INTRODUZIONE

La costruzione di un corpus è parte del progetto Hate Speech Monitoring, a cura del Dipartimento di Informatica dell'Università di Torino[1], che mira a individuare, analizzare e contrastare l'hate speech ("hs" d'ora in poi) con un approccio multidisciplinare.
Dal momento che tra le minoranze maggiormente colpite dall'hs i migranti sono quella più vulnerabile e più oggetto di attacchi, il nostro lavoro dedica a questa categoria una particolare attenzione. Tuttavia pensiamo che sia più efficace estrarre una definizione operativa di hs da un insieme di dati più vario ed eterogeneo: per questo abbiamo incluso anche dati relativi da altre categorie molto colpite, ovvero musulmani e rom.
L'identificazione dell'hs è un compito difficile e condizionato da una forte soggettività dei giudizi, soprattutto perché il fenomeno non ha un'unica definizione, ma è costituito da un insieme di variabili che andrebbero considerate caso per caso.
Consapevoli di questo, abbiamo cercato di annotare i tweet non solo in base alla presenza (o assenza) di hs, ma anche in base ad altri parametri che possono di volta in volta rafforzare o mitigare l'impatto del messaggio. I risultati sono uno schema di annotazione e delle linee guida che cercano di includere tutte queste variabili in un insieme coerente. Le categorie di annotazione, oltre all'hs, sono intensità, aggressività, offensività, stereotipo, ironia.

1. TARGET **[***religione – immigrazione - rom***]**

Religione (musulmani)

L'hs contro i musulmani può includere:

- insulti, minacce, espressioni denigratorie o di odio motivate dalla diversità di fede
- incitamento a odio, violenza o violazione dei diritti verso singoli individui o gruppi motivato dalla diversità di fede
- associazione tra fede islamica e propensione al fondamentalismo, al terrorismo, al crimine o a piani di invasione o conquista dell"Europa

Immigrazione

L'hs contro i migranti può includere:

- insulti, minacce, espressioni denigratorie o di odio motivate da differenze di etnia, provenienza, tratti somatici (colore della pelle), lingua o cultura
- incitamento a odio, violenza o violazione dei diritti verso singoli individui o gruppi motivato da differenze di etnia, provenienza, tratti somatici (colore della pelle), lingua o cultura
- associazione tra etnia/provenienza/cultura e abilità cognitive, propensione al crimine, pigrizia o altre caratteristiche antisociali
- riferimenti alla presunta inferiorità o superiorità di alcuni gruppi etnici rispetto ad altri
- delegittimazione di status sociale o affidabilità in base a etnia o provenienza
- riferimenti ad alcune etnie o provenienze come una minaccia per la sicurezza o il benessere degli italiani, o come avversari nella distribuzione delle risorse pubbliche
- espressioni disumanizzanti o associazione con animali o soggetti considerati inferiori/incivili

Rom

L'hs contro i rom può includere:

- insulti, minacce, espressioni denigratorie o di odio motivate dal riferimento all'etnia rom
- incitamento a odio, violenza o violazione dei diritti verso singoli individui o gruppi motivato

[1]http://hatespeech.di.unito.it/

dal riferimento all'etnia rom

- associazione di persone rom con una presunta predisposizione alla delinquenza, al crimine, alla sporcizia e in generale a abitudini antisociali
- uso, anche verso soggetti terzi, di epiteti offensivi o denigratori per le persone rom

2. HATE SPEECH [*sì – no*]

Per la sua identificazione prendiamo in considerazione due aspetti:

- il target: deve essere una delle tre categorie protette elencate sopra, o un individuo considerato per la sua appartenenza alla categoria
- l'azione: deve essere un messaggio che diffonde, incita, promuove o giustifica odio o violenza verso il target, o che cerca di disumanizzare, delegittimare, ferire o intimidire il target

Sì
La presenza di entrambi questi parametri è indispensabile per determinare la presenza di hs.
Ad esempio, questo tweet contiene hs:
«La prossima resistenza la dovremmo fare subito contro gli invasori islamici!»

