UNIVERSITÀ DEGLI STUDI DI TORINO

DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA IN INFORMATICA

TESI DI LAUREA DI I LIVELLO

CURRICULUM: INFORMAZIONE E CONOSCENZA

IntersHate: un corpus italiano per lo studio di misoginia e intersezionalità in Twitter

Prima Relatrice
Prof.ssa Viviana Patti

Secondo Relatore
Dott. Mirko Lai

Candidato
Ivan Spada, 861723

Anno Accademico 2019-2020

«Se si ammettono le parole dell'odio nel contesto pubblico, se si accoglie lo hate speech nella ritualità del quotidiano, si legittimano rapporti imbarbariti.
Io l'odio l'ho visto. L'ho sofferto. E so dove può portare.»

Liliana Segre

Title

IntersHate: an Italian corpus to study misogyny and intersectionality on Twitter.

Abstract (English)

This thesis contributes to computational linguistics studies. The paper explains the research carried out on *online hate speech detection* and misogyny on social media concentrating on intersectionality of hate. Given the main focus on the phenomenon of online misogyny, in this study we have developed a *corpus-based* Twitter data analysis around victims of online hate campaigns selected to study how misogyny and sexism are expressed in texts, intersecting with other categories of hate and social discrimination, such as xenophobia, racism and islamophobia. The study includes the analysis of hatred in Twitter according to targets and events and the process of developing the *IntersHate corpus*. The latter consists of several phases, from the collection in digital format of linguistic data representative of the debates around the victims of hatred on Twitter, to the annotation of the corpus according to a novel multi-layer scheme designed to assess the presence of intersectional hate. The Twitter data collection and filtering are based on a mixed methodology involving keywords, hashtags and conversational threads concerning the selected debates. The corpus analysis includes the labels distribution analysis on several layers, the analysis and discussion of the inter-annotator agreement and, lastly, an intrinsic and comparative analysis of the linguistic and lexical features of the annotated texts, relying on the *HurtLex* computational hate lexicon.

Titolo

IntersHate: un corpus italiano per lo studio di misoginia e intersezionalità in Twitter.

Abstract (Italiano)

Il contributo della tesi si colloca nell'ambito della linguistica computazionale e in particolare l'elaborato descrive la ricerca effettuata sul tema dell'*hate speech online* e della misoginia nei social media con particolare concentrazione sull'analisi dell'odio intersezionale. Posto il focus sulla misoginia, lo studio è stato condotto mediante l'analisi *corpus-based* di dati Twitter intorno a vittime di campagne d'odio online selezionate in modo da studiare come la misoginia e il sessismo affiorano nei testi intersecandosi anche con altre categorie d'odio e discriminazione sociale come xenofobia, razzismo e islamofobia. Lo studio comprende l'analisi dell'odio su Twitter in funzione dei target e degli avvenimenti che li coinvolgono e il processo di creazione del corpus *IntersHate*. Quest'ultimo si compone di diverse fasi, dalla raccolta in formato digitale di dati linguistici rappresentativi dei dibattiti intorno alle vittime dell'odio su Twitter, all'annotazione del corpus secondo un nuovo schema multilivello disegnato per valutare la presenza di odio intersezionale. La collezione e selezione dei dati Twitter si basa su una metodologia mista che coinvolge keyword, hashtag e thread conversazionali riguardanti i dibattiti scelti. L'analisi del corpus include l'analisi della distribuzione delle etichette sui vari livelli, l'analisi e la discussione dell'agreement fra gli annotatori umani e l'analisi intrinseca e comparativa relativa alle caratteristiche lessicali dei testi annotati tramite l'uso del lessico computazionale *HurtLex* di parole per ferire.

Indice

Indice delle figure

Indice delle tabelle

Introduzione

La diffusione capillare degli apparati elettronici ha favorito l'accesso di una vasta gamma di servizi web ad un numero elevato di utenti di diverse generazioni e posizioni sociali. I social network si sono integrati nel quotidiano di gran parte popolazione, non ospitano solo spiragli della vita privata degli iscritti ma fungono anche da vettore di comunicazione per attività commerciali e politiche. Il web, come mezzo libero di diffusione di notizie, è uno degli strumenti fondamentali in un sistema democratico per tenersi aggiornati velocemente e discutere criticamente le informazioni diffuse attraverso i differenti canali di comunicazione. Allo stesso tempo può fungere da collettore di contenuti violenti e potenzialmente pericolosi facilmente fruibili su smartphone, tablet e pc.

Le domande "A cosa stai pensando?" e "Cosa c'è di nuovo?", a cui gli utenti dei social network rispondono per dare spazio alle proprie opinioni, raccontare e raccontarsi, non sempre sono l'origine di un confronto utile alla crescita personale e di una comunità. Spesso la rete viene utilizzata come una piazza dove poter urlare in *caps lock* ignorando le relazioni fra le proprie azioni nel web e la vita reale. I due spazi sembrano distanti fra loro ma una reazione online, soprattutto se virale, può comportare conseguenze anche disastrose nella vita privata di chi interpreta il ruolo della vittima.

Per molti utenti consapevoli, ce ne sono altri che, per mancanza di conoscenza, responsabilità o per rabbia, diffondono contenuti d'odio. I principali bersagli sono soprattutto i soggetti appartenenti a categorie sociali divergenti rispetto allo standard assunto dalla società come "corretto", solitamente inteso come "uomo, bianco, etero e cisgender". Misoginia, xenofobia, islamofobia, omolesbobitransfobia, antisemitismo e abilismo sono le principali categorie di *hate speech online* rilevate in *La mappa dell'intolleranza* [Vox, 2019] e *Il barometro dell'odio – sessismo da tastiera* [Amnesty International, 2020]. La coesistenza di differenti forme di discriminazione pone il *focus* sul concetto di intersezionalità [Crenshaw, 1991] che riconosce chi si trova nell'incrocio come una nuova categoria. Crenshaw [1991] afferma la presenza di nuovi bersagli che, discriminati su più fronti, sono maggiormente colpiti e non hanno voce a causa dell'incapacità di definirli.

Dato l’assottigliamento dei confini fra piattaforme social e reale, questo elaborato punta ad analizzare come e quanto l’odio misogino si interseca con altre categorie sociali discriminate e quindi come possa manifestarsi su *target* specifici in Twitter, una piattaforma di *microblogging* molto utilizzata che facilmente accessibili tramite API. Il lavoro si divide in quattro fasi: monitoraggio dei flussi di discorso intorno ai soggetti *target*, raccolta e selezione dei dati da Twitter e Twita [Basile et al., 2013], annotazione del *corpus IntersHate* e analisi dei risultati orientata alla comprensione delle dinamiche dell’odio che include la discussione dell'*agreement* fra gli annotatori e l’analisi lessicale attraverso il lessico *HurtLex* [Bassignana et al., 2018].

La tesi è organizzata in sette capitoli: il primo illustra i concetti di base della ricerca relativa alle diverse tipologie di hate speech e all’intersezionalità della discriminazione su Twitter, il secondo elenca gli strumenti utilizzati, il terzo espone le considerazioni in merito alle scelte accolte per definire il *corpus IntersHate*, il quarto contiene il processo di realizzazione del nuovo schema di annotazione multilivello, i capitoli cinque e sei descrivono le analisi dei risultati rispettivamente sul piano dei *target* e dell’intersezionalità, infine il capitolo sette espone le conclusioni della ricerca.

1.Analisi del fenomeno dell'odio e basi della ricerca

In questo capitolo vengono introdotti i concetti alla base della ricerca ponendo il *focus* sul fenomeno dell'*hate speech*, le categorie sociali discriminate e come queste vengono colpite, il funzionamento e l'utilizzo dei social network in particolare di Twitter. Infine, vengono esposti il concetto di intersezionalità dell'odio e quindi le domande di ricerca.

1.1.Hate speech: descrizione del fenomeno

Nonostante sia difficile convergere verso una definizione condivisa universalmente del termine hate speech, sono riportate qui alcune definizioni come punto di partenza. Il dizionario Treccani propone la seguente definizione del termine *hate speech*: *"espressione di odio rivolta, in presenza o tramite mezzi di comunicazione, contro individui o intere fasce di popolazione (stranieri e immigrati, donne, persone di colore, omosessuali, credenti di altre religioni, disabili, ecc.)"*[1]. Una definizione molto condivisa e più focalizzata sull'odio su base etnica è quella del Consiglio d'Europa: *«The term "HS" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin»* [ECRI, 2007] che pone l'accento sugli aspetti relativi all'incitamento, promozione e giustificazione all'odio.

