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

CORSO DI LAUREA IN **INFORMATICA UMANISTICA**

RELATORI:
FELICE DELL'ORLETTA
CHIARA FAZZONE
GIULIA VENTURI

CANDIDATO:
MIHNEA SEVER MOLNAR

ANNO ACCADEMICO:
2022/2023

UNIVERSITÀ DI PISA

ABSTRACT

IL PRESENTE LAVORO CONSISTE IN UNO STUDIO SULLA SIMILARITÀ SEMANTICA TRA COPPIE DI FRASI.

PER FARE QUESTO È STATO CREATO **DATASET INNOVATIVO** CONTENENTE NON SOLO GIUDIZI NUMERICI MA ANCHE VALORI TESTUALI, SONO STATE APPLICATE DUE **METRICHE AUTOMATICHE** PER VALUTARE LA SIMILARITÀ SEMANTICA TRA COPPIE DI FRASI, SONO STATE SVOLTE **INDAGINI STATISTICHE** SUI RISULTATI OTTENUTI ED INFINE SONO STATE ANALIZZATE LE **CARATTERSTICHE LINGUISTICHE** PRESENTI ALL'INTERNO DEI DATI TESTUALI.

Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

INDICE

- CREAZIONE E ANNOTAZIONE DEL **DATASET**
- APPLICAZIONE DI DUE **METRICHE AUTOMATICHE (BLEU e SBERT)** PER STUDIARE LA PERCEZIONE DELLA SIMILARITÀ SEMANTICACON **STRUMENTI AUTOMATICI**
- UTILIZZO DI **METRICHE STATISTICHE** PER VALUTARE E STUDIARE L'**ACCORDO** E LA **CORRELAZIONE**
- ANALISI DELLE **CARATTERISTICHE LINGUISTICHE**

DOMANDE DI RICERCA

- ESISTE UN **ACCORDO** NELL'ASSEGNARE GIUDIZI DA PARTE DEGLI ANNOTATORI?
- IN CHE MISURA LE DIVERSE METRICHE AUTOMATICHE COLGONO LA SIMILARITÀ SEMANTICA TRA COPPIE DI FRASI? I LORO VALORI **CORRELANO**?
- ESISTE UNA **CORRELAZIONE** TRA I GIUDIZI NUMERICI SULLA SIMILARITÀ SEMANTICA E LE SPIEGAZIONI TESTUALI ATTRIBUITE DAI **DUE ANNOTATORI**?
- CARATTERISTICHE TESTUALI SIMILI CORRISPONDONO A UNA **PERCEZIONE** DELLA SIMILARITÀ SIMILE?

DATASET – PERCHÉ INNOVATIVO?

- LA **QUALITÀ DEL DATASET** È L'ELEMENTO CHIAVE PER OTTENERE RISULTATI AFFIDABILI: IN MACHINE LEARNING CREARE IL DATASET RAPPRESENTA UN'OPERAZIONE **COMPLESSA** E **COSTOSA**

- SI TRATTA DI UN DATASET ANNOTATO NON SOLO CON VALORI NUMERICI MA ANCHE CON **SPIEGAZIONI TESTUALI**

- **SECONDO** IN ITALIANO DOPO «The Italian e-RTE-3 Dataset»

- RENDE IL MODELLO PIÙ **HUMAN-LIKE**: DARE VALORI ESCLUSIVAMENTE NUMERICI NON BASTA PER OTTENERE RISULTATI CONCRETI!

DATASET – DATI INIZIALI

- L'ISTITUTO DI LINGUISTICA COMPUTAZIONALE «A.ZAMPOLLI» HA FORNITO:
  - **CORPUS** DI FRASI **TRADOTTE AUTOMATICAMENTE** E APPARTENENTI ALLA LETTERATURA NARRATIVA
  - **GIUDIZI NUMERICI** SULLA SIMILARITÀ SEMANTICA OTTENUTI IN MODALITÀ **CROWDSOURCING** TRAMITE LA PIATTAFORMA **PROLIFIC** CHE SI BASANO SU **SCALA LIKERT**
- TALI DATI FORNISCONO LA BASE PER LA CREAZIONE DEL **NUOVO DATASET**

