



SARCASM DETECTION

Un approccio basato sui Transformer

A cura di Poli Luca.

The image features two thick, black L-shaped brackets. One is positioned on the left side, with its vertical bar extending downwards and its horizontal bar extending to the right. The other is on the right side, with its vertical bar extending upwards and its horizontal bar extending to the left. These brackets frame the central text.

ANALISI DEI DATI

INTRODUZIONE AL DATASET

- Raccoglie commenti di Reddit, dal 2009 al 2016.
- Ogni commento è descritto da 5 variabili, di cui una fondamentale, e le altre definite contestuali.
- L'obiettivo è la previsione della label.
- Il dataset è lungo 1010 mila righe.
- 200 righe duplicate.
- 53 righe nulle.
- Verrà diviso tra al 95% tra Train e Validation.

Sarcastico	Testo	Autore	Data	Subreddit	Parent
false	i was surprised that i've never seen this either... i too am afflicted by an obsession with the her mit kingdom.	RebeccaTwatson	2013-04-01T0 0:00:00	Documentaries	i have an unhealthy obsession with north korea and i'm not sure how i haven't seen this one. i'm super excite to see thi s!
true	he must have a good reason not to give the ba ll to his top-5 rb.	ChocoPeant	2014-12-01T0 0:00:00	fantasyfootball	andy reid is stupid
false	you can run out of ammo if they plant the bo mb (and thereby extending the timer), therefor e it's not impossible.	GhostCalib3r	2015-02-01T0 0:00:00	GlobalOffensive	til it is impossible to run out of ammo in the scout in comp etitive matchmaking due to the high ammo count, low rate of fire and reload speed

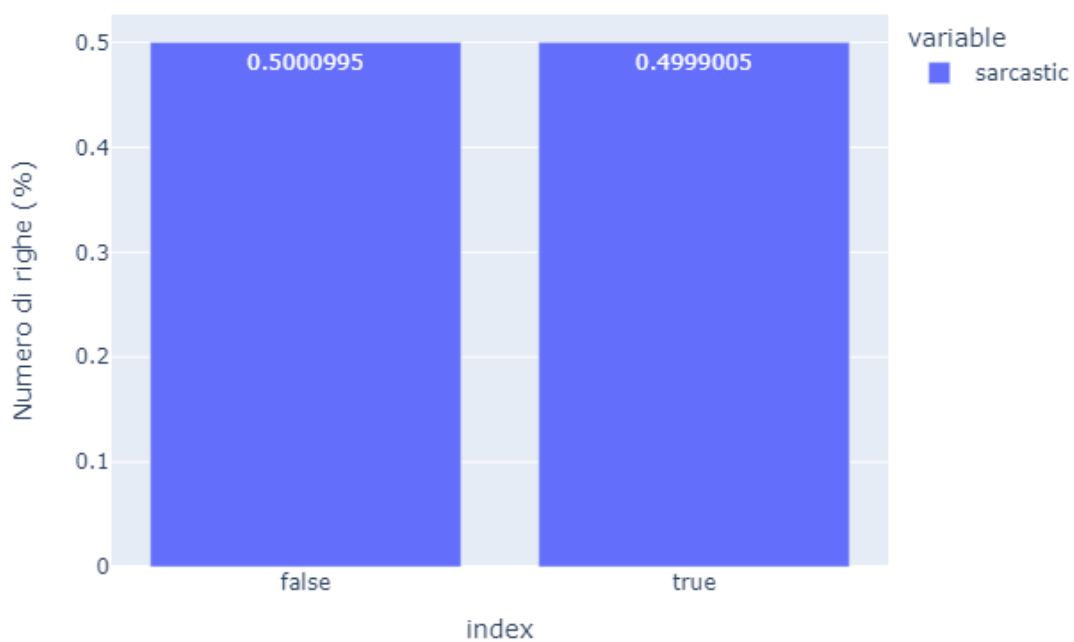
DESCRIZIONE METODOLOGIA DI ANALISI

- Il processo di analisi si divide in due fasi principali e una terza extra:
 1. Analisi degli elementi del contesto.
 2. Analisi dei tipi di testo.
 3. Analisi delle lunghezze del commento e del parent.
- In ogni fase lo strumento matematico chiave è il Rateo Informativo (IR):
 - È calcolato per ogni elemento (e) unico, per ogni feature.
 - Rappresenta quanto quell'elemento può essere rilevante per la previsione

$$IR_e = \left| \frac{n_{sc}}{n_{sc} + n_{nsc}} - S_{IR} \right| * 100$$

(Dove: n_{sc} e n_{nsc} sono il numero di commenti sarcastici e non; S_{IR} è invece la probabilità della classe più probabile)