I social network sono luoghi fertili per la proliferazione di discorsi d'odio che persistono nel tempo e nello spazio. La rete permette la diffusione veloce e ad un pubblico vasto di contenuti con *hate speech* testuali e multimediali. I discorsi d'odio si differenziano da quelli aggressivi per lessico e scopo, le due categorie possono

[1] *https://www.treccani.it/vocabolario/hate-speech_res-2f344fce-89c5-11e8-a7cb-00271042e8d9_(Neologismi)*

intersecarsi: non necessariamente l'aggressività viene espressa attraverso forme d'odio e *viceversa*. Non è semplice tracciare la linea per stabilire se un contenuto appartenga a una o entrambe le categorie, i contenuti negativi devono assumere forme d'odio verso categorie o individui per essere riconosciuti come *hate speech*. I termini *n**ro, p****na, checca* ecc. hanno una connotazione negativa facilmente identificabile come *hate speech* mentre altri come *immigrato, finocchio, vacca, befana, cozza, gallina, oca* ecc. assumono significati e intenzioni differenti a seconda del contesto. Un tema importante è quello della libertà di espressione e della libertà di esprimersi in modo offensivo, e del rischio che il monitoraggio possa aprire la strada ad azioni di censura. La visione risulta più chiara realizzando che è possibile esprimere disapprovazione senza colpire le categorie vulnerabili e quindi utilizzando un linguaggio equilibrato che non abbia come scopo l'offesa di terzi. La Corte Europea dei Diritti dell'Uomo ha ribadito che "le leggi volte a contrastare il linguaggio dell'odio e a reprimere atti ispirati dal razzismo e dalla xenofobia, rappresentano – in una società democratica – una limitazione legittima della libertà di espressione in favore della tutela necessaria della reputazione degli individui e delle libertà fondamentali" [Vox, 2019]. Sebbene esistano associazioni come *Odiare ti costa*[2], in prima linea per ostacolare le discriminazioni online, è ancora difficile stabilire il limite di cosa è concesso e il panorama della normativa europea sul tema è variegato.

L'incitamento all'odio (che spesso si accompagna alla diffusione di notizie infondate e coesiste in alcuni casi con l'incitamento alla violenza) sfrutta la persuasione per trovare nemici comuni e trasformarli in *target*, accade nei confronti di singoli e intere categorie sociali che in genere divergono dagli ideali, gli scopi e dalle caratteristiche dell'artefice. Questo tipo di retorica utilizza a proprio vantaggio dinamiche di *in-out group* di appartenenza e non-appartenenza studiate dalla psicologia sociale per influenzare l'opinione pubblica, facendo leva sui sentimenti e sul bisogno di trovare un capro espiatorio per coloro che necessitano di riconfermarsi nelle proprie certezze identitarie. Una particolare forma di incitamento è la gogna pubblica: spinge le vittime in pasto ai *leoni da tastiera* (o *conigli digitali* come li definisce la senatrice a vita Liliana Segre) avendo fra gli effetti collaterali la *spirale del silenzio* che isola il malcapitato dall'opinione di maggioranza avendo come effetti collaterali la riduzione al silenzio del soggetto che in alcuni casi abbandona lo spazio social eliminando il proprio profilo. Si è

[2] *https://www.odiareticosta.it/*

osservata inoltre un'asimmetria di genere rispetto ai bersagli di campagne d'odio. Nel *corpus* de Il barometro dell'odio [Amnesty International, 2020] appare solo un uomo posto alla gogna. Il fatto che a percentuale sia nettamente inferiore rispetto ai bersagli femminili, mette in luce una forma implicita di sessismo che compare sulle piattaforme.

Un genere di attacco programmato è il così detto *shitstorm*, che si configura come un vero e proprio attacco orchestrato online portato avanti da un gruppo di *troll* contro almeno un bersaglio, che sommergono di contenuti negativi le pagine delle vittime in massa e in modo temporalmente concentrato. Chi orchestra queste azioni conosce e sa sfruttare bene le dinamiche delle piattaforme, che si prestano ad azioni tese a sminuire un bersaglio colpendolo ripetutamente (come lapidandolo in una piazza) senza esporsi in prima persona. Molti attacchi vengono effettuati da persone senza generalità e *nickname* rintracciabili, da profili falsi e da *bot*. Tuttavia, la questione è più complessa di come sembra, e il fattore dell'anonimato, a cui spesso si fa riferimento per caratterizzare i comportamenti degli *haters*, non è sempre centrale. Vittorio Lingiardi, professore ordinario di Psicologia Dinamica presso La Sapienza di Roma, ha notato una variazione delle attitudini degli *haters* che tendono anche a voler farsi riconoscere sentendosi legittimati e parte di un gruppo [Vox, 2019].

Stereotipi e pregiudizi sono spesso alla base dell'*hate speech*, che sfrutta notizie infondate e paura. Fra questi stereotipi e pregiudizi appaiono argomenti come igiene, cibo, posizioni sociali e intenzioni fraudolente quali rubare il lavoro e commettere crimini come violenze sessuali e terrorismo. L'intenzione è di trovare i difetti e usarli contro i bersagli che possono essere persone o intere categorie sociali. Viene stabilita una netta differenza tra chi attacca e gli attaccati, i secondi comprendono anche chi li difende poiché posti in opposizione ai primi sottolineando il concetto *in-out group*. Un tipo di attacco è la criminalizzazione della vittima che appare ad esempio nel caso di Carola Rackete, i pro-immigrazione definiti anti-italiani [Libero, 2020] e coloro a favore di aborto, unioni civili e adozioni che vengono accusati di minacciare la famiglia tradizionale e la vita. Similmente, le donne sono spesso incolpate delle molestie e violenze sessuali subite.

La distribuzione degli odiatori è trasversale rispetto a età, posizione e categoria sociale, non sempre formano un'intersezione vuota con le vittime e non manca l'*hate speech* espresso da individui all'interno delle categorie colpite: donne, immigrati, islamici ecc. Al seguito del *body shaming* nei confronti di Giovanna Botteri [La Repubblica,

2020], presente in una puntata di *Striscia la notizia* che afferma di averlo fatto in maniera simpatica, alcuni utenti hanno preso l'iniziativa di sostenere la giornalista attaccando nella stessa maniera Michelle Hunziker contribuendo alla propagazione del fenomeno senza risolverlo o limitarlo. Si è osservato come le pratiche di *counter-speech* portate avanti da utenti che prendono la difesa delle vittime di *hate speech*, talvolta si sforzano di esprimere il contrasto proponendo narrative alternative a quelle che sottendono in discorso d'odio, altre volte abbracciano strategie di difesa a loro volta offensive, portando a un'amplificazione dell'intolleranza.

Le mappe termografiche dell'intolleranza di Vox [2019] sono uno dei principali indici di discriminazione sui social network in Italia. Le categorie sociali prese in considerazione sono: donne, omosessuali, migranti, disabili, ebrei e musulmani. Nell'ultima analisi di marzo-maggio 2019 (*Tabella 1*), il 70% di *tweet* sono stati di natura negativa vedendo Migranti, Donne e Islamici come le categorie più colpite. Nella versione 4.0 delle mappe dell'intolleranza [Vox, 2019] è stata aggiunta una misura relativa al livello di aggressività come fattore per l'estrazione, mediante *software*, di contenuti aggressivi a cui si accompagna un grado di virulenza. La valutazione, in fase sperimentale, ha seguito le categorie della scala MOAS (Modified Overt Aggression Scale) dimostrandosi utile per comprendere la negatività, l'aggressività, gli atteggiamenti intolleranti e discriminanti espressi nei testi [Vox, 2019].

	Tweet totali	**Tweet negativi rilevati**	**Tweet negativi rilevati**
Migranti	74.451	49.695 (32%)	22.043
Donne	55.347	39.876 (27%)	17.242
Islamici	30.387	22.537 (15%)	8.673
Disabili	23.499	16.676 (11%)	3.430
Ebrei	19.952	15.196 (10%)	6.943
Omosessuali	11.741	7.808 (5%)	3.312
TOTALI	215.377	151.783 (70%)	61.643

Tabella 1: risultati della mappa dell'intolleranza 4.0 [Vox, 2019]

Un monitoraggio periodico in Italia della diffusione dell'*hate speech* è quello effettuato dal movimento globale Amnesty International [2020], che nel 2020 ha pubblicato i risultati del monitoraggio nell'ambito del progetto denominato "Il barometro dell'odio": il 14% dei contenuti sono offensivi, discriminatori o *hate speech*; l'ultimo appare con il 0.7%. Il monitoraggio, durato cinque settimane, ha collezionato 42.143 elementi. Alcuni odiatori, definendosi vittime, accusano i loro bersagli di essere i veri odiatori, succede ad esempio con i contenuti xenofobi, razzisti e islamofobi per proteggere le tradizioni, misogini per mantenere le posizioni di genere, omolesbobitransfobici e contro l'aborto in difesa della famiglia e della vita. Con *Il barometro dell'odio* è stato valutato l'incrocio che si presenta fra le espressioni di odio nei confronti di donne, musulmani e migranti. Nei contenuti che parlano di "donne e diritti di genere" il 23% dei bersagli sono soggetti femminili mentre i migranti e i musulmani appaiono entrambi nell'11% dei casi.