Prolific

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 se
1	2	— domandò la signora Valentina, sempre con la me	— domandò Valentina Michailowna sempre con la s
1	3	Sotto la finestra, sulla sabbia del viale e sull'erba de	Malgrado il sole vivido che passava attraverso le fog
1	4	Katavasov invece, che fra le sue occupazioni scienti	Di nuovo i volontari salutarono e sporsero le teste, r
1	5	La prima cosa da fare era procurarsi almeno un po'	«Non ho a chi chiedere denaro in prestito!».
1	6	Ma in quel punto entrò in salotto un uomo di mezza	Era in giubba turchina, calze di seta, scarpini affibb
1	7	Gli davano la preminenza sugli altri la sua età rispe	A questo riguardo egli si considerava piuttosto danr
1	8	- Non è morto però, la pistola non ha sparato.	Stavrògin era in piedi con la pistola in mano, rivolta
1	9	Ivan non si lascia irretire nemmeno da migliaia di ru	E le migliaia di rubli che passano per le sue mani!
1	10	Di botto, si udì uno scoppio di voce sgradevole, e l'u	—proruppe l'ufficiale con furia da ubbriaco.
1	11	Un cameriere tolse il coperchio alla zuppiera; tutti a	«Chi ha chiesto della zuppa?» disse, entrando nella
1	12	E fu preso da una rassegnazione che gli sembrò qua	Aveva impressa in viso una devota rassegnazione al
1	13	Dall'accettazione di questo dogma deriva, come nec	– Ci credi dunque ai dogmi della chiesa?

Giudizio A	Giudizio B	Giudizio C	Giudizio D	Giudizio E
15	15	15	15	15
1	2	1	2	1
4	1	1	3	1
2	3	4	3	3
2	1	2	3	4
2	1	1	1	1
4	4	2	2	3
3	1	1	2	1

DATASET - ANNOTAZIONE

- IL PROCESSO DI **ANNOTAZIONE** SI È SVOLTO NEL SEGUENTE MODO:
  - **LETTURA** DELLA COPPIA DI FRASI
  - ASSEGNAZIONE DEL **GIUDIZIO NUMERICO** SEGUENDO LA SCALA **LIKERT**
  - **SPIEGAZIONE TESTUALE**
  - MONITORAGGIO **TEMPISTICHE** DI ANNOTAZIONE