ANALISI DEL TARGET

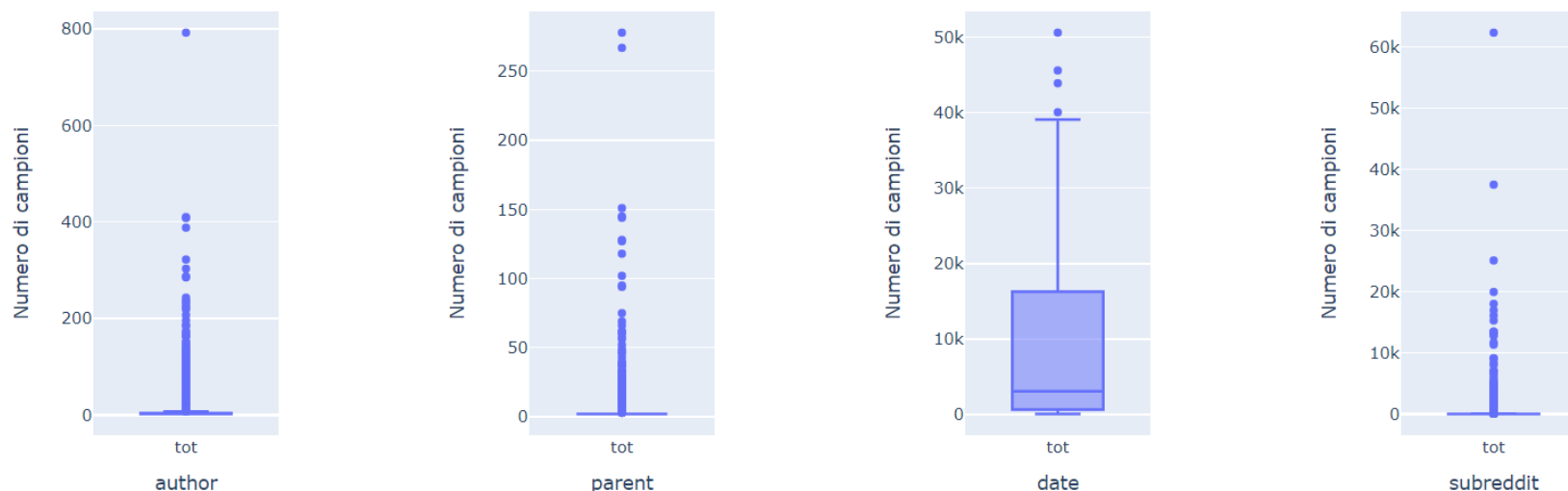


- Fase necessaria per verificare che il target sia bilanciato
- Il target è ben bilanciato
- Non è necessario effettuare bilanciamenti
- $S_{IR} \approx 0.5001$.

ANALISI DEL CONTESTO



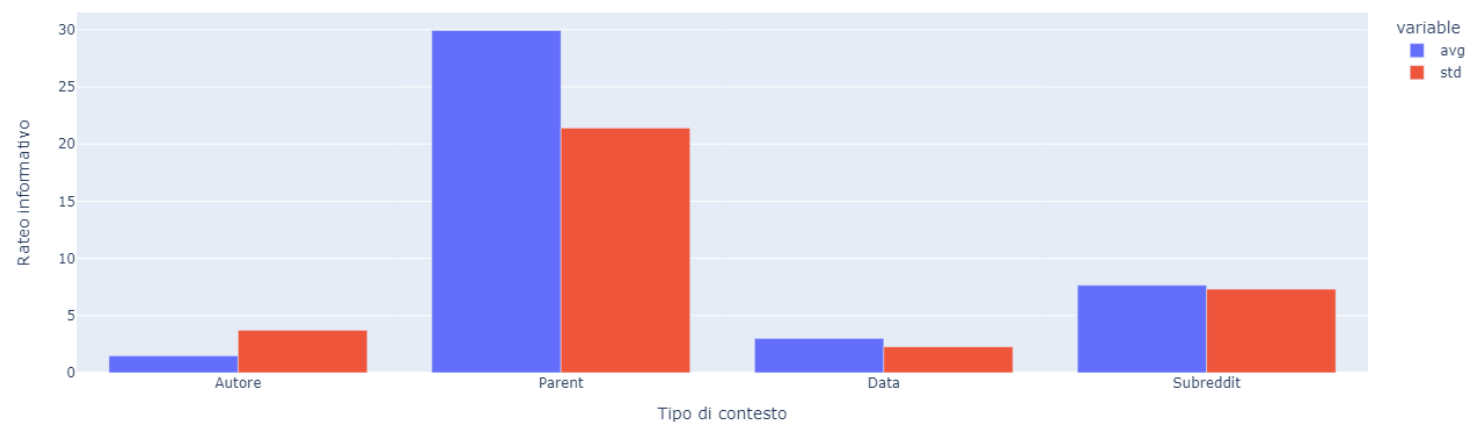
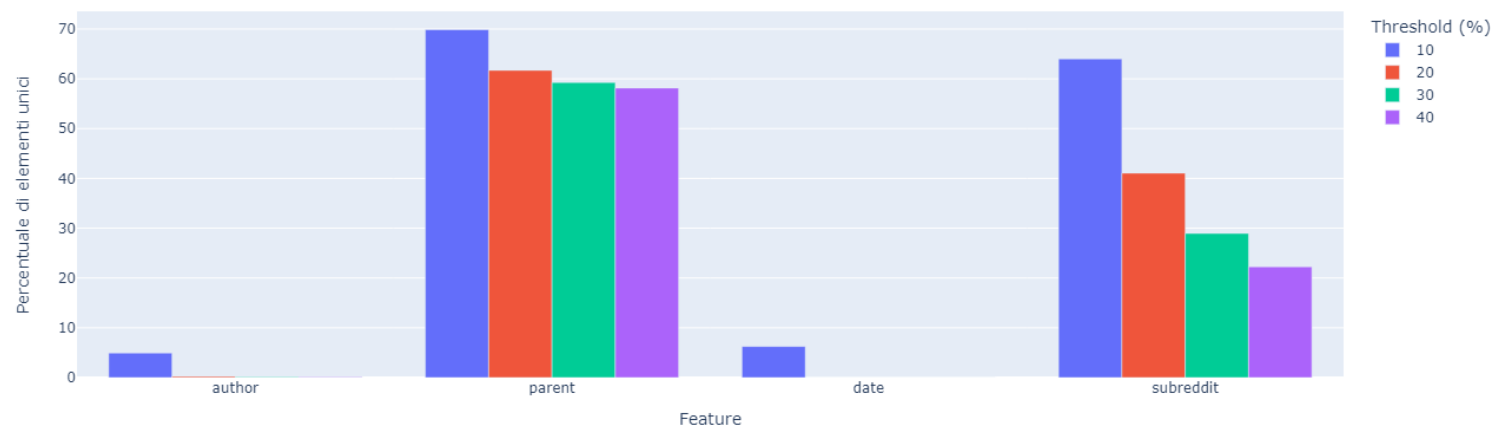
ANALISI DEL CONTESTO: NUMERO DI ELEMENTI



