DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA MAGISTRALE IN DATA SCIENCE

GESTIONE DI DATI STRUTTURATI E NON STRUTTURATI

-

PROGETTO DATI TESTUALI

Docente:

Prof. Corrado Loglisci

STUDENTE:

Walter Mangione

Anno Accademico 2024-2025

ABSTRACT

Questo lavoro si concentra sulla gestione e sull'analisi di dati testuali. Partendo da testi in forma grezza, viene effettuata una fase di pre-processing volta a trasformare i dati, seguita da un'analisi finalizzata all'estrazione di informazioni utili per distinguere tra testi generati da esseri umani e testi generati da algoritmi di Intelligenza Artificiale.

Il dataset utilizzato è composto da due colonne: la prima, denominata *text*, contiene i testi grezzi; la seconda, *generated*, indica se il testo è stato creato da un umano oppure da un modello di Intelligenza Artificiale. Alla fine del percorso, è stato osservato che vi sono, per alcune statistiche, delle differenze sostanziali tra testi scritti da umani e testi generati da modelli. Per altre statistiche, invece, le differenze si assottigliano tanto da poter essere considerate trascurabili.

Infine, presentiamo uno strumento capace di riconoscere il dominio in questione, avente l'obiettivo di categorizzare il dominio di una frase classificandola come potenzialmente generata o scritta da un essere umano. Si tratta dell'utilizzo del modello *VADER*, un semplice lessico, la cui modifica dei pesi, effettuata da noi mediante un approccio guidato dai dati, ha fornito buoni risultati, con una accuratezza pari al 78% nel riconoscimento di una categoria.

PREPROCESSING E COSTRUZIONE DEL SURROGATO

Il dataset, inizialmente, conteneva circa 30 mila osservazioni, le quali sono state ridotte a 700 per questioni legate a risorse computazionali, pur sempre mantenendo un bilanciamento fedele al dominio originale. Quindi, una volta selezionato casualmente il sottoinsieme, questo viene elaborato trasformando tutto il testo in minuscolo. Quindi, sono state espanse le negazioni, tipiche dei Paesi Anglosassoni (ad esempio, *can't* diventa *can not*).

Una volta effettuato questo, sono stati trattenuti solo caratteri che fossero lettere, numeri, spazi ed apostrofi.

A questo punto, la sequenza di token è stata trasformata in un insieme di misure quantitative che descrivono la ricchezza e la varietà **lessicale** del testo. Per comprendere meglio di cosa stiamo parlando, immaginiamo di voler capire non solo *quanti* termini ci sono, ma anche *quanti diversi* ne compaiono, quanto sono lunghi in media e con quale frequenza compaiono parole rare o ripetute. Oltre a questo, è stato effettuato uno step che include il processo di **POS-Tagging**. Da questo step, quindi, sono state estratte due informazioni fondamentali: la categoria grammaticale più grossolana e la sotto-categoria più fine. Per intenderci, potremmo avere la categoria *NOUN* per i sostantivi e *VERB* per i verbi, come prima informazione. Invece, come seconda informazione potremmo ottenere la categoria più precisa. Per esempio, *NN* e *NNS* (per distinguere un sostantivo singolare da uno al plurale), *VBD* e *VBP* (per distinguere un verbo al passato da uno al presente). Oltre ad aver salvato queste informazioni, sono state estratte le distribuzioni delle categorie stesse. In altri termini, sono state estratte le percentuali, per ogni documento, dei nomi, dei verbi, degli aggettivi e degli avverbi. Infine, sono stati estratti i bigrammi più ricorrenti all'interno di ogni documento. Si è scelto di estrarre i primi 5 bigrammi più frequenti all'interno di ogni documento.

