IMPIEGO DI TECNICHE DI NLP PER LA PROFILAZIONE AUTOMATICA DEL GENERE DEGLI AUTORI DI MESSAGGI TESTUALI.

Autore: Fabio Sposato

Relatore: Mirko Lai

L’elaborazione del linguaggio naturale

La NLP fa parte della sotto branca dell’informatica chiamata Intelligenza Artificiale

Per il suo funzionamento ha bisogno di grandi quantità di dati.

I Task di cui si occupa possono essere molti tra cui: Hate Speech Detection, Machine Translation

Author Profiling

L'Author Profiling fa parte di un insieme di Task di cui si occupa la NLP

Per profilazione si intende il cercare di capire le caratteristiche principali della persona che, come in questa tesi, ha scritto un determinato testo.

Ricerca le caratteristiche dell'autore quali: Sesso, Età, Tratti di Personalità ecc...

Gender Detection

In modo particolare, la tesi ha affrontato il Gender Detection come Task binario.

Il Gender Detection è un Sub-Task dell'Author Profiling e si occupa in modo più specifico della rilevazione del Sesso degli Autori.

È stato scelto di intraprendere solo il Task binario di Gender Detection in quanto più semplice rispetto agli altri Task multi-classe.

Obiettivo della tesi

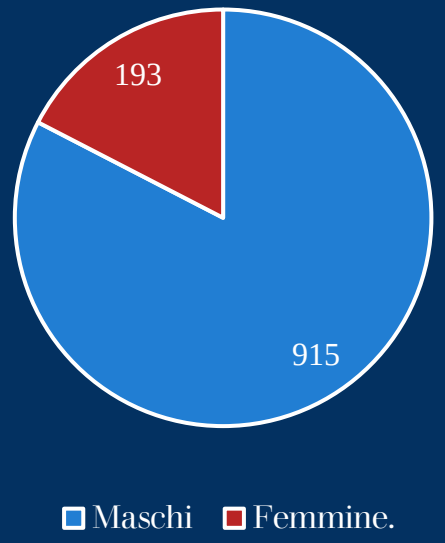
L'obiettivo della tesi era focalizzato sulla creazione di un modello automatico supervisionato per l'identificazione del genere di un autore di un testo.

Come è stato Fatto?

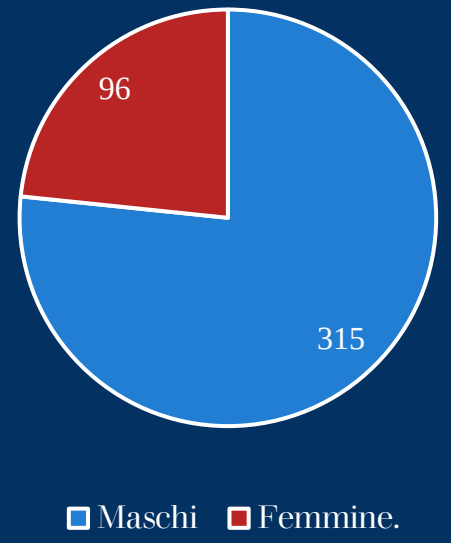
Dataset utilizzato

Per svolgere la Tesi di ricerca basata sull'Author Profiling, si è fatto riferimento al dataset di TAG-it, Shared Task organizzato da EVALITA nel 2020.
Dataset formato da testi di più autori ricavati dalla piattaforma ForumFree.

Training Set



Test Set



K Fold Validation

Esattamente come i partecipanti allo Shared Task ho affrontato la fase di sviluppo senza utilizzare il Test Set

È stata quindi utilizzata la suddivisione del Training Set creata dalla K-Fold per addestrare i modelli in modo iterativo.

BOW

Rappresentazione tramite Vettore BOW

Quante volte il token viene ritrovato nel dizionario

Id	ciao	città	casa	mare	…	nero	macchina	io	bello
1	0	0	1	0	…	4	0	0	0
…	…	…	…	…	…	…	…	…	…
12345	0	1	0	3	…	0	1	2	0