feature	Elementi unici (%) associati ad un unico testo	Elementi unici totali (%)	Elementi unici totali
author	7.03	26.67	256071
parent	98.38	97.28	933914
date	0	0.01	96
subreddit	40.65	1.52	14545

- Quasi ogni commento si riferisce ad un parent unico.
- Le date si ripetono con un'alta frequenza.
- Autore e subreddit potrebbero essere i più rilevanti.

ANALISI DEL CONTESTO: RATEO INFORMATIVO



- Data e Autore hanno un basso rateo informativo.
- Parent ha il rateo più alto, ma ciò è causato dal basso numero di elementi ripetuti.
- Subreddit presenta le caratteristiche migliori.

ANALISI DEL CONTESTO: CONCLUSIONI

- Autore presenta un numero non eccessivo di elementi unici, tuttavia essi non hanno abbastanza rateo informativo.
- Data ha ancora meno elementi unici, ma con un rateo informativo insignificante.
- Parent ha un alto rateo, ma un numero eccessivo di elementi unici.
- Subreddit presenta un numero adeguato di elementi unici e un buon rateo; verrà dunque impiegato dal modello.

ANALISI DEL TESTO

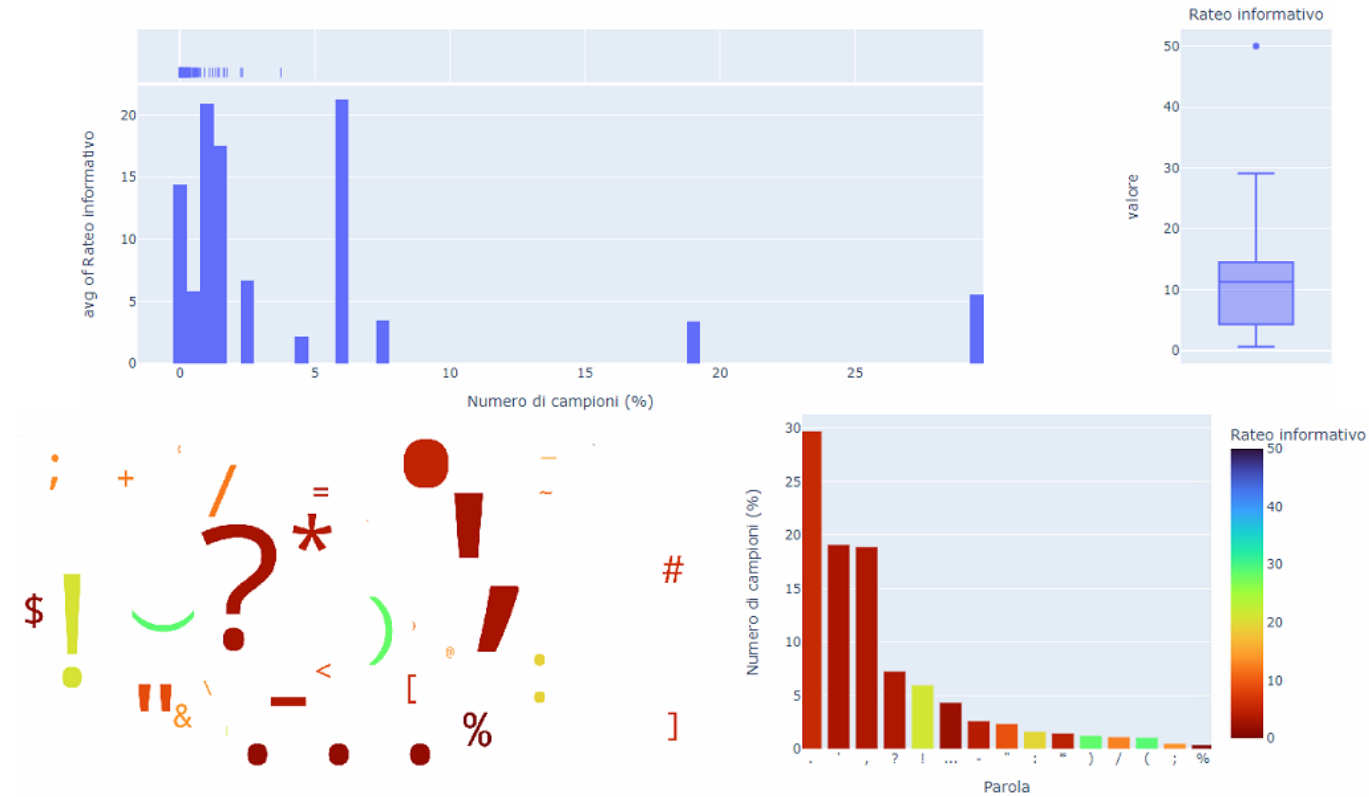


ANALISI DEL TESTO: I VARI TIPI

- 4 Tipi di testo:
 - Normale;
 - Senza stopwords;
 - Con stemming;
 - Senza stopwords e con stemming
- Prima dell'analisi bisogna effettuare la tokenizzazione.
- Ad ogni testo è stata rimossa la punteggiatura.
- Ogni testo verrà analizzato, secondo il rateo informativo dei propri token con 4 grafi: Istogramma, Boxplot, Wordcloud e Barplot.

