
# Acceptability evaluation task for Italian at EVALITA 2020

Esame di Linguistica Computazionale II

**Lucrezia Labardi - 600163**

Professori:

*Simonetta Montemagni*

*Giulia Venturi*

*Felice Dell'Orletta*

A.A. 2022-23

# Indice

# 1 Introduzione

Ogni parlante è in grado di valutare se una frase, espressa nella sua lingua madre, sia accettabile o meno. Questa è una delle competenze fondamentali dei parlanti di una lingua proprio perchè rende possibile formulare un numero illimitato di frasi corrette senza averle mai sentite prima. Addestrare un modello computazionale a svolgere questa tipologia di compito è oggi di enorme importanza, soprattutto in prospettiva dei modelli generativi: essi potranno valutare autonomamente la qualità delle proprie produzioni linguistiche e quindi evitare le frasi con un basso indice di accettabilità.

In questa relazione vengono descritti i risultati ottenuti da diversi sistemi applicati al task di valutazione dell'accettabilità proposto in occasione di AcCompl-it (Acceptability & Complexity evaluation task for Italian at EVALITA 2020) [1]. Come è scritto nel report ufficiale descrittivo dei task (Brunato et al. 2020a), questo tipo di compito è spesso impostato come una classificazione binaria che ha lo scopo di individuare e distinguere le frasi grammaticali da quelle agrammaticali. La novità introdotta dagli organizzatori di EVALITA 2020, in accordo con alcuni importanti esponenti in letteratura, si basa sull'assunto che "i giudizi sull'accettabilità siano gradienti per natura" (Lau, Clark, and Lappin 2017): il task diventa pertanto un compito di regressione per cui ad ogni frase deve essere assegnato un valore all'interno di un range (in questo caso da 1 a 7) che indica quanto essa sia accettabile. Alla base di questo si ha il fatto che all'accettabilità non concorrano soltanto fattori grammaticali ma anche altri non strettamente linguistici.

Nelle successive sezioni inserite nel capitolo 2, vengono illustrate le performance del task di regressione sfruttando diversi livelli di informazione linguistica. Nella sezione 2.1 si fa riferimento alle analisi svolte utilizzando come features soltanto informazioni linguistiche non lessicali ottenute con Profiling-UD. Nella sezione 2.2 si riportano le differenze tra analisi che sfruttano N-grammi di diverso tipo (caratteri, parole, lemmi, part-of-speech). Nella sezione 2.3 si descrivono le analisi che utilizzano i Word Embeddings ottenuti dal dataset itWaC, considerando sia il dataset nella sua interezza che in due diverse sottosezioni formate in base alle PoS. Nella sezione 2.4 si fa riferimento ai risultati ottenuti sfruttando il Neural Language Model "xlm-roberta-base". Infine, il captolo 3 contiene delle considerazioni riassuntive su tutte le analisi svolte.

[1] Sito ufficiale disponibile all'indirizzo: `https://sites.google.com/view/accompl-it/home-page?authuser=0` (visitato il 12/05/2023)

## 1.1 Informazioni preliminari

Prima di procedere con la descrizione delle analisi svolte e dei risultati ottenuti, si ritiene opportuno rendere note alcune informazioni preliminari riguardo ai dati utilizzati e alle scelte effettuate in fase di addestramento.

### 1.1.1 Il dataset

Il dataset utilizzato è quello fornito dagli organizzatori di EVALITA 2020. Esso è composto da 1683 frasi in lingua italiana estratte da quattro diversi dataset di studi psicolinguistici e annotate con un valore medio di accettabilità. Ogni frase è stata annotata da una media di 16 persone sfruttando una piattaforma di crowdsourcing. La lunghezza delle frasi va da un minimo di 4 ad un massimo di 21 parole. Le tipologie di frasi afferenti ai diversi studi sono state scelte in quanto rappresentative di fenomeni linguistici di diversa complessità. Si hanno frasi con determinanti, costruzioni copulari, frasi dichiarative o interrogative con struttura verbale minima, frasi con accordo o disaccordo tra verbo e soggetto per persona e numero, variazioni nella formazione di dipendenze sintattiche rispetto ad aggettivi o pronomi interrogativi ed infine varie frasi create inserendo variazioni, errori o mancanze in alcuni template designati a testare l'accettabilità (Brunato et al. 2020a). Il dataset è stato fornito con la divisione tra training set e test set già effettuata, e la stessa divisione è stata mantenuta durante tutte le analisi: il training set consta di 1339 frasi, mentre il test set di 344.