Quindi, è stata effettuata una analisi **sintattica** del testo, mediante il **deep parsing.** È stato costruito un albero di dipendenze, in cui ogni token (tranne la radice) dipende da un "governatore", che da ora in poi chiameremo *head*. Questo albero, dunque, riflette la struttura logica della frase: chi è soggetto di chi, chi modifica chi, quali sono le subordinate, chi è l'oggetto riferito al predicato nominale, ed altre simili informazioni sintattiche. Ad ogni token, poi, viene assegnata una *dependency label*, come, per esempio, *nsubj* (soggetto nominale), *dobj* (oggetto diretto), *amod* (aggettivo modificatore), *root* (radice

principale della frase, significa, cioè, che quel token è l'elemento centrale della frase). Questa informazione ci permette, quindi, di distinguere il ruolo grammaticale di ciascuna parola nel contesto della frase. La tabella seguente mostra un aspetto leggermente più tecnico dei campi estratti.

Campo estratto	**Descrizione**	**Case**
dep_	Etichetta di dipendenza (ruolo grammaticale).	Analisi dei ruoli (soggetto, oggetto, modificatori).
head.text	Testo della *head*.	Ricostruzione delle coppie *head–dipendente.*
head_pos	POS della *head*.	Verifica di pattern come "NOUN→VERB" o "ADJ→NOUN".
is_root	Booleano. Indica se il token è la radice sintattica	Conteggio delle frasi.
depth	Profondità del token nell'albero delle dipendenze.	Misura della complessità sintattica (vengono distinte le frasi semplici da quelle annidate).
num_children	Numero di figli **diretti** del token.	Identificazione dei nodi più ricchi (quali verbi o sostantivi complessi). Un verbo con molti figli, ad esempio, può avere soggetto, oggetto, avverbi, subordinate, suggerendo una costruzione ricca.

Infine, è stata creata una tabella denominata *named_entities*, che raccoglie alcune informazioni semantiche relative al dataset originale. La tabella, precisamente, contiene le seguenti informazioni:

Campo estratto	**Descrizione**
doc_id	Identificativo numerico univoco del documento da cui è stata estratta l'entità.
label	Etichetta assegnata al documento (ad esempio, se il testo è stato generato o meno).
entity_text	Testo esatto dell'entità riconosciuta (ad esempio, "New York", "Microsoft", "giovedì").
entity_label	Tipo o categoria dell'entità (ad esempio, PERSON per persona, ORG per organizzazione).
start_char	Posizione (indice) del carattere iniziale dell'entità all'interno del testo.
end_char	Posizione (indice) del carattere finale dell'entità all'interno del testo.
sentence_idx	Indice della frase (in termini di posizione di token) in cui si trova l'entità nel documento.

Quindi, in *Postgres*, sono state create 4 tabelle principali, il cui compito è memorizzare appunto questi stessi metadati e, su questi, effettuare delle query. I dettagli delle query sono presenti nel relativo file.

Pertanto, il nostro surrogato è composto dalle seguenti 4 tabelle, che descrivono il testo, chiaramente con una inevitabile perdita di informazione: la prima tabella, *lexical_tokens*, contiene informazioni a livello di singolo token (parola) per ogni documento analizzato ed ogni riga rappresenta un token con le sue caratteristiche **lessicali** e **sintattiche**. In questa tabella, quindi, vi sono le informazioni relative al grafo costruito dallo step di **deep parsing**. Invece, *lexical_stats* contiene statistiche aggregate a livello di documento, dove ogni riga descrive un singolo documento, con misure globali del suo contenuto lessicale. La tabella *named_entities*, invece, raccoglie tutte le **entità semantiche** riconosciute nei testi analizzati, ed ogni riga della tabella rappresenta una singola entità identificata all'interno di un documento, con informazioni semantiche e contestuali: infatti, questa contiene la posizione dell'entità all'interno del documento. Infine, *pos_bigrams* contiene la frequenza dei 5 bigrammi **POS (part-of-speech)** più diffusi nei documenti. Ogni riga rappresenta la frequenza di un bigramma ordinato di *tag* grammaticali in un determinato documento.