Normale	Senza Stopwords	Stemming	Senza Stopwords e Stemming
the old one or the old old one	old one old old one	the old on or the old old on	old on old old on
hard drives never crash on macs this is a well established fact	hard drives never crash macs well established fact	hard driv nev crash on mac thi is a wel est fac t	hard driv nev crash mac wel est fact
yeah you are probably right but i guess we will see :p	yeah probably right guess see :p	yeah you ar prob right but i guess we wil see : p	yeah prob right guess see :p
ideally we'd circle back to when employees were knowledgeable about their products they could assist people in making the best choice	ideally we'd circle back employees knowledge able products could assist people making best choices instead mindless human drones swipe code past scan take more	id we'd circ back to when employ wer knowledg about their produc they could assist peopl in mak the best cho instead of just being mindless hum dron swip upc cod past scan tak more	id we'd circ back employ knowledg produc could assist peopl mak best cho instead mindless hum dron swip upc cod past scan tak more

ANALISI DEL TESTO: LA PUNTEGGIATURA



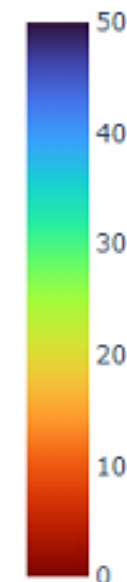
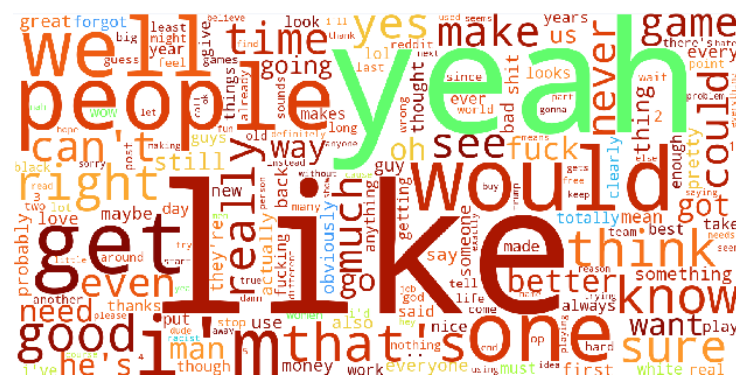
- La maggior parte dei punti, soprattutto i più frequenti, hanno un rateo informativo molto basso
- L'unico segno significativo è il '!': il 71% delle frasi in cui compare è sarcastico e la frequenza del 6%.

