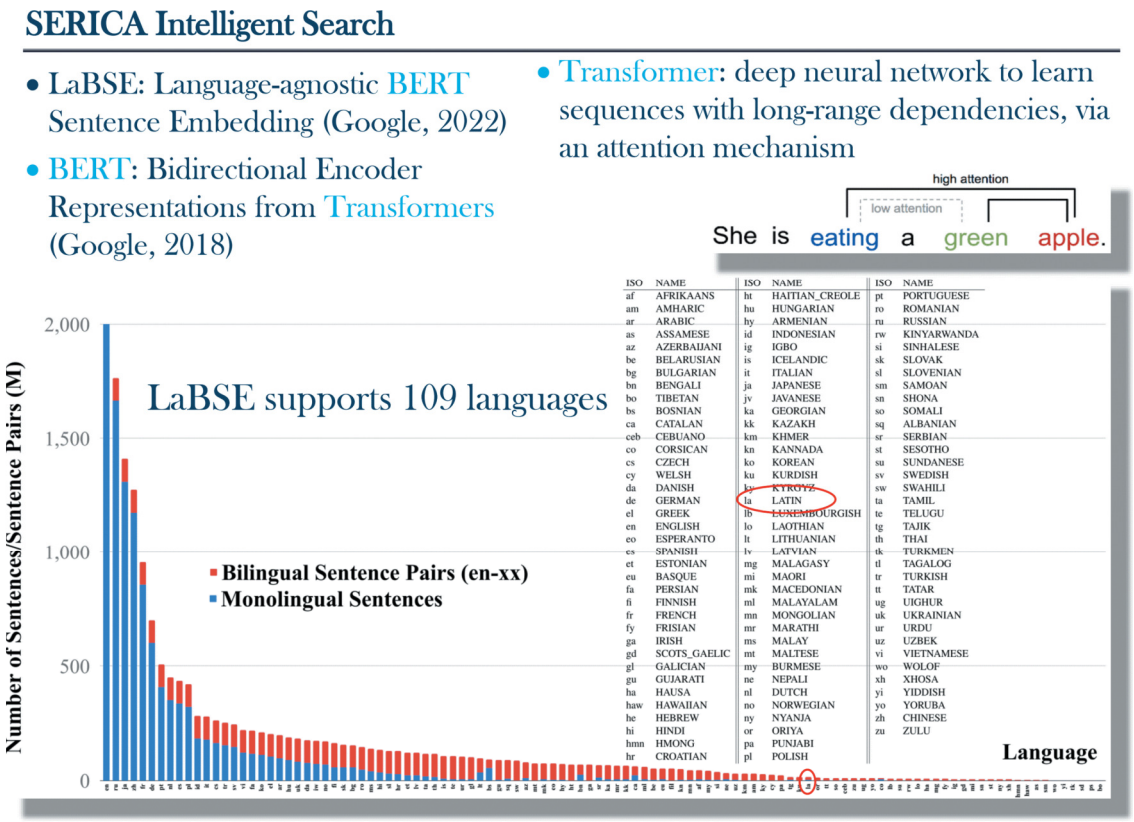

Figura 2: Fondamenti del modello LaBSE e numero di frasi adoperate per ogni lingua.

Sebbene risulti evidente la minore disponibilità di testi in lingue antiche, il fatto di creare un modello congiunto con le lingue moderne consente di avere un embedding più rappresentativo di ogni lingua, in quanto basato su un nucleo di relazioni tra parole che è "a fattore comune" in tutti i linguaggi con un'origine comune (proto-struttura linguistica).

3.3 La ricerca avanzata in SERICA: un esempio pilota

A scopo di esempio, la Figura 3 raffigura una dimostrazione della ricerca avanzata nella SERICA Digital Library (https://serica.unipi.it/ricerca-avanzata), effettuata su un sottoinsieme di 10 testi selezionati. In particolare, è stato selezionato il modello linguistico LaBSE, limitando il numero di risultati a 100, e come chiave di ricerca è stata inserita la parola *Christianitas*.

