IDEAs

Inclusive Dialogues for Equal Actions

L'Intelligenza Artificiale contro le discriminazioni

CPS CULTURE POLITICA SOCIETÀ

di.unito.it DIPARTIMENTO DI INFORMATICA

Università di Torino Dipartimento di Giurisprude

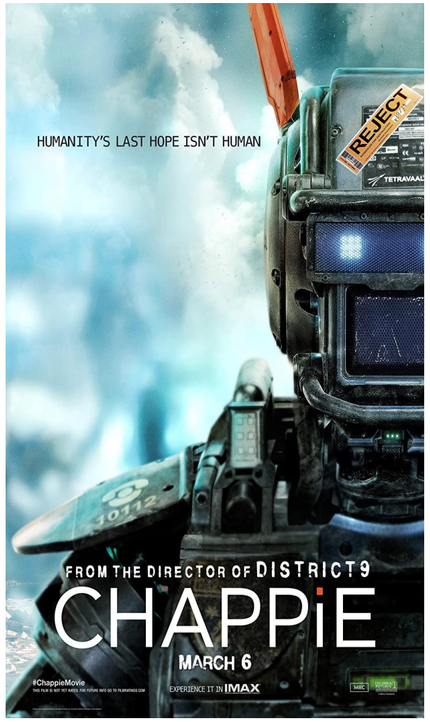
Studi um

Fisica Torino Dipartimento di Eccellenza

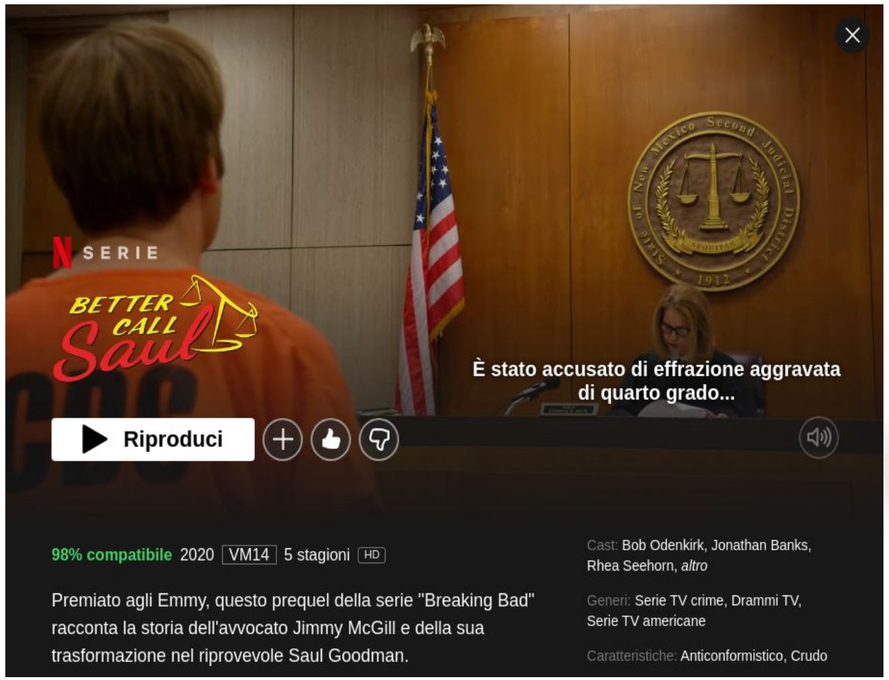
Che cos’è l’Intelligenza Artificiale

“la teoria e lo sviluppo di sistemi informatici in grado di svolgere attività che normalmente richiedono intelligenza umana”

