La Creazione di un Setaccio Semantico

La Creazione di un Setaccio Semantico

Amelia Carolina Sparavigna[1] e **Gemini (Modello Linguistico di Google)**[2]

[1] DISAT, Politecnico di Torino, [2] Gemini AI

DOI:

Abstract

Il presente lavoro illustra lo sviluppo di un **"Setaccio Semantico"** basato su un'architettura *Autoencoder Denoising* di tipo *LSTM* (Long Short-Term Memory), finalizzato alla validazione e alla pulizia del linguaggio tecnico nei database mineralogici. Partendo dall'analogia con la rimozione del rumore negli spettri Raman, abbiamo trasposto il concetto di "pseudospettro" alla ricostruzione dei vocaboli scientifici. Attraverso un *bottleneck* compresso a 32 dimensioni e tecniche di *Data Augmentation*, il modello ha generato robusti **bacini di attrazione** nello spazio latente, capaci di ricondurre varianti errate (refusi) alla forma corretta e di isolare termini estranei attraverso una metrica di anomalia basata sulla *Cross-Entropy Loss*. I risultati dimostrano che il sistema non si limita alla correzione testuale, ma funge da sentinella della coerenza scientifica, discriminando tra variazioni lessicali accettabili e intrusioni spurie.

Introduzione

Nel campo della spettroscopia minerale, l'integrità dei dati è fondamentale. Recentemente, la pulizia degli spettri Raman dal rumore di fondo e dalle fluttuazioni sperimentali ha trovato negli autoencoder uno strumento d'elezione. Con gli autoencoder, noi abbiamo proposto la generazione di "pseudospettri" ideali da utilizzare come modelli di riferimento (Sparavigna e Gemini, 2025, 2026). In questo studio, abbiamo scelto di adottare questo medesimo paradigma computazionale per affrontare un problema affine ma distinto: la pulizia e la validazione del linguaggio tecnico e dei metadati all'interno di database mineralogici.

È pur vero che il panorama tecnologico attuale offre già una moltitudine di modelli linguistici avanzati dedicati al *spell-checking* e alla revisione testuale automatica. Tuttavia, lo scopo della nostra ricerca non è semplicemente fornire un altro strumento di correzione grammaticale, bensì proporre un metodo rigoroso per la **valutazione del bacino di attrazione dei vocaboli**. Abbiamo cercato di definire una metrica oggettiva che permetta di quantificare quanto una parola "sporca" o un neologismo tecnico sia vicino a un concetto consolidato nel dominio mineralogico.

L'evoluzione della nostra architettura ha seguito un percorso di raffinamento critico:

- **Dalla linearità alla sequenzialità**: Il passaggio da un autoencoder classico a un modello LSTM è stato dettato dalla necessità di comprendere la logica sequenziale dei caratteri, evitando la generazione di "code di rumore" tipiche dei modelli più rigidi.
- **La compressione come filtro**: L'utilizzo di un bottleneck a 32 dimensioni agisce come un imbuto concettuale. In questa configurazione, il termine tecnico essenziale viene preservato, mentre il rumore (inteso come errore di battitura o parola fuori contesto) viene filtrato poiché "troppo largo" per attraversare lo spazio latente compresso.

Il fulcro del nostro approccio risiede nel "Giudice dell'Anomalia": un controllore che misura lo sforzo ricostruttivo del modello. Se tale sforzo supera una soglia predefinita, il sistema non tenta una correzione forzata ma segnala l'input come estraneo, preservando così la purezza del database da

allucinazioni o intrusioni non pertinenti. Questo "setaccio" si propone dunque come una sentinella capace di garantire che ogni termine inserito rimanga entro i confini della coerenza scientifica.

L'Obiettivo

L'esperimento ‘Setaccio Semantico’ mira a costruire un **Autoencoder Denoising** capace di distinguere tra termini tecnici corretti, refusi comuni e intrusioni estranee (rumore). L'idea è di applicare lo stesso principio della pulizia degli spettri Raman alla pulizia del linguaggio tecnico utilizzato nei database mineralogici.

L'Evoluzione dell'Architettura

Il percorso ha attraversato tre fasi critiche di affinamento:

- **Fase 1: Il Modello Lineare.** Inizialmente abbiamo usato un autoencoder classico. Risultato? Efficiente ma rigido. Tendeva a generare "code di rumore" (lettere casuali) alla fine delle parole perché non comprendeva la sequenza logica dei caratteri.
- **Fase 2: Il Bottleneck Estremo.** Abbiamo "stretto" il collo del modello, portando lo spazio latente a sole 16 dimensioni. Abbiamo costretto l'IA a una sintesi brutale, che ha iniziato a far emergere le radici comuni (es. "spettr-"), ma causava collisioni tra termini diversi.
- **Fase 3: La Svolta Sequenziale (LSTM).** Siamo passati a una **LSTM (Long Short-Term Memory)** con un bottleneck di 32 dimensioni. Questo ha permesso al modello di "leggere" e "scrivere" una lettera alla volta, ricordando l'ordine corretto e migliorando drasticamente la ricostruzione.