No
Se anche uno solo dei due parametri non è presente, non possiamo parlare di hs. Qui di seguito una lista di cosa NON è hs, anche se può sembrarlo.

- linguaggio offensivo
- blasfemia e offese alla religione in quanto tale (ovvero rivolte all'insieme di credenze e non ai fedeli)
- negazionismo storico
- incitamento al terrorismo
- diffamazione
- offese a pubblici ufficiali e rappresentanti dello stato in quanto tali

3. INTENSITÀ [*1 – 2 – 3 – 4*]

Tra i tweet che contengono hs è possibile stabilire una scala di intensità. Pur essendo tutti un modo per incitare odio, discriminazione o violenza, alcuni lo fanno in modo velato, implicito e prudente, mentre altri sono espliciti e richiamano apertamente azioni violente. Nel determinare l'intensità hanno un ruolo sia la scelta delle parole, sia l'atteggiamento di chi scrive, sia il possibile effetto su chi legge. Individuiamo quattro gradi di intensità: i gradi 1 e 2 suscitano odio o avversione verso il target, mentre i gradi 3 e 4 incitano violenza o discriminazione.

1
Non c'è incitamento esplicito, ma il messaggio attribuisce caratteristiche negative al gruppo target e a volte suggerisce che questo costituisca una minaccia per chi legge. Il messaggio somiglia ad un insulto o un giudizio basato su uno stereotipo negativo
«Anche il PD se ne accorge: "I migranti sanno solo ostentare l'ozio. La gente è stufa."»

2
Non c'è incitamento esplicito, ma il messaggio ha un effetto delegittimante o disumanizzante verso il gruppo target, oppure mette in dubbio i diritti fondamentali del gruppo trattandoli come privilegi ingiusti. Il messaggio non invoca apertamente la violenza ma incoraggia chi legge a odiare il target e a vederlo come una minaccia o un nemico.
«La polizia i controllori fermano solo italiani, rom e immigrati non li avvicina nemmeno rischiano

la vita»

3
Il messaggio incita esplicitamente ad azioni violente o discriminatorie, ma chi scrive evita di assumersi responsabilità riguardo a quelle azioni. Si limita a giustificarle o sperare che accadano, senza mettersi in gioco in prima persona. Spesso presenta queste azioni in modo distaccato, come qualcosa di oggettivamente necessario o giusto.
«Quella schifosa rom prende anche in giro, speriamo che cn i loro fuochi tossici si brucino e crepino tutti alla svelta, TOLLERANZA 0.»

4
Il messaggio incita esplicitamente ad azioni violente o discriminatorie e chi scrive incita apertamente ad agire, mettendosi in gioco in prima persona e assumendosi responsabilità dirette.
«Hanno rotto il cazzo con tutti questi atti terroristi. Io sono pronto alla guerra.»

4. AGGRESSIVITÀ [*assente – debole – forte*]

Riguarda l'intenzione dell'utente di essere aggressivo, di ferire, o anche di incitare, in diverse forme, azioni violente verso un certo target. Se presente, si può distinguere tra aggressività lieve e aggressività forte.

Debole
Un messaggio è considerato poco o lievemente aggressivo se:

- implica o legittima comportamenti o politiche discriminatorie;

«*Gli italiani prima di tutto!*»

- allude a potenziali minacce alla popolazione italiana poste dalla presenza o esistenza del target, in ragione del loro numero o di qualche loro caratteristica;

«Una nuova invasione di migranti in Europa, la minaccia fa tremare anche l'Italia»

- veicola un senso di frustrazione o insoddisfazione dovuto al (presunto) trattamento privilegiato garantito al target da parte di soggetti pubblici o privati;

«Bisogna diffondere il fatto che il governo vuole requisire le case sfitte per darle ai migranti. Atto antidemocratico»

- esprime atteggiamenti di aperta ostilità, anche se espressa con toni misurati;

«Milva e la "sua" Goro: "Se vivessi ancora li', i migranti li avrei ospitati io" ...dacci l'indirizzo...te li porto io...almeno una dozzina»
«@DrunkyBorghy Fai una cosa; portateli tutti a casa tua!!! Io a casa mia non li voglio, terroristi o non terroristi che siano!!!»

