Multi-Label Classification of Customer Reviews in the Energy Sector with BERT Base Models

1 Introduzione

Questo report descrive lo sviluppo e l'ottimizzazione di un classificatore *Multi-Label Multi-Class* applicato alle recensioni di clienti nel settore energetico italiano. L'obiettivo principale era automatizzare l'assegnazione di una o più etichette tematiche a ciascuna recensione, per consentire l'analisi su larga scala del feedback dei consumatori.

La sfida principale era rappresentata da due fattori: (1) la forte asimmetria nella distribuzione delle etichette e (2) il numero ridotto di recensioni etichettate disponibili. Per garantire che anche le classi meno frequenti fossero adeguatamente rappresentate sia nel training set (80%) che nel validation set (20%), è stato utilizzato il metodo `MultilabelStratifiedShuffleSplit`.

Per affrontare lo sbilanciamento, sono state combinate tecniche di data sampling e funzioni di loss personalizzate. In particolare:

- **WeightedRandomSampler**, per aumentare la frequenza di campionamento delle classi di minoranza durante l'addestramento;
- **Distribution-Balanced Loss (DBLoss)**, per pesare dinamicamente le classi in base alla loro frequenza e forzare il modello a prestare maggiore attenzione a quelle meno rappresentate.

Per migliorare la capacità di comprensione del dominio energetico, sono state esplorate strategie di *pre-training adattivo*:

- **Domain-Adaptive Pretraining (DAPT)**, applicato al modello di base `Musixmatch/umberto-commoncrawl-cased-v1`;
- **Task-Adaptive Pretraining (TAPT)**, per adattare ulteriormente il modello ai testi del dataset annotato.

È stato inoltre testato il modello multilingua `microsoft/mdeberta-v3-base`, dotato di *disentangled attention*, per valutare le performance di un'architettura alternativa.

L'obiettivo finale era identificare la combinazione ottimale di architettura, pre-training e tecniche di bilanciamento per massimizzare le performance sul compito di classificazione multi-label.

2 Training Dataset

Per lo sviluppo del progetto sono stati impiegati diversi dataset di recensioni in lingua italiana specifici del dominio energetico, utilizzati nelle fasi di pre-training e fine-tuning.

2.1 Corpus Etichettato

Il dataset di partenza (`training_dataset.xlsx`) è composto da 1.742 recensioni etichettate manualmente. È stato suddiviso in:

- **Training set**: 1.392 recensioni
- **Validation set**: 350 recensioni

La distribuzione delle etichette evidenzia un forte sbilanciamento, con frequenze variabili da 1.014 occorrenze (classe più popolosa, Label-7) a sole 8 (classe più rara, Label-13). In media ogni recensione possiede 2.17 etichette.

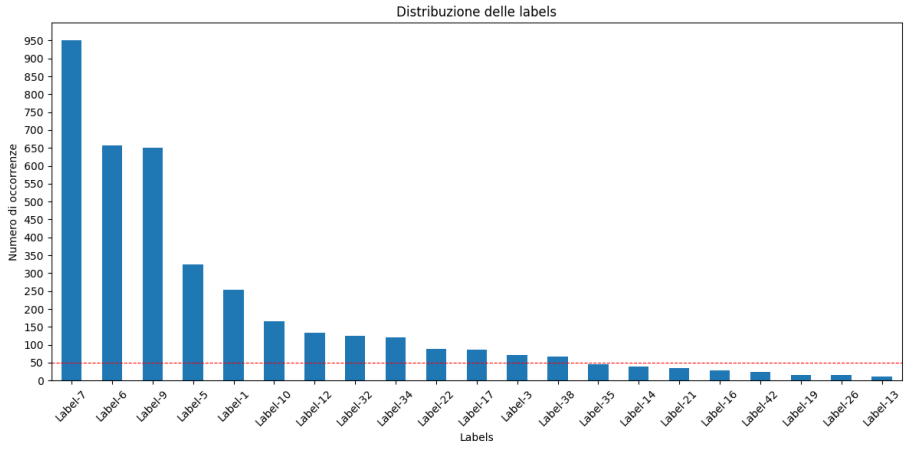

Figura 1: Distribuzione Long-Taile delle labels (placeholder immagine).

Per la tokenizzazione è stata scelta una lunghezza massima di 128 token, sulla base di un'analisi della distribuzione delle lunghezze.