ANALISI SEMANTICA: UN PICCOLO TOOL PER IL RICONOSCIMENTO DEL DOMINIO

L'ultimo task, invece, riguarda la classificazione del dominio in questione, con l'obiettivo di costruire un piccolo strumento capace di categorizzare il dominio di una frase classificandola come potenzialmente generata o scritta da un essere umano. Per affrontare il compito di classificazione, è stato impiegato ***VADER*** (*Valence Aware Dictionary and sEntiment Reasoner*), un modello di analisi del sentiment progettato specificamente per testi informali come quelli presenti sui social media, tipicamente, *basato su regole,* il quale assegna punteggi di intensità a parole e simboli comuni in contesti testuali. Quindi, possiamo immaginare di avere un nuovo **surrogato**, con ogni parola associata ad un peso che ne descrive la *polarità,* in un certo senso. Tuttavia, abbiamo preferito aggiornare manualmente il vocabolario associato ai pesi del modello, in quanto questo ha fornito dei risultati migliori; infatti, è stato effettuato un *fine-tuning* del lessico: abbiamo analizzato la distribuzione delle parole all'interno dei testi appartenenti alle due classi (label 'HUMAN' e 'AI'), ed abbiamo calcolato la frequenza media normalizzata dei termini per ciascuna classe. Tali frequenze vengono poi combinate in un dizionario che rappresenta la differenza tra le due distribuzioni, ponderata tramite gli *iperparametri alpha* e *beta*, e questa viene poi utilizzata per aggiornare dinamicamente il dizionario di ***VADER***. In questo modo, il modello viene adattato al dominio specifico, migliorando la sua sensibilità lessicale e la capacità di discriminare tra contenuti umani e generati; infatti, non è più un semplice modello general-purpose. Per ulteriori dettagli, si rimanda al codice allegato.

Pertanto, il processo di analisi lessicale è stato effettuato nuovamente, in funzione del nuovo task. Illustriamo di seguito i passi effettuati in questa fase (si noti che non è stato effettuato un passaggio di analisi sintattica):

1. **Rimozione del simbolo hash (#)**
   Eventuali hashtag sono stati privati del simbolo #, in modo da isolare le parole chiave in essi contenute e renderle più facilmente interpretabili dal lessico;
2. **Filtraggio dei caratteri non alfabetici**
   È stata mantenuta solo una selezione ristretta di caratteri: lettere dell'alfabeto (maiuscole e minuscole) e alcuni segni di punteggiatura.

Tutti gli altri simboli (inclusi numeri, caratteri speciali e simboli di punteggiatura non specificati) sono stati rimossi per ridurre il rumore nel testo;

3. **Rimozione degli URL**
   Tutti i collegamenti ipertestuali (URL) che iniziavano con la stringa http sono stati eliminati, poiché ritenuti non informativi ai fini della classificazione;
4. **Rimozione delle stopword**
   Sono state eliminate le stopword, in modo da migliorare la rilevanza semantica delle caratteristiche estratte.

Alcune operazioni tipiche di *preprocessing*, seppur implementate, non sono state utilizzate, in quanto si è osservato empiricamente che comportavano un peggioramento delle performance del modello. In particolare, la conversione di parole dal plurale al singolare, la rimozione di parole non riconosciute come inglesi e lo *stemming*. Queste trasformazioni sono state quindi escluse dalla pipeline finale al fine di preservare il contenuto informativo originario dei testi e ottimizzare l'efficacia del classificatore.

RISULTATI

Utilizziamo adesso qualche visualizzazione per renderci conto dei risultati di alcune queries. La figura 1, visualizzata mediante il risultato di una query, mostra come vi sia una percentuale più elevata di coppie *aggettivo-nome* nei testi generati da AI. Invece, nei testi scritti da umani è più frequente la presenza di coppie del tipo *determinante-nome*, più raro nei documenti generati da modelli. Più frequenti, invece, sono le coppie *nome-preposizione* nei documenti generati da modelli, rispetto ai documenti scritti da esseri umani.