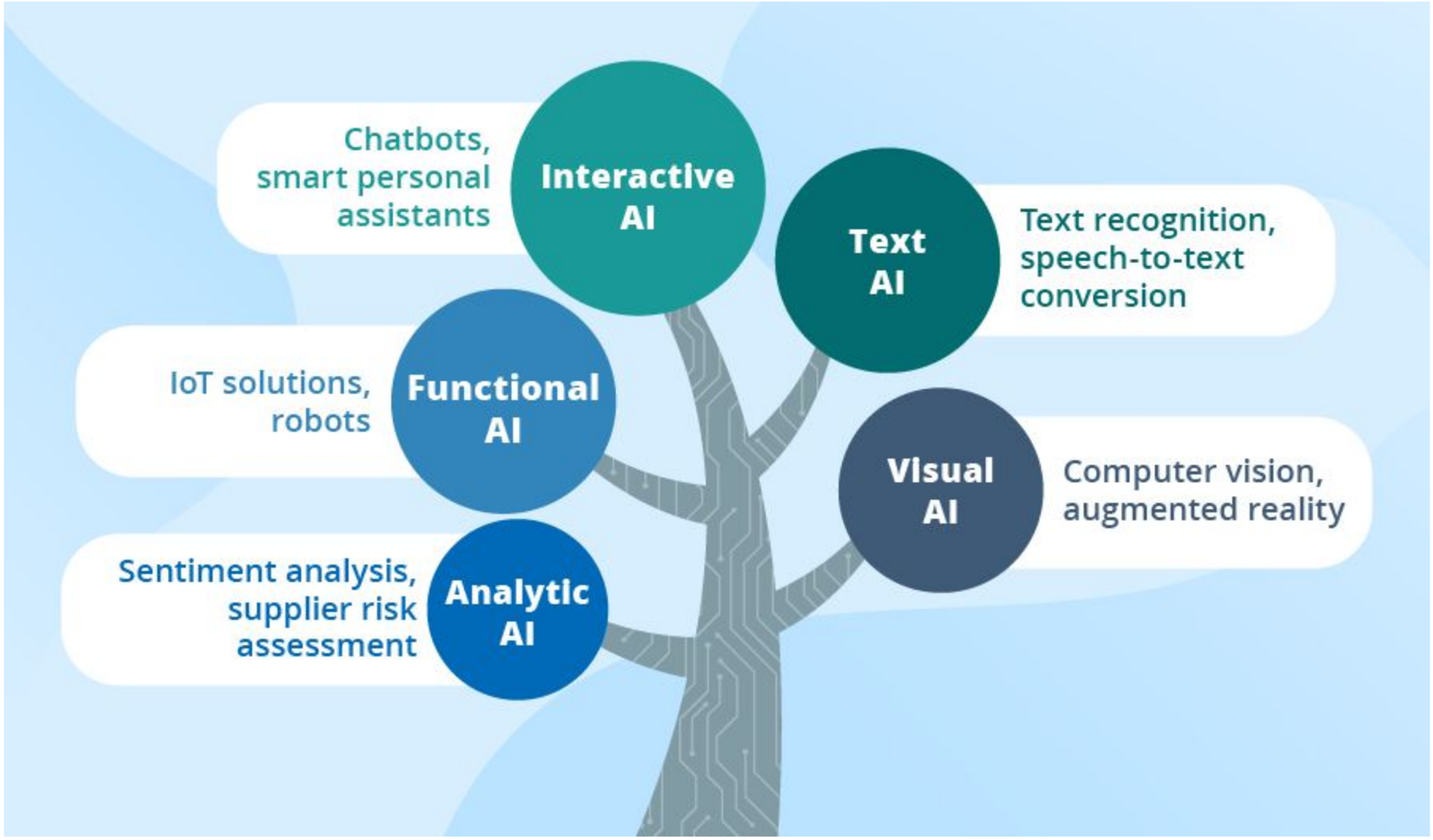
Come la immaginiamo nei film di fantascienza



Com’è in realtà



Tanti modelli per diversi task

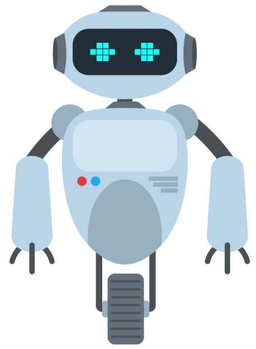


Natural Language Processing

Definizione di Natural Language Processing

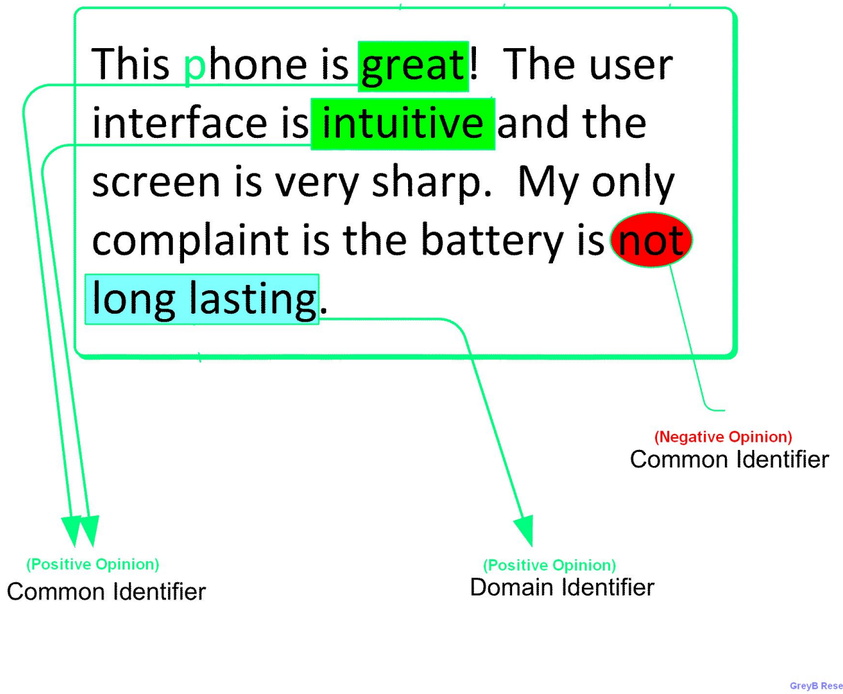
Il Natural Language Processing (NLP) è un insieme di tecniche computazionali per l'analisi e la rappresentazione di testo naturale.

