FRONTESPIZIO

1. Introduzione

Per questo progetto di Data Analytics affronterò il problema espresso nella traccia 3, ovvero la rilevazione del sarcasmo. Questo problema è uno dei più interessanti nell’ambito di NLP (Natural Language Processing), per la sua complessità; causata della sua natura soggettiva e delle sfumature linguistiche coinvolte; infatti, non è semplicissimo neanche per un essere umano. In particolare, il sistema dovrà esaminare un testo, e aiutandosi con degli elementi esterni (il contesto), per capire se esso è sarcastico o no.

Nella ricerca sono state adottate diverse strategie per questo task, divisibili in due categorie:

* Basati sulle caratteristiche linguistiche: in cui si analizzano i testi alla ricerca di pattern ligustici legati al sarcasmo; come l’uso di più parole contradditorie nella stessa frase, o rispetto al contesto, l’uso di contrasti tra parole con connotazione positiva e negativa oppure l’uso di esagerazioni.
* Basati su modelli di Machine Learning o Deep Learning: in cui si sfruttano tecniche di apprendimento supervisionato (o semi) per addestrare modelli statistici con dati etichettati per rilevare i testi sarcastici e non; generalmente i modelli posso essere più “semplici” come SVM, Random Forest Naive Bayes, o semplici reti neurali, fino ad arrivare ai modelli di Deep Learning, come RNN (Recurrent Neural Network) o i Transfomer che possono raggiungere e superare i milioni di parametri.

In questo progetto, la strategia risolutiva adottata si baserà sul training supervisionato con i Transformer; in quanto sono un modello che sta rivoluzionando il settore, e sarà interessante applicarli in primo piano (seppur in una scala ridotta) a questo difficile task.

La relazione sarà divisa in tre capitoli principali: nel primo descriverò l’analisi dei dati effettuata, da cui si baseranno le scelte adottate; nel secondo tratterò il modello usato e il relativo addestramento; nel terzo discuterò i risultati prodotti.

1. Capitolo 1: Analisi dei dati

In questo capitolo descriverò l’analisi dei dati di addestramento, esaminando gli elementi del dataset, il target della classificazione, i testi e gli elementi del contesto. Con cui poter trarre delle conclusioni per poter processare i dati e addestrare il modello con una maggior accortezza.

* 1. Descrizione generale dei dati
     1. Introduzione ai dataset e Splitting

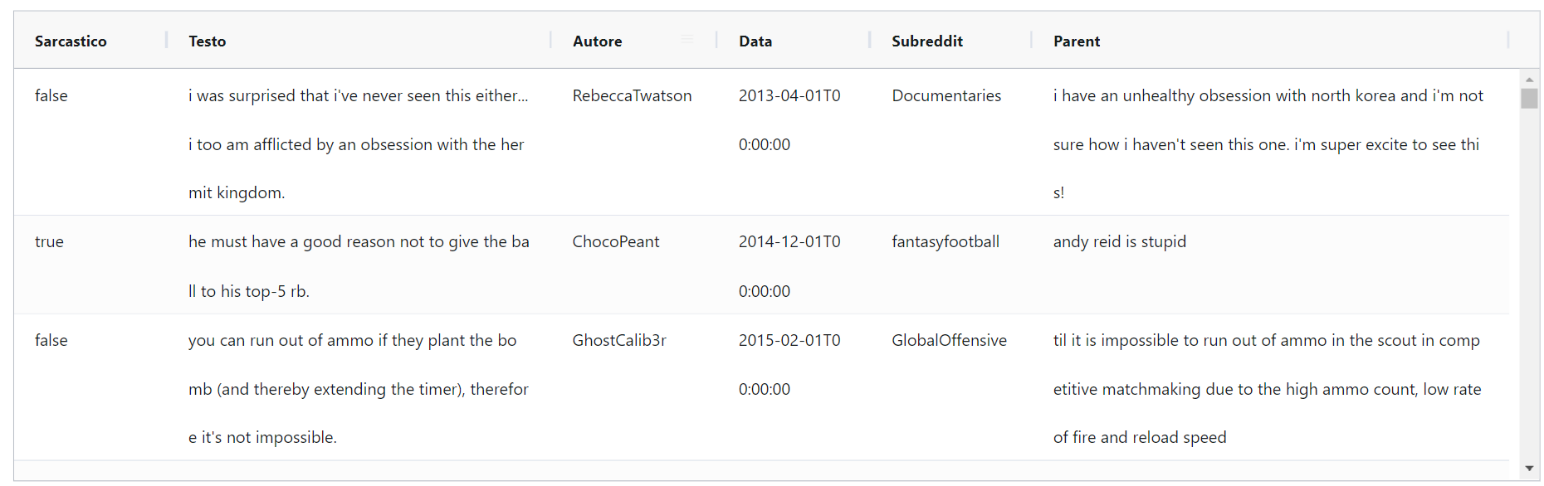
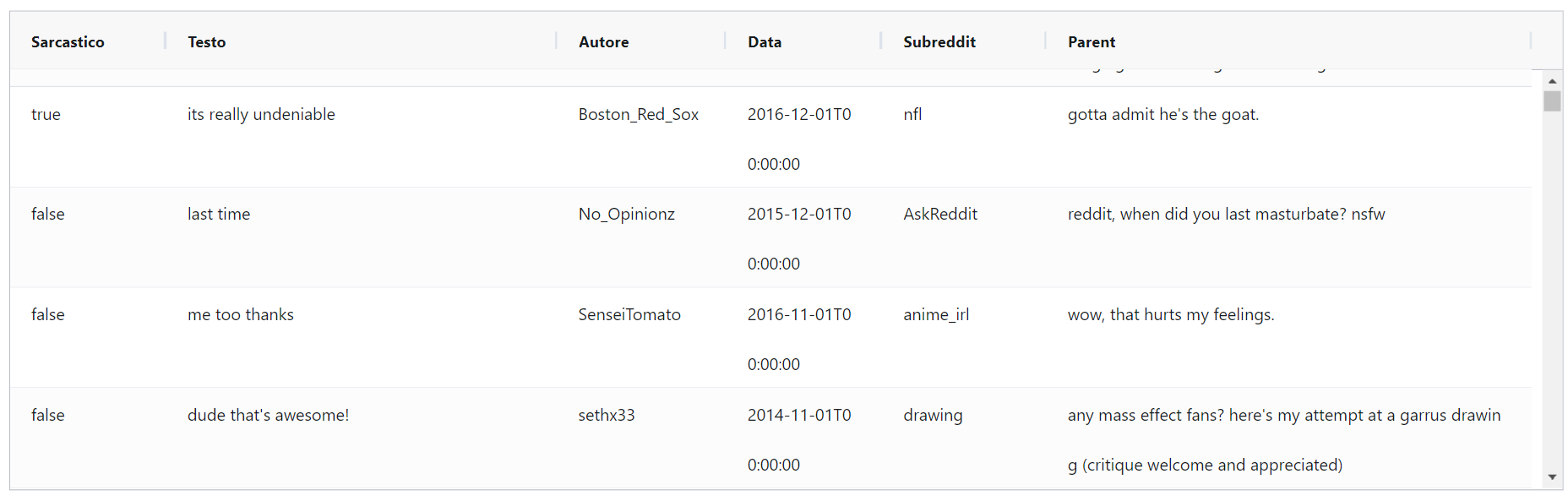
Alla traccia, sono associati due dataset, uno con i dati di addestramento (train) e uno per il testing. Essi sono due insiemi di commenti di Reddit dal 2009 al 2016; in particolare, per ogni istanza abbiamo:

* Text: Testo del commento da classificare;
* Sarcastic: Label booleana che indica se il commento è sarcastico (True) o no (False);
* Author: Nome dell’autore del commento;
* Subreddit: Nome del subreddit in cui è stato pubblicato il commento;
* Date: Data di pubblicazione del commento;
* Parent: Testo della discussione a cui il commento si riferisce;

Il dataset di training originale, con un milione di righe (precisamente 1010822), sarà diviso casualmente in due parti: quello di training definitivo (con il 95%), su cui verranno effettuate tutte le analisi e assunzioni; quello di validation (con il restante 5%), usato durante il processo di training per misurare le performance del modello.

Il dataset di training è molto completo, perché il numero di righe duplicate e non valide è relativamente irrilevante (rispettivamente 200 e 53); dunque non è necessario attuare alcuna strategia particolare per recuperare i dati mancanti (come l’imputazione) o nulli, e si può procedere con la loro rimozione. Inoltre, sia la codifica dei testi che delle date rispetta gli standard e non sono dunque presenti errori di conversione.

Figura 1: Alcune righe di training di esempio



* + 1. Descrizione della metodologia di analisi

Ogni riga è composta da una label e dalla relativa istanza, i cui attributi possono essere divisi nel testo fondamentale da classificare (l’attributo Text), e una serie di attributi contestuali che aggiungono informazioni per il testo (i restanti attributi).

Vista questa divisione ho strutturato il processo di analisi in due fasi:

* Nella prima fase esaminerò gli attributi del contesto, calcolando per ogni elemento unico di ogni attributo qual è il rateo con cui compare in testi sarcastici e non (detto Rateo Informativo), in modo da individuare gli elementi (e quindi i relativi attributi) più rilevanti ad un modello
* Nella seconda fase esaminerò i diversi modi processare il testo, calcolando quale sistema produce il testo con un maggior contento di rateo informativo.
  1. Analisi del target

Il target, ovvero l’attributo Sarcastic, è la Label che identifica la natura del commento, e sarà dunque l’obbiettivo del modello di NLP.

Per la buona riuscita del training è importante che la sua distribuzione sia bilanciata, ovvero che il numero di istanze sarcastiche e non sia approssimativamente simile. Come sostenuto nell’articolo “A Large Self-Annotated Corpus for Sarcasm” (allegato alla traccia), e dopo una fase di verifica, posso affermare che il dataset presenta questa caratteristica e quindi non è necessario applicare tecniche per migliorare il bilanciamento.

Immagine che contiene testo, schermata, diagramma

Descrizione generata automaticamente

Figura 2: Distribuzione del target ottenuta dopo la fase di verifica, come si può notare il bilanciamento è quasi perfetto

* 1. Analisi delle feature contestuali
  2. Analisi del testo
  3. Preparazione dei dati