

# IDEAs

## Inclusive Dialogues for Equal Actions

L'Intelligenza Artificiale contro le discriminazioni



di.unito.it  
DIPARTIMENTO DI INFORMATICA



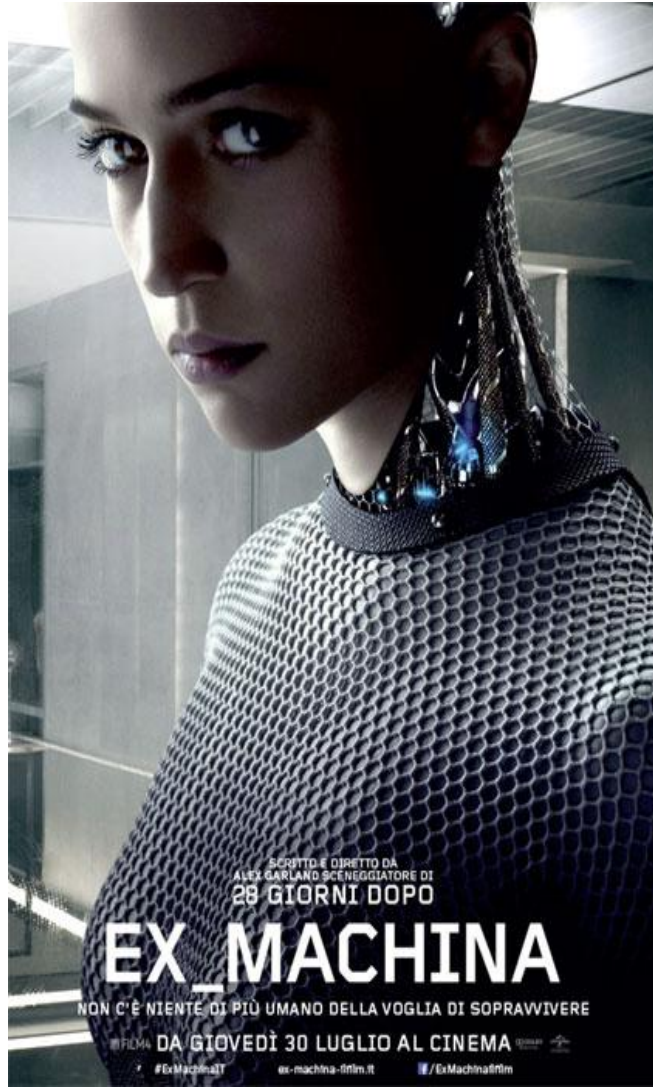
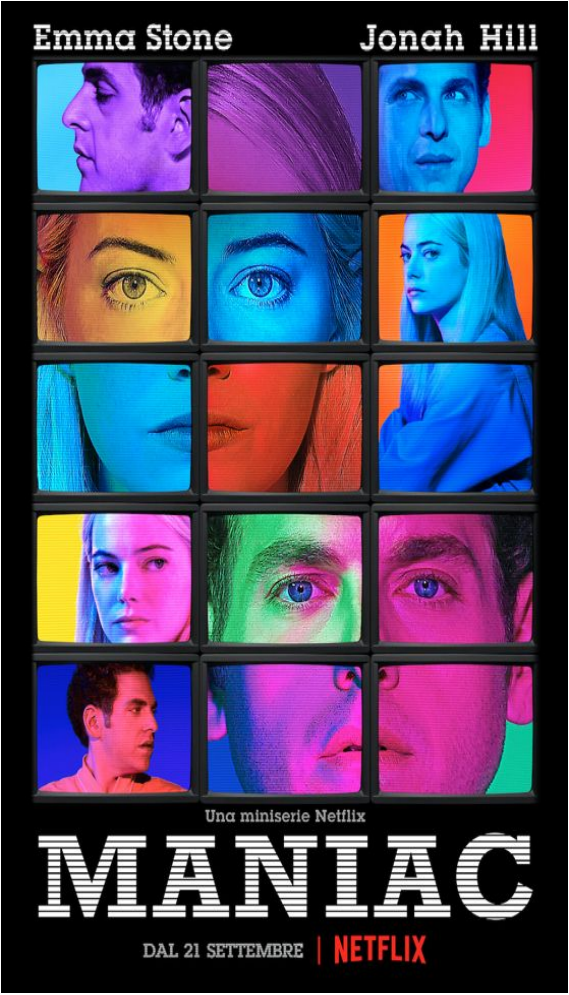
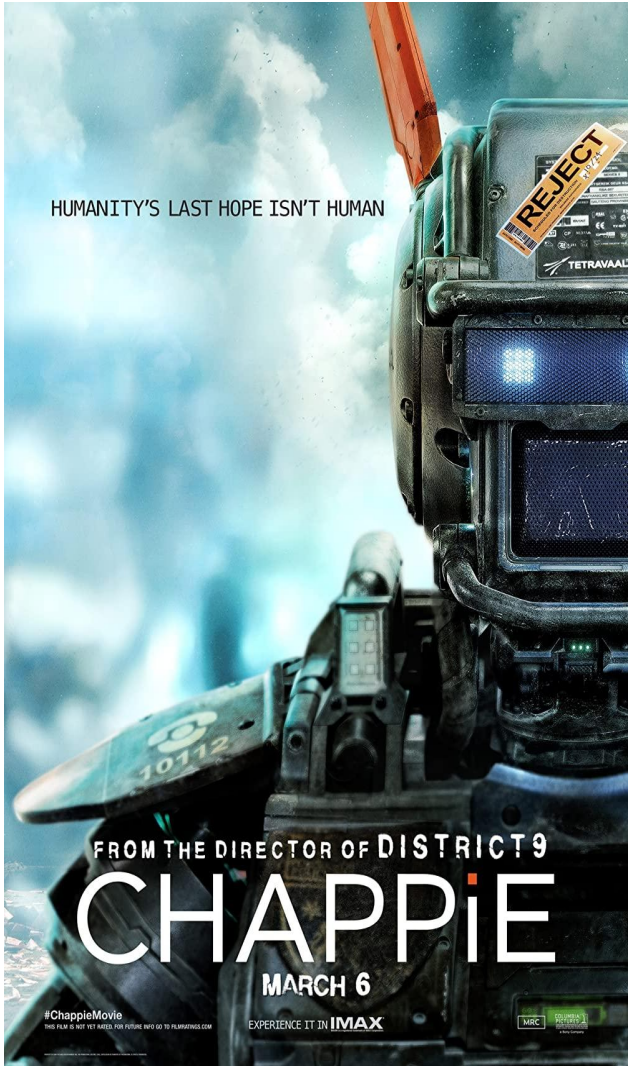
# Che cos'è l'Intelligenza Artificiale