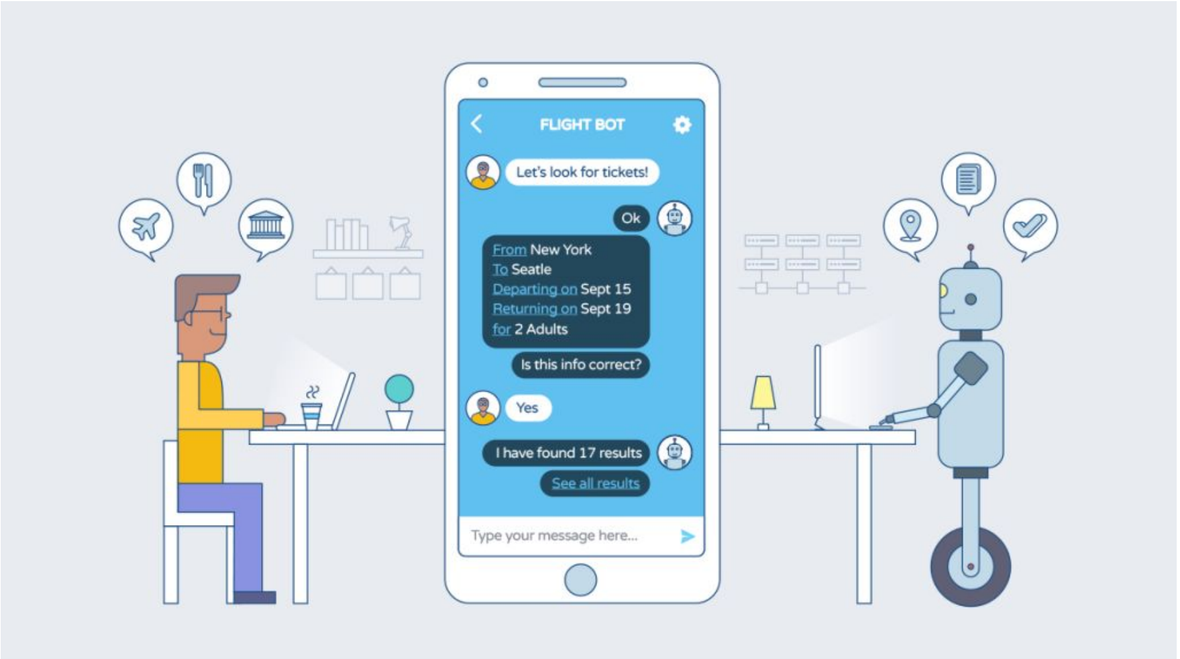
Marco mangia una mela



Marco: soggetto
mangia: predicato verbale
una mela: complemento oggetto

Alcuni task di NLP





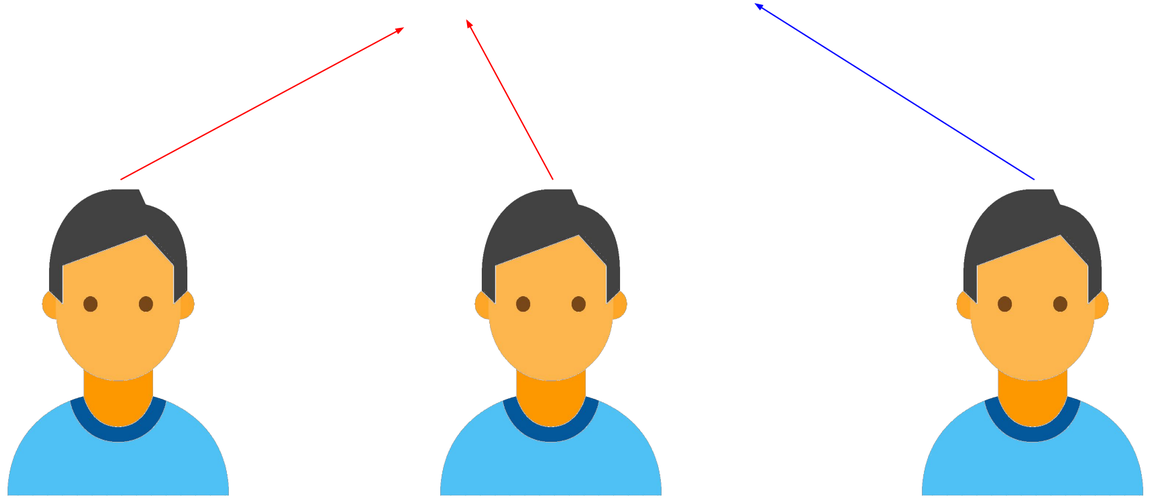
Costruire un modello di NLP: creazione del corpus