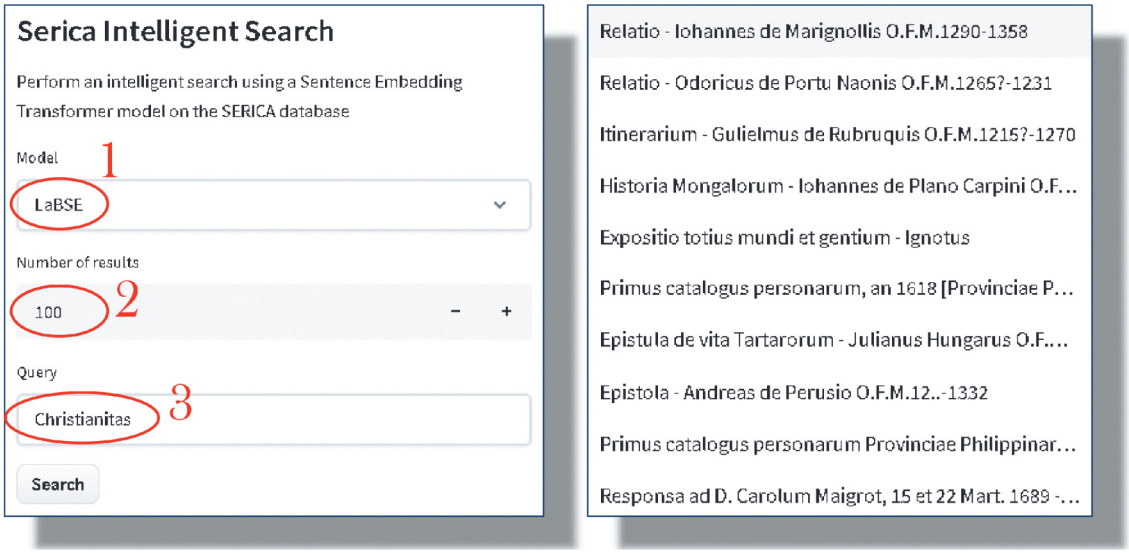

Figura 3: Dimostrazione di ricerca avanzata (a sinistra) su un sottoinsieme di testi (a destra).

La Figura 4 mostra i primi tre risultati della ricerca. Ogni risultato ha un titolo del testo, un punteggio (score), e un estratto del testo dove un termine inerente a quello cercato appare, con evidenza in grassetto il paragrafo che ha determinato il punteggio. È interessante notare che tra i primi risultati appaiono termini legati a *Christianitas* ma non esattamente coincidenti con la parola di ricerca: *Christianus*, *Christiani*, e anche il nome proprio *Christoforus* (letteralmente "colui che porta Cristo [*scil.* nel cuore]"). Il concetto importante nel paradigma di ricerca tramite neural NLP è che, da una parte, si ha molta flessibilità nell'esprimere le chiavi di ricerca, ossia non occorre preoccuparsi di inserire le varianti sintattiche del termine. Dall'altra però non si ha un controllo diretto della relazione semantica, in quanto essa discende dal processo di embedding. Di conseguenza, lo studioso si aspetta di trovare – tra i risultati – dei testi inerenti alla sua chiave di ricerca, ma anche alcuni testi non del tutto attinenti al suo studio. Tecnicamente si dice che lo studioso si aspetta una buona *precisione* e una buona *sensibilità* (detta anche *recupero* o *richiamo*). Più precisamente, per una data chiave di ricerca, consideriamo il numero totale di documenti correttamente inerenti: la precisione è il rapporto tra il numero di documenti trovati corret-

