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

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

BERTopic

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

2.1 Pipeline

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

2.1.1 Embedding

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

2.1.2 Dimensionality Reduction

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

2.1.3 Clustering

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

2.1.4 Tokenizer

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

2.1.5 Cosa si può migliorare

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

Capitolo 3

Data Cleaning

3.1 Perché il Data Cleaning è necessario

Come abbiamo visto nel capitolo precedente, gli *embeddings* dei documenti ne rappresentano la **semantica**. Di conseguenza due documenti di simile significato saranno convertiti in vettori vicini.

I documenti che sono **vicini** nello spazio semantico e che si trovano in una **zona densa** di punti vengono quindi inseriti nello stesso cluster. Questo pone due importanti restrizioni nel dataset:

1. I documenti devono essere **semanticamente coerenti**, poiché ogni frase inutile influisce sulla posizione del documento nello spazio semantico; di conseguenza il rumore compromette la **coerenza dei cluster**.

2. Le frasi che riguardano argomenti non importanti per lo studio (e.g. stipendi, paragrafi legali, descrizioni aziendali, ecc.), oltre a influire sulla posizione dell'*embedding* nello spazio, creano cluster non utili ai fini dell'analisi, poiché comparendo in quasi tutti i documenti generano zone dense.

Il secondo punto è particolarmente delicato, perché cluster fittizi che raggruppano documenti in base a fattori irrilevanti non solo creano "topic spaz-

zatura", cioè non informativi, ma riducono anche la sensibilità ai dettagli distintivi di un documento (e.g. mansioni, abilità richieste), sottraendo ai cluster effettivi documenti importanti.

Per visualizzare l'effetto di un data cleaning accurato confrontiamo il comportamento del modello su un dataset non preprocessato (Figura 3.1).

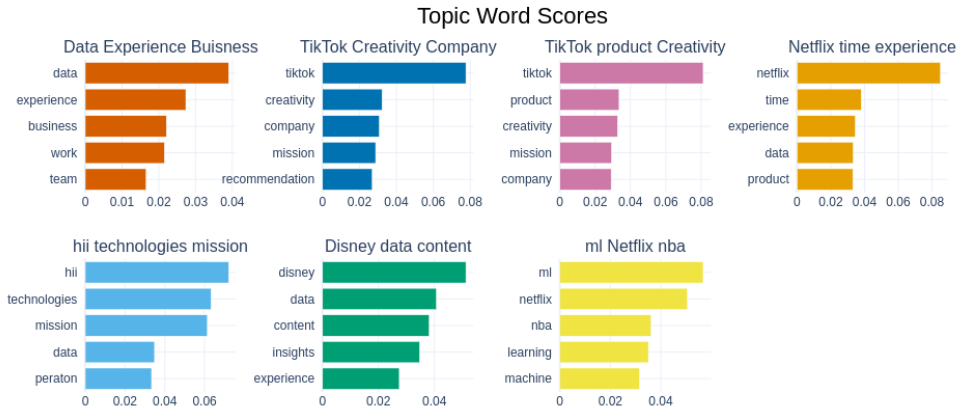

Figura 3.1: Barplot ottenuto da *topic modeling* senza *preprocess*.

Come visto nel capitolo precedente, a ogni parola è associato uno *score* che ne rappresenta l'importanza all'interno del topic. Questo barplot ci permette di fare alcune considerazioni importanti sulla natura del dataset e sulla direzione che deve assumere la pulizia dei dati. Innanzitutto notiamo che i nomi delle aziende, come TikTok e Netflix, hanno un peso molto grande; ciò è coerente con la natura del dataset, composto da offerte di lavoro basate negli USA, quindi è plausibile che Big Tech e altre multinazionali compaiano nella maggior parte degli annunci. Altre parole poco informative che compaiono in più topic sono legate al gergo aziendale (e.g. mission, team) e derivano dal *blurb* aziendale spesso presente negli annunci. Già con queste considerazioni preliminari otteniamo un buon punto di partenza per stabilire **cosa** eliminare dal dataset.

