UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

Similarità semantica tra frasi: un'indagine comparativa tra giudizi umani, spiegazioni associate e metriche automatiche

Relatori:
Felice Dell'Orletta
Chiara Fazzone
Giulia Venturi

Candidato:
Mihnea Sever Molnar

ANNO ACCADEMICO 2022/2023

Indice

Elenco delle figure

Introduzione

I training set un tempo erano formati solo da coppie evento-categoria, classe o numero di giudizio. Oggi invece, grazie ai nuovi sistemi che si basano su reti neurali, è possibile prendere dati in input non solo con uno specifico valore numerico ma è possibile anche inserire il motivo per il quale lo studioso ha scelto di dare quel determinato valore a quel preciso dato. All'interno di questo scenario si colloca il presente lavoro, nato dall'esperienza di tirocinio svolto presso l'Istituto di Linguistica Computazionale "A.Zampolli" del CNR di Pisa, con lo scopo di svolgere un'indagine semantica sulla similarità di coppie frasali per come viene percepita sia da un punto di vista umano che da quello relativo a metriche automatiche.

Nello specifico il presente lavoro è stato svolto in diverse fasi di cui la prima, descritta nel Capitolo 1, riguarda la creazione e l'annotazione di un dataset.
All'interno di questo dataset sono contenute coppie frasali ricavate da romanzi di narrativa e tradotte automaticamente in italiano sulle quali il sottoscritto (Mihnea Sever Molnar) e la collega Irene Poli hanno dato i loro giudizi di similarità basandosi su una scala Likert [13] e fornendo anche delle spiegazioni sul motivo per il quale tali frasi, costituenti la coppia, sono simili tra di loro.
É molto importante precisare che oltre al nostro dataset, per l'italiano e con le spiegazioni ne esiste solamente un altro, cioè, "The Italian e-RTE-3 Dataset" [24]

Il passo successivo è stato quello di eseguire BLEU (Bilingual Evaluation Understudy)[7][11], ovvero, un algoritmo il cui funzionamento si basa sugli n-grammi e che nasce per valutare la qualità delle traduzioni automatiche e l'algoritmo SBERT (Sentence-BERT) [14][15], estensione di BERT [6], che rispetto a BLEU è più moderno, si basa su reti neurali profonde e coinvolge modelli di linguaggio avanzato per creare delle rappresentazioni vettoriali che codificano l'informazione semantica necessaria a svolgere il task richiesto.

Sono state date in input a tali algoritmi le coppie di frasi e le spiegazioni sia di Molnar che di Poli per ottenere i punteggi di similarità delle metriche automatiche. Di questo parleremo nel Capitolo 3.

A questo punto sono stati svolti diversi calcoli statistici utilizzando metriche come l'α di Krippendorff [21], il coefficiente di correlazione di Spearman [22] ed il p-value [23][8] per rispondere a diverse domande:

- Esiste un accordo tra gli annotatori?
- Quanto correlano i risultati ottenuti con BLEU e SBERT sulle coppie frasali e sulle spiegazioni testuali?
- Quanto correlano i giudizi sulla similarità semantica tra Molnar e Poli?
- Quanto correlano i giudizi di similarità attribuiti dagli annotatori e le spiegazioni testuali sulle stesse coppie di frasi?

Tutto questo lo vedremo nel Capitolo 2.

Per concludere, il Capitolo 4 è dedicato all'analisi stilistica del testo, più precisamente è dedicato al tentativo di comprendere se caratteristiche testuali simili corrispondono a una percezione della similarità simile. Per rispondere a tale domanda abbiamo utilizzato la piattaforma web Profiling-UD [1][10] sviluppata da Italian NLP Lab di Pisa e, ancora una volta, le metriche statistiche p-value e coefficiente di correlazione di Spearman indagando sui dati relativi alle features linguistiche ottenute.

1. Costruzione del Dataset

Nel presente capitolo sarà possibile comprendere il percorso ed il processo che sta alla base della creazione del dataset utilizzato per svolgere il lavoro descritto nella presente tesi di laurea.
Tutto prende origine da una tesi precedente di cui parleremo nel paragrafo 1.1, risalente all'anno 2023, dalla quale sono stati raccolti dei dati che successivamente sono stati annotati, elaborati ed utilizzati come base per le successive analisi. Tale lavoro di annotazione è stato svolto dal sottoscritto (Mihnea Sever Molnar) e dalla collega Irene Poli.

1.1 Tesi di Partenza

La costruzione del dataset presente all'interno di questo lavoro pone le proprie basi nella tesi di laurea magistrale in Informatica Umanistica del dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa dell'anno 2023 di Tommaso Pelagatti con relatrice Giulia Venturi e correlatrice Chiara Alzetta.
Tale tesi affronta il tema della percezione dei parlanti della similarità semantica tra frasi, e la sensibilità di modelli computazionali di valutazione della similarità rispetto a tratti linguistici rintracciabili nelle frasi.
Per similarità semantica si intende la somiglianza tra entità linguistiche tenendo conto del significato di ciascuna di esse.
Questo argomento nell'ambito del Natural Language Processing (NLP) si affronta con sistemi di apprendimento automatico che permettono di sfruttare le informazioni presenti nei diversi livelli linguistici.
Per poter realizzare lo studio descritto nella tesi di partenza è stato necessario costruire un dataset annotato con giudizi umani utilizzando come base un corpus di frasi che è stato fornito dall'Istituto di Linguistica Computazionale "A.Zampolli" del CNR [1].
Tale corpus è costituito da coppie di frasi estratte da romanzi di narrativa e tradotte auto-

[1] Consiglio Nazionale delle Ricerche

maticamente in italiano.
Da qui sono poi state estratte le coppie di frasi che condividevano almeno una parola piena ovvero un nome, un verbo oppure un aggettivo. Infine, con lo scopo di ridurre il numero di frasi totali da presentare agli annotatori e quindi ridurre il numero di frasi da sottoporre al giudizio umano, è stato utilizzato l'algoritmo SBERT, selezionando solamente le coppie con un punteggio di similarità di almeno 0.65 in un range compreso tra 0 ed 1. Così facendo, è stato possibile ottenere 2144 coppie di frasi.
L'annotazione di queste coppie frasali è stata svolta in modalità crowdsourcing utilizzando la piattaforma Prolific [12].

1.1.1 Prolific

Per poter raccogliere ed avere quindi a disposizione un numero adeguato di giudizi umani è stata utilizzata la piattaforma Prolific.
Tale piattaforma viene solitamente adoperata nel campo della ricerca poiché permette ai ricercatori di entrare in contatto con un elevato numero di collaboratori attivi e controllati i quali possiedono le specifiche caratteristiche utili e necessarie alla ricerca stessa; tutto questo sotto compenso.
Il vantaggio principale di Prolific è che tutto avviene online permettendo quindi di ottenere una grande quantità di dati utili in poco tempo. Tale piattaforma è stata utilizzata nella creazione del dataset della tesi di partenza per entrare in contatto con utenti adeguati ed ottenere i giudizi numerici di similarità semantica.
Durante la fase di trial, dopo aver sottoposto 100 coppie di frasi agli annotatori, è stato deciso di avvalersi della scala Likert con un range di valori tra 1 e 5 per esprimere i giudizi di similarità tra le coppie frasali. Nel momento dell'annotazione dell'intero corpus invece è stato deciso di suddividere le circa 2000 coppie di frasi in 67 questionari contenenti ciascuno 32 coppie. Ogni questionario contiene al suo interno due coppie di frasi di controllo, queste si presentano uguali in ogni questionario, di conseguenza dovrebbero ottenere sempre le stesse valutazioni.

1.1.2 Scala Likert

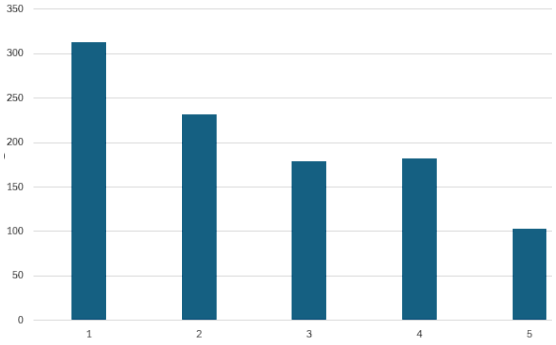

Figura 1.1: Distribuzione giudizi annotatore A di Prolific

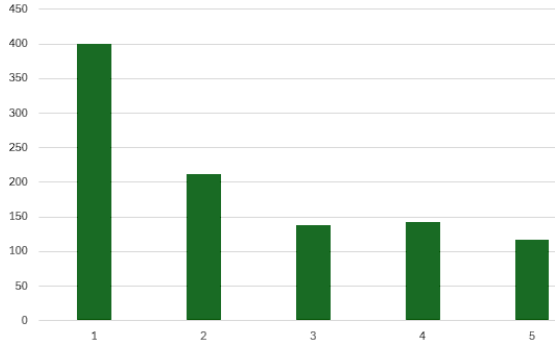

Figura 1.2: Distribuzione giudizi annotatore B di Prolific

Una scala Likert è un metodo di misurazione che spesso viene utilizzato in ricerca per valutare atteggiamenti, opinioni o percezioni.
Il vantaggio di questo metodo è la possibilità di usare domande applicabili in tanti ambiti diversi per diverse tipologie di ricerca.
Per ogni domanda che viene posta all'intervistato, questo ha a disposizione una serie di risposte standard fra cui scegliere.
Generalmente queste risposte possono essere del tipo "Pienamente d'accordo" che corrisponde al valore massimo oppure "Per niente d'accordo" che corrisponde al valore minimo.
Nel caso degli annotatori di Prolific che hanno contribuito all'annotazione del corpus di partenza, essi hanno usufruito di una scala Likert organizzata nel seguente modo: ad ogni numero compreso nel range 1-5 corrisponde un determinato giudizio semantico da associare a ciascuna coppia di frasi presente nel rispettivo questionario al quale l'annotatore viene sottoposto.

- 1 - "Completamente diverse"
- 2 - "Poco simili"
- 3 - "Abbastanza simili"

- 4 - ”Molto simili”
- 5 - ”Pressoché uguali”

Infine i risultati ottenuti da tali questionari sono stati inseriti all’interno di un file Excel dove due colonne sono dedicate alle frasi da valutare mentre le altre contengono i giudizi numerici dei vari annotatori per ogni coppia frasale. Nei grafici 1.1 e 1.2 è mostrata la distribuzione dei giudizi di alcuni degli annotatori di Prolific mentre nel grafico 1.3 è mostrata la distribuzione dei giudizi del sottoscritto.

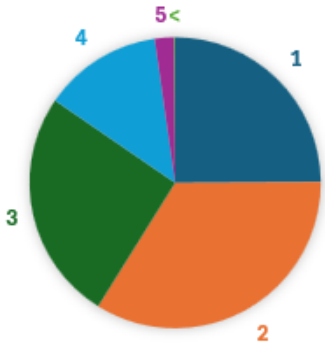

Figura 1.3: Distribuzione dei giudizi Molnar

1.2 Il Dataset

Il lavoro descritto all’interno di questa tesi ha origine da un corpus contenente coppie di frasi tradotte automaticamente appartenenti alla letteratura narrativa.
Tale dataset è stato fornito dall’Istituto di Linguistica Computazionale ”A.Zampolli” del CNR.
L’ulteriore lavoro di annotazione descritto nella sezione a seguire è stato svolto insieme alla studentessa Irene Poli.

Num_questionario	Num_domanda	Frase_1	Frase_2
1	1	— Sì, ma a che è diretta la sua attività?	— Dunque tu dividi la tua sostanza fra le tue figlie senza riservarti nulla?
1	2	— domandò la signora Valentina, sempre con la medesima soavità nello sguar	— domandò Valentina Michailowna sempre con la stessa soavità di voce.
1	3	Sotto la finestra, sulla sabbia del viale e sull'erba del prato si profilava l'ombra	Malgrado il sole vivido che passava attraverso le foglie lavate, l'aria era fredda.
1	4	Katavasov invece, che fra le sue occupazioni scientifiche non aveva avuto occa	Di nuovo i volontari salutarono e sporsero le teste, ma Sergej Ivanovic non prestò loro attenzione
1	5	La prima cosa da fare era procurarsi almeno un po' di soldi a prestito.	«Non ho a chi chiedere denaro in prestito!».
1	6	Ma in quel punto entrò in salotto un uomo di mezza statura, in costume inglese	Era in giubba turchina, calze di seta, scarpini affibbiati, profumato e impomatato.
1	7	Gli davano la preminenza sugli altri la sua età rispettabile, l'esperienza, l'abilit	A questo riguardo egli si considerava piuttosto danneggiato e maltrattato nel lavoro, ed era semp
1	8	- Non è morto però, la pistola non ha sparato.	Stavrògin era in piedi con la pistola in mano, rivolta verso il basso, e aspettava senza batter cigli
1	9	Ivan non si lascia irretire nemmeno da migliaia di rubli.	E le migliaia di rubli che passano per le sue mani!
1	10	Di botto, si udì uno scoppio di voce sgradevole, e l'ufficiale, pallido e con le lab	—proruppe l'ufficiale con furia da ubbriaco.
1	11	Un cameriere tolse il coperchio alla zuppiera; tutti avvicinarono meglio le sedie	«Chi ha chiesto della zuppa?» disse, entrando nella stanza, con una zuppiera di minestra di cavo
1	12	E fu preso da una rassegnazione che gli sembrò quasi una generosa galanteria	Aveva impressa in viso una devota rassegnazione alla volontà di Dio.
1	13	Dall'accettazione di questo dogma deriva, come necessaria conseguenza, la ve	– Ci credi dunque ai dogmi della chiesa?
1	14	Sì, ma prima di dirvi il nome di questo vantaggio mi voglio compromettere in pr	Il proprio volere, libero, personale e autonomo, il proprio capriccio personale, foss'anche il più s
1	15	— Prendi.... Non sei in collera con me, non è vero?...	Sentite un po', signor mio, com'è che vi siete messo a chiocciare così, adesso, eh?
1	16	Albeggiava; la pioggia era cessata, le nuvole si dileguavano.	Il cielo era limpido, meno una nuvola verso oriente.
1	17	La neve non si depositava sul suolo, il vento la faceva mulinare e ben presto si	Sul terreno ghiacciato durante la notte era caduta un po' di neve secca e il vento "secco e taglient
1	18	Gli avvenimenti della giornata erano stati troppo interessanti per Elizabeth perc	Elizabeth però non sapeva conservare a lungo il broncio, e benché tutte le sue prospettive di felic
1	19	Preferisco rimanere con le mie sofferenze non vendicate e nella mia indignazio	In me non ci potranno mai essere indignazione e vergogna; quindi, neanche disperazione.
1	20	– rispose con calma il sergente.	— Sarà, disse il sergente, – che mangiano da signori.