Infine, per completare il quadro del contesto, è importante notare che l'85% dei giovani, fra gli 11 e i 17 anni, possiede uno smartphone e il 72% naviga su internet tutti i giorni, di questi il 33% delle ragazze e il 24% dei ragazzi ammettono di provare ansia legata alla pubblicazione dei contenuti [Amnesty International, 2020].

1.2. Tipologie d'odio

In questo paragrafo sono esposte le discriminazioni valutate durante la ricerca traendo ispirazione dagli studi di Vox [2019] e Amnesty International [2020].

Misoginia

La misoginia, come atteggiamento di avversione generica per le donne, appare in maniera più o meno esplicita e afferma la dominanza maschile presentandosi in forme come sessismo, molestie sessuali e discredito del genere. Secondo il barometro dell'odio [Amnesty International, 2020], le donne vengono colpite per caratteristiche quali immoralità, autonomia, libertà di scelta e sostenimento di altre categorie odiate come musulmani e migranti, quindi principalmente quando esprimono opinioni e sostengono diritti.

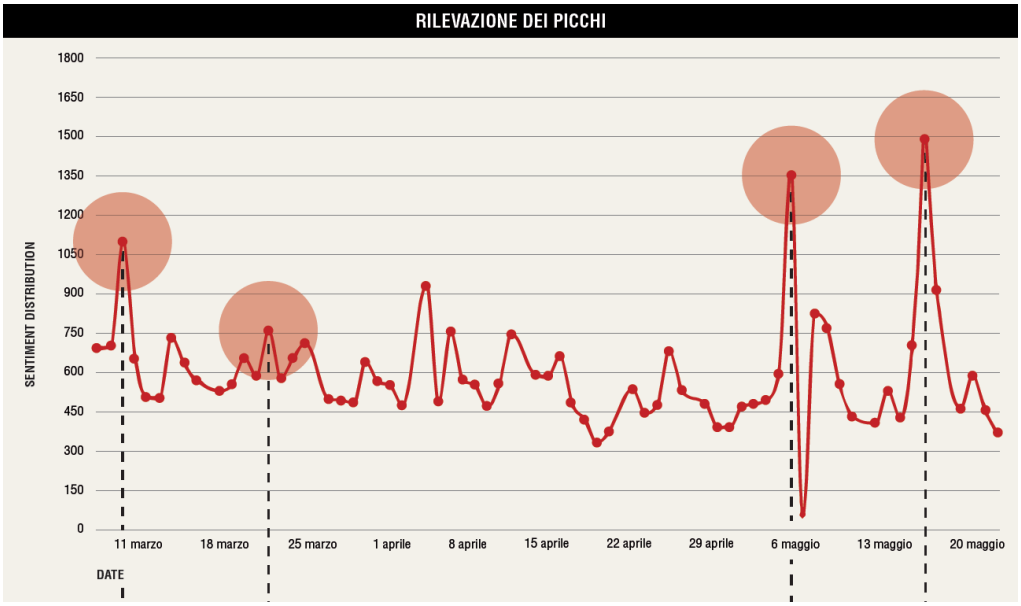

Figura 1: misoginia in corrispondenza degli eventi durante marzo-maggio 2019 [Vox, 2019]

Gli stereotipi, che ricoprono il genere, coinvolgono i diritti e doveri dei facenti parte. Il *soffitto di cristallo* è la principale metafora che viene utilizzata per indicare la difficoltà di fare carriera per le donne. Le mansioni più importanti e rappresentative sono a maggioranza maschile e resiste lo stereotipo di una figura femminile che si colloca principalmente fra le mura domestiche, dedita alla cura della famiglia, screditandone le capacità e alimentando i *bias cognitivi* causati dalla scarsa omogeneità di rappresentanza. La distinzione fra sesso forte e sesso debole pone metà della popolazione in una posizione di supposta inferiorità, a fronte di una mascolinità tossica che vuole essere dominante e si sente legittimata nell'esercitare potere sul prossimo. Si stima che il 43,6% delle donne fra i 14 e i 65 anni abbiano subito qualche forma di molestia sessuale nel corso della loro vita [Istat, 2018]. Il fenomeno non si limita al mondo reale, anche le piattaforme social sono ambienti di espressione di *sexual harassments* sottolineando quanto lo spazio digitale non sia distante dal quotidiano.

Durante il periodo marzo-maggio 2019 le donne sono risultate la terza categoria più odiata dopo musulmani e migranti, il 33% degli attacchi personali a loro rivolti sono sessisti [Amnesty International, 2020]. Silvia Brena, giornalista e Co-fondatrice di Vox - Osservatorio Italiano sui Diritti, afferma che durante gli ultimi due mesi del 2019 è aumentato l'odio contro le donne su Twitter: quasi un utente su due parla di donne e il

70% manifesta odio nei loro confronti. Dalla rilevazione di marzo-maggio 2019 a novembre-dicembre 2019, le donne sono diventate la categoria più soggetta a *hate speech* su Twitter aumentando di dodici punti percentuale arrivando al 39% di tweet negativi [Amnesty International, 2020].

Da *Il barometro dell'odio* emerge come il sessismo compaia anche in maniera implicita considerando la differenza fra le gogne pubbliche nei due sessi: solo un uomo ne è stato vittima. Fra le principali tipologie di attacchi ci sono il *body-shaming* per fare leva su eventuali fisicità dei *target*, lo *slut-shaming* anche nei confronti delle vittime di *revenge porn* e il *victim-blaming* che è costituito dall'attribuzione di colpa alle vittime come "te la sei cercata", "eri troppo scoperta", "perché non ti sei rifiutata?" ecc.

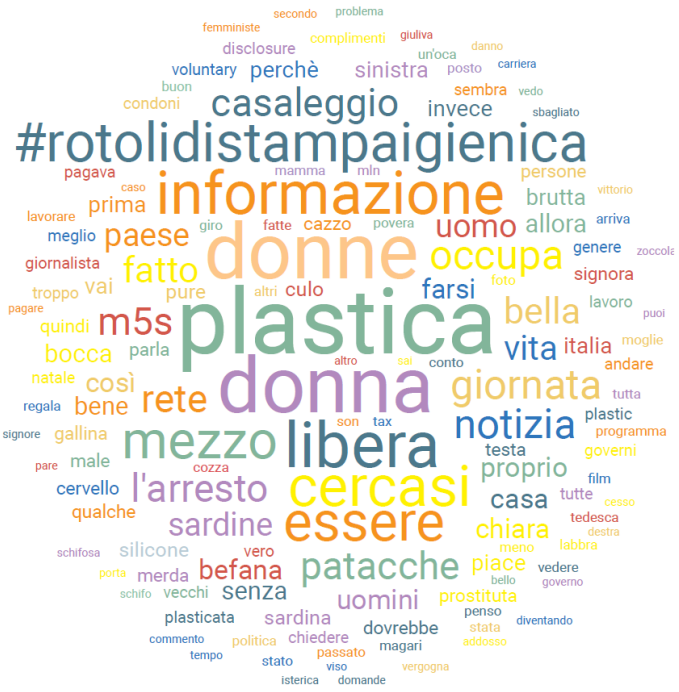

Figura 2: le 150 parole più frequenti nei contenuti negativi di matrice sessista [Amnesty International, 2020]

Xenofobia, razzismo e minoranze religiose

La xenofobia e il razzismo, rivolti maggiormente ai migranti, vogliono sostenere la diversità tra popoli provenienti da altre aree geografiche riconoscendo il proprio come superiore e corretto. Fra gli insulti più comuni, c'è la criminalizzazione della categoria sociale e la deumanizzazione dei *target* che non risparmia nemmeno la politica vedendo

Cecile Kyenge, ministra dell'integrazione negli anni 2013-2014, associata a una scimmia a causa della sua provenienza [La Repubblica, 2014].

"Aiutiamoli a casa loro" è uno degli *slogan* usati per esordire la propria disapprovazione nei confronti dell'immigrazione e si colloca nel contesto del fenomeno sociale *Nimby*: dirotta il problema al di fuori del proprio giardino volendo virare gli attracchi verso altri Stati fino ad estendere il proprio giardino all'Europa geografica e quindi sostenere l'interruzione del flusso migratorio. Questo atteggiamento appare contradditorio quando, oltre ad affermare l'aiuto nel loro Paese come soluzione, vengono attaccate le associazioni e i volontari che sostengono le popolazioni locali e le infrastrutture nei Paesi meno sviluppati, come è successo per esempio nel caso di Silvia Romano.