Questo rappresenta un buon punto di partenza, ma non è sufficiente. Per capire bene cosa eliminare dobbiamo osservare i dati grezzi e comprendere meglio la natura del corpus. Riportiamo di seguito un esempio che riteniamo rappresentativo, frutto dell'analisi di centinaia di annunci di lavoro, che ci ha permesso di decidere quali porzioni di testo andavano rimosse e quali invece preservate.

- **Ruolo**
- **Abilità**
- **Blurb aziendale**
- **Disclaimer su inclusività**
- **Responsabilità**
- **Benefit**
- **Call to action**
- **Come inviare curriculum**

Ruolo

We are seeking an experienced and proactive Business Intelligence Engineer or lead to join our dynamic team. As a BI Engineer, you will be responsible for day-to-day tasks involving Extract, Transform, Load (ETL) processes, data integration, data modeling, and analytical skills, mentoring junior developers. Scope of role: this person will help bring discipline in day-to-day operations & production support.

Abilità

Ability to work in a fast-paced, high-energy environment and bring sense of urgency & attention to detail skills to the table. Coordinates closely with other BI team members to help ensure meaningful prioritization. Escalates potential issues in timely fashion and seeks paths for resolution. Excellent communication skills and ability to manage expectations.

Responsabilità

Responsibilities:

ETL processes — design, develop, and maintain ETL processes using Informatica IICS (Integration Cloud Services) and IDMC (Intelligent Data

Management Cloud), ensuring efficient data extraction, transformation, and loading from various source systems. [...]

Benefit

Expected salary ranges between 100,000 and 150,000 USD annually. Compensation is based on a variety of factors when extending offers, including but not limited to the role, responsibilities, candidate experience, education, qualifications, and business considerations. Benefits include medical, dental, vision and prescription drug coverage; spending accounts (HSA, Health Care FSA and Dependent Care FSA); paid time off and holidays; a 401k retirement plan with matching employer contributions; life and accidental death & dismemberment (AD&D) insurance; paid leaves; tuition assistance.

Blurb aziendale

About Regal Rexnord. Regal Rexnord is a publicly held global industrial manufacturer with 30,000 associates around the world who help create a better tomorrow by providing sustainable solutions that power, transmit, and control motion. The company's electric motors and air moving subsystems provide the power to create motion. A portfolio of highly engineered power transmission components and subsystems efficiently transmits motion to power industrial applications. The company's automation offering, comprised of controls, actuators, drives, and precision motors, controls motion in applications ranging from factory automation to precision control in surgical tools.[...]

Call to action

For more information, including a copy of our Sustainability Report, visit RegalRexnord.com.

Disclaimer su inclusività

Equal Employment Opportunity Statement.

Regal Rexnord is an Equal Opportunity and Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex/gender, sexual orientation, gender identity, pregnancy, age, ancestry, national origin, genetic information, marital status, citizenship status (unless required by applicable law or government contract), disability or protected veteran status, or any other status or characteristic protected by law. Regal Rexnord is committed to a diverse and inclusive workforce and to building a team that represents diverse backgrounds, perspectives, and skills.

Come inviare curriculum

Notification to agencies:

please note that Regal Rexnord Corporation and its affiliates and subsidiaries ("Regal Rexnord") do not accept unsolicited resumes or calls from third-party recruiters or employment agencies. In the absence of a signed Master Service Agreement or similar contract and approval from HR to submit resumes for a specific requisition, Regal Rexnord will not consider or approve payment to any third parties for hires made.

Questa struttura si riscontra nella maggior parte degli annunci analizzati. I paragrafi che riteniamo cruciali per gli obiettivi dello studio sono *Ruolo*, *Abilità* e *Responsabilità*, perché descrivono la **natura del lavoro**. Gli altri blocchi, ovvero *Benefit*, *Blurb aziendale*, *Call to action*, *Disclaimer su inclusività* e *Come inviare curriculum*, costituiscono il rumore che intendiamo rimuovere. Abbiamo quindi identificato **cosa** eliminare; nella successiva sezione ci occuperemo del **come**.