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

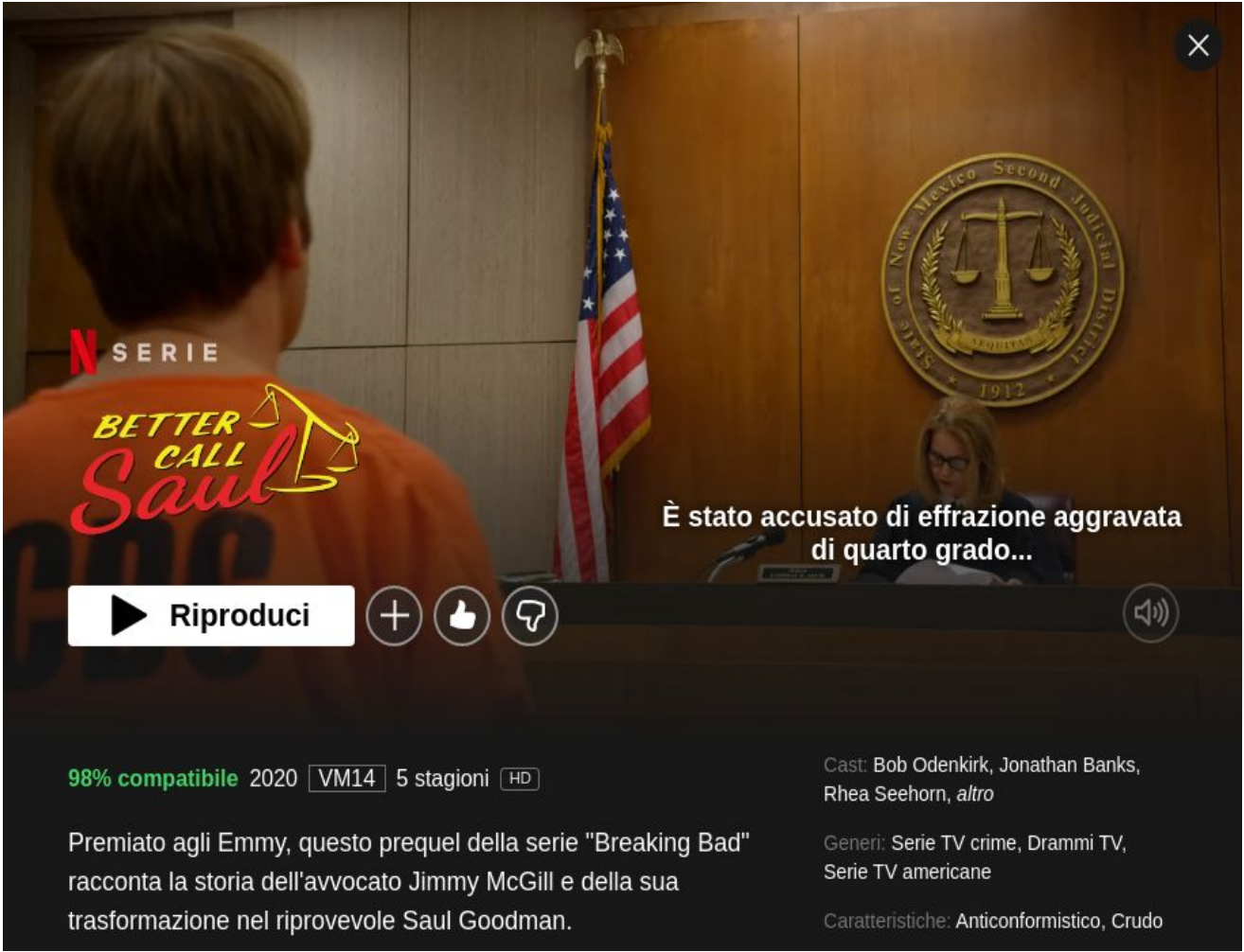
TURING TEST EXTRA CREDIT:  
CONVINCE THE EXAMINER  
THAT HE'S A COMPUTER.



# Come la immaginiamo nei film di fantascienza



# Com'è in realtà



**N SERIE**  
**BETTER CALL Saul**

È stato accusato di effrazione aggravata di quarto grado...