24 MINUTI A QUESTIONARIO, **45** SECONDI A COPPIA DI FRASI

1 - COMPLETAMENTE DIVERSE
2 - POCO SIMILI
3 - ABBASTANZA SIMILI
4 - MOLTO SIMILI
5 - PRESSOCHÉ UGUALI

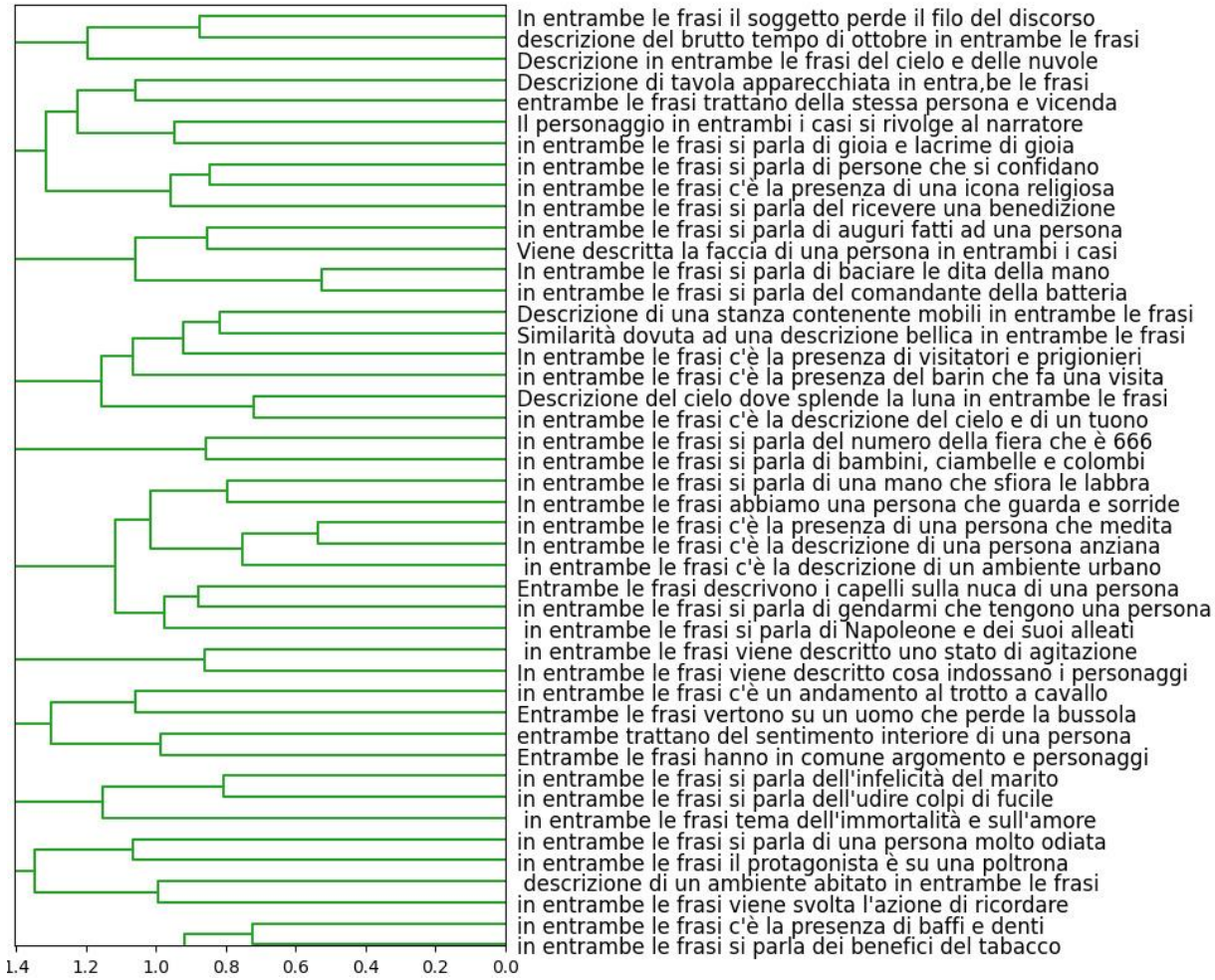
Frase 1	Frase 2
Come se avessi paura della vostra maledizione	La paura è la maledizione dell'uomo…

Giudizio Molnar	Giudizio Poli
2	1

Spiegazione Molnar	Spiegazione Poli
in entrambe le frasi si parla di paura	Completamente diverse

CLUSTERING

- TECNICA UTILIZZATA NELL'AMBITO DELL'APPRENDIMENTO AUTOMATICO E ANALISI DI DATI
- TRAMITE IL CLUSTERING SI COSTRUISCONO **GRUPPI «CLUSTER» DI DATI** ALL'INTERNO DEI QUALI GLI ELEMENTI CONDIVIDONO CARATTERISTICHE COMUNI
- TALE TECNICA È STATA APPLICATA ALLE **SPIEGAZIONI TESTUALI** SULLA SIMILARITÀ SEMANTICA CON LO SCOPO DI CREARE **GRUPPI SEMANTICAMENTE OMOGENEI**
- APPLICANDO TALE TECNICA IL RISULTATO OTTENUTO SEGUE UN ORDINE ESCLUSIVAMENTE **LESSICALE** E NON SEMANTICO A CAUSA DEI DIVERSI **ELEMENTI CIRCOSTANZIALI**



BLEU e SBERT – METRICHE AUTOMATICHE APPLICATE AL TESTO

BLEU

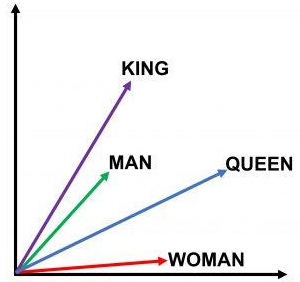
- ALGORITMO SVILUPPATO DA IBM NEI PRIMI ANNI 2000
- NASCE COME STRUMENTO PER **VALUTARE** LE TRADUZIONI AUTOMATICHE
- SI BASA SUL CALCOLO SULLA SOVRAPPOSIZIONE DEGLI **N-GRAMMI** (NEL NOSTRO CASO 1-GRAMMI)

«il cane corre veloce» ⟶ [il, cane, corre, veloce]

«il gatto corre svelto» ⟶ [il, gatto, corre, svelto]