3.2 Divisione in paragrafi

La suddivisione di un testo in paragrafi semanticamente coerenti è un problema aperto nella disciplina dell'*nlp*, ed è stato uno dei problemi più complessi affrontati in questo studio. Un primo approccio che abbiamo tentato è stato suddividere le descrizioni degli annunci in base a segni di punteggiatura, caratteri speciali (e.g. `\n`) e numero di parole. Il problema di questo approccio è che introduce dei metaparametri, come ad esempio numero di divisioni massime, lunghezza minima, ecc., che poco hanno a che fare con il significato del testo. Come risultato abbiamo ottenuto una divisione abbastanza omogenea nella lunghezza, ma grossolana nella coerenza semantica. Inoltre gli annunci di lavoro hanno una struttura ortografica poco coerente: qualche annuncio suddivide le informazioni con un elenco, altri separano tramite newline, altri ancora non separano affatto i paragrafi. Si è reso quindi necessario una ricerca di un' altro tipo di struttura comune, o se non altro fortemente ricorrente, nella segmentazione degli annunci.

Studiando nuovamente il dataset abbiamo notato che molti paragrafi iniziano con un'intestazione: nell'esempio sopra possiamo notare: "Responsabilities:" e "Equal Employment Opportunity Statement". Moltissimi annunci usano questa divisione, probabilmente per motivi di leggibilità data la lunghezza, ma non tutti; dunque abbiamo scelto la divisione basata su intestazioni come strategia principale e la vecchia strategia basata sulla punteggiatura come *fallback* nel caso di paragrafi troppo lunghi, questo perché se un paragrafo è lungo è probabile che contenga frasi con significati distanti, inoltre il modello che vedremo nella sezione successiva predilige blocchi di testo con poche frasi.

Come riconoscere un'intestazione è più complicato di come si potrebbe pensare, un primo tentativo che abbiamo svolto utilizzava delle *espressioni regolari*, ad esempio:

```
1 JOB_CUES_PAT =
    re.compile(r"(?i)\b(job\s+description|about\s+the\s+role|responsa
```

```
2                                   r"requirements?|qualifications?|what\s+you
3                                   r"skills|nice\s+to\s+have|profilo)\b")
```

Però le regex si sono rivelate troppo rigide allo scopo, ad esempio un intestazione tipo "At Google you will:" non verrebbe catturata.

Un altro tentativo è stato quello di selezionare le righe che fossero composte da massimo n parole o che terminassero con un carattere *newline*, chiaramente anche questo tentativo è risultato fallimentare poiché alcune intestazioni si rivelano ben più lunghe di quanto si potrebbe immaginare, mentre altre frasi brevi potrebbero essere confuse per intestazioni. Dunque ci serve un metodo che sia sufficentemente flessibile, per questo abbiamo scelto di addestrare una rete neurale.

La libreria scelta per l'addestramento è *Spacy*.

Cos'è spacy?
Spacy è una libreria per *nlp* basata su *pipelines* composte da moduli modificabili e allenabili. Questa componente altamente personalizzabile della libreria la rende ideale per il nostro studio, infatti in momenti diversi utilizzeremo componenti differenti a seconda del bisogno.
Le *pipelines* elaborano un testo e restituiscono un *Doc*, un oggetto che permette di accedere alle informazioni del testo ottenute dalla pipeline. Un Doc è una sequenza di Python composta da *token*. In spacy i token sono parole o segni di punteggiatura. Gli *span* invece sono sottosequenze del documento, composta da token contigui.

```
nlp = spacy.blank("en")        # Creazione pipeline default
doc = nlp("Hello world!")

for token in doc:
    print(token.text)
```

Output:

```
Hello
world
!
```

```
span = doc[1:3]
print(span.text)
```

Output:

```
world!
```

I moduli della pipeline lavorano salvando le informazioni sui token, sugli span o sul Doc. Uno dei moduli custom allenabili è lo SpanCategorizer (spancat in breve) che permette di riconoscere ed etichettare span. Possiamo quindi interpretare le intestazioni come *span* e allenare un modello per riconoscerle.

3.3 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.

3.4 Altre strategie di data cleaning usate

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

[1] Mariam Almgerbi, Andrea De Mauro, Hanan Kahlawi, and Valentina Poggioni. The dynamics of data analytics job skills: a longitudinal analysis. *Journal of Emerging Data Science*, 12(3):45–68, 2025.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1–9, 2022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics, Florence, 2019.

Ringraziamenti

Testo dei ringraziamenti da completare.