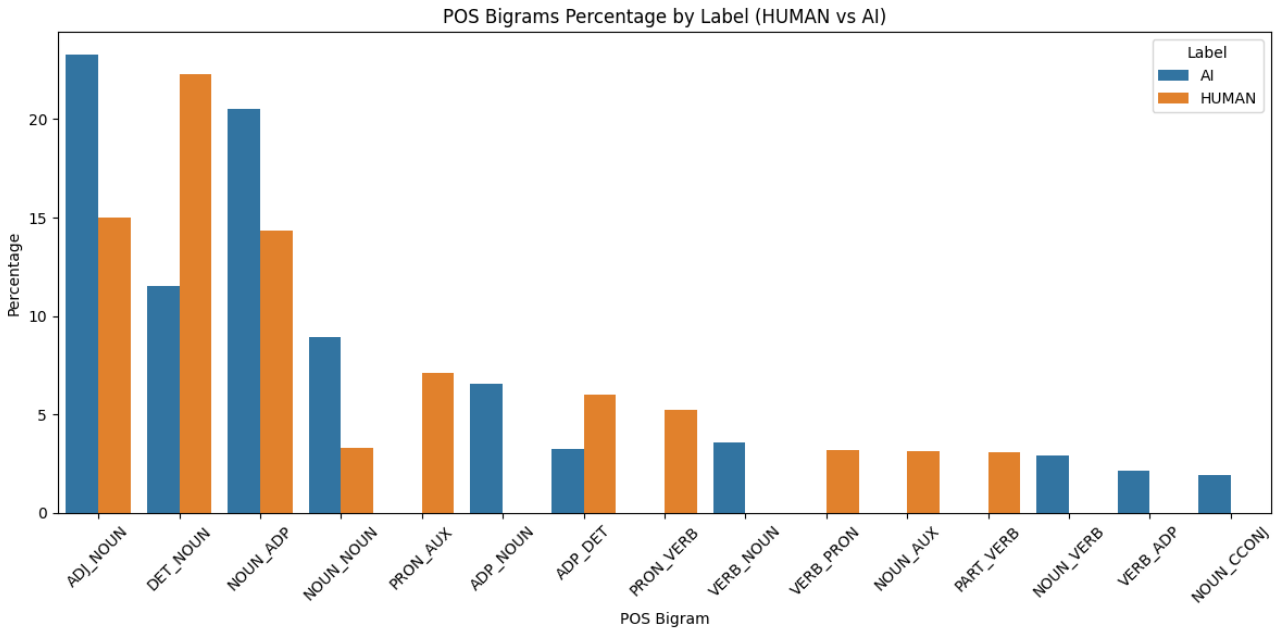

Fig.1 – Bigrammi per le due classi

I primi bigrammi della visualizzazione, quindi, mostrano delle differenze significative tra testi generati e testi redatti da umani.

Inoltre, è emerso che i testi generati da modelli di Intelligenza Artificiale siano meno ripetitivi, in quanto hanno un rapporto più alto tra unicità di token e token totali. Anche, questi tendono a fare uso di parole più rare all'interno del testo stesso.