ANALISI DEL TESTO: WORDCLOUDS

Testo normale



Testo senza stopwords



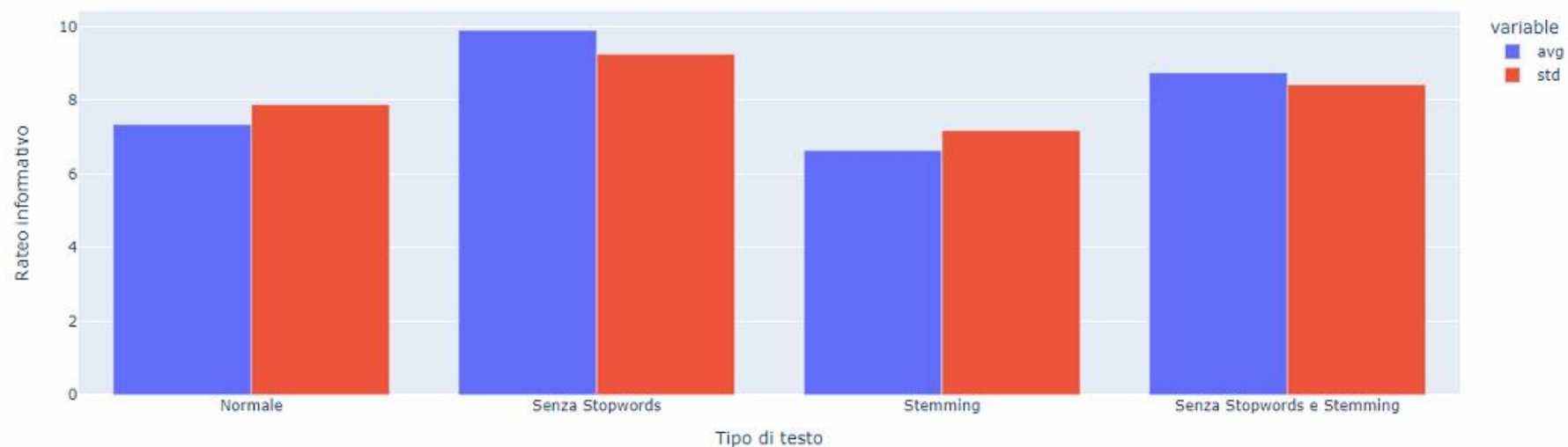
Testo con stemming



Testo senza stopwords e stemming

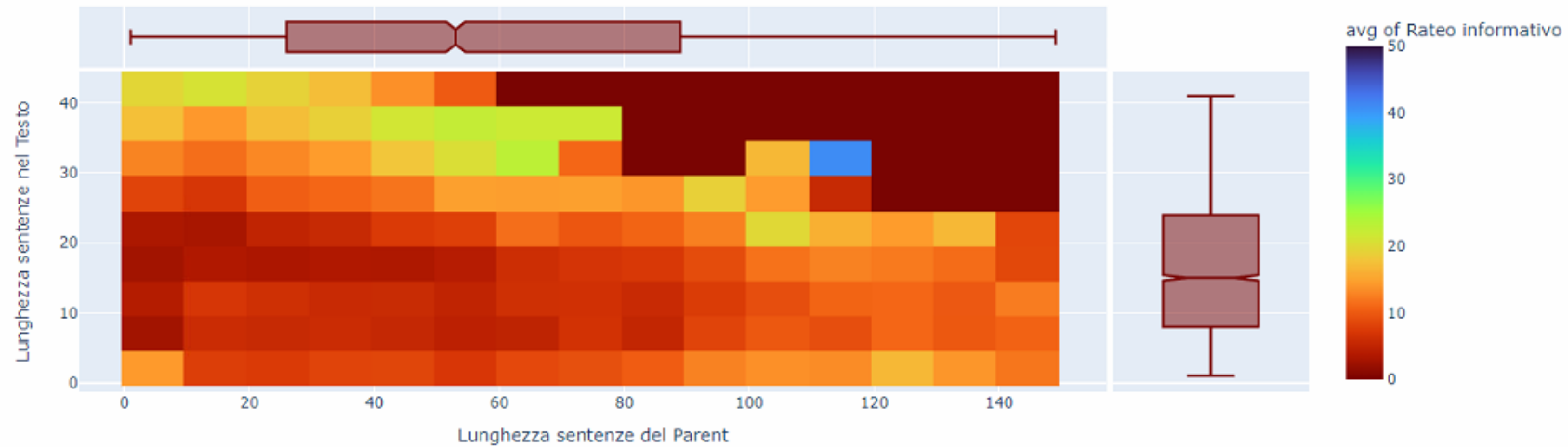


ANALISI DEL TESTO: CONCLUSIONI



- In media il rateo informativo fra i vari tipi non varia di molto.
- La rimozione delle stopwords ha l'impatto più alto.
- Lo stemming ha una rilevanza discutibile, in quanto porta ad un lieve peggioramento.
- Il testo scelto è quello «Senza Stopwords», ha il miglior rateo senza un'elevata perdita di informazione