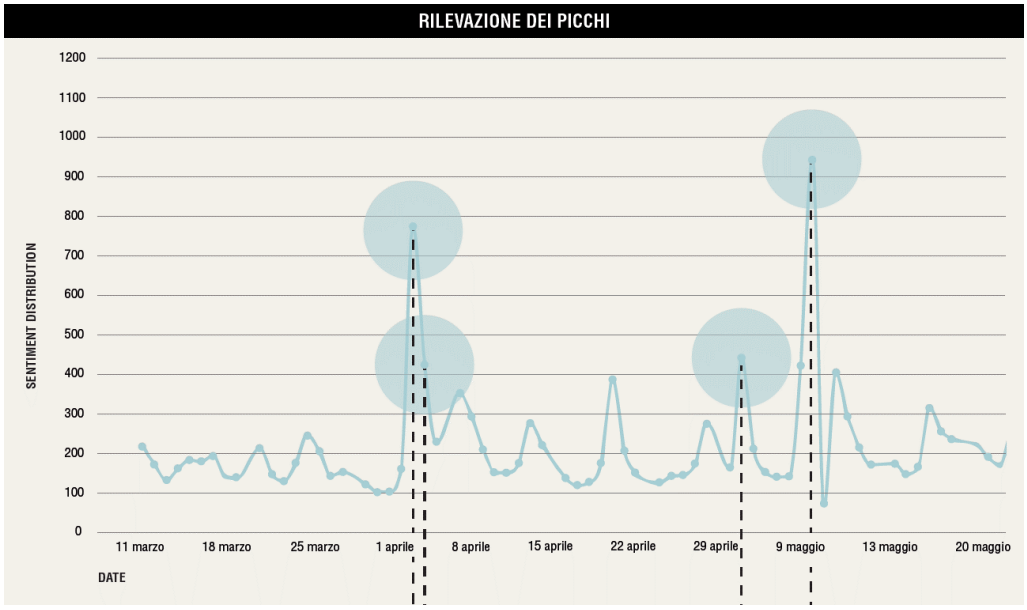

Figura 3: xenofobia in corrispondenza degli eventi durante marzo-maggio 2019 [Vox, 2019]

L'Islam, spesso associato al terrorismo, si colloca fra i primi tre bersagli rilevati delle analisi effettuate da Vox [2019] e Amnesty International [2020] vedendo la categoria colpita nel 15% dei casi e notando una particolare correlazione tra la narrativa politica e la sequenza degli eventi.

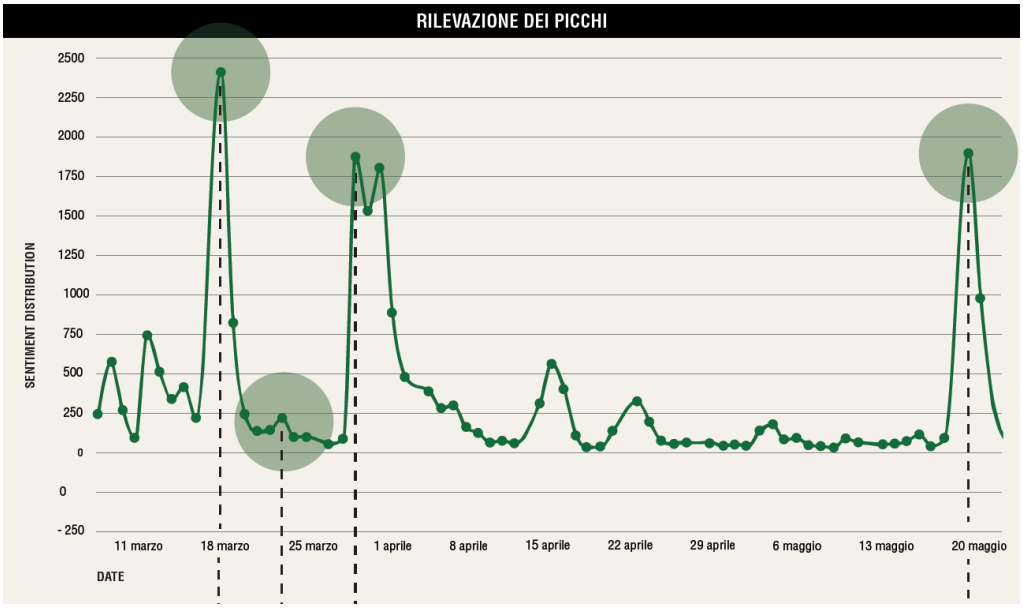

Figura 4: islamofobia in corrispondenza degli eventi durate marzo-maggio 2019 [Vox, 2019]

Durante il periodo di novembre-dicembre 2019, in concomitanza con le minacce rivolte a Liliana Segre e l'istituzione della commissione contro l'odio [La Repubblica, 2019], l'*hate speech* antisemita su Twitter ha raggiunto il 25%, 15 punti in più rispetto a marzo-maggio dello stesso anno [Amnesty International, 2020] ed è importante ricordare che il 38% degli ebrei europei ha pensato di lasciare il proprio Paese a causa dell'antisemitismo crescente [Vox, 2019].

1.3. Social network e Twitter

I social network sono il riflesso della nostra società, la mancanza di inclusività e la legittimazione del linguaggio d'odio danneggiano il quotidiano passando dalla parola ai fatti. Silvia Brena riconosce una correlazione tra *hate speech* e *hate crime*. Sostiene che "i social media sono diventati una corsia preferenziale di incitamento all'intolleranza e al disprezzo nei confronti di gruppi ritenuti socialmente più deboli. […] Tendiamo a incontrare persone che la pensano come noi, il che aumenta l'effetto di polarizzazione delle opinioni" [Amnesty International, 2020]. Fomentare odio all'interno dei *feed* può alterare la concezione della realtà dei *follower*, o amici, che vedono l'informazione sobbalzare fra i contenuti dell'*infinite scroll* dando origine a un'*echo chamber* e

alimentando il *confirmation bias* [Zollo, 2018], quindi influenzano la percezione del pericolo e del peso dei fatti quotidiani a causa della mancanza di confronto.

Gli algoritmi dei *feeds* somministrano informazioni in maniera mirata e controllata. L'informazione erogata sotto forma di *status*, gestita da codici basati su tendenze, interessi e dati degli utenti, comporta una visione ristretta della realtà applicando i principi di *agenda setting/cutting* [McCombs et al., 1972] che danno risonanza a una ridotta varietà di contenuti ponendo in una bolla gli utilizzatori (*filter bubble*). In genere non viene garantita una eterogeneità di contenuti tale da mantenere una visione equilibrata del mondo reale.

Twitter è un social network, diffuso a livello globale, che pone il suo punto di forza nella facile condivisione di contenuti organizzati per discussioni attraverso *thread* conversazionali e *hashtags*. La più forte restrizione della piattaforma è il limitato numero di caratteri concessi per *tweet* che, nonostante l'incremento effettuato negli ultimi anni, non lascia spazio a molte argomentazioni e favorisce un linguaggio conciso e senza compromessi. Tutti gli utenti sono posti sullo stesso piano concedendo la possibilità di interfacciarsi facilmente con le figure notorie che diversamente risulterebbero troppo distanti dai cittadini comuni. A differenza di Facebook, la condivisione dei contenuti altrui si effettua con il *retweet* e sarà visibile non solo agli amici ma anche in tutte ricerche per *trends* e *hashtags* allargando lo spettro del pubblico raggiungibile. La facile comunicazione, il potenziale pandemico, l'assottigliamento del *gap* sociale, la possibilità di far esprimere le opinioni di tutti ed evidenziare i *trends* del momento, hanno favorito la politica e i giornali che hanno spesso scelto la piattaforma come vettore principale per la fruizione di informazioni e iniziative. È importante capire quanto un tweet possa alterare la borsa, i mercati e le decisioni politiche: i *tweet* del ministro dell'Interno 2018 con l'*hashtag #chiudiamoiporti* hanno impedito l'attracco, senza alcun atto formale da parte dei Ministeri competenti, di navi che trasportavano migranti richiedenti asilo e potenzialmente titolari del diritto costituzionale all'asilo (art. 10, comma 3 Cost.) [Vox, 2019].

La facile canalizzazione dei contenuti rappresenta un punto di forza per le varie polarizzazioni rafforzando il concetto di comunità secondo cui gli utenti tendono a schierarsi esprimendo le proprie cause. L'assenza di gerarchia fra le figure notorie e il popolo ha concesso ai secondi di indignarsi, senza filtri, con i primi alimentando la

disapprovazione e la sfiducia nelle istituzioni. Appaiono le correnti *#facciamorete* che sfruttano la visibilità sul social per fare opposizione e mostrarsi uniti di fronte a un'idea, un gruppo o una persona. Utilizzano simbologie, come *emoji*, per identificarsi parte di una comunità e rafforzano l'eco delle informazioni all'interno dei *feeds* polarizzando le narrative online. Nascono anche i disturbatori seriali che, ignorando la possibilità di dialogo, interagiscono con i *target* commentando con contenuti fastidiosi e anche non pertinenti.

