



UNIVERSITÀ DEGLI STUDI DI  
PERUGIA  
Dipartimento di Matematica e Informatica



TESI DI LAUREA TRIENNALE IN INFORMATICA

Approccio Deep Learning  
al Topic Modeling:  
Analisi di annunci di lavoro tramite BERT

*Relatore*

**Prof. Valentina Poggioni**

*Laureando*

Tommaso Bagiana

---

Anno Accademico 2024-2025

*Lorem ipsum dolor sit amet, consectetur adipiscing elit.*

# Indice

<b>1</b>	<b>Introduzione</b>	<b>5</b>
1.1	Cos'è il Topic Modeling . . . . .	5
1.2	Analisi del paper <i>The Dynamics of Data Analytics Job Skills: a Longitudinal Analysis</i> . . . . .	5
1.2.1	Obiettivi e motivazioni . . . . .	6
1.2.2	Novità nella letteratura . . . . .	6
1.2.3	Metodologia . . . . .	6
1.2.4	Risultati . . . . .	7
1.2.5	Evoluzione del vocabolario . . . . .	8
1.2.6	Conclusioni . . . . .	8
1.2.7	Rilevanza per questo progetto . . . . .	8
1.3	Approccio con BERT . . . . .	9
<b>2</b>	<b>Data Cleaning</b>	<b>10</b>
2.1	Strategie di data cleaning usate . . . . .	10
2.2	Divisione in paragrafi . . . . .	10
2.3	Classificazione paragrafi . . . . .	10
<b>3</b>	<b>BERTopic</b>	<b>11</b>
3.1	Pipeline . . . . .	11
3.1.1	Embedding . . . . .	11
3.1.2	Dimensionality Reduction . . . . .	11

3.1.3	Clustering . . . . .	12
3.1.4	Tokenizer . . . . .	12
3.1.5	Cosa si può migliorare . . . . .	12
<b>4</b>	<b>Risultati finali</b>	<b>13</b>
<b>5</b>	<b>Conclusioni</b>	<b>14</b>
	<b>Ringraziamenti</b>	<b>16</b>

# Capitolo 1

## Introduzione

Panoramica introduttiva del progetto, motivazione della ricerca e obiettivi principali.

### 1.1 Cos'è il Topic Modeling

Descrivere in termini generali l'obiettivo del topic modeling, le tecniche classiche e i vantaggi rispetto ad approcci basati su keyword.

### 1.2 Analisi del paper *The Dynamics of Data Analytics Job Skills: a Longitudinal Analysis*

*Almgerbi, De Mauro, Kahlawi, Poggioni (2025)* presentano uno studio longitudinale sull'evoluzione delle competenze richieste negli **annunci di lavoro in ambito Data Analytics** tra il 2019 e il 2023. La loro ricerca mira a catturare le dinamiche temporali nella domanda di competenze applicando il topic modeling basato su **Latent Dirichlet Allocation (LDA)** a grandi corpora di annunci di lavoro online (Online Job Advertisements, OJA).

### 1.2.1 Obiettivi e motivazioni

Gli autori sottolineano come la rapida trasformazione digitale e la diffusione dei Big Data abbiano generato una crescente domanda di professionisti competenti in statistica, programmazione, tecnologie cloud e intelligenza artificiale. Tuttavia, la crescita dell’offerta formativa accademica e professionale non procede allo stesso ritmo delle esigenze dell’industria. Poiché ruoli come *Data Scientist*, *Business Analyst* e *Big Data Engineer* presentano competenze sovrapposte, lo studio adotta il termine più ampio **Data Analytics** per includere tutti questi ambiti professionali. Gli annunci di lavoro online vengono quindi trattati come una fonte affidabile e in tempo reale per monitorare l’andamento del mercato del lavoro.

### 1.2.2 Novità nella letteratura

Sebbene numerosi studi abbiano applicato tecniche di text mining o topic modeling agli annunci di lavoro, questo lavoro è il **primo a condurre un’analisi longitudinale su più anni**. La letteratura precedente si è concentrata su settori specifici (ad esempio marketing o IT) oppure ha impiegato strumenti proprietari e corpora statici. La metodologia proposta offre invece un quadro replicabile per **monitorare l’evoluzione delle competenze nel tempo**, fornendo indicazioni operative per le risorse umane e per le politiche formative.

### 1.2.3 Metodologia

Lo studio segue un processo articolato in quattro fasi:

1. **Raccolta dati:** gli annunci sono stati estratti da diversi siti web nel 2019 e nel 2023 utilizzando le stesse sei keyword (*big data*, *data science*, *business intelligence*, *data mining*, *machine learning*, *data analytics*), ottenendo un dataset bilanciato di 16 060 annunci (8 030 per anno).

2. **Pre-processing:** accurata pulizia del testo (rimozione di HTML e punteggiatura, filtraggio di stopword e n-gram), stemming, esclusione dei testi non in lingua inglese o troppo brevi e costruzione del dizionario/corpus per il topic modeling.
3. **Topic modeling:** esecuzione di più run LDA con  $k = 5-20$  topic; il numero ottimale ( $k = 12$ ) è stato selezionato combinando punteggi di coerenza e valutazione qualitativa degli esperti.
4. **Analisi:** interpretazione dei topic, confronto longitudinale (2019 vs. 2023) e studio dell'evoluzione del vocabolario.

### 1.2.4 Risultati

Sono stati individuati dodici topic distinti: *Financial Applications*, *Sales and Marketing Applications*, *Foundational Statistics*, *Cybersecurity Applications*, *Project Management*, *Business Intelligence*, *Databases*, *Scientific Research Applications*, *Cloud and Big Data Engineering*, *Machine Learning*, *Software Engineering*, *Senior Management*.