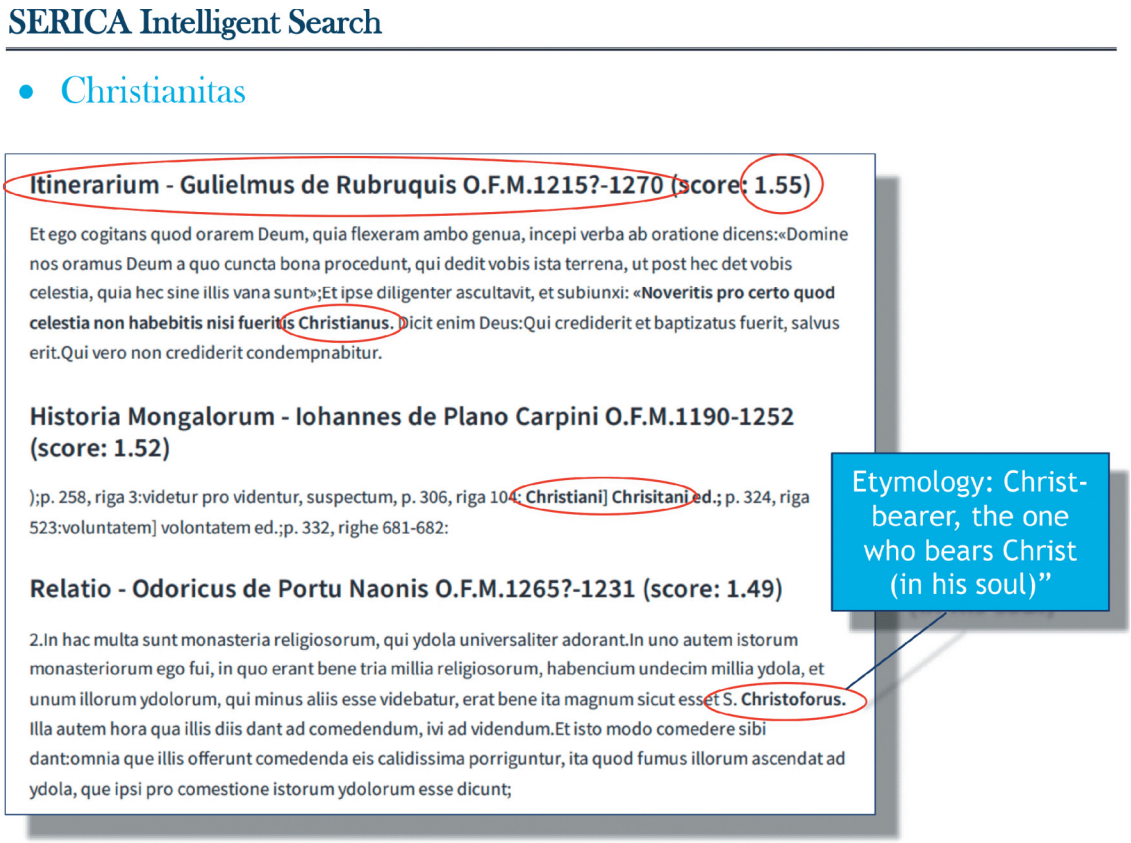

Figura 4: I primi tre risultati della ricerca di *Christianitas* tramite il modello LaBSE.

tamente e il numero totale di documenti trovati (correttamente o meno), mentre la sensibilità è misurata come il rapporto tra il numero di documenti corretti trovati sul numero totale di documenti corretti.

La Figura 5 mostra alcuni dei successivi risultati: è sorprendente trovare molte varianti della parola chiave, ma anche risultati non significativi. Per esempio, il risultato numero 5 associa al termine *Christianitas* il termine *Latin* perché si tratta di una parola latina: fatto scontato per lo studioso, quindi non rilevante. Un altro risultato, riportato come ultimo nella figura, evidenzia i termini *religionis* ed *Evangelium*, sicuramente collegati a *Christianitas*, ma sta allo studioso capire se il risultato risulta utile ai fini del suo processo di analisi.

Infine, la Figura 6 mostra altri risultati da collocare nel contesto della ricerca corrente.

Figura 5: I successivi risultati della ricerca di *Christianitas* tramite il modello LaBSE.

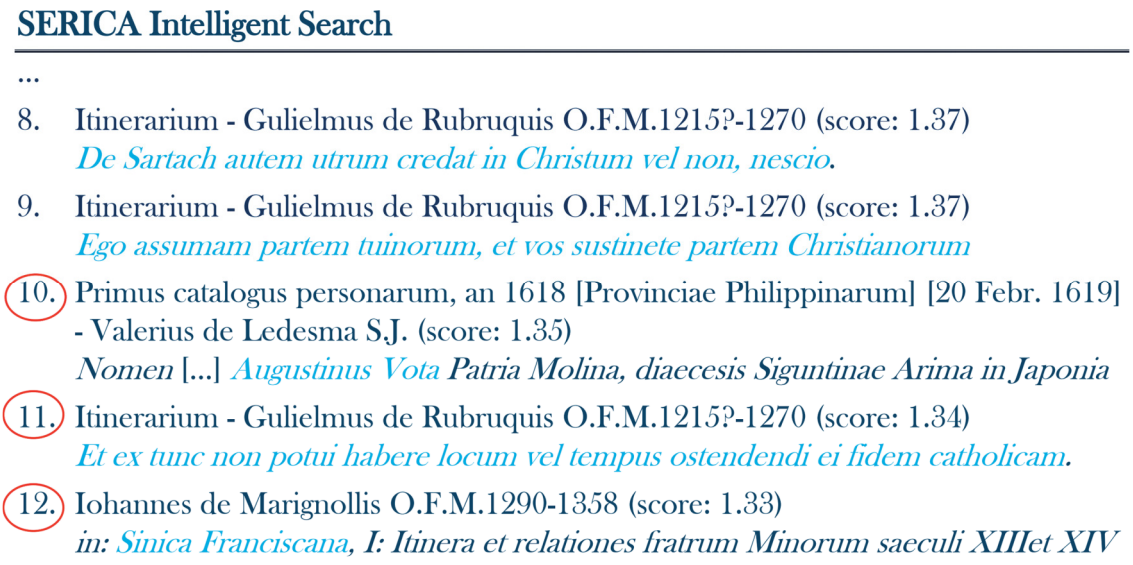

Figura 6: ulteriori risultati della ricerca di *Christianitas* tramite il modello LaBSE.