Figura 1.4: Corpus di partenza

1.2.1 La nostra annotazione

Giudizio Molnar	**Giudizio Poli**	**Spiegazione Molnar**	**Spiegazione Poli**
2	3	Entrambe parlano di proiettili che cadono	Scene di guerra in cui volano in alto proiettili
3	2	In entrambe le frasi c'è una persona che dice qualcosa alzandosi dal divano	In entrambe qualcuno si alza dal divano
2	2	Entrambe parlano di una ambulanza	In entrambe si parla di spostarsi verso un'ambulanza
4	2	Entrambe descrivono l'abbigliamento di una persona	Entrambe descrizioni di abbigliamento diversi

Tabella 1.1: Tabella contenente esempi di giudizi e spiegazioni di Molnar e Poli

L'operazione che è stata svolta nella fase iniziale e che dunque rappresenta i primi passi nel compimento del lavoro descritto nella presente tesi è stata quella di costruire il dataset per poi analizzarne i risultati con le diverse metriche statistiche e stilistiche.
Come prima cosa è stato preso un ridotto numero di questionari, circa cinque, contenenti le rispettive trentadue coppie di frasi ciascuno. Partendo dunque dal primo questionario sono state annotate separatamente sia dalla collega che dal sottoscritto le diverse coppie e

sono stati utilizzati, per i giudizi, i medesimi valori della scala Likert utilizzati anche dagli annotatori di Prolific con lo scopo quindi di valutare quanto le frasi sono simili semanticamente tra di loro.
Tuttavia è necessario fare una importante precisazione:
sebbene nella fase di annotazione sono stati utilizzati i medesimi valori numerici degli annotatori di Prolific, per noi è stato necessario fornire anche una spiegazione sul motivo per il quale quella determinata coppia frasale ha ricevuto quella precisa valutazione. Nella Tabella 1.1 è possibile visualizzare qualche esempio di annotazione sia dei giudizi che delle descrizioni ed è possibile osservare alcune divergenze sul piano semantico tra i due annotatori. Per altri esempi confrontare l'Appendice, Tabella 5.1.
Nella Tabella 1.2, invece, è possibile osservare il processo di annotazione dal punto di vista del sottoscritto:
per prima cosa vengono lette le coppie di frasi, viene poi dato un giudizio numerico rispettando i valori della scala Likert prestabiliti ed infine viene data una spiegazione su cosa accomuna semanticamente "Frase 1" e "Frase 2".

1.2.2 Tempistiche di annotazione

Sempre durante la prima fase di questo lavoro è stato necessario anche misurare il tempo di annotazione.
Per i primi sei questionari, ovvero per le prime 198 coppie frasali, il tempo di annotazione è stato chiaramente maggiore rispetto ai successivi. Infatti, il tempo medio si è aggirato tra i trenta ed i quaranta minuti per questionario.
Successivamente, a partire dal questionario sette in poi la media personale è scesa gradualmente arrivando a circa 24 minuti a questionario. Compiere tale procedimento è stato necessario al fine di verificare che il lavoro di entrambi sia stato svolto in maniera giusta dedicando un tempo adeguato per ogni coppia.

Frase 1	**Frase 2**	**Giudizio Molnar**	**Spiegazione Molnar**
Ehi, Fetin'ja, porta un piumino, dei cuscini e un lenzuolo.	Va bene, signora! disse Fetin'ja, stendendo il lenzuolo sopra il piumino e sistemando i cuscini.	4	Le frasi sono successive, medesimo argomento e personaggi
Io la guardai con stupore e con ripugnanza.	L'ho trattata con odio e ripugnanza, poi mi sono ricordato che ho commesso tante volte, almeno col pensiero, il peccato che me la rendeva odiosa; e ad un tratto, e nello stesso tempo, mi sono disprezzato, e l'ho compatita, e mi sono sentito meglio.	3	In entrambe le frasi si descrive un sentimento di odio
Ma da allora, credo che tutti e due abbiamo fatto progressi».	Anche qui io agivo in nome del progresso, ma ormai mi rapportavo criticamente al progresso stesso.	2	In entrambe le frasi si parla di progresso
– suonò una voce nella folla.	«Parlino prima gli anziani!», si gridò nella folla.	2	In entrambe le frasi c'è una voce nella folla
Non c'era nel suo sguardo il minimo turbamento, forse vi traspariva soltanto una certa meraviglia, e anche quella sembrava riferirsi unicamente al principe.	Il principe n'era spaventato e, nel suo turbamento, non sapeva che decidere.	2	In entrambe le frasi c'è la figura del principe

Tabella 1.2: Tabella contenente coppie frasali, giudizi e spiegazioni di Molnar

1.3 Clustering

Con il termine clustering[16] vogliamo rappresentare ed identificare quella tecnica di analisi che viene prevalentemente utilizzata nell'ambito della data analysis e nel mondo dell'apprendimento automatico. Lo scopo di tale tecnica è quello di costruire un gruppo di dati omogenei (un "cluster") ovvero un insieme di oggetti simili tra di loro in base a determinate caratteristiche oppure precisi criteri.
Attualmente esistono diversi metodi e tecniche di clustering come per esempio lo spectral, l'agglomerativo oppure quello gerarchico. In particolare, per questo lavoro è stato utilizzato il clustering gerarchico di cui si parlerà nel successivo paragrafo.

1.3.1 Clustering gerarchico

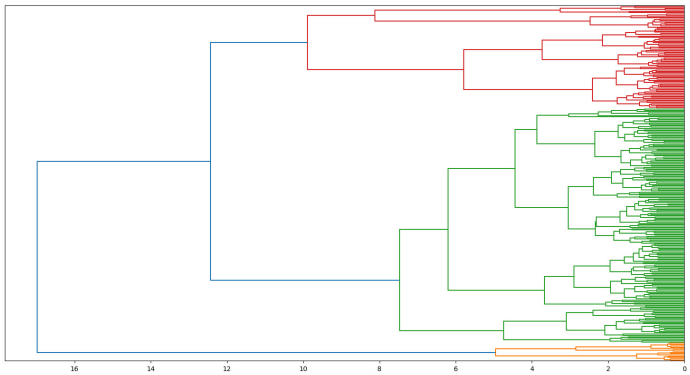

Figura 1.5: Dendogramma delle spiegazioni di Molnar

Questa tecnica mira a individuare le caratteristiche comuni a gruppi di dati raggruppandoli, nel nostro caso, in base alla loro similarità lessicale e semantica in modo tale da permetterci di ottenere gruppi di dati omogenei e distinti da altri gruppi con lo scopo di facilitare l'analisi.
Generalmente il clustering gerarchico mira a rappresentare i dati sotto forma di albero o dendogramma dove la radice è l'unico cluster che raccoglie tutti i campioni mentre le foglie sono i cluster con un solo campione.
Il nostro scopo, infatti, è stato proprio quello di riuscire a creare dei cluster dando in in-

put al programma le nostre spiegazioni sulla similarità delle coppie frasali ed ottenere in output gruppi di dati organizzati semanticamente. Tuttavia, come vedremo nel paragrafo successivo, non abbiamo ottenuto il risultato sperato.

1.3.2 Problematiche

Abbiamo visto come è possibile rappresentare l'insieme dei cluster ottenuti sotto forma di dendogramma e ovviamente l'ispezione visiva è spesso utile per comprendere la struttura stessa dei dati, tuttavia, è consigliato e più efficace concentrarsi su dimensioni del campione ridotte al fine di avere un risultato più chiaro e comprensibile. Dall'osservazioni dei risultati dell'algoritmo di clustering è emerso un limite nell'efficacia di tale tecnica applicata al testo: i gruppi, infatti, sono definiti lessicalmente e non semanticamente. Questo vuol dire che le spiegazioni sono state raggruppate per parole sovrapposte, restituendo dei cluster in cui le frasi condividono le prime parole ma non necessariamente il significato.

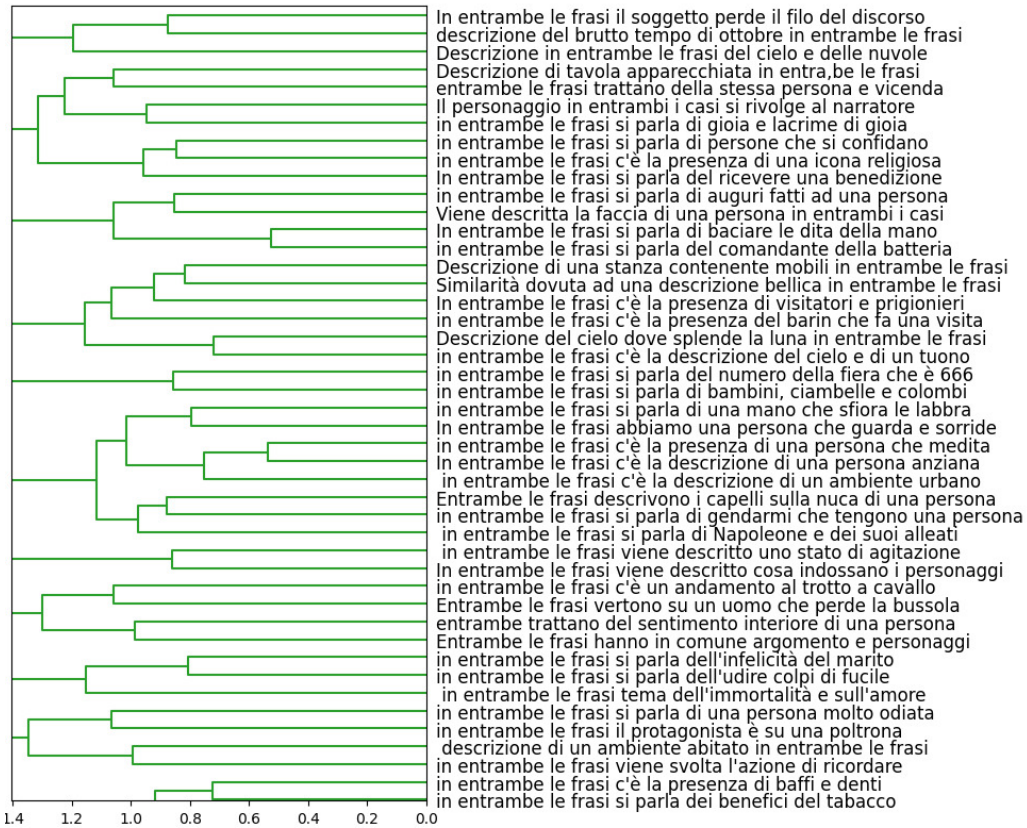

Figura 1.6: Risultato del clustering delle spiegazioni di Molnar

2. Strumenti di analisi statistica

In questo capitolo verranno mostrate e spiegate le diverse metriche statistiche utilizzate all'interno del presente lavoro e verranno anche discussi i risultati ottenuti.
Nella sezione 2.1 verrà introdotta l'Alpha di Krippendorff con le sue differenze rispetto alla K di Cohen [20] (2.2).
Nelle sezioni 2.3.1 e 2.3.2 verranno descritti, rispettivamente, la correlazione di Spearman ed il p-value. Infine nelle sezioni 2.1.2 e 2.3.4 parleremo dei risultati ottenuti da tali metriche statistiche.

2.1 Alpha di Krippendorff

l'Alpha di Krippendorf, indicata con il simbolo α e denominata tale in onore di Klaus Krippendorff è una mistura statistica che calcola l'accordo tra gli annotatori raggiunto durante una determinata operazione all'interno della quale vengono da questi assegnate delle valutazioni o categorie ad un insieme di dati. L'alfa di Krippendorff è applicabile a qualsiasi numero di codificatori e può essere utilizzabile per misurare la concordanza tra gli annotatori anche quando la quantità di categorie assegnabili è estremamente grande e anche quando gli annotatori possono essere in disaccordo tra di loro in modi diversi. Il valore dell'α di Krippendorff può assumere i seguenti valori:

- $\alpha = 1$ indica una affidabilità perfetta.
- $\alpha = 0$ indica la completa assenza di affidabilità.
- $\alpha < 0$ quando i disaccordi sono sistematici e superano quanto ci si potrebbe aspettare per caso.