### 1.1.2 I modelli utilizzati

Per ciascuna delle analisi presentate nelle sezioni da 2.1 a 2.3 è stato utilizzato il modello di regressione lineare LinearSVR[2]. Il tuning degli iperparametri è stato effettuato una sola volta utilizzando *GridSearch* ed i valori ottenuti sono stati mantenuti per ciascun fit del modello. I valori ottenuti sono stati: "C" = 0.1, "epsilon" = 0.5, "loss" = squared_epsilon_insensitive. Gli iperparametri non citati sono stati mantenuti con il loro valore di default, ed il massimo delle iterazioni (max_iter) è stato impostato a 10000.

Per le analisi della sezione 2.4 è stato utilizzato come modello "xlm-roberta-base", importato con HuggingFace[3]. Un modello di questo tipo viene pre-trainato su un compito di masking, quindi allenato a predire una parola token dato il suo contesto.

Come baseline ufficiale del task è stato indicato il punteggio ottenuto somministrando come features al regressore lineare unigrammi e bigrammi di parole. Si è scelto quindi di mantenere questo indice per effettuare i confronti.

[2]Documentazione disponibile all'indirizzo `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html`

[3]Link alla pagina ufficiale: `https://huggingface.co/`

# 2 Report delle analisi

In questa sezione vengono descritte le analisi di valutazione dell'accettabilità. Ad ogni sezione corrisponde un diverso livello di conoscenza linguistica sfruttato per addestrare e testare un modello di regressione lineare oppure un Neural Language Model.

## 2.1 Profiling-UD

In un primo momento sono state sottoposte al regressore delle features linguistiche estratte tramite Profiling-UD. Questo è uno strumento di analisi linguistica basato su Universal Dependencies che permette di estrarre da testi o sequenze di frasi fino a più di 130 features a diversi livelli di descrizione linguistica ma non lessicale (Brunato et al. 2020b). L'input della risorsa è stato costituito da una cartella di file di testo, ciascuno dei quali corrispondeva ad una frase: il nome del file era l'id associato alla frase ed il contenuto era la frase stessa. In output sono state ottenute 115 features per ciascuna frase, quindi 115 diverse informazioni sulle quali si è basato lo studio di regressione. Le frasi del training e del test set sono state analizzate insieme e successivamente separate. Entrambi i dataset di features sono stati poi normalizzati con *MinMaxScaler*. Per rendere più accurati i risultati, è stata effettuata una 5-Fold Cross-Validation sul training set ed è poi stato estratto il valore di $R^2$-Score su tutti gli esempi testati. Il valore ottenuto è stato 0.36. Successivamente sono state effettuate le predizioni sul test set ufficiale e calcolate diverse metriche di valutazione: i coefficienti di correlazione di Spearman e Pearson, gli errori e di nuovo $R^2$-Score. I risultati sono riportati in tabella 2.1.

Table 2.1: Confronto sul test set tra Profiling-UD e la baseline ufficiale

| | **Profiling-UD** | **Baseline** |
|---|---|---|
| *Corr. di Spearman* | **0.57** (p.value 2.67e-31) | 0.56 (p.value 3.94e-30) |
| *Corr. di Pearson* | **0.56** (p.value 6.74e-30) | **0.56** (p.value 4.34e-30) |
| *MAE* | 1.07 | 1.10 |
| *MSE* | 2.11 | 1.97 |
| *RMSE* | 1.45 | 1.40 |
| *$R^2$-Score* | 0.26 | **0.30** |

I valori di correlazione ottenuti sono piuttosto buoni, nel caso della correlazione di Spearman anche leggermente superiori alla baseline. Il valore di $R^2$-Score ottenuto è 0.26, leggermente inferiore alla baseline ma non sufficientemente buono da rendere affidabile il regressore per il task di accettabilità.