BOP

Rappresentazione tramite Vettore BOP

Quante volte la Part Of Speech viene ritrovata nel dizionario

Dizionario più piccolo

Id	ADJ	ADV	INTJ	NOUN	…	ADP	PRON	VERB	PUNCT
1	0	0	1	0	…	4	0	0	0
…	…	…	…	…	…	…	…	…	…
12345	0	1	0	3	…	0	1	2	0

GDF o Gender Function

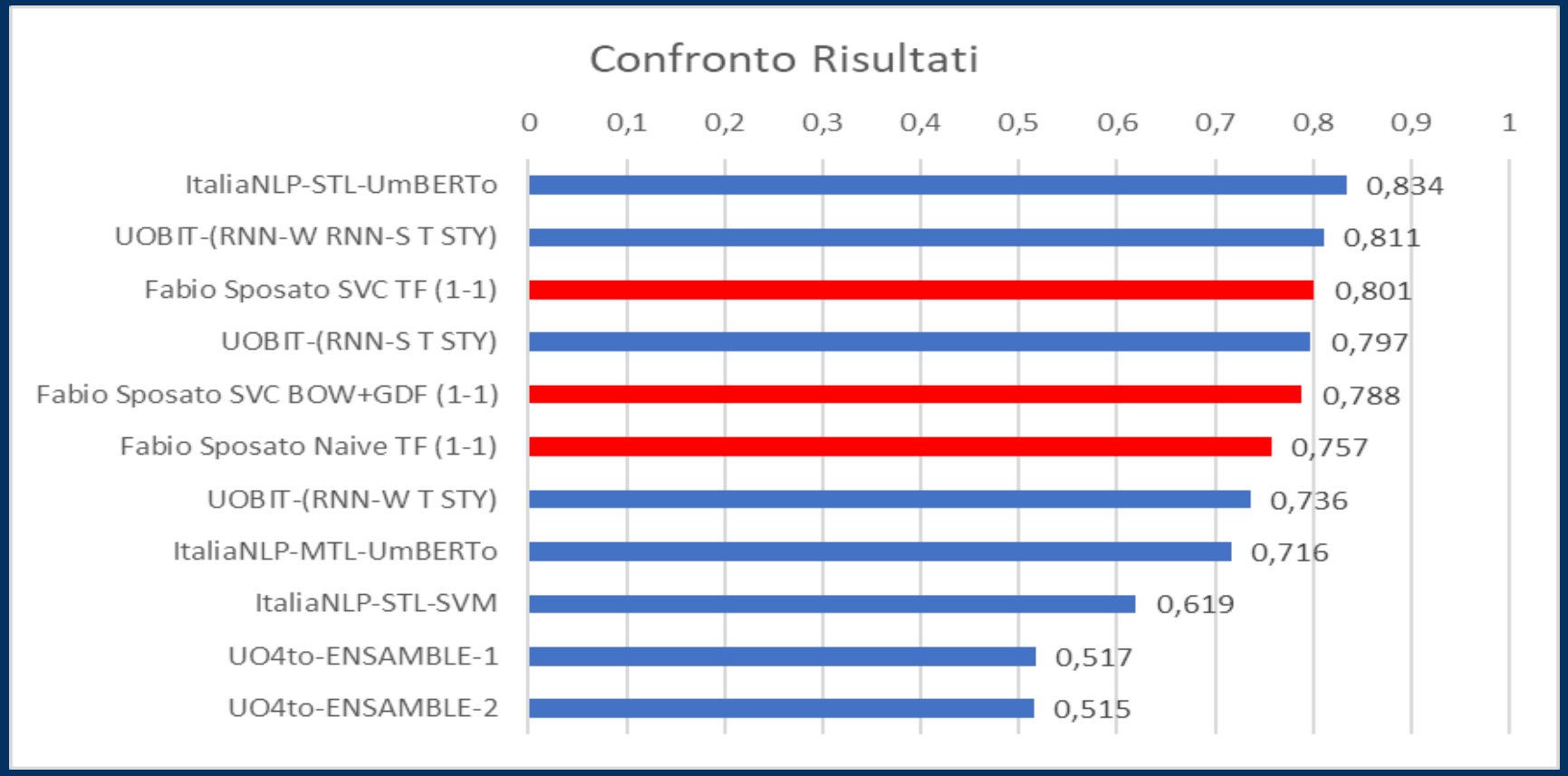
È la feature da me creata

matchea(train[i]), contaFem(train_tokens[i]), contaMasc(train_tokens[i]), len(train[i]), len(train[i].split()), len(train[i]) / len(train[i].split())

Modelli più prestanti ottenuti

Nome Run	Modello	F-Macro Gender K-Fold sul Training	F-Macro Gender Test
Fabio Sposato SVC BOW+GDF (1-1)	BOW + GDF SVC (1-1)	0,801	0,788
Fabio Sposato SVC TF (1-1)	TF SVC (1-1)	0,789	0,801
Fabio Sposato Naive TF (1-1)	TF Naïve (1-1)	0,789	0,757

Confronto con i partecipanti al Task



Limiti

Primo Esempio

“Fra tutte le cose, quelle che mi hanno sorpresa di più sono che: - il gap age fra lui e lei non mi sta dando minimamente fastidio.”

Secondo Esempio

“Benvenuto bar!!!!!”

Consigli per future ricerche

1. Bilanciamento del dataset
2. Apprendimento d'insieme

Grazie per l’attenzione