Solitamente i valori iniziano ad essere accettabili da $\alpha = 0.67$ in su mentre da $\alpha = 0.80$ in su vengono considerati molto buoni.
Tale metrica è ampiamente utilizzata nei diversi studi di ricerca soprattutto quando si lavora con dati categorici o ordinati permettendo di ottenere una misura robusta della coerenza

tra le diverse fonti di annotazione.
Il calcolo dell'α coinvolge due componenti principali: la discrepanza osservata (D_o) e la discrepanza attesa (D_e). La discrepanza osservata rappresenta la differenza tra la concordanza effettivamente osservata e quella attesa per caso, mentre la discrepanza attesa riflette la variabilità che ci si aspetterebbe di trovare tra gli annotatori nel caso in cui le annotazioni fossero distribuite casualmente. In sostanza abbiamo la seguente formula:

$$\alpha = 1 - \frac{D_o}{D_e}$$

dove:

- $D_o = \frac{1}{n} \sum_{j=1}^{N} m_j \mathbb{E}(\delta_j)$
- $D_e = \frac{1}{P(n,2)} \sum_{c \in R} \sum_{k \in R} \delta(c,k) P_{ck}$

Esistono diverse versioni di Krippendorf che variano in base alla natura delle misurazioni da effettuare o alle annotazioni su cui vengono applicate.
Infatti esiste la versione per dati nominali, ordinali, intervalli e altre ancora.

2.1.1 Applicazione dell'Alpha di Krippendorff nella presente tesi

L'alpha di Krippendorff nel presente lavoro è stata calcolata sia "ordinal" che "interval" nei seguenti casi:

- α sui valori dei giudizi annotati da Molnar e da Poli
- α sui valori dei giudizi solamente degli annotatori di Prolific forniti dall'istituto di Linguistica Computazionale "A.Zampolli"
- α sui valori dei giudizi degli annotatori di Prolific insieme a quelli di Molnar e Poli

La differenza tra la versione "ordinal" e "interval" risiede nella natura della scala di misurazione dei dati:

- **Krippendorff ordinal** α:
  tale versione è consigliata quando si lavora con dati ordinali ma non si presume che

ci sia una distanza tra le diverse categorie. Un esempio è proprio la scala Likerta che è ordinale poichè ha un ordine e la differenza semantica tra il valore associato a 1 o a 2 non è la stessa differenza tra 4 e 5. Dunque, consideriamo solo l'ordine delle categorie ma non la distanza.

- **Krippendorff interval α**:
  tale versione è invece utilizzato quando si presume che le differenze tra categorie siano di uguale intervalli. Ciò significa che questa misura è adatta per dati che hanno una scala in cui le distanze tra i punti sono significative numericamente.

Riassumendo, entrambe le versioni sono utilizzate per valutare l'affidabilità dell'accordo tra annotatori umani, ma la scelta tra di esse dipende dalla natura dei dati che vengono analizzati.
L'intervallo di valori all'interno del quale è contenuto il risultato dell'α di Krippendorff varia tra -1 e 1. Più un valore si avvicina ad 1 e più vi è concordanza perfetta per gli annotatori, 0 significa concordanza casuale mentre valori inferiori a 0 significa una concordanza minore di quella che ci si aspetterebbe per caso.

2.1.2 Risultati ottenuti

Precedentemente è stato detto che un risultato accettabile si verifica quando abbiamo un' α che va da un valore di 0.67 a salire e che un buon risultato si verifica da 0.8 in su.
Chiaramente questi tipi di risultato sono più sensati quando abbiamo a che fare con dati risultanti da misure prevalentemente basate su fenomeni di tipo oggettivo. Nel nostro caso è stata ottenuta un' α di Krippendorff mediamente con un valore intorno a 0.40 che apparentemente potrebbe essere considerato un valore relativamente basso e potrebbe indicare una concordanza molto moderata tra gli annotatori. Tuttavia, in questo caso stiamo valutando l'aspetto semantico il quale fa riferimento alla soggettività individuale e all'impossibilità di definire se un determinato dato o una determinata risposta è giusta o sbagliata, dunque, di conseguenza, un risultato come quello ottenuto è sicuramente da considerarsi accettabile ed adeguato. Nella tabella 2.1 è possibile osservare quelli che sono i dati ottenuti:

Giudizi Molnar-Poli	Giudizi Annotatori Prolific 2	Giudizi Prolific, Molnar, Poli
Ordinal metric: 0.43	Ordinal metric: 0.45	Ordinal metric: 0.42
Interval metric: 0.35	Interval metric: 0.43	Interval metric: 0.39

Tabella 2.1: Tabella contenete i risultati dell'α di Krippendorff

2.2 K di Cohen

Il K di Cohen è un coefficiente statistico che rappresenta il grado di accuratezza e affidabilità in una classificazione statistica che permette di valutare il grado di accordo tra due valutazioni qualitative effettuate sulle stesse unità.
Si tratta di un indice di concordanza che tiene conto della probabilità di concordanza casuale, infatti, questo indice si calcola come rapporto tra l'accordo in eccesso rispetto alla probabilità di concordanza casuale e l'eccesso massimo ottenibile. In questo modo è possibile stabilire quanta parte della concordanza totale osservata è veramente dovuta al reale accordo tra i due valutatori.
La formula per ottenere la K è la seguente:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{2.1}$$

dove:

- $\Pr(a)$: probabilità di concordanza osservata tra gli annotatori
- $\Pr(e)$: probabilità di concordanza attesa tra gli annotatori in assenza di accordo reale

Per quanto riguarda il risultato, un valore della Kappa vicino a 1 indica una forte concordanza mentre un valore vicino allo 0 indica una concordanza casuale. Se il valore invece è negativo indica una concordanza peggiore di quella attesa casualmente.
Inizialmente nel presente lavoro si è pensato di utilizzare anche la K di Cohen come misura statistica tuttavia ciò non si è rivelato una scelta giusta e possibile per due principali motivi:

in primis perché il K di Cohen ammette solamente due valutatori e di conseguenza sarebbe stato necessario escludere gli annotatori di Prolific e limitarsi solamente al risultato ottenuto applicando questa metrica ai giudizi del sottoscritto e della collega Irene Poli.
Come secondo motivo abbiamo il fatto che sono necessarie valutazioni qualitative mentre nel nostro caso le valutazioni seguono dei precisi valori adottati come convenzionali e definiti attraverso la scala Likert. Comunque sia, il risultato ottenuto considerando solamente i giudizi di Molnar e Poli risulta un valore modesto rispetto all' α di Krippendorff ed è di 0.24.

2.3 Correlazione di Spearman e p-value

2.3.1 Correlazione di Spearman

La correlazione di Spearman è una tecnica statistica non parametrica utilizzata per valutare la relazione tra due variabili quantitative, qualitative o ordinali. In sostanza serve per misurare il grado di relazione tra due variabili. Il coefficiente di correlazione di Spearman è denominato tale in onore dello psicologo Charles Spearman e risale ad inizio Novecento.
Al fine di poter applicare tale metrica è necessario rispettare le seguenti indicazioni:

- Le variabili devono essere di tipo o qualitativo o quantitativo oppure dobbiamo avere a che fare con dati ordinali, infatti, questo metodo è fortemente consigliato per misurare le relazioni tra dati che si basano su scala Likert
- Le variabili devono essere appaiate sugli stessi casi ovvero deve essere stato misurato un valore per ognuna di esse
- Deve esistere una relazione monotona tra le due variabili

Quando parliamo di correlazione di Spearman è necessario anche introdurre il concetto di rango.
L'uso di ranghi è motivato dalla natura ordinale dei dati, infatti, quando lavoriamo con dati ordinali i valori sono ordinati ma la distanza tra tali valori potrebbe non essere uniforme o significativa; adoperando quindi i ranghi si risolve questo tipo di problema.

In questo modo, i valori di ogni variabile verranno ordinati in modo crescente e ad ogni valore sarà associato il rango corrispondente: ovviamente il valore più basso avrà rango 1 mentre quello più alto avrà il rango maggiore. La formula matematica per calcolare il coefficiente di correlazione di Spearman (ρ) è la seguente:

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

Dove:

- ρ è il coefficiente di correlazione di Spearman.
- D rappresenta la differenza tra i ranghi di ciascuna coppia di osservazioni.
- $\sum D^2$ è la somma dei quadrati delle differenze dei ranghi.
- n è il numero totale di osservazioni.

Il risultato che si può ottenere da questa formula è un numero compreso tra -1 che indica una perfetta relazione negativa tra i ranghi e +1 che indica una perfetta relazione positiva tra i ranghi.
Un valore invece pari a 0 indica che non c'è alcuna relazione tra i ranghi dunque più l'indice è vicino allo zero e più la relazione sarà debole, al contrario, più si avvicina a -1 oppure a +1 e più sarà forte.

Figura 2.1: Interpretazione indice di correlazione di Spearman, (fonte: paolapozzolo.it)

2.3.2 P-value

In statistica inferenziale il p-value indica il grado di significatività del campione. Questo valore è chiamato anche livello di significatività osservato e rappresenta la probabilità per

una ipotesi che supponiamo vera, di ottenere dei risultati egualmente compatibili oppure meno compatibili rispetto a quelli osservati durante il test effettuato con la suddetta ipotesi.
In altri termini, il p-value aiuta a capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità introdotta dal campionamento o se tale differenza è statisticamente significativa cioè difficilmente spiegabile mediante la casualità dovuta al campionamento.
Questo strumento viene utilizzato in molti campi che spaziano dalla fisica alla biologia alla psicologia.
Quando si effettua un determinato test si fissa un'ipotesi nulla ed un valore soglia (α) che per convenzione viene posto a 0.05.
Tale valore serve a determinare il valore di significatività del test e a seconda di questo è possibile ottenere i seguenti risultati:

- Se il valore $p > \alpha$ significa che l'ipotesi campionata è nulla
- Se il valore $p < \alpha$ questo lascia pensare che l'ipotesi dei dati osservati siano statisticamente significativi
- Se invece $p = \alpha$ oppure è circa uguale ovvero vicino al valore soglia è necessaria attenzione: potrebbe essere opportuno esaminare altre informazioni, considerare il contesto e valutare se esistono altre prove che supportano o contraddicono i risultati del test.

Per calcolare il p-value generalmente si applica la formula di Anderson-Darling:

$$A^2 = -N - \frac{1}{N}\sum_{i=1}^{N}\left[\frac{2i-1}{2N}\ln(F(X_i)) + \frac{2(N-i)+1}{2N}\ln(1-F(X_i))\right]$$

Dove:

- A^2: Statistica di Anderson-Darling.
- N: Dimensione del campione.
- $\sum_{i=1}^{N}$: Sommatoria su tutti gli elementi del campione.

- X_i: Valori del campione, ordinati dal più grande al più piccolo.
- $F(X_i)$: Funzione di distribuzione cumulativa associata alla distribuzione in esame.

In conclusione, il calcolo del p-value può essere complesso e può dipendere dalla distribuzione specifica che viene testata in quel determinato caso.
In generale, il p-value è calcolato confrontando il valore A^2 ottenuto con una distribuzione di riferimento teorica (spesso la distribuzione di Anderson-Darling) o utilizzando metodi numerici.

2.3.3 Applicazione di Spearman e p-value nella presente tesi

Nel presente lavoro il coefficiente di correlazione di Spearman ed il p-value sono stati calcolati sulla base dei dati prelevati dalle diverse colonne del file Excel che rappresenta il dataset costruito nel corso del tempo e mano a mano completato con nuovi dati che sono frutto di annotazioni oppure dell'applicazione degli algoritmi descritti nel successivo capitolo.
In particolare tale file Excel attraverso la programmazione in Python è stato trasformato in un dataframe e da qui sono state prelevate le colonne necessarie contenenti i valori che ci interessavano per compiere tali calcoli. Tra questi dati, a seguire, vengono citati anche quelli ottenuti con BLEU e con SBERT di cui parleremo nel capitolo successivo.
Infine, tali metriche sono state utilizzate anche nel quarto capitolo con lo scopo di svolgere un'analisi stilistica sul testo analizzando i risultati ottenuti dalle features delle coppie frasali.
Dunque, l'applicazione di Spearman e p-value riguarda i seguenti casi:

- Calcolo di Spearman e del p-value tra i risultati ottenuti con l'algoritmo BLEU e con l'algoritmo SBERT sulle spiegazioni
- Calcolo di Spearman e del p-value tra i risultati ottenuti con l'algoritmo BLEU e con l'algoritmo SBERT sulle frasi
- Calcolo di Spearman e del p-Value sui giudizi di Molnar e di Poli

- Calcolo di Spearman e del p-value sulla differenza in valore assoluto dei giudizi di Molnar e di Poli, con BLEU applicato sulle nostre spiegazioni di similarità
- Calcolo di Spearman e del p-value sulla differenza in valore assoluto dei giudizi di Molnar e di Poli, con SBERT applicato sulle nostre spiegazioni di similarità
- Calcolo di Spearman e del p-value sulla differenza in valore assoluto delle features linguistiche ottenute con Profiling-UD sulle frasi ed il giudizio sulla similarità di Molnar
- Calcolo di Spearman e del p-value sulla differenza in valore assoluto delle features linguistiche ottenute con Profiling-UD sulle frasi ed il giudizio sulla similarità di Poli

Attenzione: gli ultimi due casi elencati verranno trattati in maniera approfondita nel quarto capitolo.