SBERT

- ESTENSIONE DI **BERT**, ALGORITMO SVILUPPATO PER GOOGLE NEL 2018
- COINVOLGE L'UTILIZZO DI RETI NEURALI E SFRUTTA MODELLI DI LINGUAGGIO AVANZATI
- CODIFICA L'INFORMAZIONE TESTUALE IN **VETTORI** E CALCOLA LA **DISTANZA DEL COSENO** TRA QUESTI



BLEU e SBERT - DATI

Frase 1	Frase 2	BLEU-SCORE	SBERT-SCORE
È il diavolo che ci mette la coda!	Dove non arriva l'ingegno, arriva il diavolo! pensò con uno strano sogghigno.	0,08	0,43
— Che vi costa a voi?...	— Avete fatto una bella prodezza, in verità!	0,12	0,17

Spiegazione Molnar	Spiegazione Poli	BLEU-SCORE	SBERT-SCORE
Entrambe le frasi citano la figura del diavolo	Completamente diverse	0,00	0,03
Completamente diverse	Completamente diverse	1,00	1,00

METRICHE STATISTICHE APPLICATE A DATI NUMERICI: KRIPPENDORFF

α DI KRIPPENDORFF:

- CALCOLA L'**ACCORDO** TRA GLI ANNOTATORI CHE VIENE RAGGIUNTO DURANTE UNA **OPERAZIONE DI ASSEGNAMENTO** DI VALUTAZIONI O CATEGORIE AD UN INSIEME DI DATI
- I RISULTATI DI TALE METRICA SONO COMPRESI IN UN RANGE CHE VA DA **-1** A **1**
- PERMETTE DI TENERE IN CONSIDERAZIONE UN **ELEVATO NUMERO DI ANNOTATORI**
- AMMETTE CRITERI DI VALUTAZIONE ORDINALI COME QUELLI DELLA **SCALA LIKERT**

→
- ACCORDO PERFETTO = **1**
- ACCORDO CASUALE = **0**
- DISACCORDO = **-1**

ESISTE UN ACCORDO NELL' ASSEGNARE GIUDIZI DA PARTE DEGLI ANNOTATORI?

α MOLNAR-POLI	α ANNOTATORI PROLIFIC	α MOLNAR, POLI E ANNOTATORI DI PROLIFIC
0.43	0.45	0.42

- GENERALMENTE UN RISULTATO **ACCETTABILE** SI HA DA **α = 0.67** IN SU, **BUONO** DA **α = 0.8** IN SU
- NEL NOSTRO CASO DOBBIAMO ESSERE SODDISFATTI DEL RISULTATO POICHÈ STIAMO PARLANDO DI GIUDIZI SEMANTICI QUINDI **SOGGETTIVI** ED **INDIVIDUALI**
- NON ESISTONO REGOLE PRECISE PER ASSEGNARE TALI VALORI
- **ESISTE UN ACCORDO NELL'ASSEGNARE GIUDIZI DA PARTE DEGLI ANNOTATORI!**

METRICHE STATISTICHE APPLICATE A DATI NUMERICI: SPEARMAN e P-VALUE

COEFFICIENTE DI CORRELAZIONE DI SPEARMAN

- TECNICA STATISTICA CHE PERMETTE DI VALUTARE LA CORRELAZIONE TRA DUE VARIABILI PRESE IN INPUT
- IL RISULTATO OTTENUTO SPAZIA TRA **-1** E **1** → CORRELAZIONE POSITIVA = **1**; CORRELAZIONE NEGATIVA = **-1**; ASSENZA DI CORRELAZIONE = **0**

P-VALUE:

- MISURA CHE SERVE AD INDICARE IL GRADO DI **SIGNIFICATIVITÀ** DEL RISULTATO
- SI DEFINISCE UN VALORE SOGLIA **α** CHE GENERALMENTE È POSTO A **0.05**
- SE **p > α** ALLORA IL RISULTATO OTTENUTO NON È SIGNIFICATIVO

SPEARMAN e P-VALUE: RISULTATI OTTENUTI

CASO	CORRELAZIONE DI SPEARMAN	P-VALUE
BLEU, SBERT (FRASI)	0.39	9.39 e-39
BLEU, SBERT (SPIEGAZIONI)	0.83	8.78 e-264
|MOLNAR – POLI|, BLEU	-0.47	2.06 e-57
|MOLNAR – POLI|, SBERT	-0.51	6.78 e-69
MOLNAR, POLI (GIUDIZI)	0.56	6.57 e-88

SPEARMAN e P-VALUE: RISPOSTE ALLE DOMANDE DI RICERCA

IN CHE MISURA LE DIVERSE METRICHE AUTOMATICHE COLGONO LA SIMILARITÀ SEMANTICA TRA COPPIE DI FRASI? I LORO VALORI **CORRELANO**?