Il social network è fornito di algoritmi per il rilevamento di violazioni dei termini e condizioni. I programmatori, come ogni individuo, sono soggetti a *bias* cognitivi che rischiano di discriminare alcune categorie di utenti in maniera anche involontaria. Appare la contraddizione su come sia possibile che uno strumento contro le discriminazioni vada a discriminare a sua volta sulla base di caratteristiche individuali e favoritismi politici. Dall'intersezione fra etnia e genere, è stato dimostrato che Twitter, negli Stati Uniti, tende a riconoscere molti falsi positivi nei contenuti degli afroamericani [Kim et al., 2020].

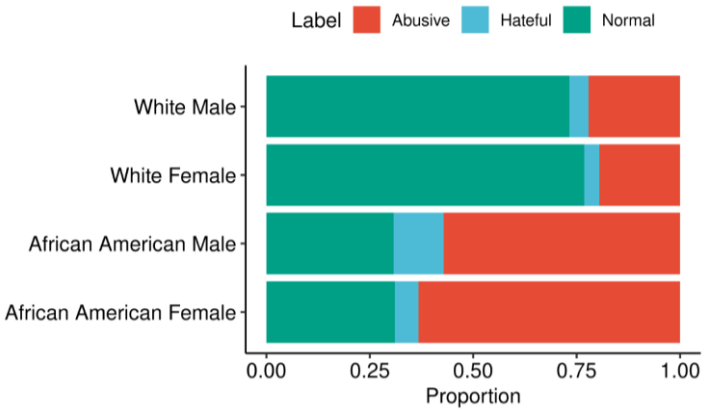

Figura 5: come l'etnia interagisce col genere [Kim et al., 2020]

Per ridurre i *bias* è strettamente necessario utilizzare campioni casuali per il *training*. Inoltre, è essenziale organizzare il lavoro in gruppi eterogenei in modo da sondare la presenza di particolari polarizzazioni negli algoritmi che andranno a interfacciarsi con la quotidianità degli utenti, questo per cercare di garantire il *safe space* alle categorie vulnerabili.

1.4. Odio intersezionale

Con il termine *intersezionalità* si intende la coesistenza di molteplici forme di disuguaglianze quali etnia, genere, orientamento sessuale, classe, disabilità, età ecc. che possono generare ostacoli nel quotidiano di coloro che vengono marginalizzati perché diversi [Crenshaw, 1991].

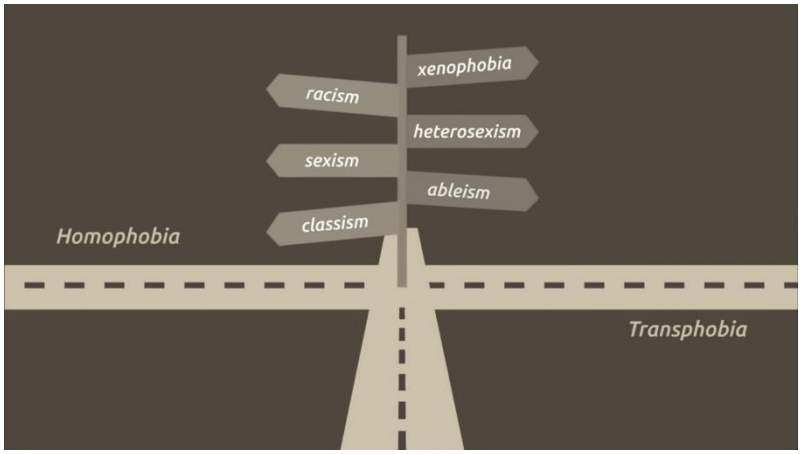

Figura 6: intersezionalità [Crenshaw, 2016]

Kimberle Crenshaw [1991] afferma che non potendo vedere il problema non lo si può risolvere e ha raccontato come, nel caso della misoginia, le discriminazioni prendano forme differenti in base ad altre dimensioni come etnia e classe sociale, non è quindi possibile analizzare il fenomeno senza metterlo in relazione con le altre caratteristiche coesistenti. Ha concettualizzato il fenomeno trattando le discriminazioni razziali e di genere delle donne nere, ha dimostrato che l'intersezione viene colpita diversamente rispetto alle dinamiche comuni di razzismo e di misoginia. La presenza di più forme di discriminazione contemporaneamente non corrisponde a una semplice somma bensì a uno sviluppo a sé stante. Un problema significativo nel contesto nord americano è stato quello delle donne nere abusate fra mura domestiche, la scelta era fra denunciare il *partner* ed essere espulse e costrette a fare ritorno nel paese d'origine oppure rimanere in silenzio [Crenshaw, 1991]. La corrente femminista e quella antirazzista sono state per molto tempo distanti. Allo stesso modo il patriarcato, radicato anche nel contesto della popolazione afroamericana, confermava la dominanza degli uomini neri sulle donne nere isolandole e non permettendo loro una sufficiente rappresentanza [Crenshaw, 1991].

Nel contesto di questi studi viene introdotto anche il concetto di *derailing*, ripreso anche da Fersini et al. [2018], che si riferisce all'intenzione di deviare il sostegno dei diritti delle donne afroamericane, indirizzando il *focus* su un discorso più confortevole agli altri manifestanti e quindi ignorando un'intera categoria soggetta a discriminazioni. Nel 2015 nasce il movimento *#SayHerName* che, dopo 25 anni, riconosce il problema della non rappresentatività delle donne nere, vittime di discriminazione negli USA, diffondendo consapevolezza sul problema.

1.5. Finalità della ricerca

Considerando le dinamiche del funzionamento di Twitter descritte nel §1.3, la correlazione fra digitale-reale e gli studi effettuati da Crenshaw [1991] descritti nel paragrafo precedente, si è deciso di analizzare come il fenomeno dell'*intersezionalità* si possa presentare nei social network. Lo scopo di questa tesi è fare luce sul fenomeno per analizzare e comprendere come la misoginia si intrecci su Twitter con altre dimensioni d'odio come xenofobia, razzismo e islamofobia. L'analisi comprende la ricerca di stereotipi, tipologie di discriminazione di genere, il confronto fra le disuguaglianze e il modo con cui gli utenti interagiscono, supportano e attaccano i *target*. Il *corpus IntersHate* è un prodotto di questa indagine ed è stato realizzato in funzione dei *target* di campagne d'odio con elementi di intersezionalità e delle reazioni intorno agli avvenimenti che li coinvolgono. In conclusione, si valuterà con quale rilevanza alcune discriminazioni prevarranno sulle altre e la tipologia di linguaggio usufruendo del lessico computazionale HurtLex [Bassignana et al., 2018].

2.Strumenti, dataset e scelte metodologiche

Questo capitolo include i *tools* collocati nel campo informatico come APIs, algoritmi, DBs, lessici, applicativi, collezioni di dati e infine vengono esposte le metodologie di ricerca applicate al fine di effettuare l'analisi.

2.1.Strumenti utilizzati

In questo paragrafo vengono descritti i *tools* utilizzati durante la ricerca per ottenere, gestire e comprendere i dati.

Standard Search API

Per collezionare i *tweet* è stata utilizzata la Standard Search API di Twitter. È risultato necessario convertire l'account personale in *developer*, la procedura ha richiesto un tempo particolarmente lungo a causa del protocollo di approvazione. I controlli sono volti a certificare il rispetto dei termini e condizioni della piattaforma per salvaguardare la *privacy* degli utenti.

La Standard Search API, essendo la versione gratuita, ha dei limiti che sono stati rilevanti durante la procedura di analisi della situazione italiana in merito a discorsi e linguaggi intorno a vicende note:

- è possibile raggiungere solo i *tweet* pubblicati negli ultimi sette giorni;
- non è possibile superare le 160 richieste ogni 15 minuti, si può impostare il numero di *tweet* raggiungibili per richiesta ma il massimo è di 100 elementi.

Il linguaggio di programmazione scelto è *python* ed è stata utilizzata la libreria *tweepy*[3]. L'API ha concesso il raggiungimento dei *tweet* filtrando per parole chiave che

[3] *https://github.com/tweepy/tweepy*

sono state stabilite al seguito di un'attenta ricerca su Twitter dei *target* considerati bersagli di discriminazioni di diverso genere. Inoltre, è stato possibile effettuare anche l'estrazione delle *timeline* per rilevare i contenuti potenzialmente sensibili che avrebbero potuto essere oggetto di analisi per il *dataset* da annotare. Di seguito alcuni codici utilizzati per l'estrazione dei contenuti:

```python
import tweepy

def __get_tweets(self, search_query: str):
    print("Filtering by {}".format(search_query))
    tweets_per_qry = 100  # API max permission
    tweets = []

    tweet_count = 0
    last_id = -1
    while tweet_count < self.num_tweets:
        try:
            new_tweets = self.api.search(q=search_query, lang='it', count=tweets_per_qry,
                                         max_id=str(last_id - 1), tweet_mode="extended")
            if not new_tweets:
                break
            tweet_count += len(new_tweets)
            print('{}: {}/{} tweets'.format(search_query, tweet_count, self.num_tweets))
            tweets += new_tweets
            last_id = new_tweets[-1].id
        except tweepy.TweepError as e:
            print("Error : " + str(e))
            break
    return tweets
```

Figura 7: codice python per l'estrazione dei tweet per parole chiave

```python
import tweepy

def get_user_timeline(self, username: str):
    return tweepy.Cursor(self.api.user_timeline, screen_name='@{}'.format(username),
                         tweet_mode="extended").items()
```

Figura 8: codice python per estrazione delle timeline

La Standard Search API restituisce i *tweet* in formato *JSON* che contengono informazioni fra cui lo *user*, il contenuto, il tipo di contenuto, se è un'interazione ad altre condivisioni, la raggiungibilità, la localizzazione se disponibile e altre[4].