2.3.4 Risultati ottenuti

Caso	Correlazione di Spearman	P-value
Bleu, Sbert - Spiegazioni	0.83	8.78e-264
Bleu, Sbert - Frasi	0.39	9.93e-39
Molnar, Poli - Giudizi	0.56	6.57e-88
$\lvert$`Molnar-Poli`$\rvert$, `Bleu`	-0.47	2.06e-57
$\lvert$`Molnar-Poli`$\rvert$, `Sbert`	-0.51	6.78e-69

Tabella 2.2: Tabella contenete i risultati di Spearman e P-Value

Nella Tabella 2.2 sono raffigurati i risultati ottenuti.
Cominciamo esaminando il risultato ottenuto calcolando il coefficiente di Spearman tra i punteggi di BLEU e SBERT applicati alle coppie di frasi prese dal corpus di partenza, che

risulta essere 0.39. Questo valore indica una correlazione positiva di intensità moderata tra i risultati dei due algoritmi. Tale correlazione può essere attribuita al fatto che, come approfondiremo più dettagliatamente nel prossimo capitolo, BLEU si basa principalmente su n-grammi, mentre SBERT utilizza reti neurali, rendendolo di conseguenza più sensibile agli aspetti semantici.
Un risultato simile, ma con un incremento leggermente maggiore, è stato ottenuto calcolando il coefficiente di correlazione di Spearman sui giudizi del sottoscritto e della collega Irene Poli, ottenendo così un punteggio di 0.56.
Tuttavia, il risultato più significativo è derivato dall'applicazione del coefficiente di Spearman ai risultati di SBERT e BLEU, entrambi applicati alle spiegazioni fornite dal sottoscritto e dalla collega. In questo caso il punteggio ottenuto è di 0.83, simbolo di una correlazione molto forte.

Continuando ad analizzare la Tabella 2.2, emergono chiaramente due risultati negativi, derivanti dall'applicazione del coefficiente di correlazione di Spearman questa volta utilizzato per esplorare la correlazione tra i giudizi numerici di similarità assegnati dai due annotatori e le corrispondenti spiegazioni semantiche testuali associate alle coppie di frasi. In questo caso la metrica statistica considera due parametri chiave: in primo luogo, il valore assoluto della differenza tra i giudizi dei due annotatori, e in secondo luogo, i valori ottenuti mediante l'utilizzo di BLEU e SBERT applicati sulle spiegazioni.
Per comprendere la natura negativa dei due risultati, è fondamentale spiegare che quanto più gli annotatori tendono ad assegnare giudizi simili, tanto più bassa sarà la differenza e, di conseguenza, tanto più elevato sarà l'accordo tra di loro. Parallelamente, gli algoritmi BLEU e SBERT produrranno valori più alti per le spiegazioni associate a tali giudizi.
La stessa regola vale anche al contrario: un basso accordo tra i due annotatori porta ad un valore alto della differenza e ad un valore basso prodotto da BLEU e da SBERT.
In sintesi, questo vuole dire che noi annotatori umani siamo più d'accordo su frasi molto simili, e tendiamo ad essere meno d'accordo su frasi meno simili.
Dunque, il segno negativo associato ai risultati ottenuti indica proprio questa relazione inversa tra i due parametri di input ed è cruciale sottolineare che il valore negativo sugge-

risce una forte correlazione tra i giudizi umani e le rispettive spiegazioni di similarità.
Più il risultato della metrica si discosta negativamente da zero (con un limite inferiore di -1), maggiore è la correlazione effettiva tra i giudizi numerici e le spiegazioni fornite dai due annotatori.
I risultati ottenuti sono i seguenti: -0.47 con BLEU e -0.51 con SBERT, risultati da considerare positivi e significativi.
Questa interpretazione assume un'importanza ancor maggiore nell'ambito dell'analisi semantica, dove la soggettività e l'individualità giocano un ruolo predominante.

Analizzando invece il p-value osserviamo che in tutti i casi il valore risulta essere sotto la soglia α di 0.05 e questo significa che il risultato ottenuto è innanzitutto significativo ma soprattutto che non è stato raggiunto per caso.

3. Metodologie di valutazione automatica della similarità semantica del testo

All'interno di questo capitolo verranno presentati e spiegati gli algoritmi che hanno come funzione quella di svolgere una valutazione automatica della similarità semantica del testo ad essi sottoposto.
In particolare qui parleremo degli algoritmi BLEU - Bilingual Evaluation Understudy nella sezione 3.1 e SBERT - Sentence-BERT, nella sezione 3.2. Infine parleremo dei risultati ottenuti rispettivamente con BLEU nella sezione 3.1.3 e con SBERT nella sezione 3.2.4.
Tali algoritmi per svolgere questo lavoro sono stati implementati tramite il linguaggio di programmazione python e sono state utilizzate le seguenti librerie:

- **Pandas** [18]: libreria per l'analisi e la manipolazione dei dati. Introduce strutture dati fondamentali come il DataFrame e la Serie. Presenta funzioni per il salvataggio dati in vari formati e facilita la gestione di dati eterogenei
- **NLTK - Natural Language Toolkit** [2]: libreria implementata per il natural language processing (NLP). Offre strumenti relativi all'analisi del testo, classificazione, stemming, lemmatizzazione e molto altro ancora
- **Scikit-learn** [4]: libreria che nasce come strumento mirato al machine-learning. Spesso viene abbreviata con sklearn
- **Numpy** [3]: libreria fondamentale per il calcolo scientifico, permette di compiere operazioni complesse su grandi quantità di dati numerici
- **Scipy** [5]: libreria di alto livello per il calcolo scientifico, in particolare si basa su Numpy ma con diverse funzionalità più avanzate

3.1 BLEU - Bilingual Evaluation Understudy

BLEU è un algoritmo sviluppato dall'IBM, presso il Watson Reserch Center nei primi anni 2000 (Papineni et al., 2002).
BLEU nasce con lo scopo di fornire risultati il più vicini possibile alla valutazione umana in relazione alla qualità delle traduzioni testuali in diverse lingue.
Questo algoritmo attualmente è uno strumento cruciale nel campo della ricerca sulla traduzione automatica e contribuisce significativamente al processo di sviluppo e valutazione dei modelli di linguaggio.
L'idea che sta alla base di BLEU è che una traduzione è tanto migliore quanto più essa è vicina alla traduzione di un traduttore professionista umano. BLEU dunque è un algoritmo nato per la valutazione automatica della traduzione automatica in un contesto dove è necessario considerare il fatto che le valutazioni delle traduzioni eseguite da umani sono estese ma estremamente costose e richiedono un notevole quantitativo di tempo. Per ovviare a questo problema BLEU propone un metodo di valutazione che allo stesso tempo è rapido, economico, indipendente dalla lingua e soprattutto correlato, in modo significativo, con la valutazione umana stessa.

3.1.1 Funzionamento BLEU

Il funzionamento di BLEU si basa sulla sua capacità di valutare la precisione delle corrispondenze tra n-grammi nella traduzione candidata e nei testi di riferimento umani. È necessario quindi precisare che il sistema richiede come ingredienti fondamentali una metrica numerica di "somiglianza di traduzione" ed un corpus di traduzioni umane di riferimento di alta qualità. La metrica dunque calcola la precisione modificata degli n-grammi, troncando i conteggi degli n-grammi candidati in base ai conteggi massimi nei riferimenti.
La formula matematica alla base di BLEU è la seguente:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Dove:

- **BP - Brevity Penality**: si tratta di un termine di penalizzazione per la brevità del testo generato rispetto ai testi di riferimento. Aiuta a non premiare eccessivamente testi troppo brevi.

- w_n: sono i pesi associati a ciascun tipo di n-gram e si calcola nel seguente modo: $w_n = \frac{1}{N}$

- p_n: è la precisione modificata per un particolare tipo di n-gramma e si calcola nel seguente modo:

$$p_n = \frac{\sum_{C' \in \text{Candidates}} \sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')}{\sum_{C \in \text{Candidates}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram})}$$

  In breve, la precisione modificata tiene conto delle occorrenze massime di un n-gramma in una singola traduzione di riferimento e quindi ne calcola la precisione nei confronti delle traduzioni candidate. Questo è un modo per penalizzare le traduzioni che generano troppi termini "ragionevoli" ma improbabili.

Il risultato della formula sopra indicata è il punteggio BLEU, che è una misura numerica della qualità di una traduzione automatica rispetto a uno o più riferimenti umani. Il punteggio BLEU varia da 0 a 1, dove 1 indica una corrispondenza perfetta tra la traduzione automatica e i riferimenti e lo 0 un totale disaccordo. Per comodità, spesso, i punteggi BLEU sono espressi in percentuale per rendere la valutazione più intuitiva.

3.1.2 Utilizzio di BLEU nella presente tesi

```
1 def bleuScore(df):
2     for index, row in df.iterrows():
3         sent1 = str(row['Frase 1']).split()
4         sent2 = str(row['Frase 2']).split()
5         print(sentence_bleu([sent1], sent2, weights=[1],))
```

Code 3.1: Codice di Bleu, unigrammi

Nella presente tesi BLEU è stato utilizzato per valutare la similarità tra le descrizioni fornite dai due annotatori e per valutare l'insieme di coppie frasali di partenza.

Il codice sopra indicato funziona nel seguente modo:
Per prima cosa viene iterata ciascuna riga del DataFrame. Vengono poi estratte le due frasi da confrontare dalle colonne ”Frase1” e ”Frase2” del file Excel. Infine viene calcolato il BLEU-Score tra le due frasi e stampato il risultato.
Ricordiamo che in questo caso il calcolo del BLEU considera solamente gli unigrammi quindi (N=1) e ignora gli altri n-grammi.

3.1.3 Risultati ottenuti

Nella Tabella 3.1 è possibile osservare qualche esempio di BLEU-SCORE calcolato sulle descrizioni del sottoscritto e della collega Irene Poli. Qualora si volessero confrontare più esempi è possibile visualizzare la Tabella 4.2 nella sezione Appendice.
Come accennato nel paragrafo precedente, nel nostro caso, vengono presi in considerazione gli unigrammi e sulla base di questi viene poi assegnato il punteggio. Dunque, se prendiamo coppie di frasi le quali tendono a non condividere nessuna parola il risultato sarà 0.00, al contrario, in frasi dove tutte le parole combaciano il punteggio sarà massimo. Tale procedimento è valido ed analogo anche per il calcolo del BLEU-SCORE sulle frasi presenti nel corpus di partenza.
Nelle immagini 3.1 e 3.2 invece sarà possibile osservare i grafici raffiguranti la distribuzione dei BLEU-SCORE sia sulle spiegazioni di similarità che sulle frasi.

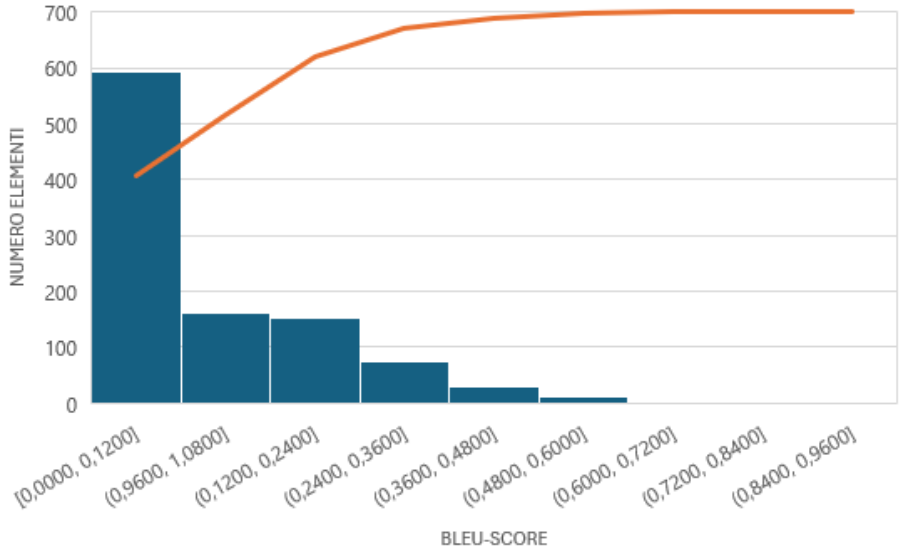

Figura 3.1: Distribuzione BLEU-SCORE delle spiegazioni di Molnar e Poli

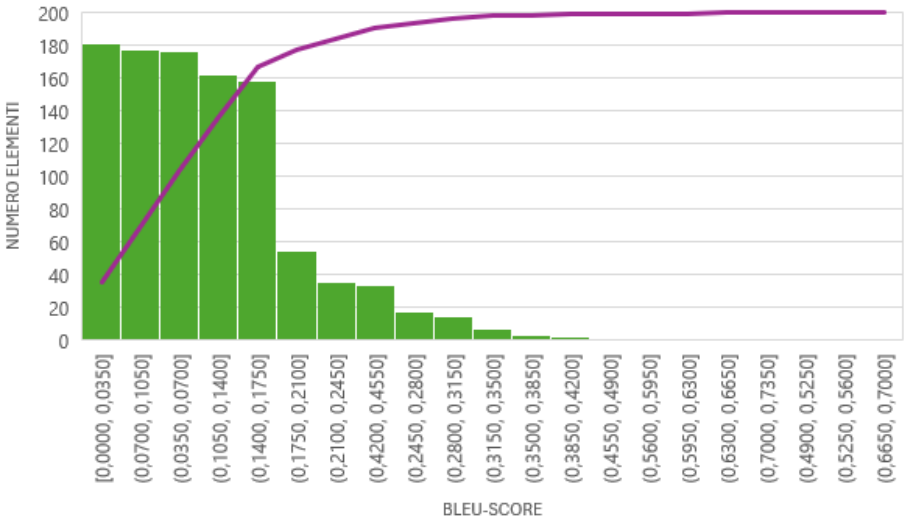

Figura 3.2: Distribuzione BLEU-SCORE delle frasi di partenza

Spiegazioni Molnar	Spiegazioni Poli	BLEU-SCORE
Presenza di un volto non bello all'interno di un quadro	Descrivono quadri diversi	0.00
In entrambe si parla di milizie	Completamente diverse	0.00
Entrambe parlano di alimenti	Entrambe menzionano cibi diversi	0.25
Entrambe parlano di una caduta in cantina	Parlano di qualcuno che cade in cantina	0.43
Pressoché identiche	Pressoché identiche	1.00
Completamente diverse	Completamente diverse	1.00

Tabella 3.1: Tabella contenente esempi di BLEU-SCORE

3.2 SBERT - Sentence-BERT

SBERT è una estensione di BERT (Bidirectional Encoder Representation from Transformers) ovvero un modello di linguaggio basato su "Transformers" che ha rivoluzionato il modo in cui le macchine comprendono il contesto linguistico nel campo di NLP (Reimers et al., 2019).