## 2.2 N-grammi

Come output della risorsa Profiling-UD (vedi sezione 2.1) è stato ottenuto un file in formato CoNLL-U dal quale è stato possibile ricavare per ciascuna frase i token che la costituivano ed i corrispondenti lemmi, Part-of-Speech (PoS) e caratteri per ogni parola. I token sono stati normalizzati tramite diverse funzioni per evitare che url, parole troppo lunghe o lettere maiuscole potessero influenzare erroneamente le successive analisi. Per questo task di regressione si è scelto di analizzare gli n-grammi a livello di caratteri, di parole token, di part-of-speech e di lemmi, in numero incrementale da 2 a 4. Questo significa che tutti gli n-grammi contengono anche le informazioni degli unigrammi e degli n-grammi di numero minore o uguale al proprio indice. Le features del training set sono state normalizzate rispetto alla tipologia di n-grammi considerati: il numero di caratteri, di parole token, di POS e di lemmi della frase analizzata. Dato l'alto numero di features estratte, per gli n-grammi di parole token e di lemmi si è scelto di scartare tutte quelle features con un indice di occorrenza inferiore a 2, mentre per gli n-grammi di caratteri tutte quelle inferiori a 3. Le features delle PoS sono state lasciate nella loro dimensione originale. Il numero delle features finali viene mostrato in tabella 2.2.

Table 2.2: Numero di features per gli n-grammi dopo il filtraggio

| | **2-grammi** | **3-grammi** | **4-grammi** |
|---|---|---|---|
| **Caratteri** | 378 | 2062 | 6159 |
| **Parole** | 3006 | 5083 | 6916 |
| **PoS** | 109 | 372 | 859 |
| **Lemmi** | 2724 | 4824 | 6747 |

Per ciascuna lunghezza degli n-grammi e per ciascun tipo di informazione contenuta, essi sono stati testati sul training set con una 5-Fold Cross-Validation. La valutazione è stata effettuata confrontando il punteggio di $R^2$-Score ottenuto su tutti gli esempi testati. In tabella 2.3 vengono mostrati i risultati per tutti gli n-grammi di caratteri. In questo caso il miglior risultato è stato ottenuto con i bigrammi, con un $R^2$-Score = 0.30.

Table 2.3: Risultati degli n-grammi di caratteri con la 5-Fold Cross-Validation

| | **2-grammi** | **3-grammi** | **4-grammi** |
|---|---|---|---|
| *$R^2$-Score sulla CV* | **0.30** | 0.27 | 0.22 |

Anche per gli n-grammi di parole il miglior risultato si è ottenuto con i bigrammi ($R^2$-Score = 0.26). Tutti i punteggi sono presentati in tabella tabella 2.4.

Per quanto riguarda gli n-grammi di Part-of-Speech, come riportato in tabella 2.5, il miglior risultato è stato ottenuto con i trigrammi, con un $R^2$-Score = 0.31.

Table 2.4: Risultati degli n-grammi di parole con la 5-Fold Cross-Validation

| | **2-grammi** | **3-grammi** | **4-grammi** |
|---|---|---|---|
| *$R^2$-Score sulla CV* | **0.26** | 0.21 | 0.17 |

Table 2.5: Risultati degli n-grammi di PoS con la 5-Fold Cross-Validation

| | **2-grammi** | **3-grammi** | **4-grammi** |
|---|---|---|---|
| *$R^2$-Score sulla CV* | 0.30 | **0.31** | 0.30 |

Infine, i risultati ottenuti con gli n-grammi di lemmi sono riportati in tabella 2.6. Il miglior punteggio è stato ottenuto anche in questo caso dai bigrammi, con un $R^2$-Score = 0.20.