3.4 Approfondimento sui modelli linguistici LaBSE e Latin BERT

Nella SERICA Digital Library è stato sviluppato un sistema DIR multilingue per testi in latino, in cui l'utente può eseguire interrogazioni utilizzando diverse lingue. Per la valutazione, sono stati selezionati campioni di interrogazioni in latino e le loro controparti in inglese. Il primo passo del metodo è la tokenizzazione, cioè la suddivisione di tutti i documenti in unità di testo più piccole. La tokenizzazione dipende dalla lingua specifica. Ogni frase viene poi sottoposta all'estrazione degli embeddings LLMs. In particolare, si sperimentano e confrontano l'alta specificità di Latin BERT e l'alta generalità di LaBSE, sebbene i due modelli siano piuttosto diversi. LaBSE parte da un modello BERT pre-addestrato in combinazione con il Multi-Lingual Model (MLM) e il Translation Language Model (TLM), e lo riaddestra per combinare gli embeddings a livello di frase in lingue diverse. Al contrario, Latin BERT è addestrato con il Masked-Language Model (MLM), che si basa sulla previsione di una parola casuale selezionata da una frase. Entrambi gli embedding di output dei due modelli hanno una dimensionalità di 768. Inoltre, sono stati calcolati l'embedding medio dei due modelli e il vettore concatenato, per verificare come la combinazione dei due modelli possa influire sulle prestazioni di recupero.

Un altro aspetto interessante del modello LaBSE è il fatto di essere multilingua. La Figura 7 mostra la ricerca, tra i medesimi testi, del termine *brother* espresso in inglese. Si noti come tra i primi risultati appaia correttamente il termine latino *frater* in diverse declinazioni.

Nella SERICA Digital Library, ogni documento latino e ogni frase correlata, insieme ai quattro embedding rappresentativi estratti, sono memorizzati in un database open source Apache Lucene che consente la ricerca indicizzata. Quando si invia una interrogazione, viene calcolato il tipo di embedding selezionato e la ricerca nel database viene effettuata utilizzando la similarità coseno, una delle metriche di similarità più utilizzate dai motori di ricerca. È noto che la similarità coseno è più robusta al problema della dimensionalità. Infine, tutti i documenti vengono ordinati per similarità.

Nella rimanente parte dell'articolo, vengono eseguiti una serie di risultati sperimentali basati su un numero di ricerche statisticamente rilevanti. Vengono quindi valutate le prestazioni della ricerca, e tratte le opportune conclusioni. L'obiettivo non è di valutare in assoluto un modello linguistico, ma di comparare più modelli linguistici.