Quando parliamo di SBERT è necessario precisare anche il seguente aspetto molto importante:

rispetto a BLEU che si basa su statistiche di n-grammi e quindi non sempre è capace di riflettere la realtà semantica effettiva, SBERT coinvolge l'utilizzo di reti neurali profonde e sfrutta modelli di linguaggio avanzati per poter svolgere il proprio compito, ovvero, quello di creare delle rappresentazioni vettoriali che codificano l'informazione semantica necessaria a svolgere un determinato task.

Una volta ottenute queste rappresentazioni per le frasi, calcoliamo la similarità tra di esse con la distanza del coseno tra questi due vettori.

In questo lavoro abbiamo scelto di utilizzare SBERT poichè rispetto a BERT è molto più leggero e semplice da far girare in locale, allo stesso tempo però restituisce embedding di alta qualità che codificano in maniera altrettanto efficace le informazioni semantiche della frase.

3.2.1 BERT - Bidirectional Encoder Representations from Transformers

BERT è un modello di apprendimento automatico il cui funzionamento consiste nell'utilizzo di una serie di Transformers impilati e viene applicato in NLP.
Questo modello nasce e viene pubblicato nel 2018 da Jacob Devlin per Google (Devlin et al., 2018).
Oggi BERT è utilizzato a livello globale.
Tale modello diventa popolare per la sua capacità di creare rappresentazioni vettoriali del testo (chiamate "embeddings") che codificano informazioni linguistiche relative al lessico, alla sintassi e alla semantica.
Questa trasformazione rappresenta un grande vantaggio nell'elaborazione automatica del linguaggio naturale poiché gli algoritmi di apprendimento automatico non possono operare direttamente su testi grezzi ma possono facilmente elaborare vettori numerici.
Ciò consente, dunque, di confrontare due rappresentazioni vettoriali del testo per similarità utilizzando una metrica standard come la distanza euclidea o il coseno.

3.2.2 Funzionamento SBERT

A livello di funzionamento SBERT introduce il concetto di "Siamese network" per la rappresentazione semantica delle frasi e ciò significa che due frasi vengono elaborate in maniera indipendente attraverso lo stesso modello BERT.
Nella maggior parte delle volte un'architettura di tipo "Siamese network" è rappresentata con diversi modelli ma in realtà, possiamo considerare e vedere l'insieme come un singolo modello con configurazione e pesi condivisi tra diversi input paralleli.
Questo significa che ogni volta che i pesi del modello vengono aggiornati per un singolo

input, vengono aggiornati successivamente e allo stesso modo per tutti gli altri input.
La differenza principale tra un modello Siamese e uno non-Siamese è che il modello non-Siamese accetta entrambi gli input contemporaneamente mentre il modello Siamese accetta entrambi gli input in parallelo.

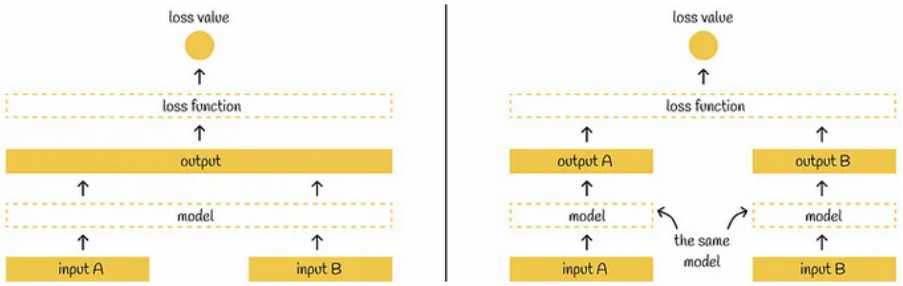

Figura 3.3: Confronto tra modello non-Siamese a sinistra e Siamese a destra, (fonte:towardsdatascience.com)

Questa configurazione permette una maggiore flessibilità nel trattare coppie di frasi, poiché le rappresentazioni non dipendono l'una dall'altra. Nel caso di SBERT, dopo il passaggio di una frase attraverso l'algoritmo BERT, viene poi successivamente applicato uno strato di pooling agli embedding di BERT per ottenere una rappresentazione di dimensione inferiore e più compatta, riducendo notevolmente la complessità computazionale.
I vettori risultanti da questo processo vengono chiamati u e v e sono poi successivamente sottoposti a diverse operazioni algebriche in primis la loro differenza in valore assoluto che permette di determinare se le frasi sono simili o meno tra di loro nel senso che più è piccola tale differenza e più queste sono simili.
Altre operazioni successive sono la classificazione, dove i vettori e la loro differenza vengono combinati e moltiplicati per una matrice di pesi addestrabile ed il risultato alimenta un classificatore softmax[1] che fornisce le probabilità per le diverse classi, e la regressione, ovvero una tecnica statistica utilizzata per analizzare la relazione tra una variabile dipendente (o di risposta) e una o più variabili indipendenti (o predittive) con lo scopo di poter fare previsioni sulla variabile dipendente basandosi sui valori di quelle indipendenti.
In particolare, per la regressione, i vettori u e v vengono utilizzati per calcolare il punteg-

[1]Funzione matematica che trasforma un vettore di punteggi in una distribuzione di probabilità

gio di similarità mediante una metrica come la similarità coseno. Allo stesso tempo, per aggiornare il modello viene utilizzata la perdita di errore quadratico medio (MSE).

3.2.3 Utilizzo di SBERT nella presente tesi

Nella presente tesi SBERT, proprio come BLEU, è stato utilizzato per valutare la similarità tra le spiegazioni semantiche fornite dal sottoscritto e dalla collega Irene Poli e successivamente per ottenere una valutazione sulle frasi di partenza.

Per quanto riguarda il codice successivamente indicato, questo funziona nel seguente modo: per prima cosa vengono inizializzate due liste al fine di contenere e memorizzare i valori delle rispettive colonne ”Spiegazioni Molnar” e ”Spiegazioni Poli” prese dal file Excel. Ricordiamo che il programma funziona in maniera analoga anche per le frasi. All'interno del ciclo for viene iterato il dataframe e vengono memorizzate le varie descrizioni in *descM* e *descI*. Ovviamente in descM ci sono le spiegazioni di Molnar mentre in descI quelle di Poli. Successivamente queste spiegazioni vengono codificate in vettori di embeddings che sono a loro volta memorizzati in *embeddings1* ed *embeddings2*.

In conclusione viene calcolata la similarità coseno tra questi vettori e memorizzata in *cosinescores*

```
1  from sentence_transformers import SentenceTransformer, util
2  def sbert(df):
3      descM = []
4      descI = []
5      for index, row in df.iterrows():
6          descM.append(row['Spiegazioni Molnar'])
7          descI.append(row['Spiegazioni Poli'])
8
9      embeddings1 = model.encode(descM, convert_to_tensor=True)
10     embeddings2 = model.encode(descI, convert_to_tensor=True)
11
12     cosinescores = util.cos_sim(embeddings1, embeddings2)
```

Code 3.2: Codice SBERT

3.2.4 Risultati ottenuti

Nella Tabella 3.2 è possibile osservare il risultato di SBERT su alcune spiegazioni del sottoscritto e della collega Irene Poli prese dal file Excel.

Per altri esempi è necessario consultare la Tabella 5.3 nella sezione Appendice.

Per essere precisi, quando parliamo dei risultati di SBERT possiamo sicuramente osservare che, rispetto a quelli ottenuti con BLEU, abbiamo dei dati più uniformi ed omogenei come ci dimostrano le immagini 3.4 e 3.5.

Tali risultati spaziano dal valore di 0 o pochissimo più per coppie di frasi completamente diverse come nel primo esempio della tabella, oppure possiamo incontrare valori poco superiori allo 0 come nel secondo esempio dove sebbene le descrizioni sono distanti tra di loro hanno comunque un minimo di similarità dovuta al fatto che entrambe, in questo caso, riguardano fenomeni che avvengono in un ambiente esterno.

Prendiamo adesso l'ultimo esempio della tabella con punteggio 0.86. Ciò implica che le due descrizioni sono quasi identiche semanticamente, utilizzando invece BLEU-SCORE e basandoci solamente sugli unigrammi il risultato è di circa 0.5 che è lontanissimo dalla realtà effettiva.

Ovviamente anche qui il punteggio massimo, ovvero 1, lo hanno le frasi completamente identiche. Dunque in conclusione possiamo dire che all'interno del nostro lavoro per valutare la semantica testuale è molto più adatto ed affidabile SBERT.

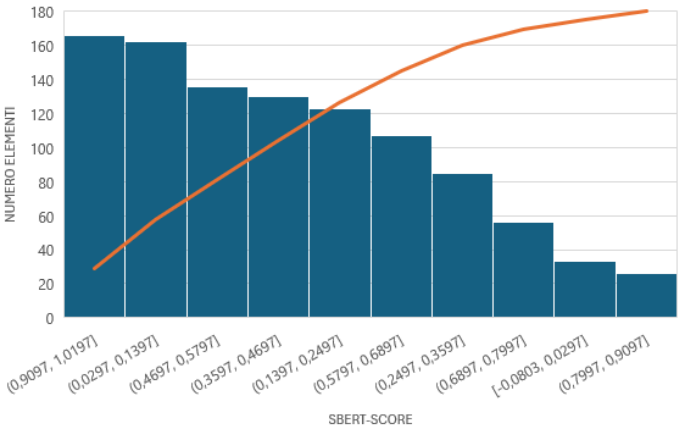

Figura 3.4: Distribuzione SBERT-SCORE delle spiegazioni di Molnar e Poli

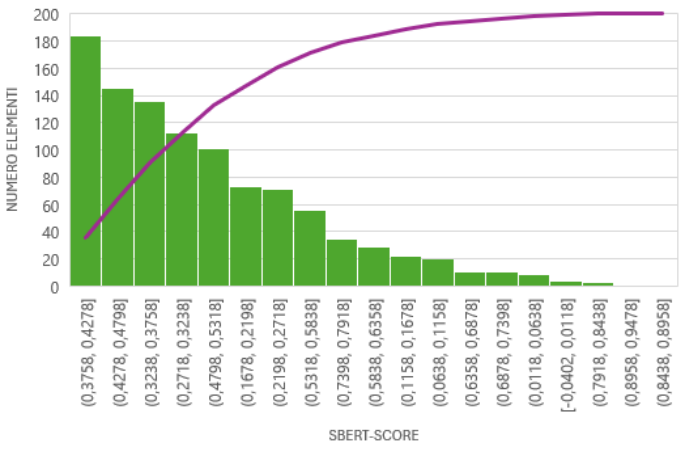

Figura 3.5: Distribuzione SBERT-SCORE delle frasi di partenza

Spiegazioni Molnar	Spiegazioni Poli	SBERT
Entrambe parlano del caos interiore di un uomo	Completamente diverse	0.02
Entrambe parlano di rumori provenienti dall'esterno	Menzionano Lampi	0.20
Entrambe descrivono una sincerità pagata a caro prezzo	Parlano del prezzo pagato per poter fare qualcosa	0.42
In entrambe le frasi si descrivono nuvole	Descrivono scenari dove c'è una nube a tratti	0.62
Entrambe parlano di un bacio	In entrambe si parla dell'azione di baciare	0.76
Entrambe parlano di leggi di natura	Entrambe fanno riferimento alle leggi di natura	0.86

Tabella 3.2: Tabella contenente alcuni risultati di SBERT

4. Analisi Stilistica

Grazie alla tecnologia in campo linguistico-computazionale è possibile accedere al contenuto linguistico intrinseco ad ogni parola, frase o testo.
Addirittura oggi possiamo studiare grandi corpora di dati con relativa facilità.
L'ultimo passo del presente lavoro è stato il tentativo di rispondere ad una precisa domanda: caratteristiche testuali simili corrispondono a una percezione della similarità simile?
Per poter rispondere a tale quesito abbiamo utilizzato lo strumento di annotazione linguistica automatica Profiling-UD di cui parleremo nella sezione 4.2.