Table 2.6: Risultati degli n-grammi di lemmi con la 5-Fold Cross-Validation

| | **2-grammi** | **3-grammi** | **4-grammi** |
|---|---|---|---|
| *$R^2$-Score sulla CV* | **0.20** | 0.16 | 0.13 |

Valutando i risultati ottenuti con tutte le tipologie e con tutti i numeri di n-grammi, si osserva che il miglior risultato in assoluto è stato ottenuto dai trigrammi di PoS. E' interessante notare come i punteggi degli n-grammi di lemmi siano risultati inferiori rispetto agli altri e come quelli degli n-grammi di caratteri e di parole abbiano punteggi molto diversi tra bigrammi e 4-grammi. Gli n-grammi di PoS, che hanno un numero di features decisamente inferiore rispetto a tutti gli altri n-grammi, ottengono risultati stabili. In seguito si è proceduto ad effettuare la valutazione sul test set ufficiale con il sistema dei trigrammi di Part-of-speech, proprio in quanto è risultato essere il miglior sistema di questa sezione (tabella 2.7).

Table 2.7: Confronto sul test set tra trigrammi di PoS e la baseline ufficiale

| | **Trigrammi PoS** | **Baseline** |
|---|---|---|
| *Corr. Spearman* | 0.54 (p.value 1.56e-27) | **0.56** (p.value 3.86e-30) |
| *Corr. Pearson* | 0.55 (p.value 7.21e-29) | **0.56** (p.value 4.34e-30) |
| *MAE* | 1.15 | 1.10 |
| *MSE* | 2.00 | 1.97 |
| *RMSE* | 1.41 | 1.40 |
| *$R^2$-Score* | 0.29 | **0.30** |

Rispetto ai risultati ottenuti con Profiling-UD, la performance del regressore che sfrutta trigrammi di PoS sul dataset di test è leggermente superiore nel valore di $R^2$-Score ed inferiore nei coefficienti di correlazione. La baseline, che si ricorda essere ottenuta con features di unigrammi e bigrammi di

parole token, ha valori migliori per entrambe le tipologie di misurazione, ma soltanto con 0.02 punti di differenza.

## 2.3 Word Embeddings

Per questo livello di analisi sono stati utilizzati i word embeddings. Essi costituiscono un metodo di rappresentazione delle caratteristiche semantiche e sintattiche delle parole tramite vettori numerici. I word embeddings utilizzati sono stati ottenuti dal dataset ItWaC e generati con word2vec. Essi sono stati resi disponibili da ItaliaNLP[1] in occasione della partecipazione a EVALITA 2018 (Cimino, Mattei, and Dell'Orletta 2018). Ad ogni parola sono associati 128 embeddings, quindi 128 informazioni che costituiscono la sua rappresentazione. In questa fase sono state valutate le performance tramite 5-Fold Cross-Validation sia sfruttando gli embeddings di tutte le parole, sia sfruttando soltanto quelli associati a diverse part-of-speech. Per ciascuna frase, infatti, sono state separate le parole lessicali, quindi quelle a cui era associata una Part-of-Speech tra nome, verbo, aggettivo o avverbio (NOUN, PROPN, VERB, ADJ, ADV), da quelle funzionali, a cui era associata una qualsiasi altra PoS. Al momento di svolgere le analisi su una delle due parti, gli embeddings delle altre parole venivano impostati come vettori di 0 e considerati allo stesso modo delle parole di cui non si avevano embeddings a disposizione. In tabella 2.8 vengono riportati i risultati di $R^2$-Score su tutti gli esempi testati in seguito alla 5-Fold Cross-Validation.