ANALISI DELLE LUNGHEZZE

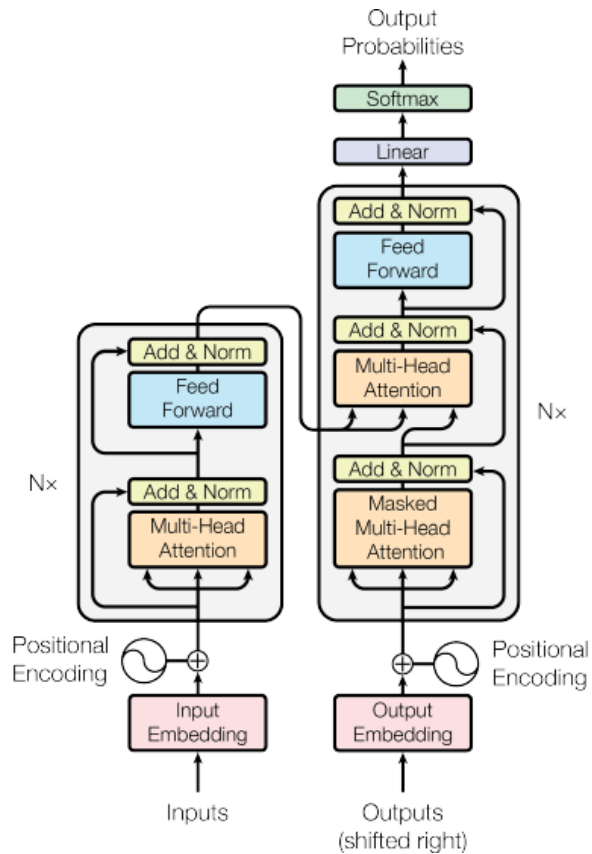


- Sono state filtrate le coppie di lunghezze poco significative.
- L'area a rateo 0% è causata dal filtraggio dei campioni.
- Il rateo informativo medio non è molto alto, tranne in alcune aree.
- Ho scelto di aggiungere, comunque, queste variabili in quanto non richiedono molti parametri al modello.



MODELLO

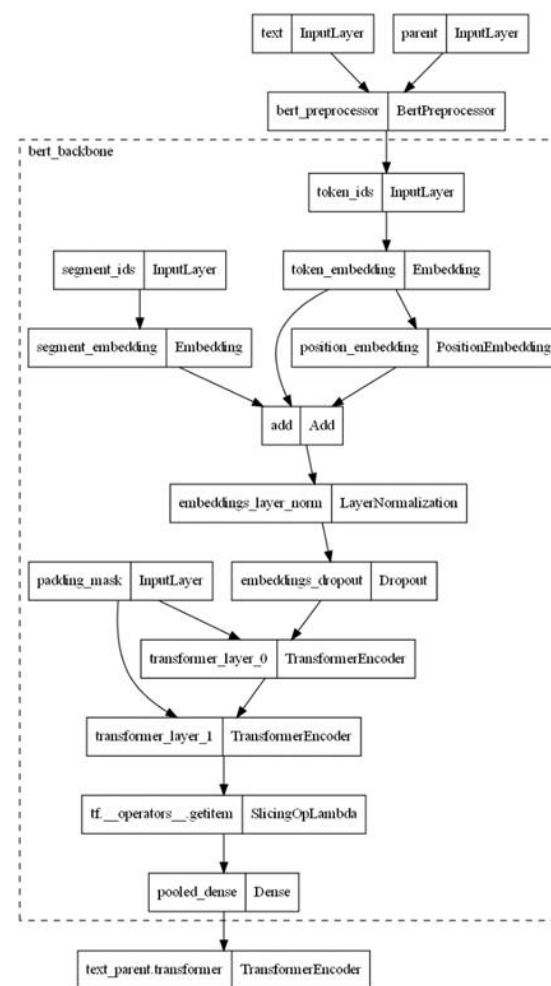
MODELLO: BASATO SUI TRANSFORMER



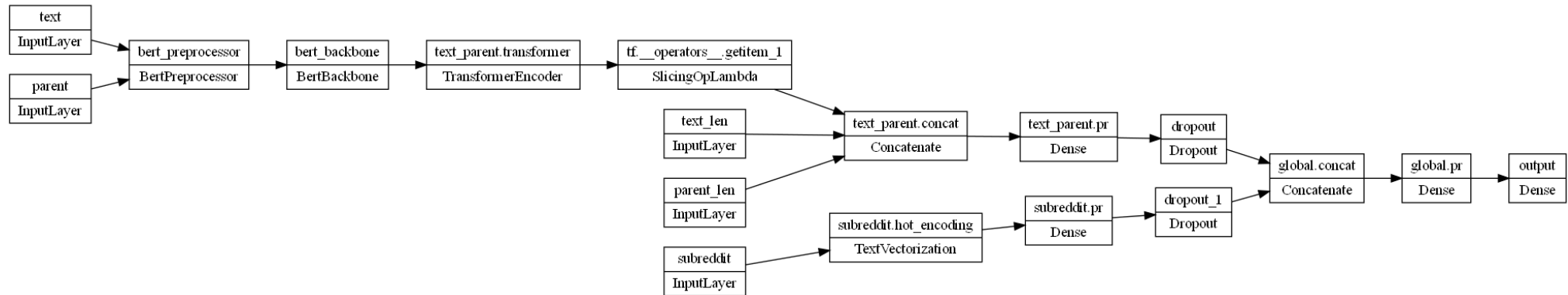
- Sono stati introdotti nel 2017 con il paper "Attention Is All You Need" da Vaswani et al.
 - Non hanno una struttura sequenziale (a differenza di RNN e LSTM).
 - Sono in grado di catturare le contestuali tra le parole.
-
- BERT è un modello Pre-Addestrato, composto da multipli TransformerEncoder.
 - Permette di estrarre relazioni tra una o due sentenze.