4.1 Annotazione linguistica automatica

L'annotazione linguistica automatica è il processo mediante il quale un testo viene annotato linguisticamente in modo automatico utilizzando algoritmi e modelli computazionali.
Al fine di poter determinare e descrivere la struttura linguistica di un testo in maniera efficace, generalmente, è fondamentale seguire una serie di passaggi essenziali dove il successivo passaggio ha sempre come input il risultato di quello precedente.
Dunque, tale sequenza è la seguente:

- **Sentence splitting**: fase iniziale dove il testo che abbiamo a disposizione viene "splittato" cioè diviso in frasi
- **Tokenizzazione**: ogni frase ottenuta precedentemente viene segmentata nelle sue unità atomiche le quali sono la base di ogni analisi successiva
- **Lemmatizzazione**: ogni token viene ricondotto al relativo lemma, ovvero, la forma con cui entra nel dizionario
- **Analisi morofo-sintattica**: processo di assegnazione ad ogni token del testo l'informazione relativa alla categoria grammaticale che detiene in quello specifico contesto

- **Parsing**: descrizione di ogni frase in termini di relazione di dipendenza tra parole

Tramite l'annotazione automatica è quindi possibile monitorare un ampio numero di caratteristiche appartenenti ai diversi livelli di descrizione linguistica: lessicale, morfologico, sintattico.
Per poter svolgere tale procedimento di annotazione linguistica automatica all'interno del presente lavoro è stata utilizzata la piattaforma web Profiling-UD, della quale parleremo in maniera più dettagliata nel prossimo paragrafo.

4.2 Profiling-UD

Profiling-UD è un'applicazione web realizzata dall'Italian Natural Language Processing Lab di Pisa (Brunato et al., 2020).
Tale strumento viene utilizzato per il monitoraggio e l'analisi linguistica e consente l'estrazione di oltre 130 caratteristiche o "features", che spaziano attraverso diversi livelli di descrizione linguistica, di un testo o di una collezione di testi.
Una delle più importanti caratteristiche di Profiling-UD è che è stato progettato per essere multilingue in quanto si basa sul framework Universal Dependencies.
L'analisi svolta tramite tale applicazione si divide in due fasi:
la prima è di annotazione linguistica mentre la seconda è di profiling linguistico. La fase di annotazione viene eseguita mediante il modello UDPipe ovvero una pipeline addestrabile per funzioni come tokenizzazione, tagging, lemmatizzazione e analisi delle dipendenze.
Il testo automaticamente annotato viene poi utilizzato come input per la fase successiva, eseguita dal componente di profiling linguistico che definisce le regole per estrarre e quantificare le proprietà formali. Per ogni elemento che viene caricato su tale piattaforma al fine di essere elaborato, e che quindi può essere una frase, un testo o una raccolta di testi, Profiling-UD produce tre distinti file scaricabili:

- un formato CoNLLU contenente i risultati della fase di annotazione automatica
- un file in formato csv contenente i risultati del profiling linguistico con ogni caratteristica monitorata

- un file in formato txt contenente la legenda delle features

L'output, di ogni testo caricato sulla piattaforma, dunque sarà visualizzabile anche nel formato tabulare CoNLL-U:
questo significa che ogni token viene trascritto su una riga appartenente alla tabella e verranno mostrate le seguenti proprietà sulle diverse colonne:

- **Id**: indice numerico progressivo ed univoco per ogni elemento, il primo Id, corrispondente al primo token e viene inizializzato al valore 1
- **Forma**: forma della parola, come si presenta il token all'interno del testo inserito
- **Lemma**: voce del dizionario alla quale corrisponde il token
- **UPOS**: indica la parte del discorso relativa alla parola quindi indica il valore morfologico come per esempio nome, verbo, aggettivo
- **XPOS**: tag di annotazione facoltativo relativo alla parte del linguaggio specifico e sensibile al contesto
- **Testa**: indicatore numerico relativo all'Id della parola dalla quale il token in questione dipende sintatticamente. La testa dunque è l'elemento frasale dal quale il token dipende
- **Tratti**: lista dei tratti morfologici appartenenti al token. I vari tratti vengono separati da barra verticale e rappresentano caratteristiche come per esempio numero, genere, modo del verbo
- **Tipi di relazione**: definisce la relazione di dipendenza tra la testa e il dipendente: se la testa ha valore pari a zero la relazione sarà di tipo root

Proseguiamo adesso facendo un esempio di funzionamento di tale strumento analizzando una frase presa dal corpus di partenza.
La frase presa in considerazione è la seguente:
"Tutta la sua figura aveva un'espressione straordinariamente dignitosa.".

Nella seguente immagine è mostrato il risultato in formato tabellare CoNLL-U ottenuta con UDPipe [17].

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = italian-markit-ud-2.12-230717									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = Tutta la sua figura aveva un'espressione straordinariamente dignitosa.									
1	Tutta	tutto	DET	T	Gender=Fem|Number=Sing|PronType=Tot	4	det:predet	_	TokenRange=0:5
2	la	il	DET	RD	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	4	det	_	TokenRange=6:8
3	sua	suo	DET	AP	Gender=Fem|Number=Sing|Poss=Yes|PronType=Prs	4	det:poss	_	TokenRange=9:12
4	figura	figura	NOUN	S	Gender=Fem|Number=Sing	5	nsubj	_	TokenRange=13:19
5	aveva	avere	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin	0	root	_	TokenRange=20:25
6	un'	uno	DET	RI	Definite=Ind|Gender=Fem|Number=Sing|PronType=Art	7	det	_	SpaceAfter=No|TokenRange=26:29
7	espressione	espressione	NOUN	S	Gender=Fem|Number=Sing	5	obj	_	TokenRange=29:40
8	straordinariamente	straordinariamente	ADV	B	_	9	advmod	_	TokenRange=41:59
9	dignitosa	dignitoso	ADJ	A	Gender=Fem|Number=Sing	7	amod	_	SpaceAfter=No|TokenRange=60:69
10	.	.	PUNCT	FS	_	5	punct	_	SpaceAfter=No|TokenRange=69:70

Figura 4.1: Tabella formato CoNLL-U

Osservando tale immagine notiamo che il token con ID 5 ”aveva” essendo verbo (VERB) corrisponde al lemma ”avere”. Dall’immagine notiamo anche che si tratta di un modo indicativo (Mood=Ind) e tempo imperfetto (Tense=Imp) la cui testa è pari a 0, ciò significa anche che si tratta dell’elemento radice.

Prendiamo invece il token con ID 9 ”dignitosa” è un aggettivo (ADJ) che è stato lemmatizzato nella forma ”dignitoso” e che ha come testa l’id 7 ovvero ”espressione” che a sua volta è un nome (NOUN) e ha come testa il token con ID 5 che è la radice.

Tale frase può essere vista anche sotto forma di albero a dipendenze tra le relazioni sintattiche.

In informatica il termine albero si riferisce generalmente ad una struttura dati utilizzata per organizzare i dati in modo gerarchico rappresentando una relazione padre-figlio tra i suoi elementi. Un albero è composto da nodi che hanno una radice in comune e sono collegati

tra loro da archi. Ogni nodo può avere o non avere dei figli e nel caso non li avesse tale nodo viene chiamato foglia.
Nel nostro caso, si tratta di una struttura ad albero, in quanto ogni parola e la sua relativa parte del discorso rappresentano un nodo e le relazioni tra le parole sono stabilite tramite degli archi.

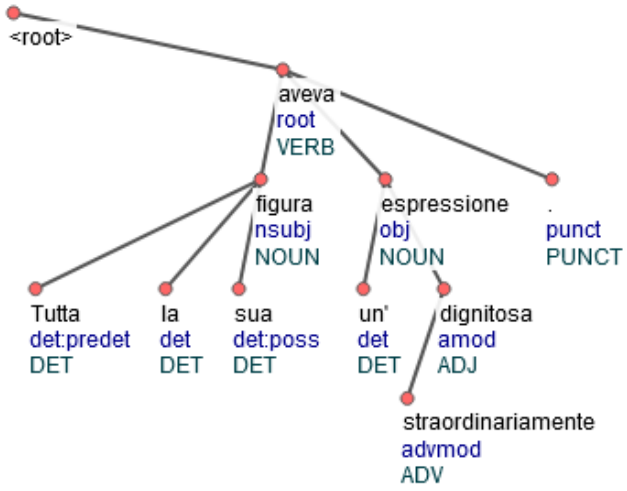

Figura 4.2: Visualizzazione grafica albero a dipendenze tra le relazioni sintattiche

4.2.1 Utilizzo di Profiling-UD nella presente tesi

Abbiamo scelto di sfruttare tale applicazione web per ricavare dei dati che ci avrebbero poi permesso di osservare se esiste uno stile di scrittura delle frasi che correla con la similarità percepita.
Per poter fare questo sono state eseguite in successione diverse operazioni.
Per prima cosa dal dataset sono state prelevate le colonne contenti le frasi di partenza ovvero ”Frasi 1” e ”Frasi 2”.
Tali colonne sono state separate e l’insieme delle frasi che esse contenevano sono state salvate su due file distinti i quali, successivamente, sono stati caricati su Profiling-UD per ottenere le features ed i rispettivi valori necessari alla nostra ricerca.
Con i risultati ottenuti, infine, sono stati realizzati due file Excel contenenti rispettivamente

tutte le features ed i rispettivi risultati per ogni frase di ”Frasi 1” e di ”Frasi 2”. Nella seguente immagine è possibile vedere alcune delle 130 features estratte grazie a Profiling-UD.

Filename	n_sentence	n_token	tokens_per_se	char_per_to
1	1	12	12	3,222222222
2	1	19	19	4,4375
3	1	29	29	3,666666667
4	1	27	27	5,75
5	1	16	16	3,8
6	1	30	30	4,538461538
7	1	32	32	4,307692308
8	1	12	12	3,777777778
9	1	11	11	4,7
10	1	39	39	4,588235294
11	2	24	12	4,909090909

Figura 4.3: Features relative alle frasi estratte con Profiling-UD

4.2.2 Elaborazione dei dati ottenuti con Profiling-UD

Come passo successivo per rispondere alla domanda che ci siamo posti in partenza, sono stati presi i due file Excel descritti nel paragrafo precedente e sono stati dati come input ad un programma realizzato in python che ha svolto la differenza in valore assoluto tra i valori delle features delle 130 colonne di ”Frasi 1” con i valori delle features di ”Frasi 2”, permettendoci così di ottenere in output un terzo file contenente il risultato di tale differenza che è sato chiamato ”risultatiF.xlsx”.

Il passo successivo è stato quello di calcolare il coefficiente di correlazione di Spearman ed il p-value tra ogni colonna di features di ”risultatiF.xlsx” ed il giudizio prima di Molnar e poi di Poli.

Per compiere tale operazione è stato implementato un programma in python che funziona nel seguente modo:

Inizialmente vengono caricati in input i due file Excel che sono ”CompletiMolnarPoli.xlsx” da dove vengono prelevate le colonne contenenti i giudizi di similarità di Molnar e di Poli, e ”risultatiF.xlsx”.

A questo punto grazie alla libreria Pandas è possibile inserire i dati di tali file in un DataFrame.

Viene quindi creata la lista, inizialmente vuota, che conterrà i risultati.

Come prossimo passo viene eseguito un ciclo for dove vengono iterate le 130 colonne di ”risultatiF.xlsx” e calcolato Spearman e p-value richiamando una funzione definita precedentemente e chiamata ”spearmanPv”, tra ”Giudizio Molnar” e la rispettiva colonna di features.

Analogo procedimento anche con ”Giudizio Poli”.

I risultati ottenuti dal calcolo vengono salvati nella lista dei risultati e successivamente viene creato un DataFrame dei risultati.

L’ultimo passaggio è il salvataggio dell’output su un file Excel generato dal programma.

In seguito è possibile osservare il codice di tale programma:

```
1  def spearmanPv(file1_col, file2_col):
2      spearman_corr, p_value = spearmanr(file1_col, file2_col)
3      return spearman_corr, p_value
4
5  def main():
6      # Percorsi dei file
7      file1_path = 'CompletiMolnarPoli.xlsx'
8      file2_path = 'risultatiF.xlsx'
9      output_path = 'output_file.xlsx'
10     # Vengono inseriti i dati relativi ai giudizi e alle features
11     file1_data = pd.read_excel(file1_path, usecols=['Giudizio Molnar
       '], nrows=1024)
12     file2_data = pd.read_excel(file2_path, nrows=1024)
13     # Creazione lista risultati
14     result_list = []
15     # Iterazione colonne file contente le features
16     for col_name in file2_data.columns:
17         # Calcolo Spearman e p-value eichiamando la funzione "
       calculate_spearmanr"
18         spearman_corr, p_value = spearmanPv(file1_data['Giudizio
       Molnar'], file2_data[col_name])
19         # Inserimento dati nella lista risultati
20         result_list.append({'nome feature': col_name, 'Spearman':
```

```
    spearman_corr, 'p-value': p_value})
21  # Creazione dataframe risultati
22  output_df = pd.DataFrame(result_list)
23  # Operazione di salvataggio dell'output del DataFrame su un file
    Excel
24  output_df.to_excel(output_path, index=False)
```

4.2.3 Risultati ottenuti

Dopo aver eseguito il programma descritto nella precedente sezione, i risultati ottenuti sono stati filtrati in base al valore del p-value.

Sono state prese in considerazione solamente le features con un p-value inferiore a 0.05.

Nella Tabella 4.1 sono contenute le features pertinenti in relazione ai giudizi di similarità di Poli, mentre nella Tabella 4.2 sono contenute le featuers pertinenti in relazione ai giudizi di Molnar.

Nella sezione 5.4 dell'Appendice invece è possibile avere degli approfondimenti sulle features presenti nelle due tabelle.

Analizziamo adesso i risultati ottenuti:

In quasi in tutti i casi, dove il p-value è pertinente, il valore del coefficiente di correlazione di Spearman si avvicina in modo significativo allo 0, sia positivamente che negativamente. Facendo una media dei risultati ottenuti con Spearman per i risultati del sottoscritto ci aggiriamo sui 0.021 mentre per Poli la media è di 0.018.

In conclusione possiamo affermare che, in entrambi i casi, i valori ottenuti con Spearman indicano che non c'è alcun tipo di correlazione tra i dati inseriti in input, dunque, possiamo affermare che non esiste uno stile di scrittura delle frasi che correla con la similarità percepita dai due annotatori.