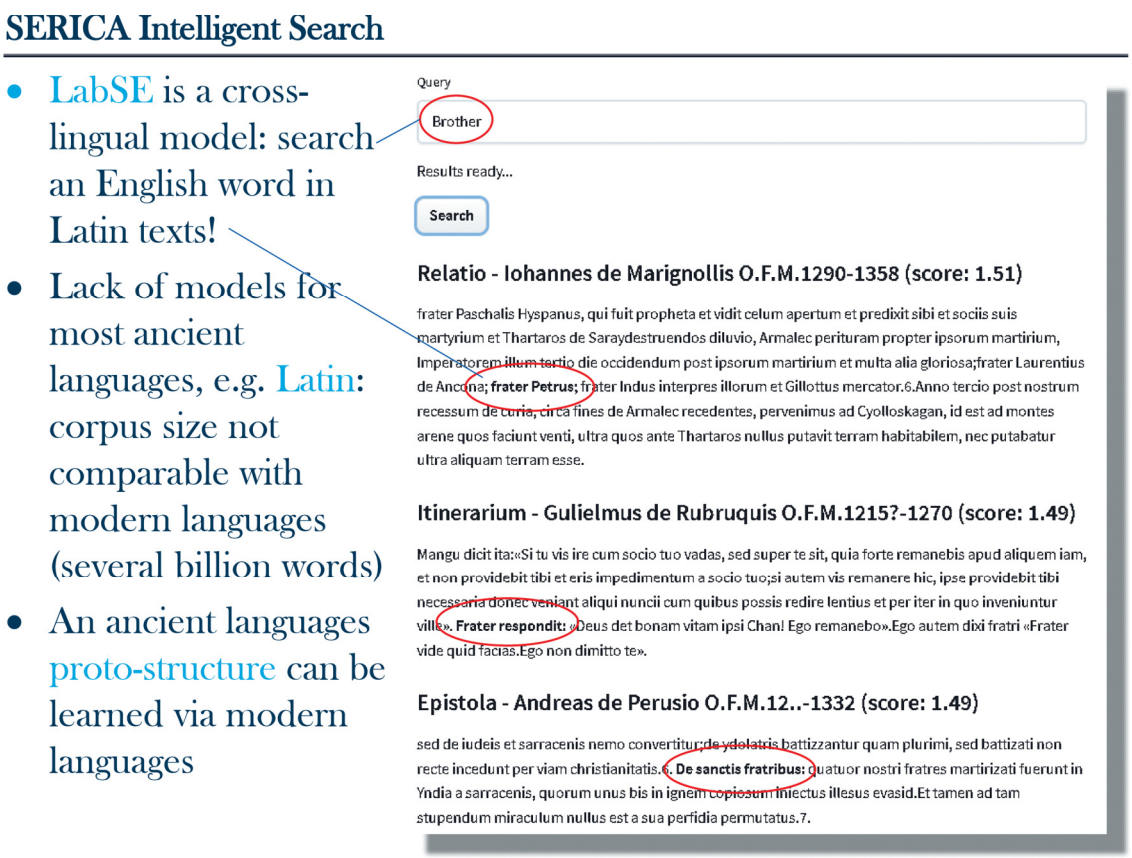

Figura 7: Primi risultati di ricerca del termine inglese *brother* in testi latini con il modello LaBSE.

3.5 Metodo di valutazione delle prestazioni

Per la valutazione del sistema DIR proposto, sono state selezionate Q frasi di interrogazioni, sia in latino che nelle rispettive traduzioni in inglese. Poi sono stati estratti i primi R risultati delle interrogazioni. Infine, ogni frase risultante è stata valutata attraverso una valutazione graduata tra 1 e 5. La valutazione è stata effettuata da un esperto di latino e inglese. La valutazione di ogni documento recuperato si basa sul rapporto di coerenza semantica tra l'interrogazione e la frase recuperata. Per valutare ogni interrogazione viene utilizzato il Discounted Cumulative Gain (Järvelin and Kekäläinen 2002):

$$DCG_q = \sum_{i=1}^{|D|} \frac{grade_i}{\log_2(i+1)}$$

dove D è il numero totale dei documenti risultanti per la interrogazione (query) q. Il grado è la valutazione nell'intervallo definito per quel risultato. Si utilizza il *DCG* normalizzato: il valore del *DCG* viene normalizzato rispetto al *DCG* ideale,

ricorrendo a tutti i risultati secondo la valutazione. Il valore risultante è quindi nell'intervallo [0, 1].

$$nDCG_q = \frac{DCG_q}{IDCG_q}$$

dove $IDCG_q$ (*DCG* ideale) corrisponde al *DCG* calcolato con tutti i documenti recuperati ordinati in ordine decrescente di grado. Infine, viene calcolato il *DCG* medio normalizzato:

$$AnDCG = \frac{1}{|Q|}\sum_{i}^{|Q|} nDCG_i$$

dove $|Q|$ è il numero totale di interrogazioni. Notiamo che l'*nDCG* confronta il ranking effettivo dei documenti della interrogazione rispetto a quello ideale, ma solo per i documenti rilevanti. Ciò comporta una complicazione nella definizione della soglia per considerare un documento rilevante o meno. Inoltre, l'*nDCG* non è adatto a considerare le prestazioni complessive di tutti i documenti recuperati, oltre al loro grado.

Per questo motivo, è stata sviluppata una nuova metrica di valutazione delle prestazioni, il DCG normalizzato penalizzato:

$$PnDCG_q = \frac{DCG_q}{IDCG_q} * \frac{MaxDist - Dist}{MaxDist}$$

dove *MaxDist* è la norma L1 tra la migliore classifica possibile e la peggiore classifica possibile per un insieme di *D* documenti recuperati:

$$MaxDist = |D| * |maxGrade - minGrade|$$

Dist è invece la norma L1 tra la migliore classifica possibile e la classifica ideale:

$$Dist = \sum_{i=1}^{|D|} maxGrade - idealGrade_i$$

Vale la pena notare che, se la classifica idealizzata è esattamente uguale al voto migliore (dove tutti i documenti recuperati hanno il voto massimo), $PnDCG_q$ è uguale a 1,0. Al contrario, se tutti i documenti recuperati hanno il voto minimo, il lato sinistro del prodotto ($nDCG_q$) è uguale a 1,0, mentre il lato destro diventa 0,0, portando l'intera metrica di valutazione è 0,0, riflettendo cioè una bassa performance. Infine, il *DCG* medio penalizzato normalizzato viene calcolato come segue:

$$APnDCG = \frac{1}{|Q|} \sum_{i}^{|Q|} PnDCG_{qi}$$

4 Studi sperimentali

Ai fini degli esperimenti, sono state selezionate Q = 10 frasi di interrogazione, sia in latino che nelle rispettive traduzioni in inglese. I primi R = 10 risultati delle interrogazioni sono stati estratti e classificati dall'esperto.