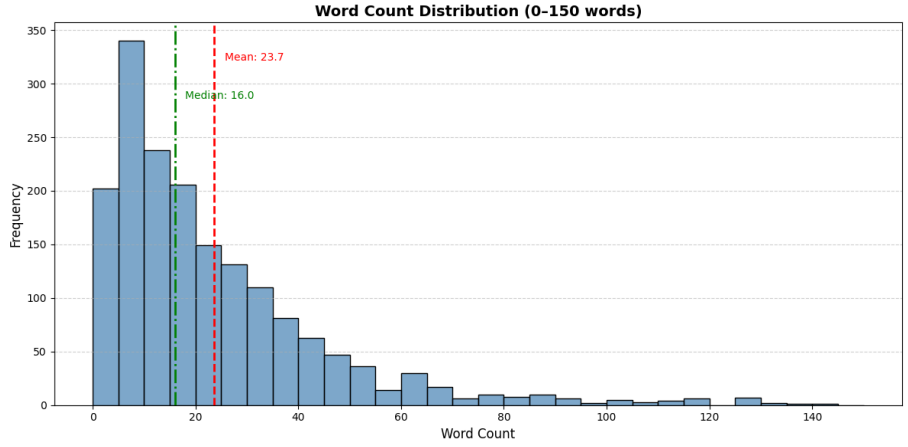

Figura 2: Distribuzione della lunghezza delle recensioni (placeholder immagine).

2.2 Corpus di Dominio (per DAPT)

Per il *Domain-Adaptive Pretraining* è stato assemblato un corpus di 34.795 recensioni non etichettate provenienti da 6 diverse fonti. Le statistiche principali sono riportate nella Tabella 1.

Fonte	**Recensioni**	**Avg Word Count**	**Median Word Count**
df_octopus	10279	27.4	20.0
df_eon	1822	60.7	36.0
df_plenitude	8636	51.3	35.0
df_sorgenia	9019	24.5	14.0
df_edison	1712	30.0	15.0
df_enel	3327	37.3	19.0

Tabella 1: Statistiche del corpus non etichettato utilizzato per DAPT.

2.3 Corpus Aumentato

Per alcuni esperimenti, il training set da 1.392 campioni è stato ampliato tramite la tecnica di *back-translation*, generando 335 nuove recensioni sintetiche. Il training set aumentato ha così raggiunto 1.727 campioni, mentre il validation set è rimasto invariato a 350 recensioni.

2.4 Preprocessing

Prima di ogni fase di addestramento i testi sono stati normalizzati tramite una funzione dedicata. Le operazioni includevano:

- rimozione di URL, emoji e caratteri speciali non pertinenti;
- normalizzazione degli spazi superflui;
- preservazione della punteggiatura e del casing originale.

3 Evaluation Measures

Per valutare le performance dei modelli sono state selezionate metriche basate sull'**F1-score**, una misura che bilancia i concetti di *precision* e *recall*. L'F1 è infatti la loro media armonica e restituisce un valore elevato solo se entrambe le metriche sono alte.

- **Macro F1 (metrica principale)**: calcola l'F1-score per ogni classe in modo indipendente e ne fa poi la media non pesata. Ha il vantaggio di dare la stessa importanza a tutte le classi, a prescindere dalla loro frequenza. In un dataset con molte classi rare, come in questo progetto, il Macro F1 è fondamentale per misurare la capacità del modello di riconoscere correttamente anche le categorie meno rappresentate.
- **Micro F1 (metrica secondaria)**: aggrega i conteggi di veri positivi, falsi positivi e falsi negativi di tutte le classi prima di calcolare lo score. In questo modo dà più peso alle classi frequenti. È stata utilizzata come metrica di supporto per valutare la performance complessiva a livello di singola predizione.

4 Methodology

L'approccio metodologico del progetto è stato strutturato in fasi sperimentali progressive, con l'obiettivo di identificare la pipeline ottimale per la classificazione multi-label in presenza di dati fortemente sbilanciati.

4.1 Modelli Considerati

Sono state prese in esame due architetture Transformer pre-addestrate:

- `Musixmatch/umberto-commoncrawl-cased-v1`, un modello basato su RoBERTa e addestrato su un vasto corpus italiano (Common Crawl), scelto come riferimento principale.
- `microsoft/mdeberta-v3-base`, un modello multilingua basato su DeBERTa, che introduce il meccanismo di *disentangled attention*, valutato come alternativa per il testo in italiano.

4.2 Strategie di Pre-Training Adattivo

Ispirandosi a *Don't Stop Pretraining* (Gururangan et al., 2020), sono state applicate due strategie di specializzazione al modello umBERTo:

1. **Domain-Adaptive Pretraining (DAPT)**: ulteriore pre-addestramento con obiettivo *Masked Language Modeling* su circa 35.000 recensioni non etichettate del dominio energetico, per adattare il modello al lessico e allo stile del settore.
2. **Task-Adaptive Pretraining (TAPT)**: pre-addestramento con lo stesso obiettivo sui testi del dataset annotato (1.742 campioni), per specializzare ulteriormente il modello ai dati esatti del task.