**Riproduci** + 👍 👎 🔊

98% compatibile 2020 VM14 5 stagioni HD

Cast: Bob Odenkirk, Jonathan Banks, Rhea Seehorn, altro

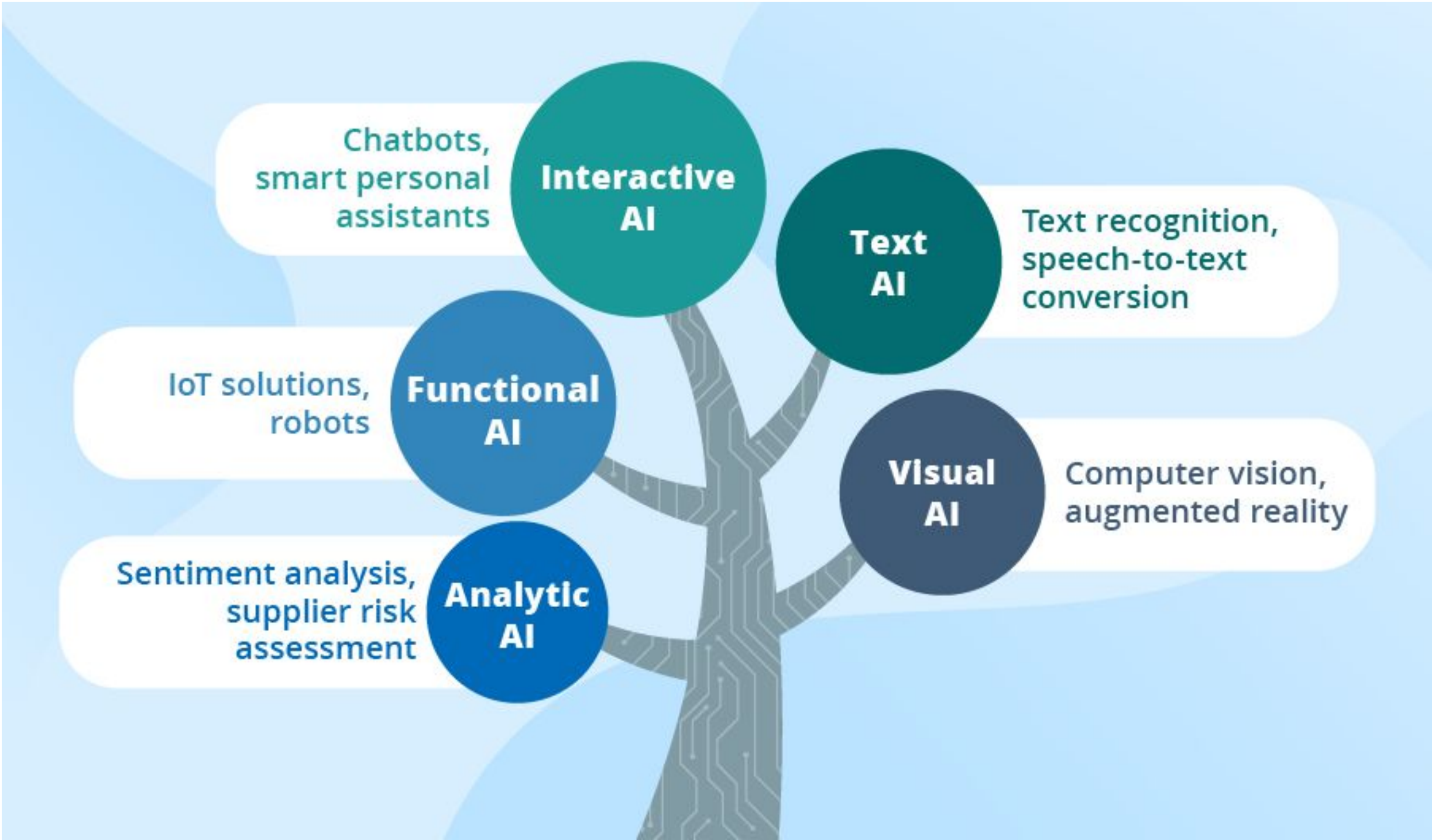
Generi: Serie TV crime, Drammi TV, Serie TV americane

Caratteristiche: Anticonformistico, Crudo

Premiato agli Emmy, questo prequel della serie "Breaking Bad" racconta la storia dell'avvocato Jimmy McGill e della sua trasformazione nel riprovevole Saul Goodman.