Table 2.8: Risultati degli word embeddings con la 5-Fold Cross-Validation

| | **Tutte le parole** | **Solo parole funzionali** | **Solo parole lessicali** |
|---|---|---|---|
| *$R^2$-Score CV* | 0.29 | **0.32** | 0.18 |

Il miglior risultato si è ottenuto sfruttando gli embeddings delle sole parole funzionali ($R^2$-Score = 0.32), e si è anche del miglior risultato ottenuto fino ad ora con la 5-Fold Cross Validation sul training set. Risulta evidente come il risultato ottenuto con gli embeddings di tutte le parole si discosti poco da quello delle sole parole funzionali. Al contrario, il risultato ottenuto con le sole parole lessicali è molto distante. Per analizzare questo fenomeno dobbiamo considerare che il dataset completo conta 15051 parole, quello di sole parole funzionali 7585 e quello di sole parole lessicali 7466. Vista la grandezza molto simile dei due dataset partizionati, la spiegazione ai risultati ottenuti può essere attribuita al fatto che l'accettabilità di una frase è maggiormente collegata alle parole funzionali che a quelle lessicali. In particolare, è probabile che le informazioni riguardanti l'accettabilità grammaticale e sintattica, contenute soprattutto nelle parole funzionali, siano più rilevanti di quelle che riguardano la semantica, contenute soprattutto nelle parole lessali. Inoltre, tra le Part-of-Speech delle parole funzionali rientra

[1]Link alla pagina ufficiale contenente la risorsa: `http://www.italianlp.it/resources/italian-word-embeddings/`

anche "X", cioè la PoS assegnata alle parole di cui non si riesce ad individuare una categoria. Dal momento che tra le parole del dataset sono state inserite anche parole troncate, quindi non formalmente corrette, è possibile che Profiling-UD non sia stato in grado di riconoscerle con la loro Part-of-Speech originaria. Questo potrebbe aver sbilanciato la suddivisione a discapito della partizione con sole parole lessicali.

La validazione sul test set ufficiale è stata effettuata sfruttando gli embeddings del dataset formato da parole solo funzionali in quanto aveva ottenuto i migliori risultati in fase di training. I risultati sono mostrati in tabella 2.9.

Table 2.9: Risultati sul test set tra Word Embeddings e la baseline ufficiale

| | **Word Embeddings** | **Baseline** |
|---|---|---|
| *Corr. di Spearman* | 0.53 (p.value 2.46e-27) | **0.56** (p.value 3.86e-30) |
| *Corr. di Pearson* | 0.55 (p.value 2.58e-29) | **0.56** (p.value 4.34e-30) |
| *MAE* | 1.19 | 1.10 |
| *MSE* | 2.06 | 1.97 |
| *RMSE* | 1.43 | 1.40 |
| $R^2$*-Score* | 0.27 | **0.30** |

Come si può vedere, l'$R^2$-Score ottenuto è pari a 0.27, anche in questo caso leggermente inferiore alla baseline. Gli indici di correlazione sono poco sopra il 50% e gli errori continuano ad essere piuttosto alti. La baseline rimane migliore del regressore lineare impostato anche a livello di word embeddings.

## 2.4 Neural Language Model

Il Neural Language Model utilizzato per questa fase delle analisi è stato il Tranformer Decoder Model "xlm-roberta-base"[2] (Conneau et al. 2019). Tale modello è una versione multilingue di RoBERTa, pre-trainato su 2.5TB di dati estratti da CommonCrawl contenenti 100 lingue diverse. A partire dal modello pre-trainato, è stato effettuato un fine-tuning per 5 epoche per il task di regressione preso in esame, dopo aver estratto dal dataset di training una parte pari al 10% per la validazione. Gli argomenti del trainer sono stati: learning_rate = 2e-5 e weight_decay = 0.01. Sono state effettuate delle prove diminuendo il valore del learning rate ma si è ritenuto superfluo riportare nella presente relazione i valori registrati in quanto la diminuzione della loss sia del training che del test risultava troppo lenta nelle prime 5 epoche. Per ciascuna epoca vengono riportati in tabella 2.10 i valori della loss sul dataset di training e di validation ed il valore del coefficiente di correlazione di Spearman. In figura **??** viene mostrato l'andamento delle curve di loss di training e validation set.