Un altro mondiale senza l'Italia sarà veramente difficile da digerire... Ma se non batti la Macedonia, non meriti di giocarlo, quel mondiale

#ItaliaMacedoniadelNord

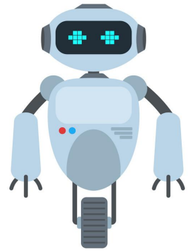
10:39 PM · 24 mar 2022 · Twitter Web App

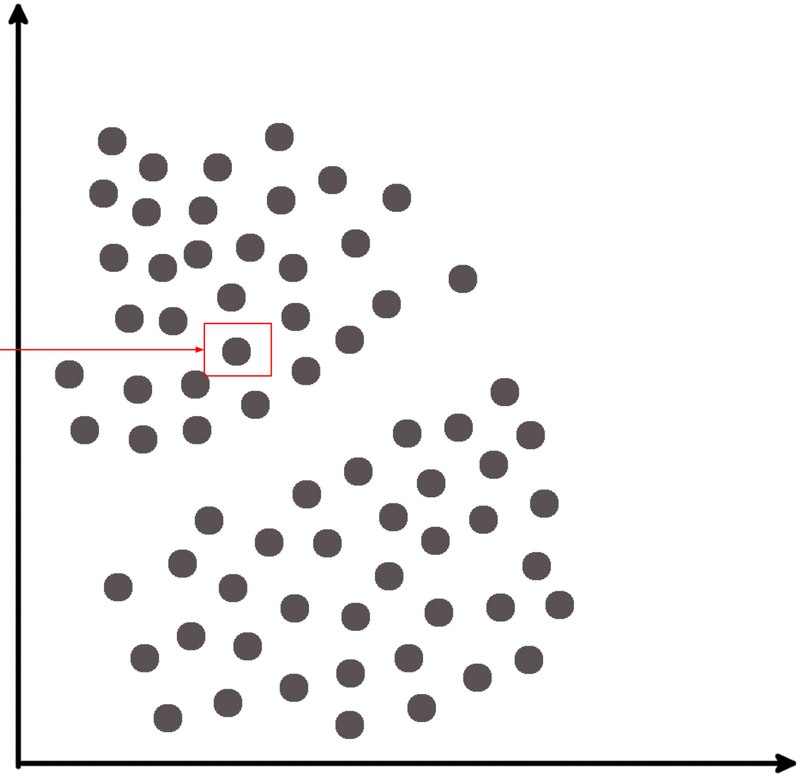
Il messaggio contiene un'opinione positiva?: **NO** **SÌ**



Costruire un modello di NLP: vettorizzazione del testo

[0.1, 0.3, 0.2, 1.5, 0.4, 1.1 … 0.8, 2.1, 0.7]