Sì, i loro valori correlano! Abbiamo ottenuto un punteggio di **0.39** sulla correlazione tra BLEU e SBERT applicati alle frasi e **0.83** applicandoli alle spiegazioni spiegazioni.

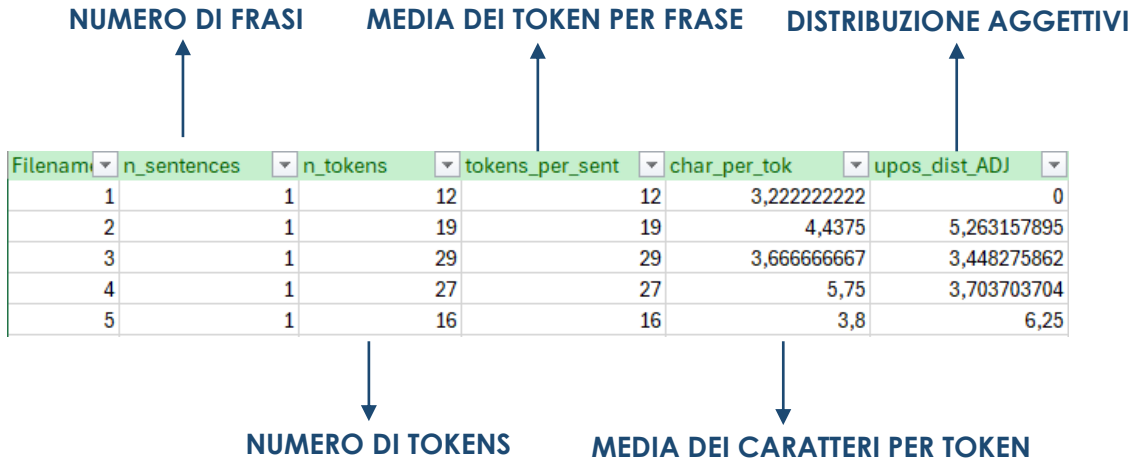
ESISTE UNA **CORRELAZIONE** TRA I GIUDIZI NUMERICI SULLA SIMILARITÀ SEMANTICA E LE SPIEGAZIONI TESTUALI ATTRIBUITE DAI **DUE ANNOTATORI**?

Sì! La risposta positiva si evince dai risultati ottenuti: **-0.47** considerando BLEU e **-0.51** considerando SBERT

ATTENZIONE: I RISULTATI NEGATIVI INDICANO UNA RELAZIONE **INVERSAMENTE PROPORZIONALE** TRA I PARAMETRI PRESI IN INPUT DALLA METRICA STATISTICA DUNQUE IN QUESTO CASO PIÙ SI AVVICINANO A -1 E PIÙ ALTA È LA CORRELAZIONE.

ANALISI LINGUISTICA

- PER ESTRARRE LE **CARATTERISTICHE LINGUISTICHE** È STATO UTILIZZATO **PROFILING-UD** (BRUNATO ET AL., 2020)
- TALE PIATTAFORMA PERMETTE L'ESTRAZIONE DI OLTRE **130** CARATTERISTICHE APPARTENENTI A DIVERSI LIVELLI DELL'ANALISI LINGUISTICA COME: **NUMERO TOTALE DI FRASI** DI CUI È COMPOSTO IL TESTO, **NUMERO TOTALE DI CARATTERI**, **TTR** PER VERIFICARE LA DIVERSITÀ LESSICALE PRESENTE NEL TESTO, **DISTRIBUZIONI** (VERBI, NOMI, AGGETTIVI), **FREQUENZE**, **MEDIE** E TANTO ALTRO
- PROFILING-UD SI BASA SUL FRAMEWORK **UNIVERSAL DEPENDENCIES** QUINDI È **MULTILINGUE**
- PROFIGLING-UD È STATO UTILIZZATO PER ANALIZZARE LE **COPPIE DI FRASI**