[2]Link alla pagina ufficiale su HuggingFace: `https://huggingface.co/xlm-roberta-base`

Table 2.10: Andamento della loss di training ed validation set per le prime 5 epoche di fine-tuning

| Epoca | Training Loss | Validation Loss | Corr. di Spearman |
|---|---|---|---|
| 1 | 5.163600 | 2.605917 | 0.539706 |
| 2 | 2.595800 | 1.706379 | 0.666639 |
| 3 | 2.020200 | 1.672983 | 0.759488 |
| 4 | 1.432400 | 0.939927 | 0.848005 |
| 5 | 0.887100 | 1.009601 | 0.864181 |

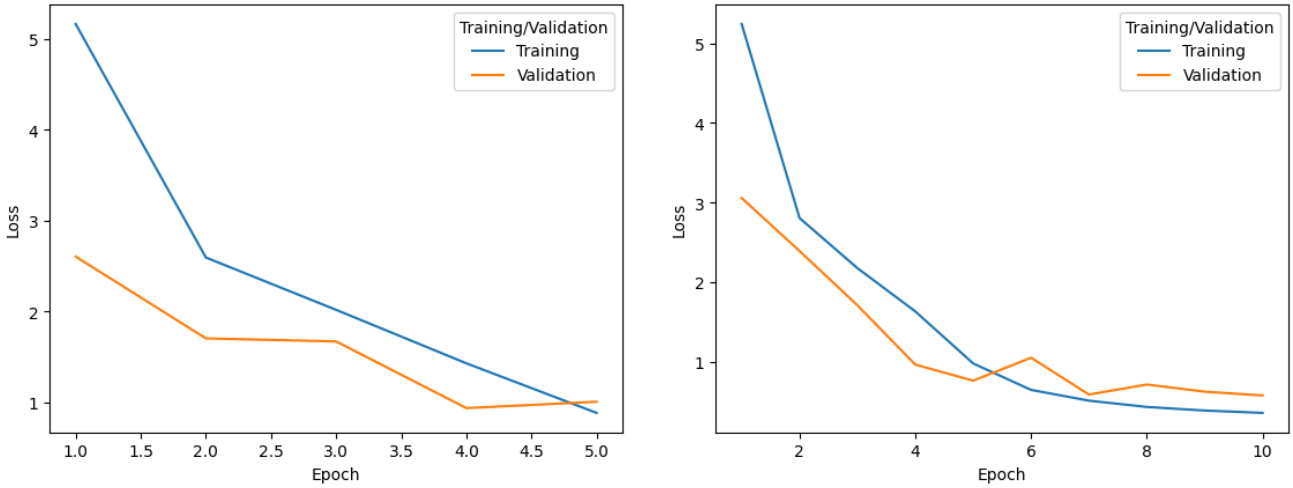


Figure 2.1: Curva di Loss di training ed validation set per 5 epoche (a sinistra) e per 10 epoche (a destra)

In figura 2.1, nella parte sinistra, viene mostrato come il valore della loss del training set rimanga a lungo superiore rispetto a quella del validation set. Circa in corrispondenza della quinta epoca però le due linee si incrociano e la loss del validation set inizia a risalire. Basandoci soltanto sulle prime 5 epoche di addestramento, si potrebbe considerare questo punto come un inizio di overfitting. Visto il basso numero di epoche ed i valori di loss ancora molto vicini a 1, si è ritenuto più probabile che si fosse in presenza di un minimo locale, e che quindi con un numero superiore di epoche si sarebbe potuto ottenere un risultato migliore in termini di loss e soprattutto più accurato riguardo all'andamento del fine-tuning. La validità di questa ipotesi era avvalorata dal fatto che la correlazione di Spearman, arrivati alla quinta epoca, non avesse smesso di aumentare. Per questo motivo è stato deciso di indagare ulteriormente procedendo con altre 5 epoche di addestramento, per arrivare dunque ad un totale di 10. Come possiamo vedere nella parte destra della figura 2.1, in effetti la curva di loss del validation set sale rispetto a quella del training, ma allo stesso tempo continua a scendere rispetto all'asse delle y. Questo non è un chiaro segnale di overfitting e potrebbero servire altre epoche di addestramento per registrare un risultato più evidente. Si riportano i risultati ottenuti dalla validazione sul test set ufficiale in tabella 2.11.