[4] *https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/intro-to-tweet-json*

Figura 9: tweet esempio[5]


```
{
    "created_at": "Thu Apr 06 15:24:15 +0000 2017",
    "id_str": "850006245121695744",
    "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API
    "user": {
      "id": 2244994945,
      "name": "Twitter Dev",
      "screen_name": "TwitterDev",
      "location": "Internet",
      "url": "https:\/\/dev.twitter.com\/",
      "description": "Your official source for Twitter Platform news, updates & events.
    },
    "place": {
    },
    "entities": {
      "hashtags": [
      ],
      "urls": [
        {
          "url": "https:\/\/t.co\/XweGngmxlP",
          "unwound": {
            "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
            "title": "Building the Future of the Twitter API Platform"
          }
        }
      ],
      "user_mentions": [
      ]
    }
}
```


Figura 10: JSON del tweet esempio in figura 9[6]

Twita

Twita [Basile et al., 2013] è una collezione di *tweet*, generata attraverso procedure automatiche, prima dell'Università di Groningen e ora di Torino. Iniziata nel

[5] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/intro-to-tweet-json
[6] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/intro-to-tweet-json

2012, contiene numerosi contenuti scritti in lingua italiana. Nel settembre 2018 la collezione ha superato i 500 milioni di *tweet* suddivisi in sette anni [Basile et al., 2018]. Da giugno 2018 è stata arricchita la collezione con altre raccolte indipendenti e viene usata la libreria *tweepy* di *python* per estrarre i contenuti e filtrarli per vocali e lingua italiana in modo da ridurre i *tweet* in altre lingue erroneamente estratti [Basile et al., 2018].

Considerati i limiti della Standard Search API di Twitter sopra citati, è stato necessario appoggiarsi a Twita per ottenere i contenuti che diversamente non avremmo potuto raggiungere. In questo modo sono stati ricostruiti i *thread* delle discussioni etichettate come interessanti ai fini della ricerca ed è stato possibile capire quanti contenuti sono stati rimossi per volontà degli utenti o della piattaforma a causa della violazione dei termini.

Hurtlex

Hurtlex [Bassignana et al., 2018] è un lessico *machine-readable*, disponibile su *GitHub*[7], contenente parole offensive, aggressive e di odio in più di 50 lingue. È basato sul lessico "Le parole per ferire" del linguista Tullio De Mauro, espanso con *MultiWordNet* e *Babel-Net*, i lemmi sono divisi in 17 categorie linguistiche (vedere *figura 11*) e una macrocategoria che indica la presenza di stereotipi.

Utilizzato come risorsa dal team di Unito durante la partecipazione ad AMI 2018, sono state rilevate le seguenti categorie come le più identificative per il riconoscimento automatico della misoginia: *Prostitution, Female and Male Sexual Apparatus* e *Physical and Mental Diversity and Disability* [Bassignana et al., 2018].

[7] *https://github.com/valeriobasile/hurtlex*

Label	Description
PS	negative stereotypes ethnic slurs
RCI	locations and demonyms
PA	professions and occupations
DDF	physical disabilities and diversity
DDP	cognitive disabilities and diversity
DMC	moral and behavioral defects
IS	words related to social and economic disadvantage
OR	plants
AN	animals
ASM	male genitalia
ASF	female genitalia
PR:	words related to prostitution
OM:	words related to homosexuality
QAS	with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
SVP	words related to the seven deadly sins of the Christian tradition

Figura 11: categorie di Hurtlex[8]

Strumenti per l'annotazione

Il *pilot set*, di cento *tweet* estratti da *IntersHate*, è stato annotato dai codificatori compilando una matrice *n x m* su un foglio di calcolo Microsoft Excel dove *n* rappresenta i *tweet* e *m* i livelli dell'annotazione. È stata utilizzata la funzione "convalida dati" per la scelta dei valori di ogni *label* nello schema di annotazione in modo da ridurre la confusione e gli errori. Considerando le dimensioni maggiori del *corpus* della seconda fase che ammonta a 1680 elementi, è stato realizzato un sito web per semplificare la compilazione dello schema.

2.2. Datasets

Inizialmente è stata effettuata un'analisi dei *target*, potenzialmente bersagli di discriminazioni e odio, attraverso programmi televisivi Rai in modo da rilevare argomenti, parole chiave e *trends*. Nella fase iniziale sono stati utilizzati i *datasets*, forniti

[8] *https://github.com/valeriobasile/hurtlex*

da Rai, contenenti le reazioni su Twitter e Instagram in merito a episodi etichettati come rilevanti. Sfortunatamente le collezioni, essendo concentrate maggiormente sui programmi, non contenevano un campione rappresentativo delle reazioni riguardanti le discussioni sui *target*.

Al seguito sono state generate altre raccolte, e quindi *IntersHate*, filtrando i contenuti di Twita in funzione delle analisi effettuate attraverso l'API ufficiale di Twitter. I campi dei contenuti collezionati sono: *id tweet, data pubblicazione, testo, tipo* (*tweet, retweet, quote* o *reply*), *id utente*, *screen name utente*, *descrizione utente, numero di stati*, *numero di followers* e *urls*.

2.3. Scelte metodologiche

Le metodologie applicate variano dall'analisi qualitativa collocata durante la fase decisionale per la costruzione del *corpus IntersHate*, fino a quella quantitativa per la classificazione dei contenuti mediante un'annotazione manuale effettuata da dodici codificatori equamente distribuiti per genere allo scopo di ridurre i *bias* che ne derivano. Al seguito del monitoraggio, l'analisi qualitativa è servita per valutare i *target* potenzialmente soggetti a più forme di discriminazione coesistenti con la misoginia. Considerando le domande di ricerca, per ottenere un campione rappresentativo sono stati applicati filtri (descritti nel §3.2) quali la scelta dei *target*, dei giorni con maggior concentrazione di tweet, le emoji apparse correlabili alle discriminazioni nei contenuti e l'eliminazione dei *record* simili mediante un algoritmo di distanza lessicale (inserito nel §3.2) impostato sul limite 0.7. Lo schema multilivello per l'annotazione delle varie dimensioni e dell'intersezionalità è costituito da un'analisi a grada larga per la rilevazione di misoginia, xenofobia, islamofobia, stereotipi, difesa del *target* e *stance* nei confronti di un eventuale *parent tweet* mentre l'analisi a grana fine è stata applicata all'intersezionalità per verificare la prevalenza di una dimensione d'odio rispetto alle altre e alla misoginia per la classificazione di molestie sessuali, *derailing*, discredito e *dominance*. L'analisi quantitativa è stata organizzata in modo da avere tre annotazioni per *tweet* e quindi poter valutare numericamente le dinamiche delle varie etichette all'interno della collezione.

3.Sviluppo del corpus IntersHate

Il capitolo spiega le considerazioni e le decisioni prese per la realizzazione del *corpus IntersHate* in funzione dei *target*, delle loro caratteristiche e di quelle delle estrazioni dei *tweet*.