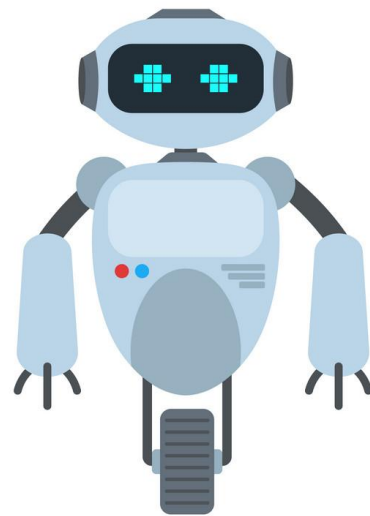
# Tanti modelli per diversi task



# Natural Language Processing

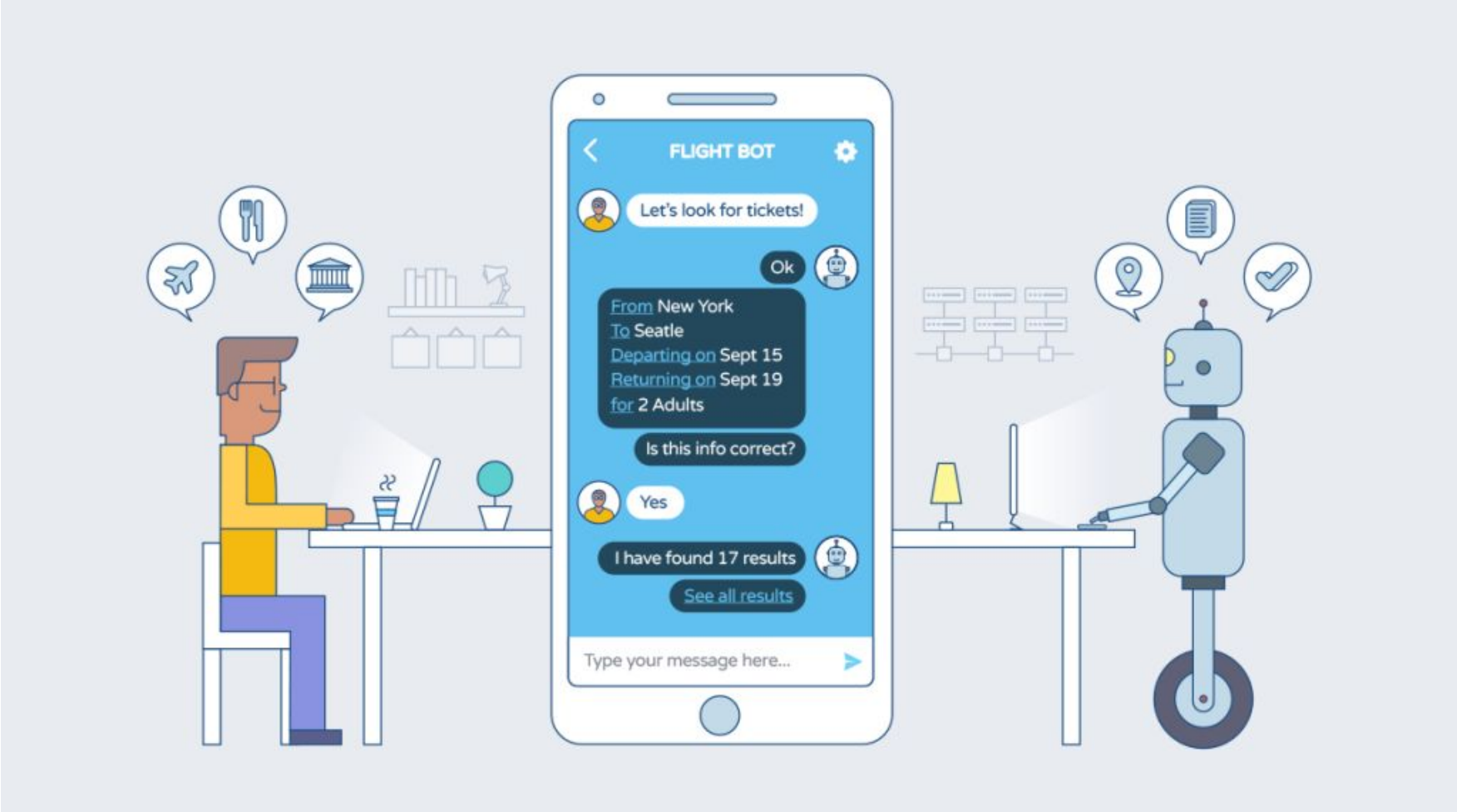
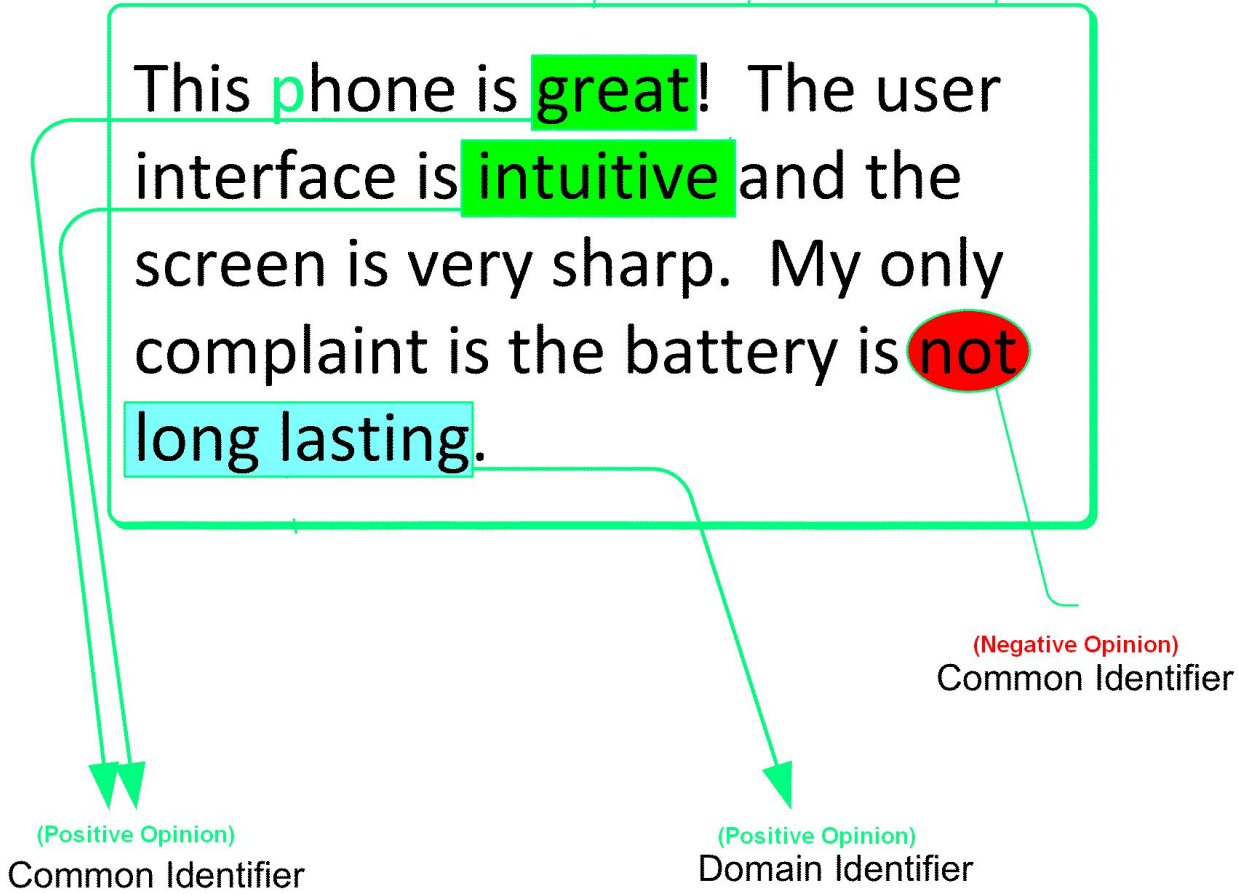
Il Natural Language Processing (NLP) è un insieme di tecniche computazionali per l'analisi e la rappresentazione di testo naturale.