La Tabella 1 mostra i risultati sperimentali. Si noti innanzitutto che gli embeddings LaBSE superano tutti gli altri embeddings estratti, anche la combinazione LaBSE e LatinBERT. D'altra parte, sembra che gli embeddings LatinBERT non siano adatti a questo compito, confermando i risultati ottenuti da Hu [Hu et al. (2020)]. In particolare, Hu et al. hanno scoperto che le prestazioni di tali modelli in compiti di recupero sono molto scarse se non sono abbinate a una messa a punto a livello di frase. Inoltre, è importante notare che la ricerca della stessa frase in inglese ha risultati migliori rispetto a quella in latino. Questa distorsione potrebbe essere attribuita all'addestramento del modello LaBSE [Feng et al. (2022)]: la maggior parte delle frasi utilizzate per l'addestramento bilingue sono in inglese, il che porta il modello ad avere una maggiore capacità di astrazione in inglese.

Tabella 1: Valutazione delle prestazioni del sistema DIR proposto.

	LaBSE	**LatinBERT**	**LaBSE-LatinBERT Mean**	**LaBSE-LatinBERT Concat**
Latin	**0.33±0.06**	0.05±0.02	0.32±0.06	0.32±0.06
English	**0.52±0.05**	NA	0.43±0.06	0.43±0.06

Per spiegare ulteriormente le prestazioni, è stata calcolata la PCA (Principal Component Analysis) degli embeddings, per ottenere una riduzione della dimensionalità a 50 dimensioni. Si è inoltre utilizzato il t-SNE (t-Distributed Stochastic Neighbor Embedding) per visualizzare lo spazio latente su un grafico bidimensionale. Per visualizzare il funzionamento del motore di ricerca, il t-SNE viene calcolato con l'uso della similarità del coseno, per rappresentare meglio la metrica di distanza utilizzata dal motore di ricerca.

La Figura 8a riporta lo spazio degli embeddings delle frasi LaBSE per diversi documenti. È evidente che la distribuzione dello spazio LaBSE è più regolare rispetto agli altri spazi. Considerando che tutti i documenti sono coerenti negli argo-

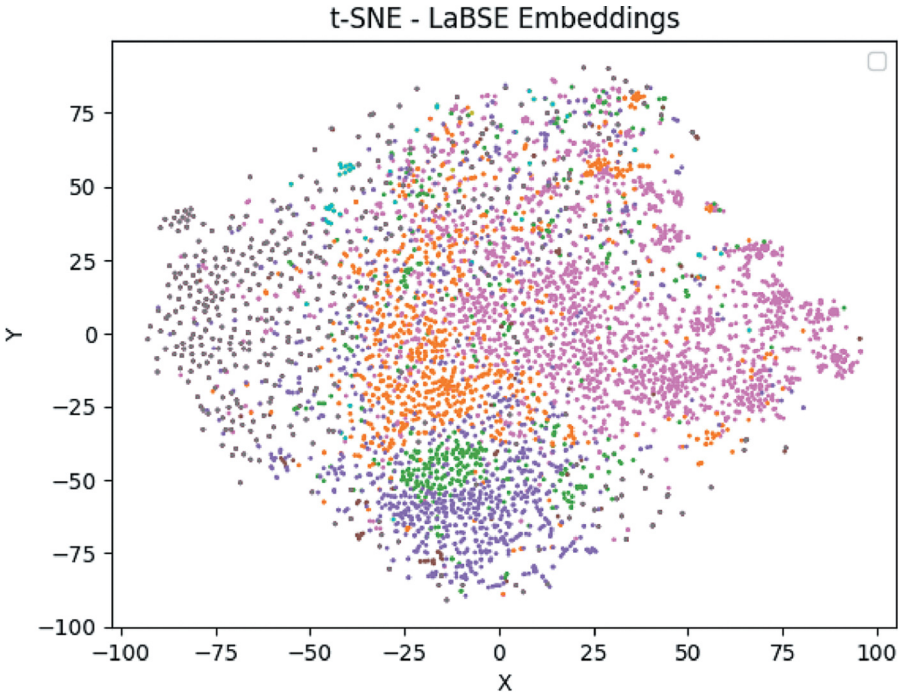

Figura 8a: LaBSE Embeddings t-SNE – proiezioni di frasi dei documenti.