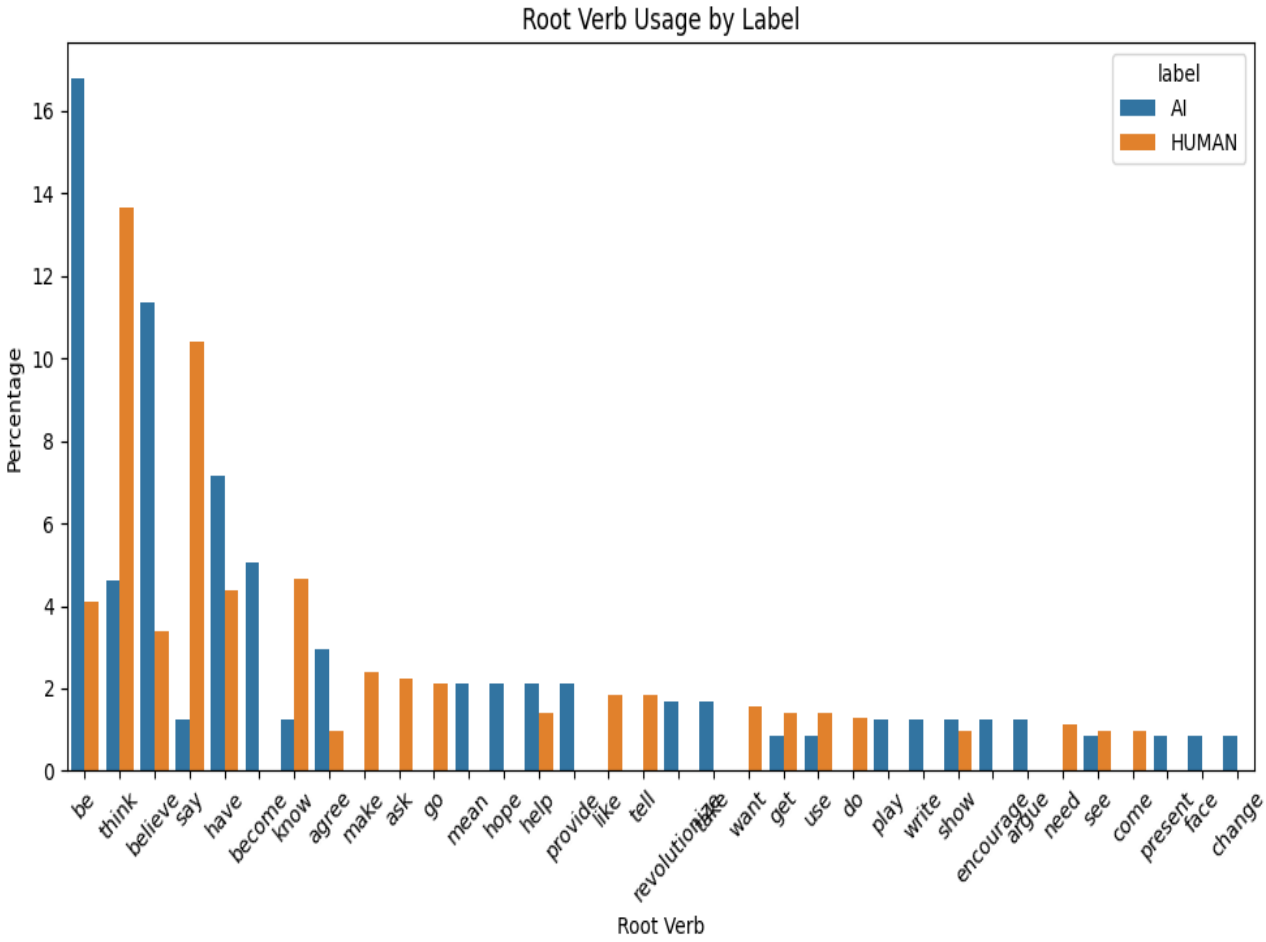

Fig.2 – Frequenza dei verbi utilizzati come radici all'interno dell'albero.

Il grafico in questione, invece, evidenzia come alcuni verbi siano utilizzati perlopiù come *radici* da testi generati anziché da testi scritti da esseri umani. È possibile visualizzare anche la situazione inversa. È curioso vedere come il verbo *say*, in tutte le sue forme declinate venga espresso come radice in testi scritti da umani e raramente in testi generati. Invece, il verbo *be* è più presente in testi generati, usato come radice, mentre il verbo *think* in testi scritti da umani.

Invece, non è stata osservata nessuna differenza sostanziale nell'uso di nomi, aggettivi, verbi ed avverbi al variare delle classi. Anche, non è stata osservata nessuna differenza significativa nell'uso semantico delle entità, al variare della classe.

Per quanto riguarda l’applicazione del modello ***VADER***, si sono ottenuti buoni risultati, con un’accuratezza del 78.5% ed un valore di F1-Score pari al 64.7%.

CONCLUSIONI

Il lavoro ha comunque permesso di poter estrarre caratteristiche capaci di discriminare un testo generato da un testo non generato, senza tuttavia una distinzione netta che permetta di tracciare un *solco* tra le due classi. Tuttavia, si potrebbe, in futuro, cercare di utilizzare strumenti più potenti (ad esempio, si potrebbe fare uso di estrattori di entità basati su modelli molto più sofisticati). Per quanto riguarda l'applicazione del modello ***VADER***, potrebbe essere interessante valutare quanto le prestazioni possano migliorare modificando gli *iperparametri*, obiettivo che, però, va al di là del nostro lavoro.