Tabella 4.1: Features pertinenti Poli

Nome Feature	Spearman	p-value
char_per_tok	−0.090 944 944	0.003 615 806
upos_dist_DET	−0.090 362 043	0.003 838 455
upos_dist_PRON	0.105 581 262	0.000 723 221
upos_dist_PROPN	−0.084 325 863	0.006 990 479
lexical_density	−0.069 298 938	0.026 735 921
verbs_tense_dist_Imp	0.081 998 063	0.008 726 582
verbs_tense_dist_Past	0.071 366 898	0.022 511 363
verbs_mood_dist_Cnd	0.065 302 494	0.036 859 982
verbs_form_dist_Ger	0.074 353 241	0.017 436 991
aux_tense_dist_Imp	0.071 478 599	0.022 300 669
aux_num_pers_dist_Sing+1	0.069 531 925	0.026 228 069
avg_verb_edges	−0.068 733 512	0.028 003 791
verb_edges_dist_0	0.061 537 335	0.049 214 649
verb_edges_dist_1	−0.110 747 438	0.000 389 587
verb_edges_dist_3	0.100 411 874	0.001 307 909
avg_prepositional_chain_len	−0.066 061 135	0.034 719 627
prep_dist_2	−0.073 703 47	0.018 446 636
obj_pre	0.087 511 45	0.005 116 928
obj_post	0.111 095 866	0.000 373 308
subj_post	0.071 948 729	0.021 432 552
dep_dist_advcl	0.064 727 447	0.038 556 216
dep_dist_case	0.073 764 986	0.018 348 908
dep_dist_det	−0.095 003 895	0.002 363 297
dep_dist_expl	0.062 741 698	0.044 931 628
dep_dist_flat:name	−0.125 591 979	$5.672\,06 \times 10^{-5}$
dep_dist_root	−0.088 155 328	0.004 798 511
principal_proposition_dist	0.083 533 095	0.007 543 434
subordinate_proposition_dist	0.086 690 466	0.005 550 541
subordinate_post	0.066 585 763	0.033 302 277
avg_subordinate_chain_len	0.069 157 186	0.027 049 035
subordinate_dist_1	0.062 273 909	0.046 556 386

Tabella 4.2: Features pertinenti Molnar

Nome Feature	Spearman	p-value
char_per_tok	−0.079 563 036	0.010 944 953
upos_dist_ADP	0.076 282 83	0.014 718 351
upos_dist_DET	−0.088 608 454	0.004 585 278
upos_dist_PRON	0.114 892 994	0.000 232 61
upos_dist_PROPN	−0.089 070 444	0.004 376 732
lexical_density	−0.065 980 614	0.034 941 657
verbs_tense_dist_Imp	0.113 883 883	0.000 264 143
verbs_tense_dist_Past	0.088 820 638	0.004 488 406
verbs_form_dist_Fin	0.065 609 194	0.035 981 543
aux_mood_dist_Ind	0.063 108 924	0.043 689 847
aux_form_dist_Inf	0.085 656 556	0.006 143 587
aux_num_pers_dist_Sing+2	−0.097 019 274	0.001 902 048
verbal_head_per_sent	0.080 500 319	0.010 037 837
verb_edges_dist_1	−0.066 259 46	0.034 177 892
verb_edges_dist_4	−0.071 631 082	0.022 015 814
verb_edges_dist_6	0.071 824 496	0.021 659 048
avg_prepositional_chain_len	−0.087 634 757	0.005 054 511
prep_dist_2	−0.092 574 281	0.003 054 256
obj_pre	0.082 837 649	0.008 060 41
obj_post	0.075 819 123	0.015 335 164
subj_post	0.087 759 043	0.004 992 295
dep_dist_case	0.096 859 843	0.001 935 28
dep_dist_det	−0.117 559 091	0.000 165 43
dep_dist_flat:name	−0.144 692 505	$3.410\,06 \times 10^{-6}$
dep_dist_nsubj	0.071 020 602	0.023 175 577
principal_proposition_dist	0.120 361 742	0.000 114 724
subordinate_proposition_dist	0.114 838 12	0.000 234 229
subordinate_post	0.107 490 453	0.000 577 205
avg_subordinate_chain_len	0.098 647 136	0.001 591 452
subordinate_dist_1	0.098 977 994	0.001 534 329

Conclusioni

Il presente elaborato si è posto l'obiettivo di riprendere un precedente progetto di ricerca che affronta il tema della percezione dei parlanti della similarità semantica tra frasi e della sensibilità di modelli computazionali di valutazione della similarità rispetto a dei tratti linguistici rintracciabili nelle frasi.
In particolare, dal precedente lavoro sono stati ripresi alcuni dati, ovvero, circa metà delle coppie frasali totali ottenute con la traduzione automatica di testi appartenenti alla letteratura narrativa ed i rispettivi giudizi sulla similarità semantica tra di esse degli annotatori della piattaforma Prolific che hanno seguito i valori indicati su una scala Likert.
Una volta ottenuti tali dati ha avuto origine il processo della creazione di un dataset innovativo poiché a differenza di un passato non troppo lontano, grazie ai nuovi sistemi che si basano su reti neurali è possibile dare come input ad un sistema non solo valori numerici relativi ai diversi dati ma anche fornire al sistema stesso le spiegazioni sui motivi per i quali l'annotatore ha scelto di attribuire quei determinati valori.
Dunque, per questo, è importante precisare che il dataset realizzato durante questo periodo di tirocinio insieme alla collega Irene Poli è il secondo in Italia, dopo "The Italian e-RTE-3 Dataset", e contiene al proprio interno delle spiegazioni fornite da entrambi che riguardano giudizi di similarità semantica.

Al fine di annotare questo dataset, il sottoscritto e la collega hanno analizzato ed annotato 32 questionari contenenti ciascuno 32 coppie di frasi per un totale di 1024 coppie ed un tempo medio di circa 24 minuti a questionario. Il processo di annotazione è consistito nel leggere le frasi costituenti la coppia, dare un giudizio numerico sulla similarità seguendo i valori nel range 1-5 della scala Likert prestabilita e successivamente scrivere una breve spiegazione sul motivo per il quale secondo il sottoscritto e secondo la collega quelle frasi effettivamente si assomigliano.
Il corpus contenente l'insieme delle coppie di frasi da analizzare ed annotare ed i giudizi degli annotatori di Prolific sulla similarità semantica sono stati forniti dall'Istituto di Lin-

guistica Computazionale ”A.Zampolli” del CNR.

Una volta eseguita l’annotazione sono stati utilizzati due algoritmi capaci di svolgere una valutazione automatica della similarità semantica del testo.
In particolare è stato utilizzato BLEU (Bilingual Evaluation Understudy) il quale funzionamento si basa sul calcolo sulla sovrapposizione di n-grammi nelle rispettive frasi, precisando il fatto che per il presente lavoro ci siamo limitati solamente agli 1-grammi, e SBERT (Sentence-BERT) che rispetto al precedente coinvolge l’utilizzo di reti neurali profonde e modelli di linguaggio avanzati. Questa differenza tra i due algoritmi si osserva nei risultati, alquanto distanti, ottenuti dando loro in input le coppie di frasi.
Comunque, tale distanza è stata misurata utilizzando il coefficiente di correlazione di Spearman fornendo come parametri i punteggi ottenuti con il calcolo di BLEU e di SBERT sulle coppie frasali.
Il risultato ottenuto da tale misura è di 0.39 su un massimo di 1 e da qui possiamo evincere che i punteggi ottenuti correlano positivamente ma che tale correlazione è moderata.

Dopo, abbiamo voluto verificare quanto i punteggi di BLEU e di SBERT, questa volta calcolati sulle spiegazioni di Molnar e di Poli, correlassero tra di loro, ottenendo un valore di 0.83 sinonimo anche di un forte accordo nelle spiegazioni semantiche sulle coppie frasali tra i due annotatori, e successivamente, anche quanto correlassero i giudizi numerici dei due annotatori, valutando la similarità semantica delle frasi di partenza. Questa volta il risultato ottenuto tra i due annotatori è stato di 0.56 sempre su un massimo di 1, ancora una volta segno di una correlazione positiva e buona.

Infine, abbiamo utilizzato Spearman per valutare la correlazione tra i giudizi di similarità assegnati dal sottoscritto e dalla collega e le spiegazioni testuali relative alle stesse coppie di frasi. Per svolgere tale compito, abbiamo applicato questa metrica considerando come parametri la differenza in valore assoluto tra i giudizi di Molnar e Poli ed i risultati ottenuti mediante BLEU e SBERT applicati alle spiegazioni di similarità.
I risultati ottenuti sono stati -0,47 con BLEU e -0,51 con SBERT.

Questi valori negativi, come descritto nella sezione 2.3.4, derivano da una relazione di proporzionalità inversa che si verifica tra i parametri in input alla metrica.
Per essere più precisi, maggiore è l'avvicinamento di tali valori a -1, migliore è da considerarsi la correlazione. In questo contesto, siamo soddisfatti dei risultati ottenuti, poiché ci troviamo sul piano semantico in cui la soggettività e l'individualità svolgono un ruolo significativo.

Per valutare la significatività di tutti gli esiti ottenuti con Spearman è sempre stato calcolato il p-value relativo ad ogni risultato prodotto e sono stati tenuti in considerazione solamente i risultati con un p-value inferiore a 0.05, ovvero, il valore standard pertinente alla soglia di accettabilità.

Altre metriche statistiche utilizzate oltre alla correlazione di Spearman e al p-value sono state l' α di Krippendorff e la K di Cohen.
Tali misure nascono con l'obiettivo di calcolare l'accordo tra gli annotatori che viene raggiunto durante operazioni che hanno lo scopo di assegnare delle valutazioni o delle categorie ad un insieme di dati.
In particolare per questo lavoro la K di Cohen, nonostante sia stata applicata, non è assolutamente indicata a causa dei motivi elencati nella sezione 2.2. Il risultato ottenuto con tale misura è stato di 0.24, risultato che non rispecchia assolutamente la realtà.
Dunque, per capire se effettivamente esiste un accordo tra i vari annotatori in relazione solamente ai giudizi numerici sulla similarità semantica attribuiti seguendo la scala Likert è stata utilizzata esclusivamente l'α di Krippendorff.
Tale metrica, infatti, permette di considerare un numero di annotatori maggiore dei due permessi dalla K di Cohen e soprattutto ammette valutazioni quantitative. L'α di Krippendorff è stata calcolata sia nella versione "ordinal" ovvero quando si assume che non ci sia una distanza tra le diverse categorie ed anche "interval" quando invece si presume che ci sia.
I risultati ottenuti dal calcolo dell'αdi Krippendorff sono i seguenti:

- α di Krippendorff considerando solamente i giudizi di Molnar e Poli:

Ordinal metric: 0.44

Interval metric: 0.35

- α di Krippendorff considerando solamente i giudizi dei cinque annotatori di Prolific:
  Ordinal metric: 0.45
  Interval metric: 0.43

- α di Krippendorff considerando i giudizi di Molnar, Poli e dei cinque annotatori di Prolific:
  Ordinal metric: 0.42
  Interval metric: 0.39

Generalmente quando parliamo di risultati ottenibili con Krippendorff possiamo dire che un risultato accettabile si ha da $\alpha = 0.67$ in su, mentre se raggiunto un risultato di almeno 0.80 il risultato è da considerarsi buono.
Nel nostro caso abbiamo dei risultati che si aggirano intorno ad un' α di 0.40.
Questo valore può sembrare relativamente basso, tuttavia, è necessario ricordare nuovamente che stiamo svolgendo un'analisi di tipo semantico dove non esistono regole precise e ci troviamo dunque sul piano del soggettivo.
Proprio per questo i risultati ottenuti, generalmente simili e vicini tra di loro, sono da considerarsi accettabili ed adeguati.

L'ultima parte del lavoro è stata dedicata al tentativo di comprendere se caratteristiche testuali simili corrispondono a una percezione della similarità simile.
Per fare questo abbiamo utilizzato lo strumento di annotazione linguistica Profiling-UD.
Attraverso la seguente piattaforma è stato possibile ottenere tutte le features linguistiche delle coppie di frasi con i rispettivi valori associati, dopodichè è stata calcolata la differenza in valore assoluto tra i valori delle features di tutte le prime frasi e tutte le seconde frasi costituenti le varie coppie.
A questo punto, sono stati calcolati i coefficienti di correlazione di Spearman tra le features e i giudizi numerici assegnati alle coppie di frasi, prima di Molnar e successivamente di Poli.

In entrambi i casi, i risultati ottenuti sono stati molto prossimi a zero. Questo suggerisce che la risposta alla domanda finale è negativa, poiché valori prossimi allo zero indicano una mancanza totale di correlazione tra i parametri presi in considerazione dalla metrica statistica.

Per concludere, riguardo agli sviluppi futuri di questo progetto di ricerca, emerge la prospettiva di poter addestrare un modello per eseguire task di annotazione che consistono nella capacità di attribuire giudizi su determinati dati e generare le spiegazioni correlate. Ciò aprirebbe la possibilità di creare dataset annotati di dimensioni considerevoli, cosa che risulta estremamente costosa in termini di tempo e di risorse appoggiandosi esclusivamente ad annotatori umani.

Bibliografia

[1] Dominique Brunato et al. «Profiling-UD: A Tool for Linguistic Profiling of Texts». In: *Proceedings of the 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France, mag. 2020. URL: http://www.italianlp.it/demo/profiling-ud/.