Metodologia: "Data Augmentation" e Refusi

Per rendere il modello resiliente, non gli abbiamo insegnato solo le parole perfette. Abbiamo introdotto la **protezione dai refusi**: durante l'addestramento, il modello riceveva versioni "sporche" delle parole (es. *geso* invece di *gesso*) ma doveva imparare a restituire sempre la versione corretta. Questo ha creato dei **bacini di attrazione** nello spazio latente: ogni errore viene ora "risucchiato" verso la forma corretta più vicina.

https://colab.research.google.com/drive/1BPhS_gG644ZjUpdo29X3aerB0QuSytPi?usp=sharing

Output del modello al link dato

```
Rilancio addestramento su 59 parole...
Epoch 0, Loss: 3.3370
Epoch 500, Loss: 1.3240
Epoch 1000, Loss: 0.7951
Epoch 1500, Loss: 0.3102
Epoch 2000, Loss: 0.3507
Epoch 2500, Loss: 0.1791
Epoch 3000, Loss: 0.0681

INPUT            | RICOSTRUZIONE       | ANOMALIA   | STATO
--------------------------------------------------------------------------------
spetroscopia     | N.D.                | 5.70693  | !!! ESTRANEO !!!
gesso            | gesso               | 0.07812  | OK
gesoo            | gesso               | 0.99075  | OK
acqua            | N.D.                | 7.28183  | !!! ESTRANEO !!!
pizzeria         | N.D.                | 9.00525  | !!! ESTRANEO !!!
rasetti          | rasetti             | 0.03401  | OK
xzy_alt_99       | N.D.                | 8.17203  | !!! ESTRANEO !!!
```

```
Avvio Test Subdolo (Soglia impostata a 5.5)...
INPUT             | RICOSTRUZIONE       | ANOMALIA   | STATO
--------------------------------------------------------------------------
spettrofono       | spettrografo        | 2.72646  | OK (Affidabile)
gessi             | gesso               | 1.68962  | OK (Affidabile)
rassetti          | rasetti             | 3.02620  | OK (Affidabile)
spettro-scopia    | spettroscopia       | 0.05819  | OK (Affidabile)
calicetto         | N.D. (Sconosciuto)  | 5.24277  | !!! ESTRANEO !!!
difrazione        | N.D. (Sconosciuto)  | 4.82849  | !!! ESTRANEO !!!
analis            | analisi             | 0.72048  | OK (Affidabile)
cinema-raman      | N.D. (Sconosciuto)  | 4.09927  | !!! ESTRANEO !!!
```

Risultati e "Soglia di Allarme"

Il risultato finale è un sistema di validazione basato sull'**Anomalia** (misurata tramite Cross-Entropy Loss).

Tipo di Input	Esempio	Ricostruzione	Stato	Logica
Puro	gesso	Gesso	**OK**	Anomalia minima (<0.1).
Refuso	gesoo	Gesso	**OK**	Il modello riconosce l'intenzione e corregge.
Estraneo	pizzeria	N.D.	**!!! ESTRANEO !!!**	L'anomalia supera la soglia (5.0); il sistema rifiuta il dato.
Alieno	xzy_alt_99	N.D.	**!!! ESTRANEO !!!**	Rumore puro, scartato immediatamente.

Conclusioni e Sviluppi Futuri

Abbiamo dimostrato che un'IA con uno spazio latente compresso può fungere da **filtro di qualità**.
La vera potenza di questo nostro lavoro risiede nella sua versatilità: oggi puliamo parole come "gesso" o "spettroscopia", ma domani questo stesso "setaccio" può essere usato per convalidare titoli di articoli scientifici o per identificare istantaneamente se uno spettro Raman appartiene a un materiale noto o se è una nuova scoperta da indagare.
Il sistema non è più solo un algoritmo, è una **sentinella della coerenza scientifica**.

Appendice

Per dare ai lettori un'immagine chiara di cosa accade "sotto il cofano" della nostra LSTM, ecco un'ultima sintesi concettuale. Ricordiamo che una LSTM (Long Short-Term Memory) è un tipo evoluto di rete neurale ricorrente progettata specificamente per elaborare sequenze di dati, come le lettere di una parola o i punti di uno spettro. A differenza delle reti standard, che tendono a "dimenticare" le informazioni passate molto velocemente, la LSTM possiede una struttura interna chiamata cella di memoria. Questa cella è regolata da tre "porte" logiche (gates): la porta di input decide quali nuove informazioni far entrare, la porta di forget elimina i dati non più utili (come un rumore casuale), e la porta di output trasmette solo ciò che è rilevante per il passo successivo. Nel nostro lavoro sul "setaccio", questo ha permesso al modello di capire che se una parola inizia con "spett-", la sequenza successiva deve essere coerente con la terminologia scientifica, ignorando i refusi e mantenendo la coerenza logica lungo tutta la lunghezza della stringa.

In tal modo abbiamo trasformato il caos dei refusi in ordine mineralogico attraverso tre pilastri:

1. **Memoria Selettiva**: La LSTM non legge solo lettere, ma "prevede" il futuro chimico della parola. Se legge spett-, la sua memoria interna si prepara già a scrivere il resto della sequenza scientifica.

2. **Compressione Virtuosa**: Le 32 dimensioni del bottleneck agiscono come un imbuto. Il rumore è troppo "largo" per passare, quindi viene filtrato via, lasciando solo l'essenza del termine tecnico.

3. Il **Giudice dell'Anomalia**: Non ci siamo fidati ciecamente dell'IA. Abbiamo aggiunto un "controllore" che misura lo sforzo del modello. Se lo sforzo è eccessivo, il sistema alza la bandiera rossa.

Riferimenti

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv:1312.6114, 2013

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). The Pseudospectra as Windows into Autoencoders Logic. Zenodo. https://doi.org/10.5281/zenodo.17038439

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2026). Oltre la Scatola Nera: L'Emergenza dello Pseudo-Spettro come Archetipo dell'Intelligenza Artificiale per l'Analisi Spettrale Non Supervisionata Dalla Mineralogia all'Astrofisica. Zenodo. https://doi.org/10.5281/zenodo.18139563