Table 2.11: Confronto sul test set tra xlm-roberta-base e la baseline ufficiale

| | **xlm-roberta-base** | **Baseline** |
|---|---|---|
| *Corr. di Spearman* | **0.79** (p.value 1.01e-76) | 0.56 (p.value 4.34e-30) |
| *MAE* | 0.76 | 1.10 |
| *MSE* | 1.08 | 1.97 |
| $R^2$*-Score* | **0.62** | 0.30 |

Come è visibile, questo modello ottiene risultati nettamente migliori rispetto agli altri, con un $R^2$-Score pari a 0.62 sul test set. La correlazione di Spearman calcolata tra le predizioni sul test set ed i valori golden è decisamente alta (79%) e gli errori, anche se non del tutto soddisfacenti, sono nettamente più bassi rispetto alla baseline e agli altri regressori testati in precedenza.

# 3 Conclusioni

In questa relazione sono stati mostrati i risultati delle analisi di un task di regressione con l'obiettivo di assegnare ad una frase un punteggio da 1 a 7 relativo alla sua accettabilità. Per tutta la prima fase si è utilizzato un regressore lineare (LinearSVR), a cui inizialmente sono state somministrate come features le informazioni ricavate da Profiling-UD, ottenendo risultati non del tutto soddisfacenti. In seguito si sono sfruttate diverse tipologie di n-grammi di cui i migliori si sono rivelati essere i trigrammi di PoS. I risultati ottenuti tramite word embeddings sono stati simili. Considerando come baseline un regressore addestrato con unigrammi e bigrammi di parole, che ha raggiunto un $R^2$-Score pari a 0.30 ed una correlazione di Spearman pari a 0.56, nessuno dei regressori testati è stato in grado di migliorare questi risultati. Nella seconda parte è stato sfruttato un modello neurale del linguaggio pre-trainato a cui è stata aggiunta una fase di fine-tuning di 5 epoche. I risultati ottenuti sono stati migliori rispetti alla baseline, con un $R^2$-Score pari a 0.62 ed una correlazione di Spearman pari a 0.79. Certamente nel caso del NLM il miglioramento è dovuto al fatto che il modello era già stato pre-addestrato su un'enorme quantità di testi, ma sarebbe interessante indagare quanta informazione effettivamente provenga dalla fase di pre-training e quanta sia dovuta al fine-tuning (effettuato, tra l'altro, su poco più di mille frasi). Questo esame potrebbe essere svolto somministrando ad un regressore lineare gli embeddings relativi a ciascuna frase che il modello neurale permette di estrarre in corrispondenza di ciascun layer. Un interessante sviluppo potrebbe inoltre comprendere la valutazione di tutti i sistemi descritti, in particolare quello basato su n-grammi, prevedendo la distinzione tra le frasi provenienti dai quattro diversi dataset di origine. Dal momento che ciascuno di loro rappresenta un costrutto linguistico specifico, sarebbe interessante osservare se nella loro valutazione di accettabilità le diverse tipologie d'informazione prendono parte differentemente.

# Bibliografia


Brunato, D., C. Chesi, F. Dell'Orletta, S. Montemagni, G. Venturi, and R. Zamparelli (2020a). *AcCompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian*. URL: https://ceur-ws.org/Vol-2765/paper163.pdf.

Brunato, D., A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni (2020b). "Profiling-UD: a Tool for Linguistic Profiling of Texts". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 7145–7151. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.883.

Cimino, A., L. D. Mattei, and F. Dell'Orletta (2018). "Multi-task Learning in Deep Neural Networks at EVALITA 2018". In: *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop*. Torino, Italy: Accademia University Press. DOI: https://doi.org/10.4000/books.aaccademia.4527.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

Lau, J. H., A. Clark, and S. Lappin (2017). "Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge". In: *Cognitive Science* 41.5, pp. 1202–1241. DOI: https://doi.org/10.1111/cogs.12414.