Forte
Un messaggio è considerato fortemente o molto aggressivo se fa riferimento (esplicito o implicito) ad azioni violente o discriminatorie di ogni tipo:
«tutto tempo danaro e sacrificio umano sprecato senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile»

No
Un messaggio può essere non aggressivo anche se contiene hs o offensività:
«@TutteLeNotizie @MediasetTgcom24 e ce l'hai ancora visto che appoggi clandestini rom iussoli e stranieri criminali liberi impuniti a zonzo»

5. OFFENSIVITÀ [*assente – debole – forte*]

Al contrario dell'aggressività, l'offensività riguarda gli effetti potenzialmente dannosi o lesivi del messaggio su un determinato target.

Debole

Un messaggio è considerato poco o lievemente offensivo se presenta almeno una delle seguenti caratteristiche:

- il target è associato a vizi spiacevoli (soprattutto pigrizia):

«Italiani sfrattati e immigrati viziati»
«E meno male che dovevano pagare le nostre pensioni... #migranti #parassiti»
«@RaiStoria @MassimoMasini Gli immigrati africani in Italia, invece, sono ospitati a oziare in alberghi a 3-4 stelle. Bella differenza.»

- o, in generale, a caratteristiche negative:

«Apparte voto subito quando iniziamo a cacciare gli usurpatori migranti anche siriani in mare si sta 'nn ci importa!»

- si mette in dubbio lo status di minoranza svantaggiata o discriminata:

«@ilmessaggeroit Quattro poveri #profughi" fra cui un minore nn accompagnato?»
«PONTIFEX....insiste: il popolo italiano deve essere sommerso e cancellato da finti profughi che vogliono ciò che non hanno mai saputo fare!»
«#dallavostraparte ok occupare non va bene però perché se rom od immigrati fanno vedere bimbi restano ed italiano no, pagava pure per di più»

- i membri del gruppo target sono descritti o considerati come persone spiacevoli da cui è meglio tenersi alla larga:

«@DrunkyBorghy fai una cosa; portateli tutti a casa tua!!! Io a casa mia non li voglio, terroristi o non terroristi che siano!!!»
«Questo vedevano gli abitanti di Roma quando aprivano gli occhi. Adesso solo immondizia, immigrati, caos e tasse»

- c'è un intento derisorio:

«Facciamo partecipare i ""leader rom"" anche al prossimo G8.»

Forte

Il messaggio è considerato fortemente o molto offensivo quando:

- ci si riferisce al gruppo target con espressioni chiaramente offensive o degradanti:

«Barletta, sgomberato mega-campo rom... #raccoltadifferenziata»
«Che strano, il terrorista cotechino si e suicidato in carcere. Impiccandosi con una t-shirt alla porta»

- sono presenti insulti o parole volgari:

«A tutti i musulmani dell'isis, vi sentite forti adesso? Bravi coglioni!»
«@Fcoglioni @erpedrini sti comunisti zecche perché non emigrano in Algeria Arabia ecc. visto che si trovano bene con i loro fratelli islamici»

No

Un messsaggio può essere non offensivo anche se contiene hs, linguaggio aggressivo o uno stereotipo:
«*Trovato in Croazia il tesoro della "regina" rom: sequestro da 6 milioni di euro*»

6. IRONIA [*sì – no*]

Usiamo il termine "ironia" come etichetta generica che include anche sfumature come l'umorismo, il sarcasmo e la satira. In alcuni casi i contenuti di odio infatti possono essere veicolati in maniera sottile, tramite una frase ironica che rende l'odio più difficile da individuare – ma non per questo meno grave.