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

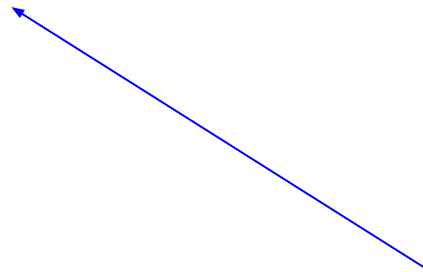
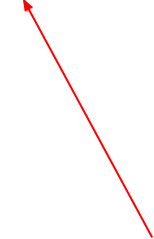
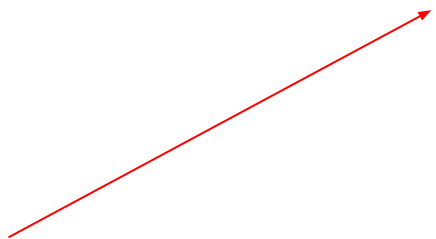
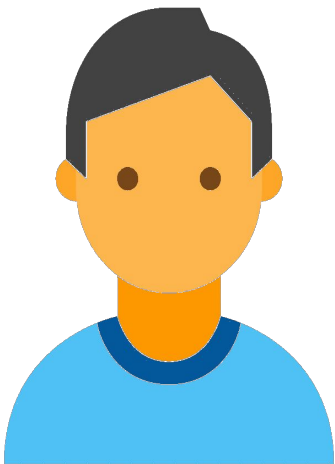
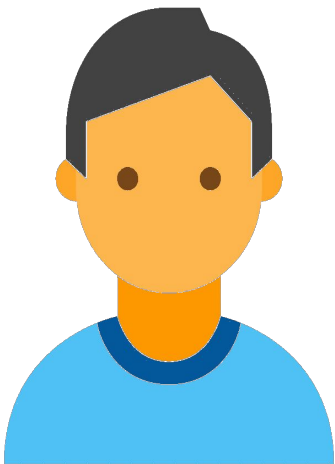
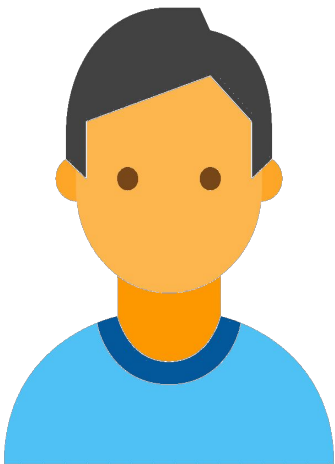
#ItaliaMacedoniadelNord

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

Il messaggio contiene un'opinione positiva?:

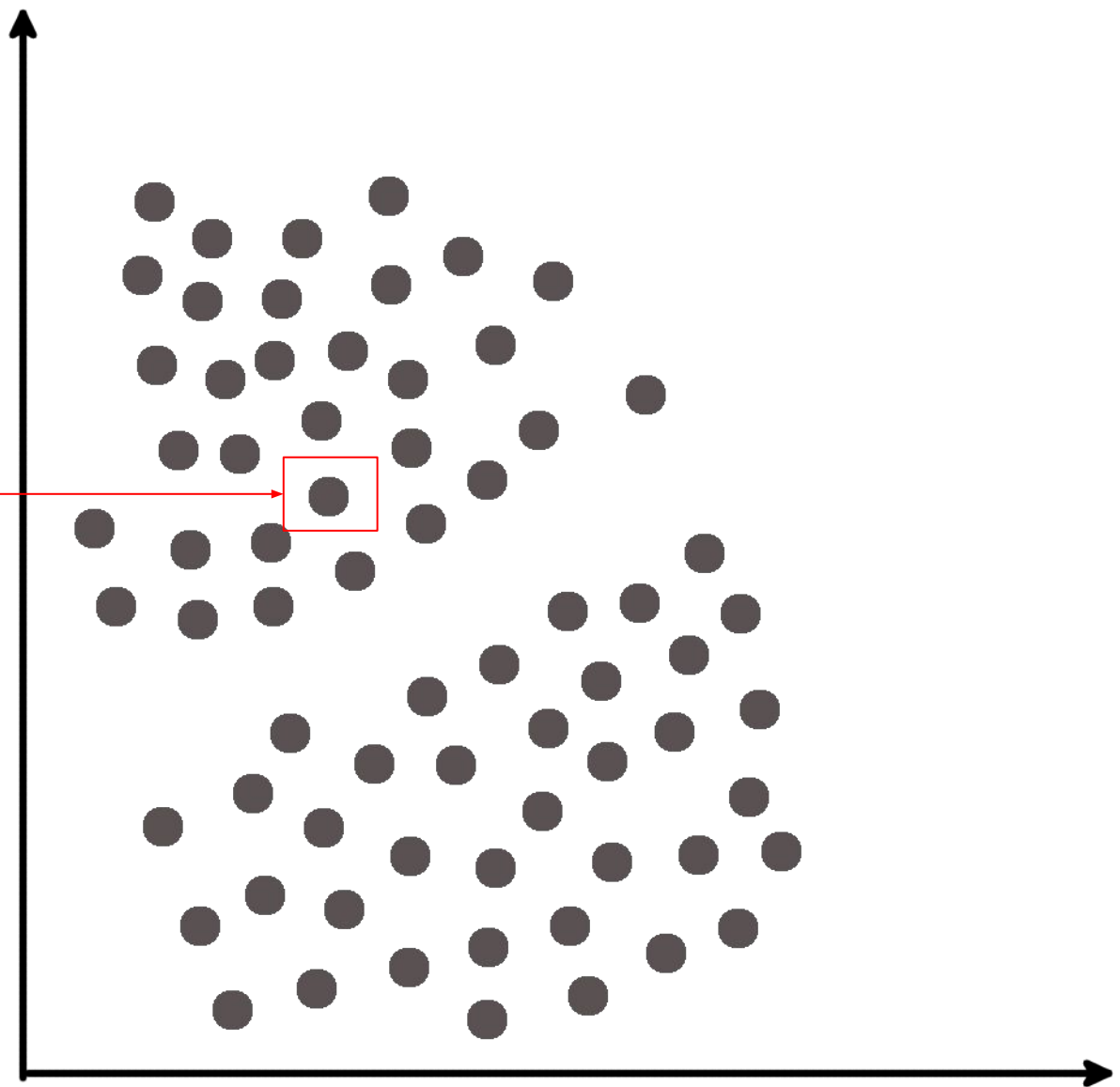
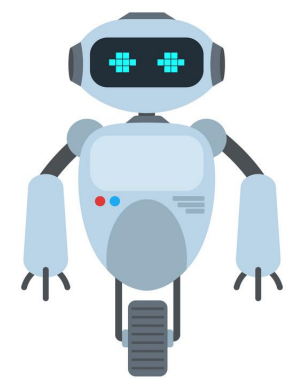
**NO**