MODELLO: PRIMA PARTE

- Sfrutta BERT.
- Processa simultaneamente testo e parent.
- Prima utilizza il preprocessor di BERT e poi il Backbone pre-addestrato.
- Ha 4.39M parametri.
- Il suo risultato viene restituito ad un altro TransformerEncoder
 - Con 131K parametri
 - Effettua il fine-tuning



MODELLO: SECONDA PARTE



- Elabora separatamente il subreddit: effettuando prima il multi_hot encoding, per poi elaborare il risultato. Ha 100K parametri.
- Concatena autonomamente il risultato della prima parte con le lunghezze del parent e del testo; per poi elaborare tale input.
- Elabora successivamente i due risultati.
- Con l'ultimo neurone restituisce la probabilità di appartenenza alla classe.

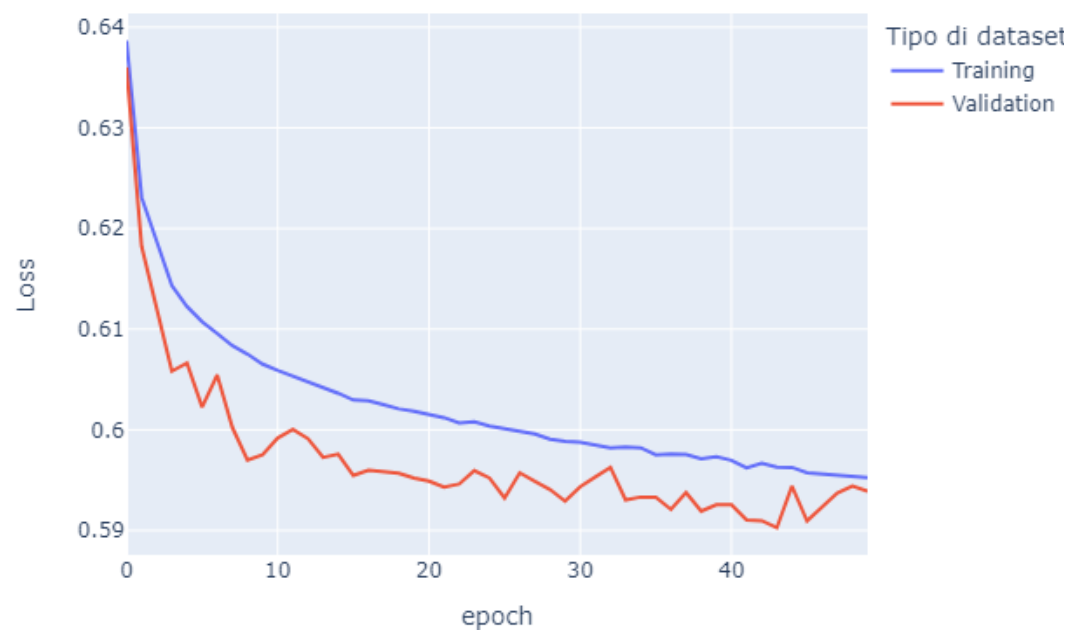


RISULTATI

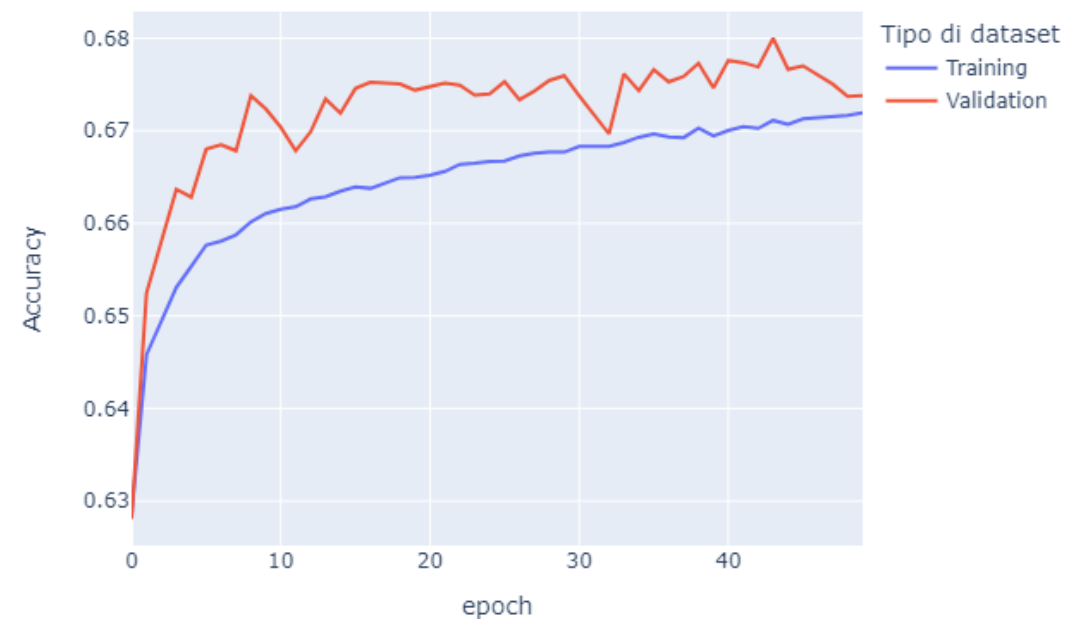