Sì
Riportiamo alcuni esempi tipici di messaggi ironici usati per attaccare un gruppo target:
«Toh, che caso: clandestino, islamico radicale e terrorista»

Tuttavia, l'ironia può essere usata anche per riferirsi ad un target con intenti umoristici e non denigratori o offensivi:
«1:"Che lavoro vorresti fare?" 2:"Il pescatore a #Lampedusa ." 1:"Ma se sei #vegano !" 2"Devo salvare i migranti mica mangiarmeli."»

Infine, l'ironia può essere presente ma non essere rivolta a uno dei target che ci interessano:
«Vi comprendiamo #USA, anche in #Italia, per decenni, una potenza straniera interferì in campagna elettorale, VOI!»

7. STEREOTIPO [*sì - no*]

Determina se il messaggio contiene riferimenti (impliciti o espliciti) a, o si basa su, idee o tratti negativi attribuiti genericamente all'intero gruppo. In alcuni casi gli stereotipi possono essere alla base dei discorsi che promuovono o veicolano odio verso alcuni gruppi di minoranza.

Sì
Un messaggio contiene uno stereotipo se presenta almeno una delle seguenti caratteristiche:

- i membri di una categoria target sono descritti come invasori (o simili attributi)

«Se gli italiani continuiamo a non fare figli ci ritroveremo presto sottomessi dai musulmani. Fermiamoli prima che sia troppo tardi»

- o come opportunisti, buoni a nulla, parassiti

«Gli immigrati non muoiono di fatica . Sono spesati di tutto.»

- o come criminali o delinquenti

«@TutteLeNotizie @MediasetTgcom24 e ce l'hai ancora visto che appoggi clandestini rom iussoli e stanieri criminali liberi impuniti a zonzo»

- o come persone sporche e ripugnanti

«Mentre la #Raggi investe 12 milioni di euro in favore dei #rom un migrante CAGA tranquillamente davanti all'altare del...»

- quando un titolo di giornale relativo a notizie di cronaca riporta la nazionalità, etnia o religione dei soggetti interessati senza che questo sia necessario per la comprensione dell'informazione, contribuendo così ad associare determinati comportamenti a determinate identità

«Identificato l'autore della rapina all'anziano di Annone Veneto: è un giovane nomade» [ad esempio, in questo caso non è necessario specificare che l'autore è un nomade, e farlo suggerisce che ci sia una correlazione tra questa caratteristica e il compimento della rapina]

- per quanto riguarda in particolare i musulmani, sono frequenti alcuni stereotipi che li rappresentano come misogini, antidemocratici e violenti

«e quelle col burqua dov'erano? Non gli hanno consentito di manifestare i maritini islamici integrati?»
«@marini_valerio il bello della democrazia a cui siamo abituati noi è questo, tu hai la tua opinione

e io la mia. Chissà nei paesi islamici?»
«#islam "religione" di pace e amore...»

No
Attenzione: tweet che criticano o smentiscono stereotipi o notizie calunniose e parziali non contengono stereotipi.
«Psicosi (e leggende) galoppano su Facebook: Attenzione al furgone pieno di rom accanto alla...»

*****NOTA*****

Nel caso di tweet che contengono un sentimento di odio diretto a un target diverso dai tre presi in considerazione (quindi, ad esempio, verso personaggi famosi, politici, rappresentanti delle istituzioni e via dicendo), l'hs deve essere segnato come assente. Per questa ricerca ci interessa evidenziare solo l'odio che colpisce immmigrati, musulmani e rom. Le altre categorie, invece, vanno annotate indipendentemente dal target.

Esempio:
«#POLITICA la ue di m.:"l'italia discrimina i migranti" sono lontani da noi, SONO GLI ITALIANI DISCRIMINATI!!PEZZI DI M»
Hate speech: **no**
Aggressività: **forte**
Offensività: **forte**
Ironia: **no**
Stereotipo: **sì**