Filenam	n_sentences	n_tokens	tokens_per_sent	char_per_tok	upos_dist_ADJ
1	1	12	12	3,222222222	0
2	1	19	19	4,4375	5,263157895
3	1	29	29	3,666666667	3,448275862
4	1	27	27	5,75	3,703703704
5	1	16	16	3,8	6,25

ANALISI LINGUISTICA - RISULTATI

CARATTERISTICHE TESTUALI SIMILI CORRISPONDONO A UNA **PERCEZIONE** DELLA SIMILARITÀ SIMILE?

PER RISPONDERE:

1. SONO STATE **ESTRATTE TUTTE LE FEATURES** DI **«FRASI 1»** E **«FRASI 2»** CALCOLANDO LA DIFFERENZA IN VALORE ASSOLUTO TRA QUESTE
2. È STATO CALCOLATO **SPEARMAN ED IL P-VALUE** TRA I **VALORI** DELLE FEATURES OTTENUTE ED I **GIUDIZI** PRIMA DI MOLNAR POI DI POLI

TUTTI I RISULTATI OTTENUTI CON SPEARMAN E AVENTI UN P-VALUE INFERIORE A 0.05 SONO PROSSIMI ALLO 0, LA RISPOSTA QUINDI È NEGATIVA

Features	Spearman_corr	P_value
principal_proposition_dist	-0,06155428	0,048930285
verbs_num_pers_dist_Sing+	-0,063081072	0,043576789
dep_dist_cc	0,066225299	0,034094891
upos_dist_CCONJ	0,06952484	0,02609668
verbal_head_per_sent	-0,069938242	0,025219025
subordinate_proposition_di	-0,070390145	0,024288929
dep_dist_nummod	0,071619151	0,021908106
avg_links_len	0,071625456	0,021896435
subj_post	-0,07340717	0,018807204
dep_dist_obj	0,073579701	0,018529325
aux_mood_dist_Ind	-0,073636897	0,018438
dep_dist_iobj	-0,07511347	0,016212686
upos_dist_VERB	0,076785512	0,013980479

SVILUPPI FUTURI

AMPLIARE IL DATASET ED IL NUMERO DI ANNOTATORI

TROVARE UN METODO EFFICACE PER FARE **CLUSTERING SEMANTICO**

ADDESTRARE UN MODELLO PER ESEGUIRE TASK DI ANNOTAZIONE CHE CONSISTONO NELLA CAPACITÀ DI ATTRIBUIRE GIUDIZI E DETERMINARE SPIEGAZIONI CORRELATE IN MANIERA AUTOMATICA

CONCLUSIONI

ESISTE UN **ACCORDO** NELL'ASSEGNARE GIUDIZI DA PARTE DEGLI ANNOTATORI?

Sì! Lo abbiamo visto applicando Krippendorff ed ottenendo un risultato positivo.

IN CHE MISURA LE DIVERSE METRICHE AUTOMATICHE COLGONO LA SIMILARITÀ SEMANTICA TRA COPPIE DI FRASI? I LORO VALORI **CORRELANO**?

Sì! I valori ottenuti applicando BLEU e SBERT sia sulle coppie di frasi che sulle spiegazioni sulla similarità semantica correlano. Lo abbiamo visto dai risultati del coefficiente di correlazione di Spearman.

ESISTE UNA **CORRELAZIONE** TRA I GIUDIZI NUMERICI SULLA SIMILARITÀ SEMANTICA E LE SPIEGAZIONI TESTUALI ATTRIBUITE DAI **DUE ANNOTATORI**?

Sì! Lo abbiamo verificato applicando Sperman dando come parametri la differenza in valore assoluto tra i giudizi numerici dei due annotatori ed i risultati di BLEU e SBERT sulle spiegazioni semantiche.

CARATTERISTICHE TESTUALI SIMILI CORRISPONDONO A UNA **PERCEZIONE** DELLA SIMILARITÀ SIMILE?

NO! Stando ai calcoli statistici eseguiti i risultati ci hanno portato a dare una risposta negativa.

GRAZIE PER LA VOSTRA ATTENZIONE!