**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]

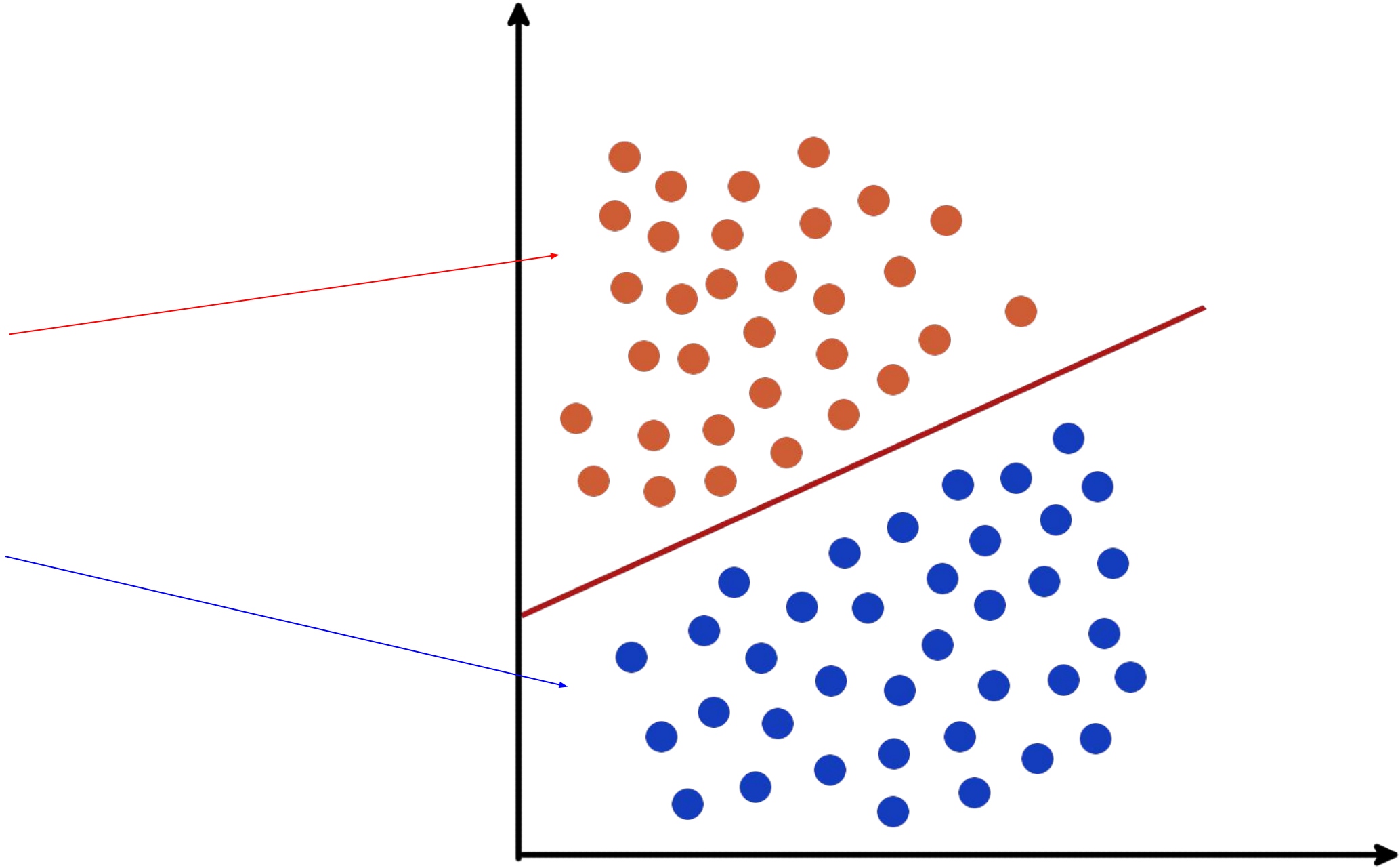
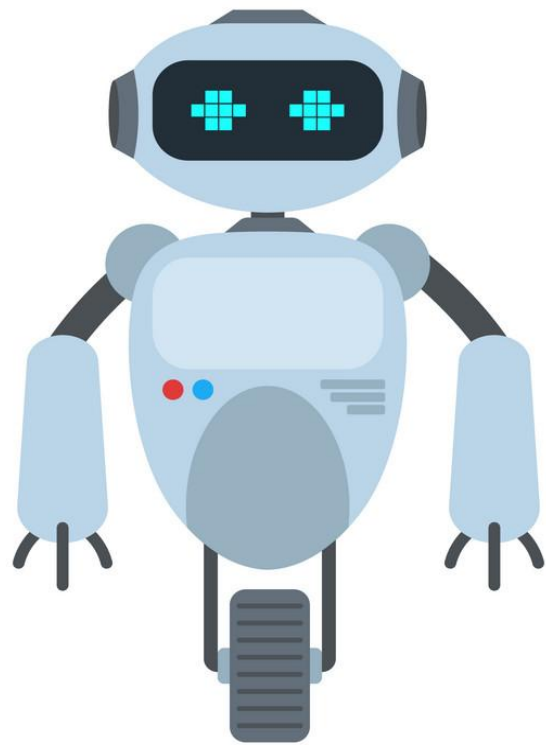


