HATE SPEECH & SOCIAL MEDIA

Verso il riconoscimento automatico dell'odio in rete

HATE SPEECH AND SOCIAL MEDIA

I ♥ HATE PREJUDICE

Imparare a riconoscere l'odio

L'odio è un **sentimento** o uno **stato d'animo**, come la gioia, la rabbia, la soddisfazione…

Possiamo usare molti modi per esprimere uno stato d'animo:

esplicito — *«Sono felicissimo di vederti!»*

implicito — (sorriso)

HATE SPEECH AND SOCIAL MEDIA

I ♥ HATE PREJUDICE

Imparare a riconoscere l'odio

Nella comunicazione online sono precluse alcune possibilità comunicative (linguaggio non verbale, prossemica, intonazione), ma ne compaiono di nuove:

- Emoticons, emoji
- Massimo utilizzo di espedienti grafici (maiuscolo, punteggiatura)
- Azioni consentite dalla piattaforma (*like, share, retweet, follow*)

Abbiamo nuovi modi per esprimere sentimenti e stati d'animo.

Senza presenza fisica può essere difficile capire quale emozione si cela dietro una frase.

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Imparare a riconoscere l'odio

«Ringraziamo l'amministrazione comunale per questi ottimi risultati!»

Entusiasmo o ironica disapprovazione?

«Barricate contro l'arrivo di 20 profughi. E alla fine i migranti vengono dirottati. Comincia la rivolta»

Constatazione neutrale o incitamento alla violenza?

«'Bastardi islamici' è il titolo di giornale più vergognoso di sempre»

Offesa ai fedeli di una religione o critica a chi ha scritto l'offesa?

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Imparare a riconoscere l'odio

Problema dell'ambiguità e dei significati impliciti. Per individuare l'odio non basta cercare le «parole d'odio», serve analisi più profonda.

Un sistema di riconoscimento automatico deve distinguere **hate speech** da contenuti **leciti**, anche se offensivi o fastidiosi.

Metodo di ricerca:

- raccolta di un corpus di tweet tramite **keywords neutre**
- definizione di **criteri operativi** per analisi approfondita
- **annotazione** di 5 giudici
- **confronto** dei risultati

Obiettivi:

- ottenere un corpus annotato affidabile per addestrare un **classificatore automatico**
- rendere disponibile lo strumento e il **bagaglio di conoscenze**

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Metodo di ricerca
Creazione del corpus

- Utilizzare **keywords neutre** per trovare tutti i tweet relativi a un argomento di discussione. L'hate speech non è veicolato solo da termini offensivi, il livello di analisi è l'intero tweet.

musulmani	immigrati	rom
terrorismo terrorist islam musulman* corano	immigrat* immigrazione migrant* stranier* profug*	rom nomad

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Metodo di ricerca
Criteri operativi

L'hate speech è un concetto complesso, composto da diverse dimensioni.

In ogni tweet abbiamo cercato:

- **hate speech**
- **aggressività**
- **offensività**
- **ironia**
- **stereotipo**

Metodo di ricerca
Annotazione

- corpus: 2100 tweet (1828 effettivi), divisi in due parti uguali
- annotatori: 5
- processo di annotazione in tre fasi:
  - ogni tweet è annotato da una coppia di giudici
  - i casi di disaccordo sono valutati dall'altra coppia di giudici
  - i rimanenti casi di disaccordo sono valutati dal quinto giudice

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Metodo di ricerca
Confronto

Confrontare le diverse valutazioni dei singoli annotatori su ogni tweet permette di capire quanto è stato difficile annotare.

Tenere conto di **4 aspetti + hate speech** permette di comprendere meglio il fenomeno, ma rende più difficile trovare un accordo.

Molteplicità di giudici è fondamentale per bilanciare gli errori e i condizionamenti individuali. Un sistema di riconoscimento automatico deve essere il più possibile imparziale e non condizionato dalle diverse 'tare' dei giudici umani.

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Criteri operativi
hate speech

valori: sì - no

«Sarebbe da VIETARE il culto dell'islam, bisognerebbe DISTRUGGERE le moschee, DEPORTARE tutti gli islamici e dichiarare l'islam FUORI LEGGE!»