RISULTATI: HISTORY DI TRAINING

Funzione di Loss per epoca

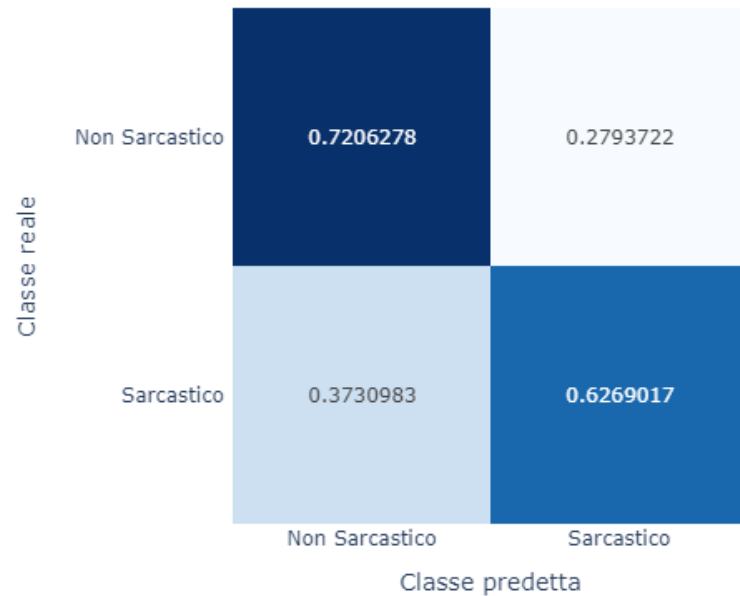


Precisione per epoca



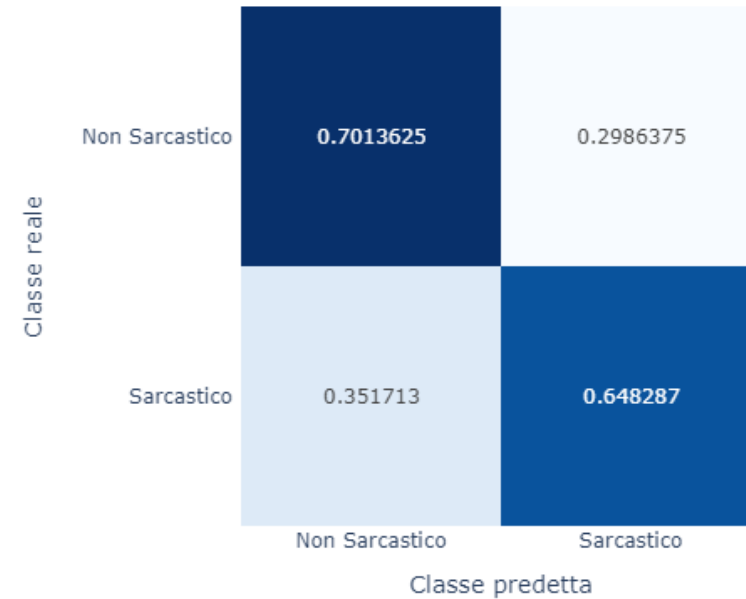
RISULTATI: HISTORY DI TRAINING

Confusion Matrix sul Validation



Accuracy di validation: 67.4%

Confusion Matrix sul Test



Accuracy di test: 67.5%

- Il modello ha un alto tasso di falsi positivi, rispetto a quelli di falsi negativi; cioè, tende maggiormente a classificare positive istanze negative rispetto al contrario

CONCLUSIONI FINALI

- Le analisi effettuate, nel primo capitolo, sono state uno strumento utile per migliorare la precisione del modello.
- Il modello mostra ottime capacità di generalizzazione, ma anche un certo grado di underfitting.
- Nel futuro, con migliori risorse computazionali, si potrebbe effettuare tuning degli iperparametri.