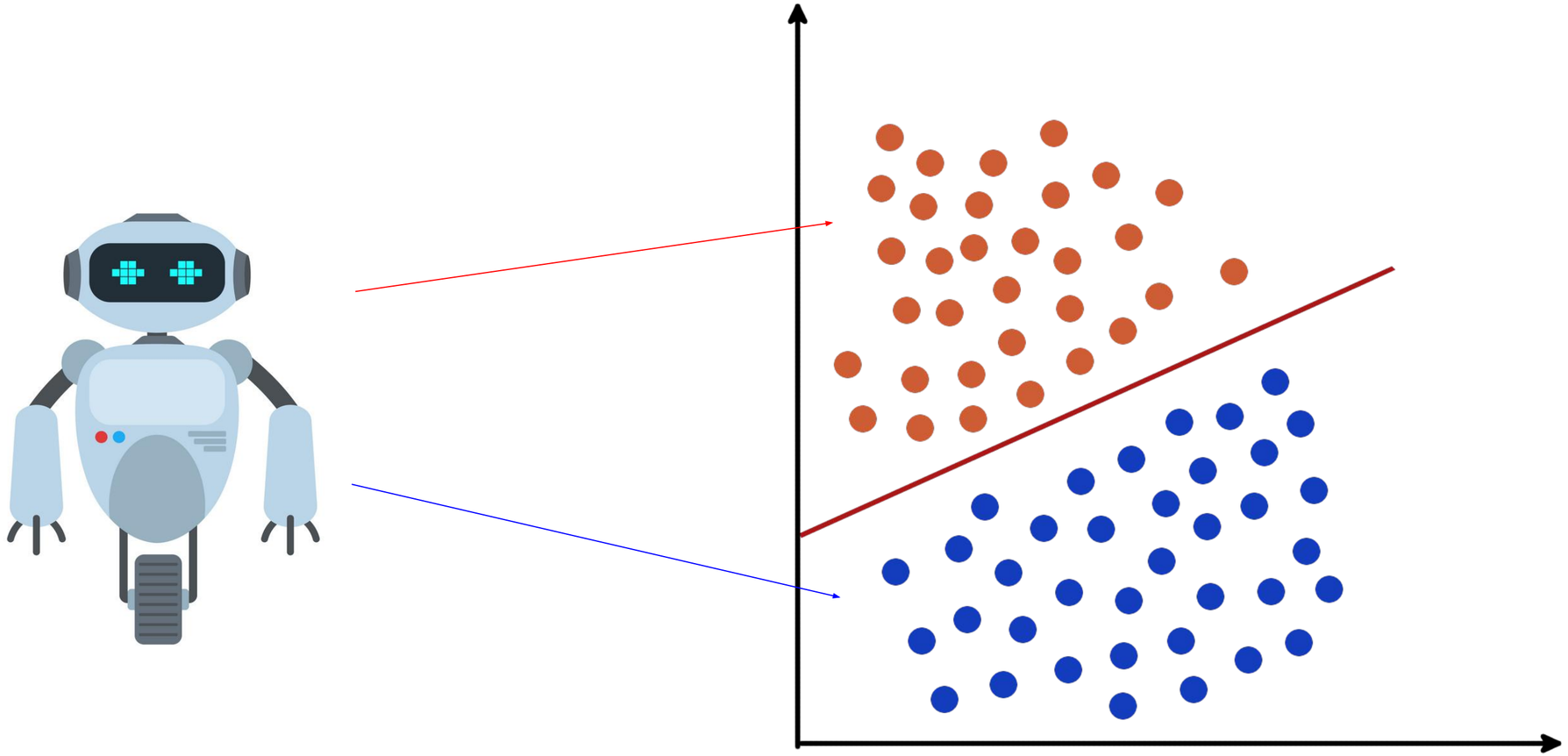


Un altro mondiale senza l'Italia sarà veramente difficile da digerire... Ma se non batti la Macedonia, non meriti di giocarlo, quel mondiale