L'analisi comparativa ha messo in evidenza alcune tendenze chiave:

- **Crescente specializzazione settoriale:** l'aumento della domanda in finanza (+6%) e marketing (+10%) evidenzia il passaggio da ruoli generalisti ad analisti focalizzati su domini specifici.
- **Dalle competenze teoriche a quelle applicative:** il calo di *Foundational Statistics* (-13%) e la crescita di *Machine Learning* (+12%) indicano una transizione verso competenze pratiche in ambito AI e NLP.
- **Commoditizzazione dell'infrastruttura:** la diminuzione di richieste per *Database* (-39%) e *Cloud Engineering* (-7%) suggerisce l'impatto di automazione e servizi cloud sempre più user-friendly.

- **Maggiore bisogno di software e governance:** l'incremento di *Software Engineering* (+13%) e *Senior Management* (+8%) segnala l'importanza crescente di capacità di leadership, governo dei processi e integrazione software.

### 1.2.5 Evoluzione del vocabolario

L'analisi delle frequenze (termini con più di 500 occorrenze) mostra un chiaro cambio tecnologico:

- **Termini in crescita:** *Databricks* (+551%), *PyTorch* (+226%), *NLP* (+130%), *Modelling* (+129%), *GCP* (+114%), a testimonianza del ruolo crescente di framework cloud e machine learning.
- **Termini in diminuzione:** *Hadoop* (−50%), *Java* (−51%), *SSRS* (−61%), *CSS* (−51%), che rappresentano il declino di tecnologie legacy orientate all'infrastruttura.

### 1.2.6 Conclusioni

Il paper mostra come il mercato del lavoro in ambito Data Analytics stia evolvendo verso ruoli **specializzati, guidati dall'AI e orientati al software**. Le competenze di leadership, governance e integrazione dei sistemi acquistano peso accanto alle competenze tecniche. Gli autori propongono inoltre una **metodologia longitudinale di topic modeling replicabile** applicabile anche ad altri domini professionali. Tra i limiti figurano il bias intrinseco degli annunci online e la rapida obsolescenza dei trend tecnologici. Per il futuro si suggerisce l'integrazione di ulteriori fonti (sondaggi, curricula formativi) per anticipare meglio le competenze emergenti.

### 1.2.7 Rilevanza per questo progetto

Questo studio fornisce le basi concettuali e metodologiche per il nostro lavoro:



- stabilisce il topic modeling longitudinale basato su LDA come benchmark per il confronto con i metodi moderni basati su embedding come **BERTopic**;
- conferma l'importanza di analizzare le **dinamiche temporali** nei corpora di annunci di lavoro;
- enfatizza la combinazione tra metriche quantitative di coerenza e interpretabilità umana, principio mantenuto nel nostro approccio tramite UMAP, HDBSCAN e probabilità di topic.

## 1.3 Approccio con BERT

Introdurre il modello BERT, motivare la sua scelta per il topic modeling e spiegare a livello concettuale la pipeline prevista.

# Capitolo 2

## Data Cleaning

Introdurre le fasi preliminari di pulizia dati e il razionale delle scelte effettuate.

### 2.1 Strategie di data cleaning usate

Elencare le tecniche applicate (rimozione stopwords, normalizzazione, gestione dei duplicati, anonimizzazione, ecc.) indicando per ciascuna motivazioni e strumenti.

### 2.2 Divisione in paragrafi

Descrivere come i testi degli annunci vengono segmentati in paragrafi o blocchi tematici, includendo eventuali regole euristiche o soglie adottate.

### 2.3 Classificazione paragrafi

Spiegare il processo di etichettatura o clustering preliminare dei paragrafi, specificando se è supervisionato o meno e come si valida la coerenza delle classi.

# Capitolo 3

## BERTopic

Descrivere l'architettura di BERTopic e le motivazioni per cui è stata scelta per l'analisi degli annunci.

### 3.1 Pipeline

Presentare la pipeline end-to-end, evidenziando gli input, le trasformazioni intermedie e l'output finale del modello.

#### 3.1.1 Embedding

Indicare come vengono generati gli embedding con BERT (o varianti), eventuali fine-tuning e impostazioni rilevanti.

#### 3.1.2 Dimensionality Reduction

Illustrare l'algoritmo di riduzione dimensionale adottato (es. UMAP), i parametri principali e l'impatto sulla qualità dei topic.

### **3.1.3 Clustering**

Spiegare il metodo di clustering (es. HDBSCAN), i criteri di scelta dei parametri e come viene determinato il numero di topic.

### **3.1.4 Tokenizer**

Dettagliare il tokenizer utilizzato, le strategie di pre-processing e qualsiasi personalizzazione per il dominio degli annunci di lavoro.

### **3.1.5 Cosa si può migliorare**

Discutere criticità note, possibili ottimizzazioni della pipeline e idee per estensioni future di BERTopic nel progetto.

## Capitolo 4

### Risultati finali

Descrivere i topic ottenuti, le metriche di valutazione e le principali evidenze emerse dall'analisi degli annunci.

Inserire tabelle, grafici e commenti qualitativi che illustrino chiaramente l'efficacia dell'approccio BERTopic applicato al dataset.

## Capitolo 5

## Conclusioni

Sintetizzare i risultati principali e delineare i possibili sviluppi futuri.

## Bibliografia

# Ringraziamenti

Testo dei ringraziamenti da completare.