[2] NLTK Contributors. *Natural Language Toolkit (NLTK) Documentation*. 2024. URL: https://www.nltk.org/.

[3] NumPy Contributors. *NumPy Documentation*. 2024. URL: https://numpy.org/.

[4] scikit-learn Contributors. *scikit-learn Documentation*. 2024. URL: https://scikit-learn.org/stable/.

[5] SciPy Developers. *SciPy Documentation*. 2024. URL: https://scipy.org/.

[6] J. Devlin et al. «BERT: Pre-training of deep bidirectional transformers for language understanding». In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Ketan Doshi. *Foundations of NLP Explained: BLEU Score and WER Metrics*. 2021. URL: https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b.

[8] HeaDvisor. *P-value*. n.d. URL: https://www.headvisor.it/p-value.

[9] ISST-TANL. *Part-of-Speech Tagset for Italian*. Rapp. tecn. Recuperato il giorno, mese, anno. Italian Society of Speech Sciences e Technologies (ISST), year. URL: http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf.

[10] Italian Natural Language Processing Lab. *Profiling-UD*. 2020. URL: http://www.italianlp.it/demo/profiling-ud/.

[11] Kishore Papineni et al. «BLEU: A method for automatic evaluation of machine translation». In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Lug. 2002, pp. 311–318.

[12] Prolific. *Prolific - Where Trust and Quality Matter*. 2024. URL: https://www.prolific.co/.

[13] Qualtrics. *Scala Likert: Guida completa alla creazione e all'uso*. n.d. URL: https://www.qualtrics.com/it/experience-management/ricerca/scala-likert/.

[14] Nils Reimers e Iryna Gurevych. «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks». In: *arXiv* (2019). URL: https://arxiv.org/pdf/1908.10084.pdf.

[15] Nils Reimers e Iryna Gurevych. «Sentence-BERT: Sentence embeddings using Siamese BERT-networks». In: *arXiv preprint arXiv:1908.10084* (2019).

[16] Scikit-learn. *Clustering - scikit-learn documentation*. n.d. URL: https://scikit-learn.org/stable/modules/clustering.html.

[17] Milan Straka e Jana Straková. «UDPipe: Trainable Pipeline for Processing Universal Dependency Trees». In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). 2016, pp. 3348–3355. URL: https://www.aclweb.org/anthology/L16-1264.

[18] Pandas Development Team. *Pandas Documentation*. 2024. URL: https://pandas.pydata.org/.

[19] *Universal Dependencies - Part-of-Speech Tags*. URL: https://universaldependencies.org/u/pos/.

[20] Wikipedia contributors. *Kappa di Cohen*. n.d. URL: https://it.wikipedia.org/wiki/Kappa_di_Cohen.

[21] Wikipedia contributors. *Krippendorff's alpha*. n.d. URL: https://en.wikipedia.org/wiki/Krippendorff%27s_alpha.

[22] Wikipedia contributors. *Spearman's rank correlation coefficient*. n.d. URL: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[23] Wikipedia contributors. *Valore p*. n.d. URL: https://it.wikipedia.org/wiki/Valore_p.

[24] Magnini Zaninello Brenna. ≪Textual Entailment with Natural Language Explanations: The Italian e-RTE-3 Dataset≫. In: *CLiC-it 2023: 9th Italian Conference on Computational Linguistics*. 2023, pp. 1–5. URL: https://ceur-ws.org/Vol-3596/short21.pdf.

5. Appendice

5.1 Esempi di assegnazione di giudizi e spiegazioni

Giudizio Molnar	Giudizio Poli	Spiegazione Molnar	Spiegazione Poli
3	3	in entrambe le frasi c'è la presenza di una persona che medita	Descrivono persone che riflettono su qualcosa
4	5	Le frasi descrivono i medesimi luoghi	Pressochè identiche
3	2	Entrambe le frasi parlano di amore	Parlano del sentimento dell'amore
2	2	Entrambe parlano di ragni	Menzionano ragni
2	2	Entrambe citano Versilov	Menzionano lo stesso personaggio
3	3	Entrambe parlano di una nave e la sua direzione	Parlano dei movimenti di una nave
4	3	Entrambe descrivono il solito personaggio come un codardo	Menzionano lo stesso personaggio, entrambe dicono che è codardo
3	2	Entrambe parlano di libertà e necessità	Parlano di libertà e necessità
2	2	Entrambe condividono la figura del lacchè	Menzionano il lacchè
3	2	Entrambe le frasi riguardano il dormire	Parlano di qualcuno che si è svegliato dal sonno

Tabella 5.1: Tabella contenente esempi di giudizi e spiegazioni di Molnar e Poli

5.2 BLEU-SCORES

Spiegazioni Molnar	Spiegazioni Poli	BLEU-SCORE
In entrambe le frasi si descrivono nuvole	Descrivono scenari dove c'è una nube a tratti	0.00
In entrambe si parla di tempi molto antichi	Si fa riferimento a un tempo passato	0.00
In entrambe le frasi viene mangiata della zuppa	Fanno riferimento a qualcuno che ha mangiato una zuppa	0.01
Viene menzionato lo stesso personaggio	Entrambe citano il personaggio chiamato Basarow	0.16
In entrambe si parla di avere ragione	Chi parla da ragione a qualcun altro	0.28
In entrambe siamo ad un processo e si richiede una lettura	In entrambe si richiede di leggere qualcosa	0.32
Entrambe parlano di vestiti e di persone sedute su un divano	In entrambe sono presenti persone sedute su un divano che sono vestite bene	0.38
completamente diverse	Completamente diverse	0.50
In entrambe le frasi si parla di mani e faccia	In entrambe le frasi il protagonista si tocca la faccia con le mani	0.53
Entrambe parlano di una benedizione	In entrambe parlano di dare una benedizione	0.57
Entrambe parlano di leggi di natura	Entrambe fanno riferimento alle leggi di natura	0.57
Entrambe parlano di odore	Entrambe parlano di odori	0.75
Completamente diverse	Completamente diverse	1.00

Tabella 5.2: Tabella contenente esempi di BLEU-SCORE

5.3 Risultati SBERT

Spiegazioni Molnar	Spiegazioni Poli	SBERT
Completamente diverse	Parlano di acqua che cade	0,05
Completamente diverse	Parlano di dogmi cristiani	0,13
In entrambe le frasi vi è un ufficiale che urla, solo in una delle due si specifica che è ubriaco.	In entrambe sono coinvolti gradi militari, stessa scena descritta con dettagli diversi	0,30
in entrambe le frasi si parla di gioia e lacrime di gioia	Descrizioni di persone che piangono	0.30
In entrambe le frasi c'è un foglietto con delle scritte	Fanno riferimento a dei foglietti che sono stati scritti	0.46
Descrizione di una stanza contenente mobili in entrambe le frasi	Descrizioni di stanze	0.54
In entrambe le frasi si parla di un agente della polizia	Il soggetto è un agente di polizia	0.67
in entrambe le frasi si parla di una persona molto odiata	Chi parla nutre odio verso qualcuno	0.67
In entrambe le frasi si parla di un argomento militare relativo alla sconfitta dei francesi	Parlano di battaglie francesi avvenute nello stesso luogo	0.71
Entrambe parlano di un debito	Parlano di debiti	0,76
Entrambe parlano di campane che suonano	In entrambe ci sono le campane che suonano	0.84
Entrambe parlano di compiere il proprio dovere	In entrambe chi parla dice di compiere i propri doveri	0.87
Completamente diverse	Completamente diverse	1.00

Tabella 5.3: Tabella contenente risultati di SBERT

5.4 Descrizione delle features ottenute con Profiling-UD

All'interno di questa sezione saranno raggruppate e descritte le features contenute nelle Tabelle 4.1 e 4.2 relative alle guidelines di UD-Pipe [9][19].

- **char_per_tok**: media del numero di caratteri presenti in ciascun token
- **lexical_density**: proporzione tra numero di parole piene rispetto al numero totale di parole nelle frasi. Le parole piene includono tipicamente sostantivi, verbi, aggettivi e avverbi.

Elenco delle distribuzioni:

- **upos_dist_ADP**: distribuzione delle apposizioni
- **upos_dist_DET**: distribuzione degli articoli determinativi
- **upos_dist_PRON**: distribuzione dei pronomi
- **upos_dist_PROPN**: distribuzioni delle proposizioni

Elenco delle distribuzioni, medie e frequenze verbali:

- **verbs_tense_dist_Imp**: distribuzione dei verbi coniugati all'imperfetto
- **verbs_tense_dist_Past**: distribuzione dei verbi coniugati ad un tempo passato
- **verbs_tense_dist_Fin**: distribuzione dei verbi coniugati in forma finita
- **verbs_mood_dist_Cnd**: distribuzione dei verbi al modo condizionale
- **verbs_form_dist_Ger**: distribuzione dei verbi al modo gerundio
- **verbs_form_dist_Fin**: distribuzione delle forme finite dei verbi nel testo
- **aux_tense_dist_Imp**: distribuzione delle ausiliari coniugate all'imperfetto
- **aux_num_pers_dist_Sing+1**: frequenza con cui sono state utilizzate le ausiliari in situazioni in cui il soggetto è nella forma singolare alla prima persona

- **aux_num_pers_dist_Sing+2**: frequenza con cui sono state utilizzate le ausiliari in situazioni in cui il soggetto è nella forma singolare alla seconda persona
- **aux_mood_dist_Ind**: distribuzione delle ausiliari al modo Indicativo
- **aux_form_dist_Inf**: distribuzione delle ausiliari al modo Infinito
- **avg_verb_edges**:media delle connessioni o relazioni tra i verbi in un determinato contesto linguistico.
- **avg_verb_edges_dist_0**: la media delle connessioni dirette a distanza zero tra i verbi
- **avg_verb_edges_dist_1**: la media delle connessioni dirette a distanza uno tra i verbi
- **avg_verb_edges_dist_3**: la media delle connessioni dirette a distanza tre tra i verbi
- **avg_verb_edges_dist_4**: la media delle connessioni dirette a distanza quattro tra i verbi
- **avg_verb_edges_dist_6**: la media delle connessioni dirette a distanza sei tra i verbi
- **verbal_head_per_sent**: frequenza dei nuclei verbali (verbi principali) in relazione al numero di frasi

Elenco dei valori relativi a preposizioni:

- **avg_prepositional_chain_len**: lunghezza media di catene preposizionali nelle frasi
- **prep_dist_2**: valore della distanza preposizionale
- **obj_pre**: Indica che l'oggetto è preceduto da preposizione
- **obj_post**: Indica che l'oggetto è postposto cioè seguito da una preposizione
- **subj_post**: Indica che il soggetto è seguito da una preposizione

Elenco delle distanze tra elementi:

- **dep_dist_case**: distanza tra un elemento dipendente e il suo caso

- **dep_dist_det**: distanza tra un elemento dipendente e il suo determinante
- **dep_dist_expl**: distanza tra un elemento dipendente e una parte espletiva
- **dep_dist_advcl**: distanza tra un elemento dipendente e una proposizione avverbiale
- **dep_dist_root**: distanza tra un elemento dipendente e la radice della struttura sintattica

Elenco delle features relative alle proposizioni:

- **subordinate_post**: proposizione subordinata postposta, indica che la proposizione subordinata segue la proposizione principale.
- **avg_subordinate_chain_len**: lunghezza media delle catene subordinate
- **subordinate_proposition_dist**: distanza o relazione tra proposizioni subordinate
- **subordinate_dist_1**: distanza della proposizione subordinata a livello 1 rispetto alla proposizione principale
- **principal_proposition_dist**: distanza o relazione tra proposizioni principali

5.5 Repository GitHub

All'interno della repository relativa a questo progetto, disponibile al seguente link: `https://github.com/holymolny/Tesi-Molnar-Poli`, è possibile trovare il dataset realizzato in collaborazione con la collega Irene Poli, i programmi utilizzati per svolgere i calcoli e i risultati delle features ottenuti con la piattaforma Profiling-UD.

6. Ringraziamenti

Desidero esprimere la mia sincera gratitudine al Prof. Felice Dell'Orletta e a Chiara Fazzone i quali hanno fornito preziosi consigli, guidato questo percorso di tirocinio ed il successivo sviluppo della tesi con grandissima professionalità.
In particolare ringrazio il Prof. Dell'Orletta per avermi fatto appassionare a tale disciplina durante il corso di Linguistica Computazionale; per la prima volta studiare è stato veramente "divertente".

Un ringraziamento speciale va alla mia collega, nonché grandissima amica, Irene.
Insieme abbiamo preparato e superato, quasi sempre con ottimi voti, la maggior parte degli esami.
Ricorderò sempre e con grande affetto tutte le centinaia di ore passate sulla piattaforma Teams a ripetere e a programmare, ricorderò sempre anche i viaggi in macchina prima di ogni esame e la classica frase "Oioi fra, questo lo boccio" ... Mai ci fu frase più ipocrita.
Grazie Ire, ti voglio bene e spero davvero che il nostro infallibile duo continui anche durante la magistrale.

Infine, dedico un sentito ringraziamento alla mia famiglia che ha reso possibile tutto questo, in particolar modo a mamma che mi ha sostenuto sempre ed in tutto, a Letizia, alla Proff.ssa Simonetta Simone per la sua capacità di trasformare un noioso manuale di letteratura in qualcosa di affascinante, ai miei colleghi che spero tanto riescano a raggiungere i propri obiettivi e soprattutto ai miei amici che conosco da una vita e non cambierei per niente al mondo.
Grazie per il vostro amore, incoraggiamento e comprensione durante questo percorso accademico.

Con affetto,
Mihnea Sever Molnar