3.1.Considerazioni e indagini preliminari per una raccolta target-based

La ricerca è iniziata analizzando i possibili *target* soggetti a più forme di odio discriminatorio, le categorie prese in considerazione sono un'estensione di quelle di Vox [2019]. Nel periodo marzo-giugno 2020, Twitter e i programmi Rai sono stati utilizzati per le ricerche che hanno permesso di trovare caratteristiche, argomenti, discorsi ed avvenimenti su diciassette soggetti. Un documento di testo contiene per ogni bersaglio: *username* Twitter ufficiale (se disponibile), caratteristiche soggette a odio, *hashtags* rilevati nei contenuti Twitter che lo riguardano, periodo analizzato, programmi Rai con relativi *links* alle puntate a cui ha partecipato. È stato compilato anche un foglio di calcolo con una matrice *n x m*, dove *n* corrisponde ai bersagli e *m* alle tipologie d'odio, inoltre un'altra colonna serve per indicare i possibili soggetti con caratteristiche comuni e divergenti in modo da identificare possibili confronti su cui effettuare le successive analisi. È stata valorizzata inserendo una *x* nelle caselle che intersecano il bersaglio con le categorie che lo rappresentano. La *tabella 2* mostra un estratto della matrice *n x m* compilata secondo i riscontri ricevuti dalle ricerche effettuate su Twitter. La matrice è utile per riconoscere velocemente le intersezioni fra *target*. Essendo la misoginia al centro dei confronti, è stata usata come punto fermo su cui procedere con la ricerca. Monitorando i *trends* su Twitter, è risultato noto come alcuni bersagli venivano più discussi rispetto ad altri e quindi è stata fatta una selezione per avere una quantità di materiale maggiore su cui filtrare il campione rappresentativo per l'annotazione. Il controllo delle tendenze è servito anche per trovare le gogne pubbliche, fenomeno molto importante per lo studio. Vox [2019] mostra come si sviluppano i picchi dei contenuti negativi in corrispondenza degli eventi (vedere *figura 1*, *figura 3* e *figura 4*), le tendenze sui social sono correlate sia

ad eventi nel digitale che nel reale e in questo modo è stato possibile raccogliere le reazioni pubbliche nei confronti di ciò che succedeva al passare delle settimane.

	Misog.	**Xenof.**	**Antisem.**	**Islamof.**	**Omof.**	**Opinioni politiche**	**Body-shaming**
Cathy La Torre[9]	X				X	X	X
Carola Rackete[10]	X	X					
Elodie[11]	X	X				X	
Laura Boldrini[12]	X					X	X
Liliana Segre[13]	X		X				
Silvia Romano[14]	X			X			

Tabella 2: estratto della matrice n x m sui target

Target

L'analisi preliminare si è basata sulla valutazione delle discussioni su Twitter in merito alle vicende intorno a diciassette *target* raggruppandoli per la categoria predominante:

- antisemitismo: Liliana Segre e Aldo Rolfi;
- misoginia e xenofobia (comprendendo i pro-immigrazione): Laura Boldrini, Cécile Kyenge, Carola Rackete, Silvia Romano, Giovanna Botteri, Michelle Hunziker, Elodie, Michela Murgia, Rula Jebreal e Diletta Leotta;

[9] *https://it.wikipedia.org/wiki/Cathy_La_Torre*
[10] *https://it.wikipedia.org/wiki/Carola_Rackete*
[11] *https://it.wikipedia.org/wiki/Elodie_(cantante)*
[12] *https://it.wikipedia.org/wiki/Laura_Boldrini*
[13] *https://it.wikipedia.org/wiki/Liliana_Segre*
[14] *https://www.agi.it/cronaca/news/2020-05-09/chi-e-silvia-romano-8565666/*

- omolesbobitransfobia: Cathy La Torre, Vladimir Luxuria, Alessandro Zan, Tiziano Ferro e Achille Lauro.

3.2. Collezione dei dati

Dal 25 luglio al 31 ottobre 2020, sono stati estratti i *tweet* attraverso l'API ufficiale di Twitter descritta nel capitolo 2. Come effettuato da Vox [2019], il filtraggio è avvenuto per parole chiave strettamente legate alle discussioni sui *target* in modo da ridurre il più possibile il rumore bianco. L'estrazione è servita per capire l'interfacciamento degli utenti verso i bersagli tenendo conto dell'aggressività, dell'odio e del tipo di linguaggio, questa fase è stata utile in modo da filtrare i soggetti per l'annotazione. Grazie all'estrazione e alla ricerca manuale sulla piattaforma Twitter, è stato possibile collezionare gli identificatori dei contenuti più rilevanti riguardanti i malcapitati che al seguito sono serviti per estrarre i *thread* conversazionali dal *database* di Twita. Considerando i *tweet* rilevati dall'estrazione, la situazione appare simile a quella di Le mappe dell'intolleranza [Vox, 2019] e Il barometro dell'odio – sessismo da tastiera [Amnesty International, 2020] mostrando a prima vista una certa presenza di misoginia, xenofobia, razzismo e islamofobia. Traendo ispirazione dai picchi d'odio in Le mappe dell'intolleranza [Vox, 2019], sono stati scelti due eventi per la realizzazione della raccolta *IntersHate* ai fini dell'annotazione: la liberazione e il ritorno in Italia di Silvia Romano del 9-10 maggio 2020 e il dibattito fra Elodie e Lega Salvini iniziato l'11 agosto 2020.

Nel caso di Silvia Romano, la collezione è composta dai contenuti estratti dal 9 al 24 maggio 2020 contenenti sia il nome che il cognome del *target*, è stato deciso in modo da limitare il più possibile il rumore bianco. Inoltre, è stata arricchita con le reazioni a 32 *tweet* di giornali e politici identificati come rappresentanti del dibattito. La raccolta, che si aggira intorno ai 250.000 elementi, non è completa in quanto non considera tutte le condivisioni che contengono solo il cognome, il nome, il nuovo nome (Aisha) oppure quando non la citano in maniera diretta. Pertanto, l'argomento risulta essere il più discusso fra i *target* in un periodo molto concentrato (la maggior parte dei contenuti sono stati pubblicati nei primi sei giorni). Alcune testate giornalistiche hanno pubblicato molti *tweet* accentuando la bufera intorno all'accaduto, sono stati compresi nel *dataset* per valutare l'incidenza sul fenomeno.

La raccolta sul dibattito riguardante Elodie è composta dalle reazioni a quattro *tweet* di giornali e politici. Sono state escluse le iterazioni al contenuto pubblicato della cantante il 14 agosto 2020, poiché prevalentemente di supporto nei confronti del *target*, in modo da ridurre il *dataset* da assegnare agli annotatori. Sono stati collezionati 731 elementi.

Il *corpus* per l'annotazione contiene le seguenti etichette: *id* (numero progressivo scollegato da quello all'interno di Twitter), *parent text* (se esiste) e *tweet text.*

Emoji

Come sostiene Vox [2019], appare rilevante la correlazione fra odio sui social e politica nel caso di migranti, ebrei e musulmani. Visionando le estrazioni è apparsa frequente l'associazione fra xenofobia, islamofobia e la presenza del tricolore nel *name*, *screen_name*, *description* e *tweet* degli utenti. La bandiera italiana è un'*emoji* spesso usata per esporre il proprio sostegno ai partiti di destra, di solito nazionalisti e anti-immigrazione. Altri simboli, seppur con rilevanza inferiore, sono presenti all'interno della raccolta (vedere *figura 12* e *figura 13*). La bandiera europea, l'arcobaleno e la bandiera arcobaleno appaiono in percentuali particolarmente basse negli utenti che usano il tricolore, questo lascia ipotizzare una polarizzazione degli *users* conforme prevalentemente alle opinioni di destra. Inoltre, le proporzioni in *figura 12* e *figura 13* risultano simili indicando una correlazione fra i due *datasets* nonostante le dimensioni che sono rispettivamente di 248.240 e 731 elementi. Sono state notate nove *emoji* richiamanti simbologie di vario genere associabili a correnti di pensiero con polarizzazione sia negativa che positiva (vedere *figura 12* e *figura 13*). La bandiera italiana, che rappresenta il patriottismo nei confronti del Paese, viene utilizzata generalmente dagli utenti anti-immigrazione e antieuropeisti, appare con una proporzione sempre molto alta ed è soggetto di analisi nell'annotazione. La bandiera UE è generalmente utilizzata dagli europeisti mentre l'arcobaleno e la bandiera arcobaleno per indicare il sostenimento alla comunità LGBTQ+. In *figura 12* e *figura 13* è visibile la scarsa coesistenza della bandiera italiana con quella UE e l'arcobaleno rafforzando il significato del messaggio. La bandiera nera, il cuore nero, il falco e la mano alzata possono essere associati a posizioni estremiste ma appaiono in percentuali basse, pertanto

il cuore nero potrebbe essere confuso col tifo calcistico nel tentativo di comporre i colori delle squadre. Sono anche presenti la croce che indica la fede religiosa e le stelle trovate online nei movimenti per fare rete associati al *made in Italy*.