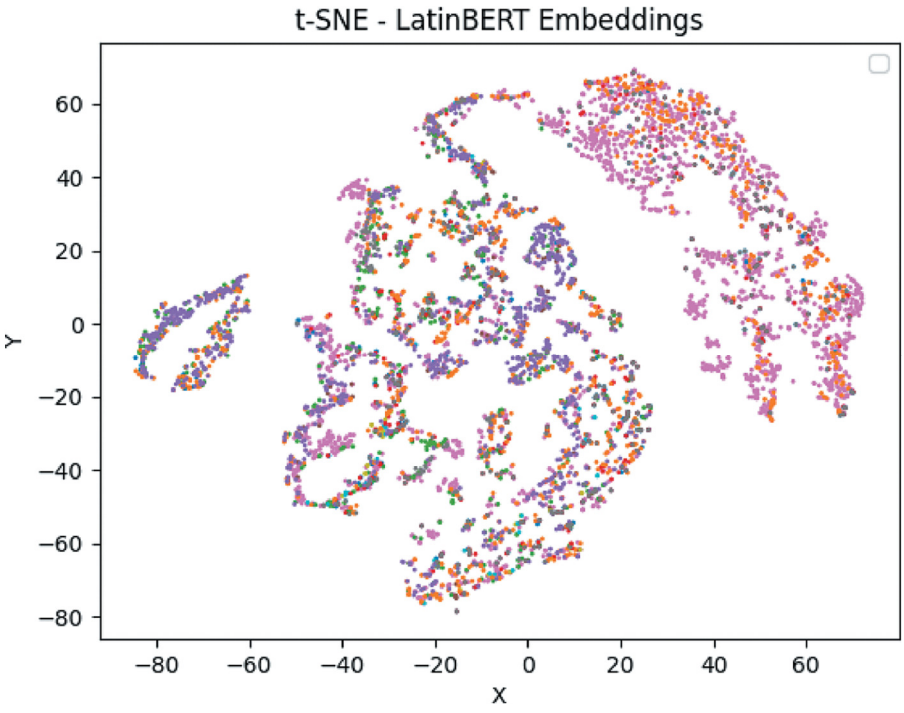

Figura 8b: LatinBERT Embeddings t-SNE – proiezioni di frasi dei documenti.

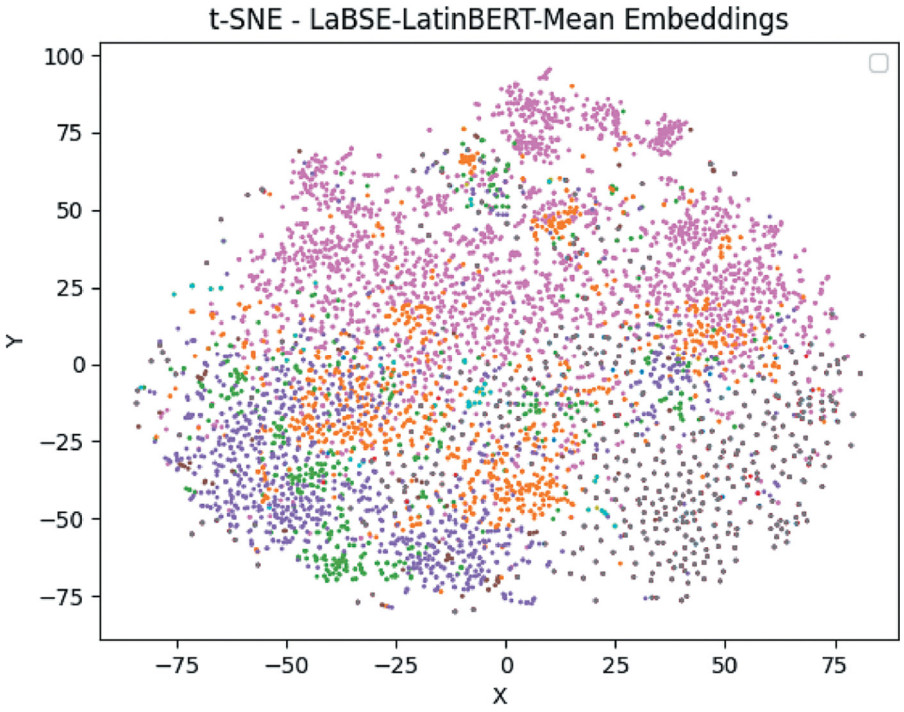

Figura 9a: LaBSE-LatinBERT Mean Embeddings t-SNE – proiezione di frasi di documenti.