Un altro mondiale senza l'Italia sarà veramente difficile da digerire... Ma se non batti la Macedonia, non meriti di giocarlo, quel mondiale

[#ItaliaMacedoniadelNord](#)

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

# Costruire un modello di NLP: l'addestramento del modello



# NLP contro le discriminazioni online



Every Muslim is a potential terrorist.

- Hate message

“Are you suggesting that their very existence is an excuse for your hate?”

- HateMeter Bot

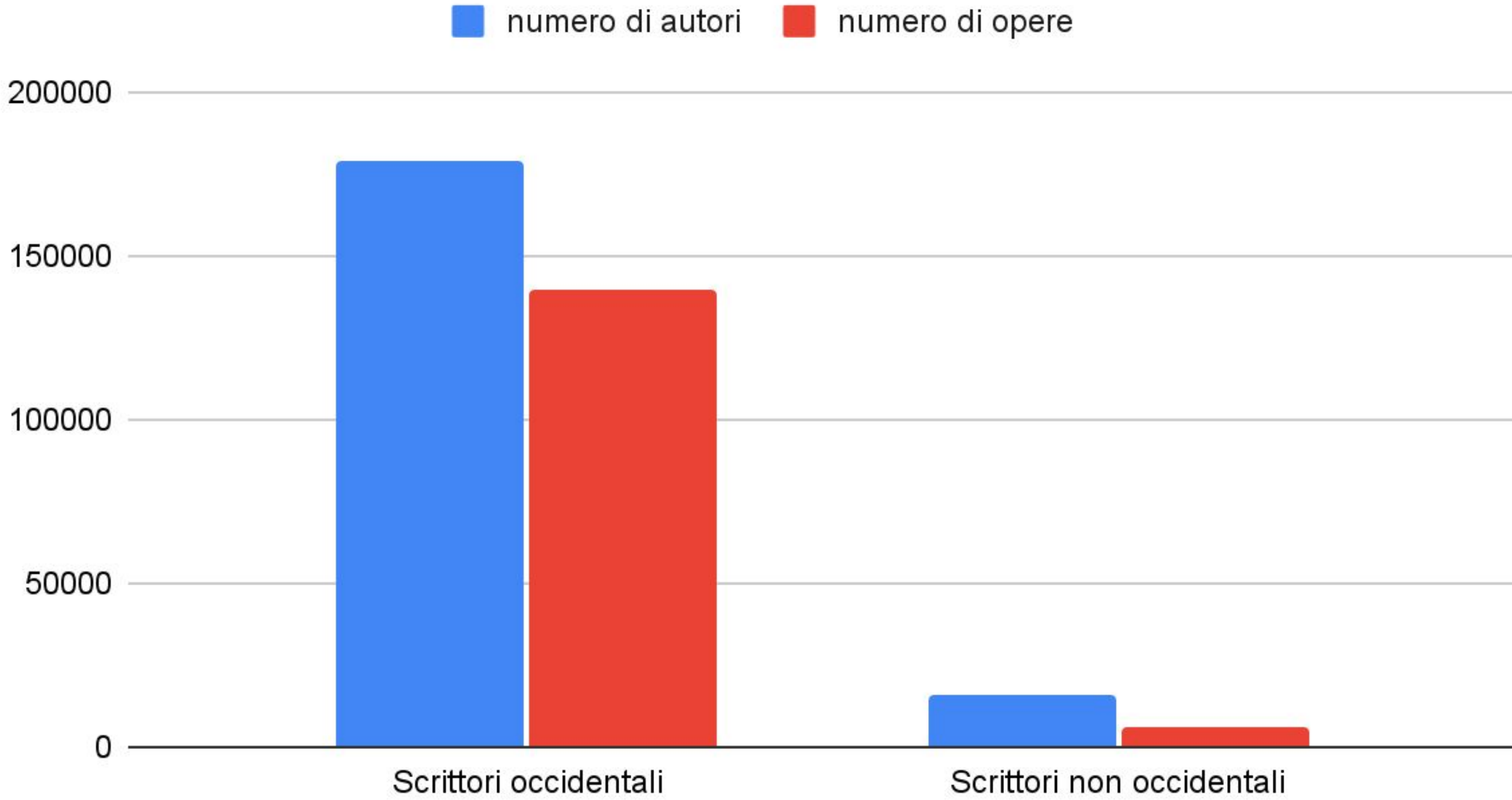
“Surely it is our 'actions' that make us criminals, not whichever faith we follow, whether than be by choice or tradition or culture?”