- **azione** (intenzionale):
  - diffonde, incita, promuove o giustifica odio o violenza
  - svilisce, delegittima, offende o minaccia

- **bersaglio** (caratteristica protetta)
  - caratteristica protetta
    - religione (musulmani)
    - provenienza (immigrati)
    - etnia (rom)
  - l'intero gruppo, o un individuo considerato per la sua appartenenza al gruppo

HATE SPEECH AND SOCIAL MEDIA

I ♥ HATE PREJUDICE

Criteri operativi
bersagli

Scelti in base all'attuale contesto socio-culturale italiano.

Il discorso contro questi gruppi possiede:

- **tratti generici:**
  - insulti, minacce, espressioni denigranti o deplorevoli, epiteti offensivi
  - incitamento a odio, violenza o negazione di diritti
- **tratti specifici**:
  - **musulmani:** associare la fede islamica a estremismo, terrorismo, inferiorità culturale o biologica, piani di conquista dell'europa
  - **immigrati**: associare la provenienza a inferiorità culturale o biologica, propensione al crimine, comportamenti anti-sociali, minaccia per la sicurezza o il benessere
  - **rom**: associare l'etnia a comportamenti anti-sociali, propensione al crimine, minaccia per la sicurezza e il benessere

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Criteri operativi

aggressività

valori: assente – debole - forte

«Gli italiani prima di tutto!»

«Senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile»

Intenzione del parlante di nuocere al bersaglio, di creare una situazione di conflitto o di promuovere violenza.

- **assente:** contenuti offensivi o di odio non sono per forza aggressivi
- **debole:**
  - legittima atteggiamenti o politiche discriminatorie
  - rappresenta il bersaglio come una minaccia da cui difendersi
  - rappresenta il bersaglio come un gruppo che gode di ingiusti privilegi a nostro discapito
  - esprime un'ostilità misurata
- **forte:**
  - si riferisce in modo implicito o esplicito ad azioni violente

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

Criteri operativi

offensività

valori: assente – debole - forte

«Gli immigrati africani in Italia sono ospitati a oziare in alberghi a 3-4 stelle»

«Barletta, sgomberato mega-campo rom… #raccoltadifferenziata»

Intenzione del parlante di ferire il bersaglio, insultarlo o promuoverne un'immagine negativa.

- **assente:** contenuti aggressivi o di odio non sono per forza offensivi
- **debole:**
  - associa il bersaglio a caratteristiche negative o spiacevoli, lo rappresenta come un soggetto da evitare
  - mette in dubbio la legittimità dello status del bersaglio
  - disprezza o deride il bersaglio
- **forte:**
  - usa insulti, espressioni dispregiative o degradanti verso il bersaglio

HATE SPEECH AND SOCIAL MEDIA

I ♥ HATE PREJUDICE

Criteri operativi

ironia

«Toh, che caso: clandestino, islamico e terrorista»

valori: sì - no

Termine ombrello che include sfumature come umorismo, sarcasmo e satira.

Dall'osservazione emerge come possibile strategia per veicolare hate speech velatamente.

- Importante non scambiare la satira per discorso d'odio.

Criteri operativi **stereotipo**

valori: sì - no

«Gli immigrati non muoiono di fatica, sono spesati di tutto»

«Identificato l'autore della rapina all'anziano di Annone Veneto: è un giovane nomade»

Qualunque generalizzazione o approssimazione negativa su un gruppo di persone, che viene rappresentato in base a una caratteristica reale o presunta di alcuni suoi membri, spesso eteroattribuita. L'osservazione mostra che la maggior parte dei casi di hate speech si basano su uno stereotipo.

Alcuni esempi:

- i membri del gruppo sono rappresentati come
  - terroristi, invasori, culturalmente arretrati, misogini, nemici del progresso e dei diritti (**musulmani**)
  - oziosi, opportunisti, dediti alla delinquenza (**immigrati**)
  - sporchi, incapaci di vivere all'interno delle strutture sociali (**rom**)
- i titoli dei quotidiani relativi a fatti di cronaca mettono in risalto la nazionalità/etnia delle persone coinvolte solo se diverse dalla nostra