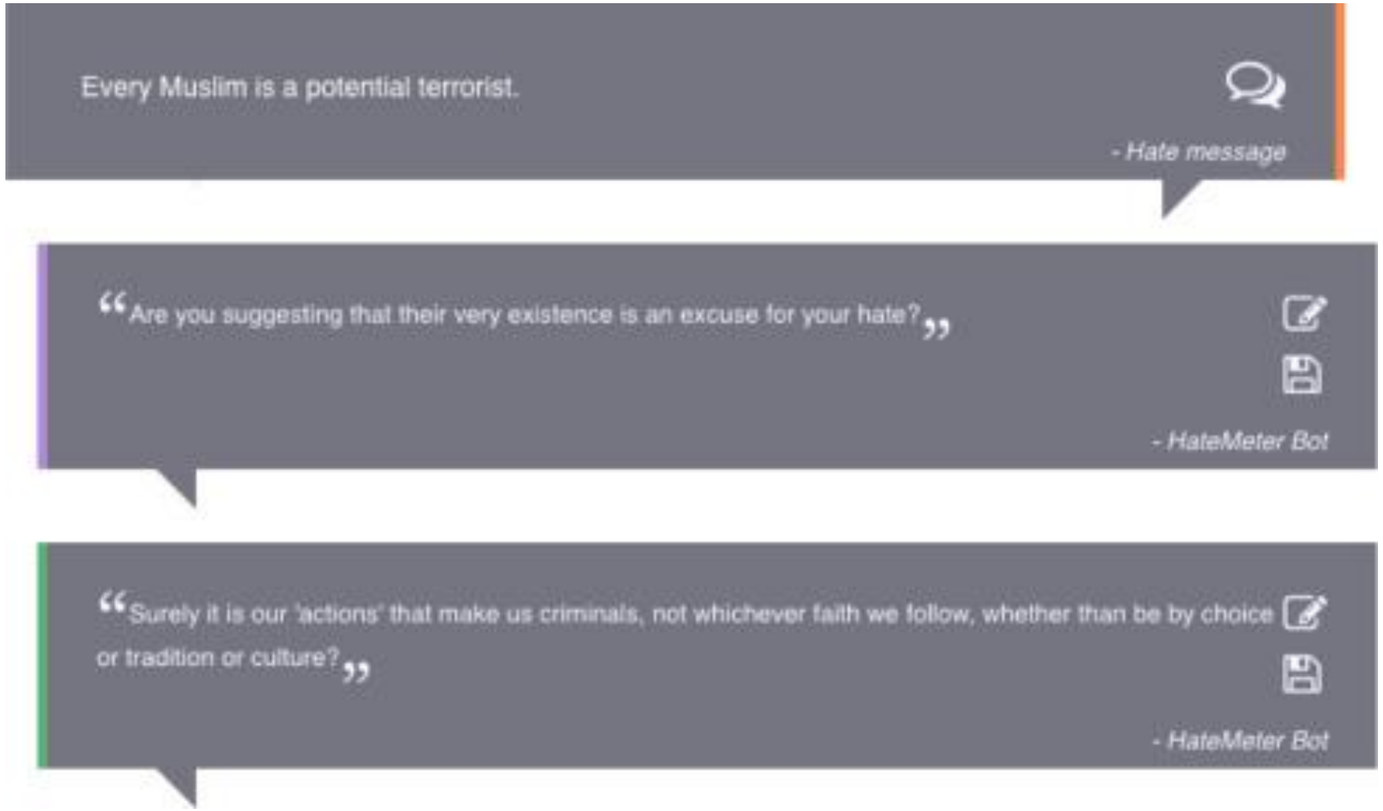
#ItaliaMacedoniadelNord

10:39 PM · 24 mar 2022 · Twitter Web App

Costruire un modello di NLP: l’addestramento del modello



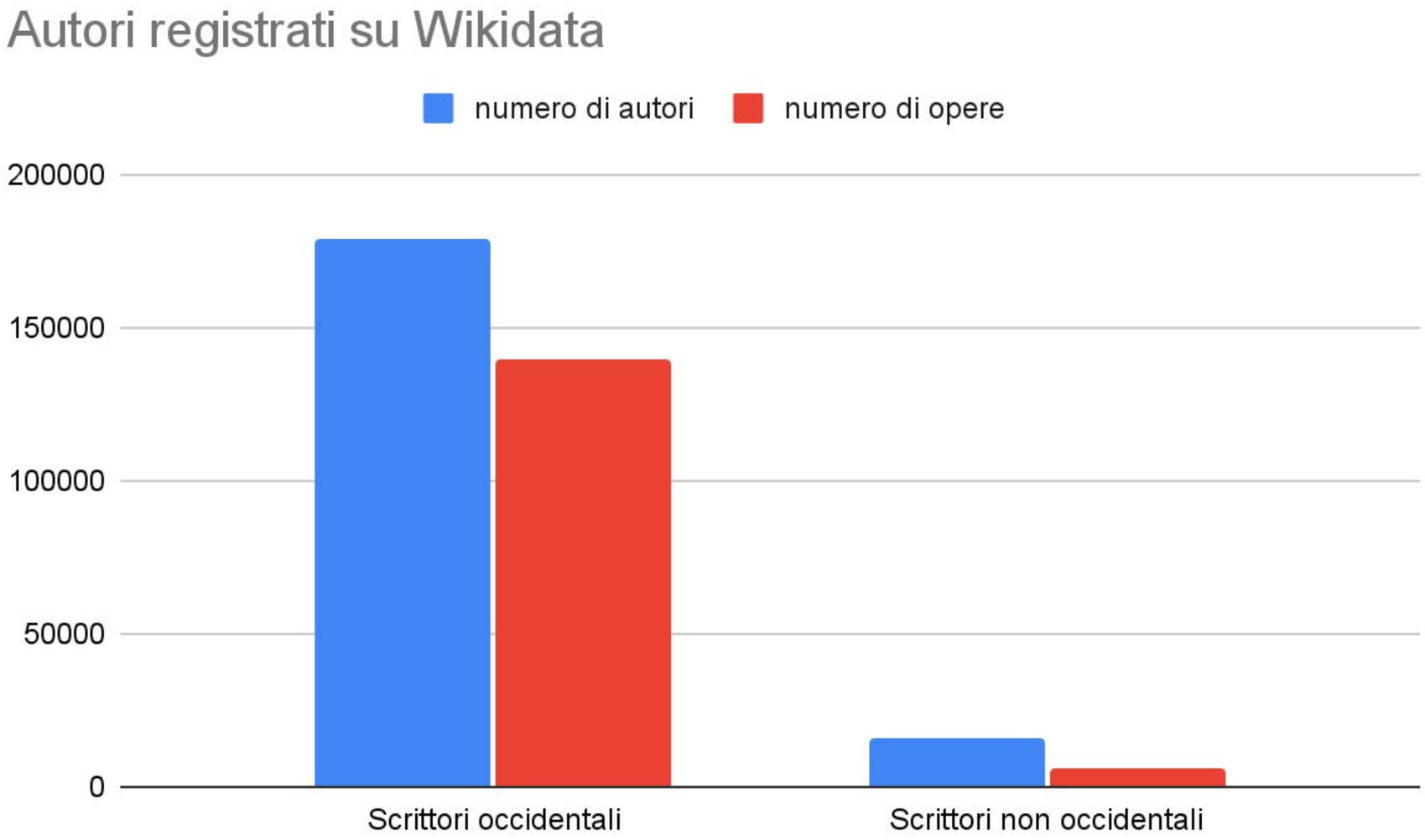
NLP contro le discriminazioni online



NLP e bias sociali

I bias sono causati da pregiudizi nei dati usati per addestrare i modelli e possiamo raccogliere dati solo dal mondo che abbiamo, che ha una lunga storia di discriminazione, quindi la tendenza predefinita di questi sistemi sarà quella di riflettere i nostri pregiudizi