- HateMeter Bot

The banner features the Amnesty International logo on the left and a 'Conta fino a 10' logo on the right. The text in the center reads 'ELEZIONI 2018' and 'BAROMETRO DELL'ODIO'.

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

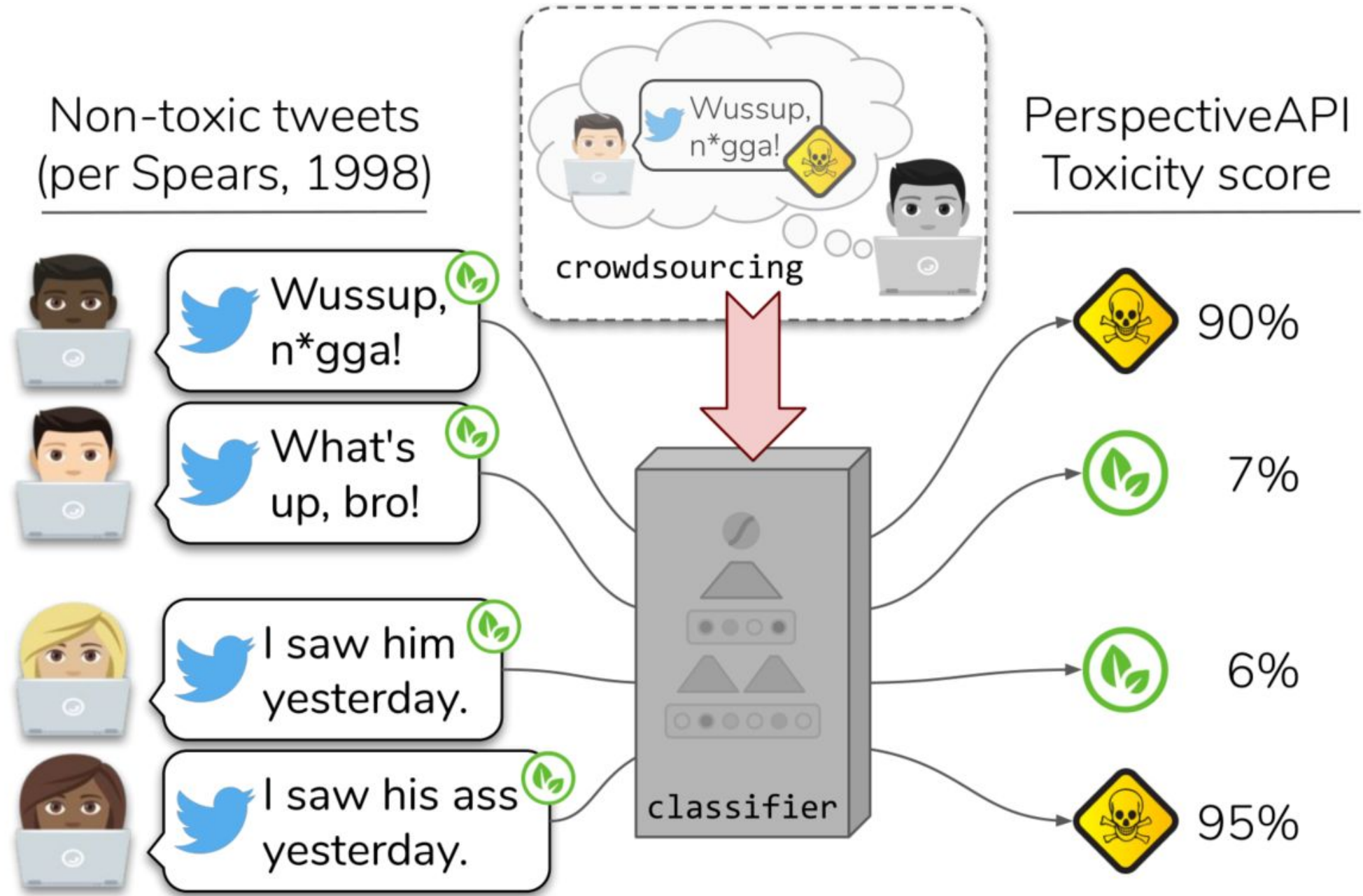
### Autori registrati su Wikidata



# 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



## Il nostro progetto

Le tecnologie di rilevazione automatica degli **Hate Speech**, sviluppate dall'**Università di Torino** per **Contro l'odio**, imparano da testi valutati da esseri umani.

Con il trascorrere del tempo è però necessario avere un numero sempre maggiore di testi annotati manualmente per fare in modo che il sistema continui a rilevare in modo preciso gli Hate Speech.

### Per questo motivo abbiamo bisogno del tuo aiuto!

Accedendo alla piattaforma di annotazione, potrai valutare dei tweet indicando il livello di odio che, secondo te, ogni tweet esprime nei confronti di tre gruppi vulnerabili alle discriminazioni: minoranze etniche, minoranze religiose e rom.

Ti verranno proposti **15** tweet diversi alla volta; potrai interrompere e riprendere in qualsiasi momento la tua sessione di annotazione e partecipare quante volte vorrai.

Abbiamo scelto come strumento per l'annotazione una scala di colore dove il bianco è associato alla totale assenza di odio mentre la tonalità più scura di rosso si abbina al livello massimo.

Il tasto fuori tema serve invece a segnalare i testi che non parlano dell'argomento.



### Accedi tramite e-mail

Inserisci il tuo indirizzo di posta

Ti verrà inviata una mail contenente un link che ti permetterà di continuare a contribuire al nostro servizio conservando le tue vecchie sessioni. ⓘ

Non memorizzeremo in alcun modo la tua e-mail.

### Accedi tramite Cookies

Nel caso tu non voglia inserire il tuo indirizzo di posta puoi comunque accedere al servizio, ma non sarà garantito il salvataggio delle tue vecchie sessioni. ⓘ

## 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