4.3 Gestione dello Sbilanciamento

Data la forte disparità nella frequenza delle etichette, sono state implementate diverse tecniche per ridurre l'impatto dell'imbalance:

- **Suddivisione Stratificata**: con `MultilabelStratifiedShuffleSplit`, per preservare la distribuzione multilabel sia nel training set (80%) che nel validation set (20%).
- **Data Augmentation (Back-Translation)**: generazione di nuove recensioni sintetiche traducendo testi rari in inglese (`Helsinki-NLP/opus-mt-it-en`) e nuovamente in italiano (`Helsinki-NLP/opus-mt-en-it`). Questo ha fornito frasi sintatticamente diverse ma semanticamente equivalenti, aumentando la rappresentatività delle classi rare.
- **WeightedRandomSampler**: durante il fine-tuning, le recensioni del training set sono state campionate con probabilità inversamente proporzionali alla frequenza delle etichette in esse contenute.
- **Distribution-Balanced Loss (DBLoss)**: in sostituzione della classica *Binary Cross-Entropy*, con pesi dinamici proporzionali alla rarità delle classi, per penalizzare maggiormente gli errori sulle categorie meno rappresentate.

5 Fine-Tuning e Valutazione

Tutti gli esperimenti di fine-tuning per la classificazione sono stati condotti utilizzando il `Trainer` della libreria `transformers`. L'addestramento è stato monitorato tramite la metrica **Macro F1** sul validation set.

Per prevenire overfitting e garantire il salvataggio del miglior modello, è stato utilizzato un `EarlyStoppingCallback`, che interrompeva l'addestramento se il Macro F1 non migliorava per un numero prefissato di epoche.

Tutti i training dei modelli sono stati documentati su *Weights & Biases*.

5.1 Post-Processing (Threshold Tuning)

È stata implementata una fase di post-processing per ottimizzare le soglie di decisione sul miglior modello.

Invece di adottare una soglia fissa di 0.5 per tutte le classi, è stata utilizzata una procedura di *threshold tuning*: per ciascuna delle 21 etichette è stata ricercata iterativamente (in un range 0.1–0.9) la soglia che massimizzava l'F1-score individuale.

L'applicazione di queste soglie ottimizzate ha portato a un miglioramento significativo delle metriche complessive:

- **Macro F1**: 0.615
- **Micro F1**: 0.783

Il threshold tuning tende ad aumentare il recall, ma comporta inevitabilmente anche un incremento dei falsi positivi. In un contesto di etichettatura di recensioni al fine di ocndurre indagini statistiche, i falsi positivi (etichette assegnate erroneamente) sono

particolarmente dannosi, poiché introducono dati errati che rischiano di compromettere la correttezza delle analisi e portare a decisioni di business fuorvianti.

Al contrario, i falsi negativi (etichette mancanti) riducono la numerosità del campione, ma non ne compromettono l'affidabilità complessiva.

Per queste ragioni, nel deployment si è preferito mantenere una soglia conservativa e uniforme, favorendo la **precision** rispetto al **recall**, così da garantire insight più puliti e affidabili.

6 Results and Conclusions

Modello / Strategia	**Macro F1**	**Micro F1**
BT-umBERTo-DAPT	0.524	0.770
umBERTo-DAPT	0.519	0.756
umBERTo (baseline)	0.502	0.757
umBERTo-TAPT	0.500	0.760
BT-umBERTo-TAPT	0.491	0.757
umBERTo-DAPT+TAPT	0.472	0.758
mDeBERTa-v3	0.465	0.760
umBERTo-TAPT	0.429	0.740
umBERTo-DAPT+TAPT	0.410	0.735
umBERTo-BCE	0.274	0.743

Tabella 2: Risultati comparativi dei modelli testati.

Dall'analisi comparativa degli esperimenti, il modello con le migliori performance in termini di Macro F1 è stato **BT-umBERTo-DAPT**.

Questo modello è stato sviluppato seguendo una pipeline in due fasi principali:

1. **Pre-training Adattivo (DAPT)**: ulteriore addestramento (Masked Language Modeling) su un corpus di circa 35.000 recensioni del settore energetico, per specializzarne la comprensione linguistica.

2. **Fine-Tuning con gestione dello sbilanciamento**: addestramento sul task di classificazione multi-label utilizzando un training set ampliato con back-translation. Sono state adottate strategie specifiche per l'imbalance, come *WeightedRandomSampler* e *Distribution-Balanced Loss* ($\beta = 0.95$, $\gamma = 2.5$, $\lambda = 2.0$). L'addestramento è stato guidato da una learning rate di $3e^{-5}$ e da un weight decay di 0.05, con monitoraggio tramite `EarlyStoppingCallback`.

In sintesi:

- Il modello baseline con `BCEWithLogitsLoss`, che non considera la distribuzione long-tail delle etichette, ha mostrato forti limiti (Macro F1 = 0.274).

- La pipeline ottimizzata ha portato a un miglioramento consistente, raggiungendo Macro F1 = 0.615 e Micro F1 = 0.783 dopo threshold tuning.

Sebbene le performance non siano ancora ottimali, i risultati indicano che l'integrazione di **pre-training di dominio**, **tecniche di bilanciamento** e **data augmentation** rappresenta una direzione promettente per affrontare la classificazione multi-label in scenari caratterizzati da distribuzioni long-tail e dataset limitati.

Link alla repository : GitHub Repository,