Assenza di dati su determinate categorie

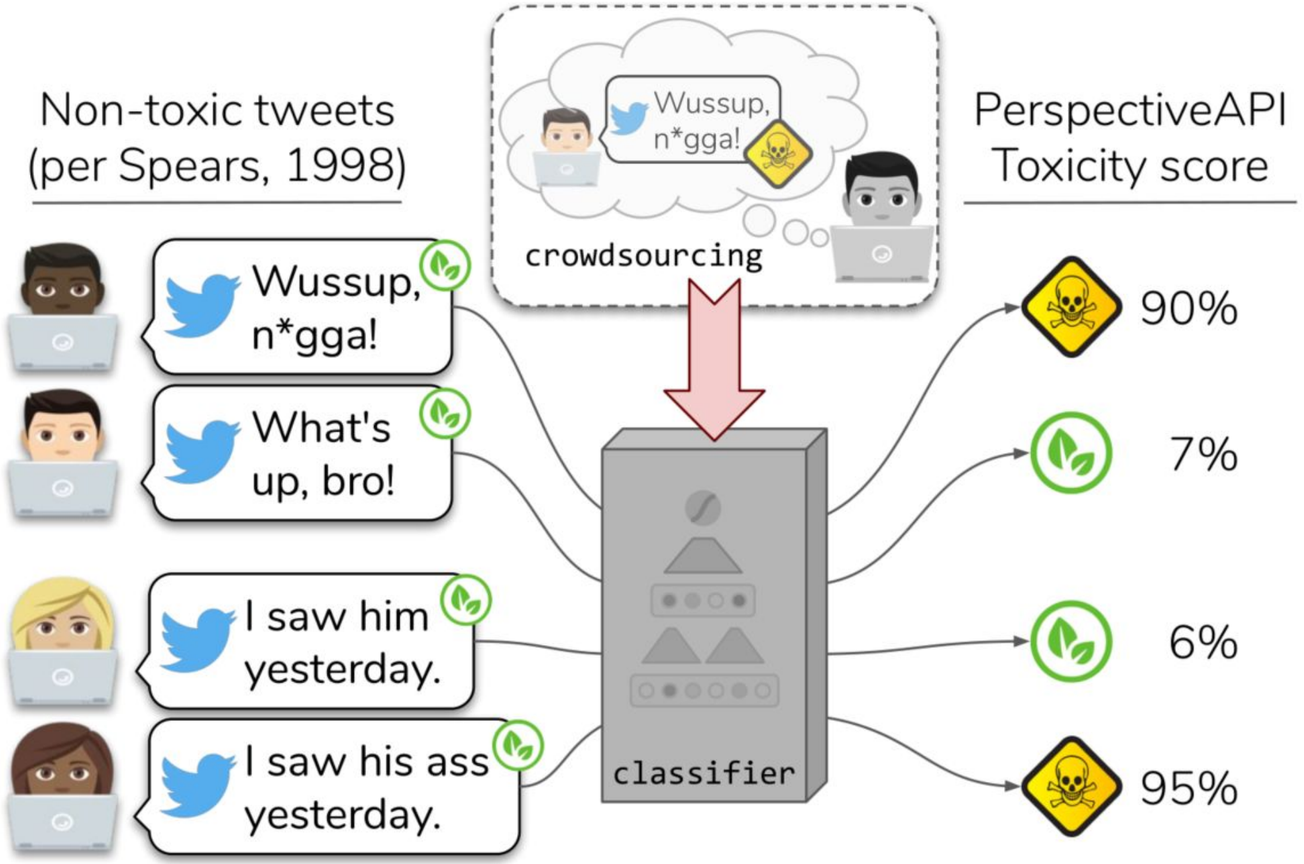


Stereotipi impliciti su categorie sottorappresentate

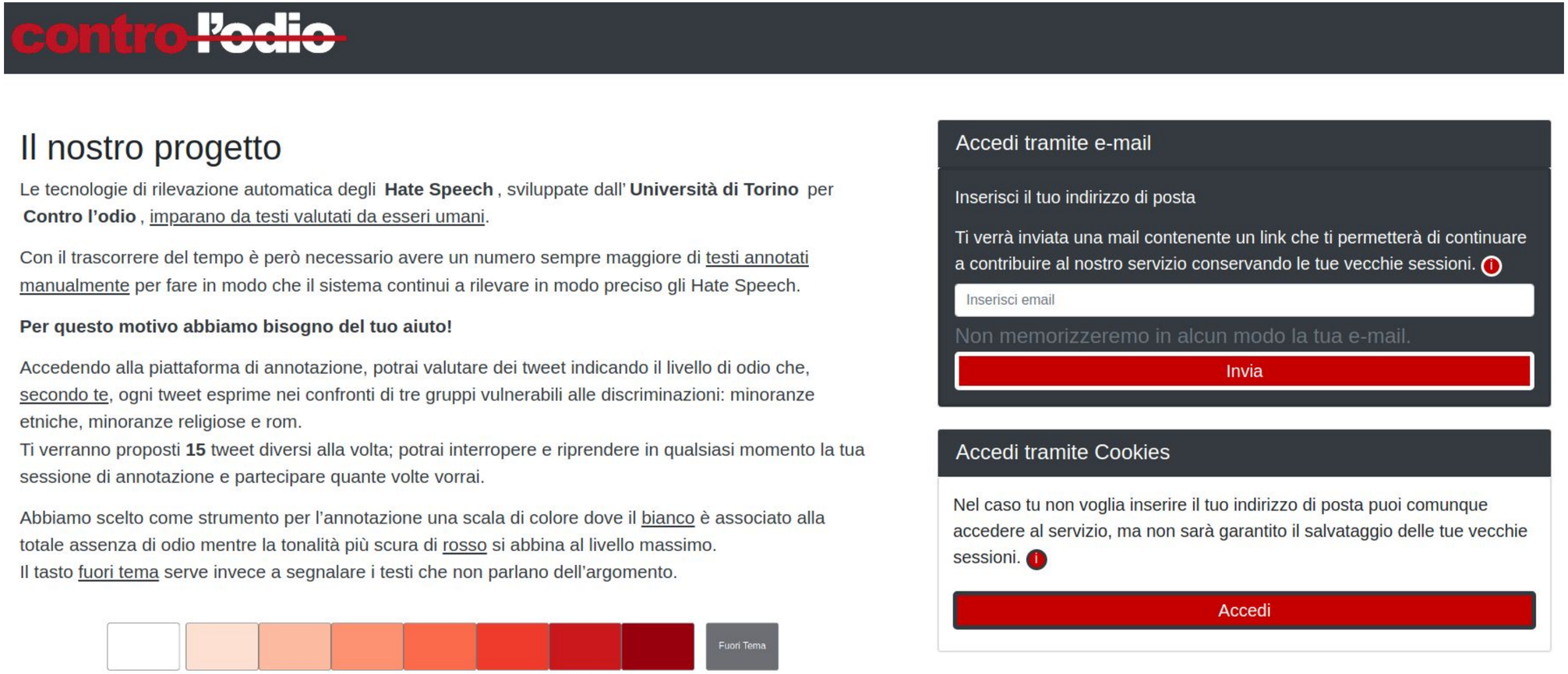
Occupation	Events in Female Career Description	Events in Male Career Description	WEAT∗	WEAT
Writer	♦ divorce, ♦ marriage, involve, organize, ♦ wedding	argue, ⊕ election, ▲ protest, rise, shoot	-0.17	1.51
Acting	♦ divorce, ♦ wedding, guest, name, commit	support, ▲ arrest, ▲ war, ■ sue, trial	-0.19	0.88
Comedian	♦ birth, eliminate, ♦ wedding, ♥ relocate, partner	enjoy, hear, cause, ● buy, conceive	-0.19	0.54
Podcaster	♥ land, interview, portray, ♦ married, report	direct, ask, provide, continue, bring	-0.24	0.53
Dancer	♦ married, ♦ marriage, ♥ depart, ♥ arrive, organize	drop, team, choreograph, explore break	-0.14	0.22
Artist	paint, exhibit, include, ♥ return, teach	start, found, feature, award, begin	-0.02	0.17
Chef	⊕ hire, △ meet, debut, eliminate, sign	include, focus, explore, award, ● raise	-0.13	-0.38
Musician	run, record, ♦ death, found, contribute	sign, direct, produce, premier, open	-0.19	–0.41

Annotations: ♦ Life ♥ Transportation ⊕ Personell ▲ Conflict ■ Justice ● Transaction △ Contact

Annotazione errata di alcune varietà linguistiche sottorappresentate



Costruiamo un corpus insieme



Costruiamo un corpus insieme

- Identifica un tweet o un aspetto dell'annotazione che ti ha messo in difficoltà
- Fai uno screenshot
- Inseriscilo qui: https://bit.ly/3Djvt8O

Grazie per l’attenzione