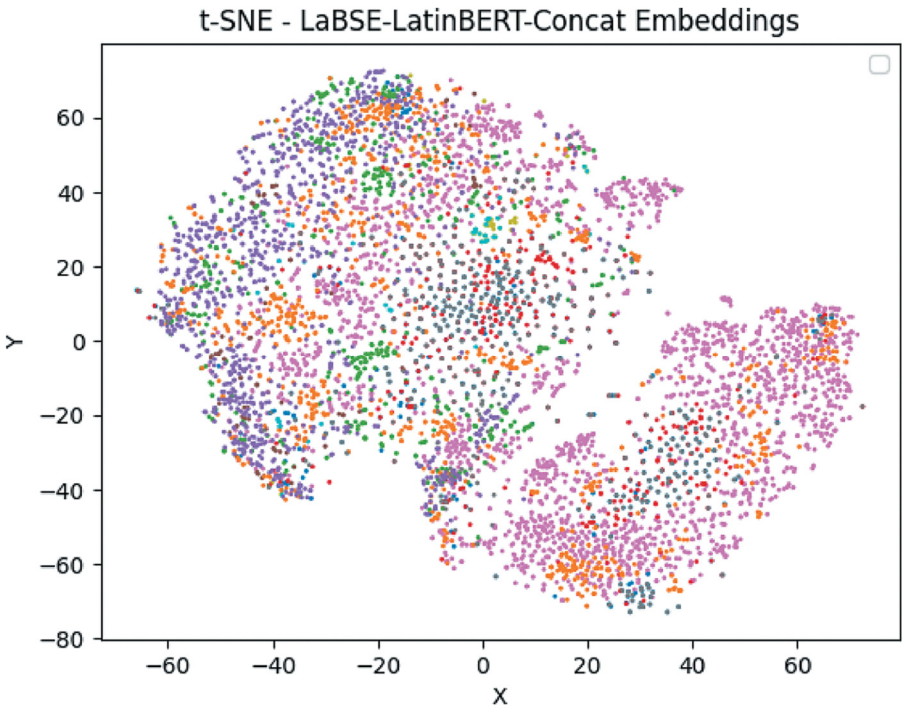

Figura 9b: LaBSE-LatinBERT-Concat Embeddings t-SNE – proiezioni di frasi di documenti.

menti che trattano, nonostante lo spettro lessicale, questa regolarità dello spazio potrebbe spiegare le migliori prestazioni ottenute dal modello LaBSE, così come le minori prestazioni del modello Latin BERT, rappresentato nella Figura 8b. Infine, gli embeddings Mean e Concat, rappresentati nella Figura 9, riflettono una diminuzione della regolarità dello spazio.

5 Conclusioni

Nonostante gli ampi progressi dei sistemi di Dense Information Retrieval (DIR), il recupero di testi in lingue antiche, come il latino, deve essere ulteriormente investigato, viste le sfide aggiuntive. Questo lavoro di ricerca illustra lo sviluppo di un sistema DIR e la sua sperimentazione sulla biblioteca digitale latina SERICA, effettuando interrogazioni multilingue in latino e in inglese. Viene inoltre proposta una metrica per tali motori di ricerca, al fine di valutare le prestazioni del sistema a partire da un insieme di documenti classificati. I primi risultati mostrano il potenziale di questo quadro comparativo, incoraggiando ricerche future. In particolare, è stato dimostrato che il modello LaBSE supera il modello Latin BERT, e che le interrogazioni in inglese hanno risultati migliori di quelle in latino. Il codice sorgente è stato rilasciato pubblicamente, insieme a una dimostrazione in-browser.

Ringraziamenti

Lavoro parzialmente supportato dal Ministero dell'Università e della Ricerca (MUR), nell'ambito di: (i) programma FISR 2019, nell'ambito del Grant n. 03602 del progetto "SERICA"; (ii) progetto FoReLab (Dipartimenti di Eccellenza); (iii) progetto "Reasoning", Programma PRIN 2020 LS, Numero progetto 2493 04–11-2021. Ricerca parzialmente finanziata dal PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR – Future Artificial Intelligence Research" – Spoke 1 "Human-centered AI", finanziato dalla Commissione Europea nell'ambito del programma comunitario NextGenerationEU. Questo lavoro è stato parzialmente svolto all'interno del Centro Nazionale per la Mobilità Sostenibile MOST/Spoke10, finanziato dal Ministero dell'Università e della Ricerca nell'ambito del Piano Nazionale di Ripresa e Resilienza.