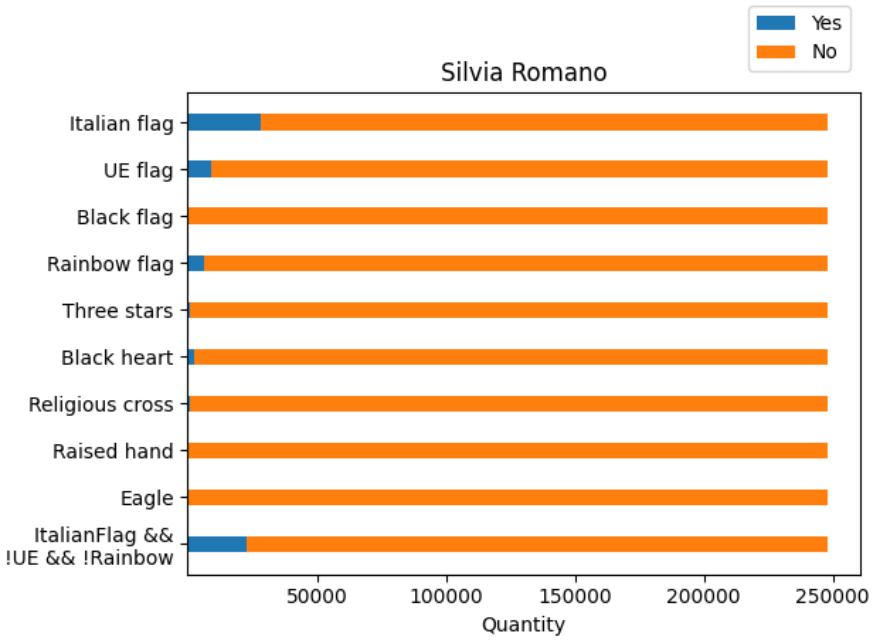

Figura 12: proporzioni dei tweet con le emoji rispetto al totale raccolto nel periodo 9-24 maggio 2020, i contenuti sono stati filtrati per le chiavi "Silvia" e "Romano" coesistenti

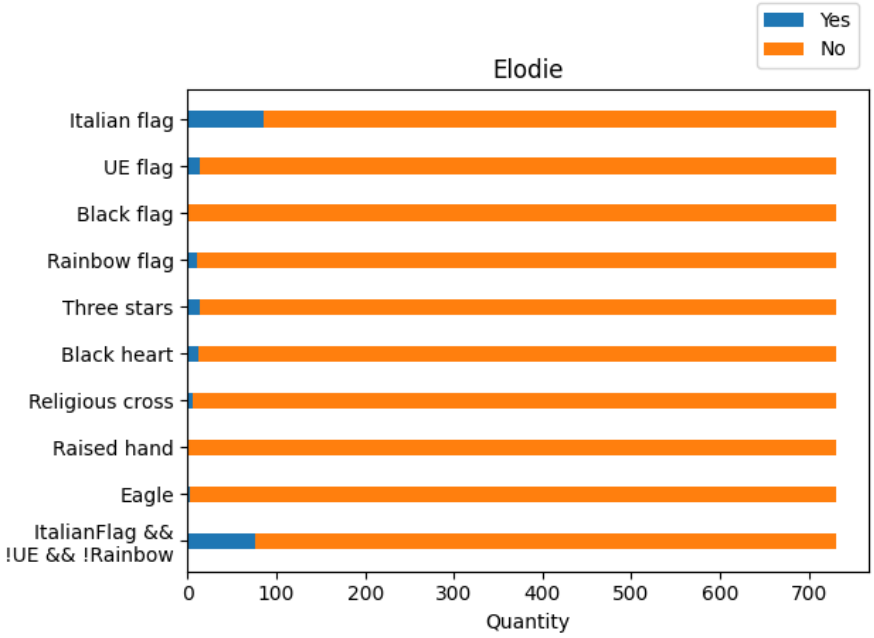

Figura 13: proporzioni dei tweet con le emoji rispetto al totale raccolto nel periodo 15-17 agosto 2020, i contenuti sono stati prelevati da quattro tweet provocanti che riguardavano la discussione fra Elodie e Lega Salvini iniziata l'11 agosto 2020

Decisioni e filtri

Fra i *target* sono state selezionate Elodie e Silvia Romano per l'annotazione, entrambe coinvolte in gogne pubbliche esposte precedentemente nel §3.2. Sono stati raccolti molti contenuti riguardanti i due avvenimenti, citati nel §3.2, per visionare il comportamento negativo degli utenti nei loro confronti.

La raccolta su Silvia Romano è stata ridotta a causa della sua enorme dimensione non sostenibile per l'annotazione manuale. Sono stati rimossi i *retweet* ed essendo non trascurabile la presenza della bandiera italiana all'interno di *name*, *screen_name*, *description* e *tweet* (vedere *figura 12*), la collezione è stata filtrata in modo da contenere elementi con e senza il tricolore con una proporzione del 50% mantenendo la stessa distribuzione per i giorni e le ore (vedere *figura 15*). È stata presa questa decisione per cercare di annotare un campione contenente elementi di entrambi gli insiemi per poterli quindi confrontare. La distribuzione per giorni e ore è necessaria poiché gli elementi con la bandiera si concentrano in determinati *slot* temporali (solitamente pomeriggio e sera). Inoltre, sono stati selezionati solo i *tweet* pubblicati nei giorni più concentrati, ovvero 10-12 maggio 2020.

Considerato che la cardinalità della raccolta continuava a essere alta, è stato deciso di rimuovere i *tweet* simili utilizzando la Cosine Similarity (vedere *figura 14*), un algoritmo per il calcolo della distanza lessicale che restituisce la somiglianza in percentuale secondo un valore da 0 a 1. Il valore limite per l'accettazione dei contenuti simili è stato stabilito a 0,7. In questo modo si tenta di effettuare ipotesi sulle polarizzazioni dei contenuti non annotati manualmente ma poco differenti da quelli analizzati.

```
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
def get_cosine_sim(*strs):
    vectors = [t for t in get_vectors(*strs)]
    return cosine_similarity(vectors)

def get_vectors(*strs):
    text = [t for t in strs]
    vectorizer = CountVectorizer(text)
    vectorizer.fit(text)
    return vectorizer.transform(text).toarray()
```

Figura 14: Cosine Similarity algorithm [Towards Data Science, 2018]

La collezione su Silvia Romano è arrivata a 3006 *tweet*. Le 180 reazioni ai 32 *tweet* selezionati sono state la risultante del filtro per bandiera italiana all'interno di *name*, *screen_name*, *description* o *tweet*, mentre i 2826 contenuti rimanenti sono stati filtrati estraendo in maniera casuale 1500 elementi usando la funzione python random.sample(population, k, *, counts=None)[15]. Al termine di questo processo di selezione sono stati ottenuti 1680 *tweet*.

La raccolta in merito al dibattito su Elodie è stata mantenuta intatta confermando i 731 elementi estratti dalle reazioni ai contenuti più rilevanti sul tema ai fini della tesi.

[15] *https://docs.python.org/3/library/random.html*

4.Realizzazione del nuovo schema di annotazione multilivello

Questo capitolo descrive il processo di creazione del nuovo schema di annotazione multilivello comprendendo la valutazione di altri, utilizzati in altre ricerche, in funzione delle forme discriminatorie che questa tesi punta ad analizzare. Inoltre, vengono descritte nel dettaglio le etichette a grana fine e larga affiancando alcuni esempi tratti dal *corpus IntersHate*.

4.1.Annotazione dei dati sul piano dell'intersezionalità

L'annotazione è stata realizzata sulla base di altri schemi già presenti in letteratura in modo da poterli analizzare e quindi trarne ispirazione per questo studio.

Lo schema utilizzato per AMI (Automatic Misogyny Identification) [Fersini et al., 2018] si concentra sulla rilevazione della misoginia verso singoli e gruppi annotando le seguenti categorie: *Stereotype & Objectification*, *Dominance*, *Derailing*, *Sexual Harassment & Threats of Violence* e *Discredit*. Sono state considerate per annotare la misoginia all'interno del *corpus IntersHate*: se i *tweet* contengono discorsi misogini, viene verificata la presenza di molestie sessuali e derailing, forme di discredito e dominance mentre gli stereotipi vengono rilevati per qualunque categoria presa in analisi in quanto si è interessati a valutarne l'intersezione. In merito a individualità o genericità del *target*, trattata anche da Basile et al. [2019], non è stato necessario applicare questa distinzione in quanto i dibattiti, intorno ai due bersagli, risultano essere particolarmente specifici. Inoltre, Basile et al. [2019] hanno annotato l'*hate speech* nei confronti di donne e immigrati, la tesi punta a volerne valutare anche le intersezionalità ed è risultato utile per capire come avviene la distribuzione di odio e aggressività all'interno delle due categorie.

Sanguinetti et al. [2018], annotando un *corpus* italiano sull'immigrazione, hanno misurato l'intensità dell'*hate speech* con una scala da 0 a 4. La misurazione del livello d'odio ha posto le basi per il ragionamento sulla comparazione delle categorie intersecate

nei *target*. Presa la misoginia come punto fermo per il confronto, viene ordinata per intensità con le altre categorie in modo da valutarne l'incidenza e la predominanza all'interno di *IntersHate*.

L'analisi delle *stance* di Pamungkas et al. [2019] è stata considerata per verificare come le *reply* interagiscono ai *tweet*. Sono state incluse le quattro etichette: *AGREE-ACCEPT (support)*, *REJECT (deny)*, *INFO-REQUEST (question)* e *OPINION (comment)*. Il modello di Pamungkas et al. [2019], essendo basato sulle reazioni ai *rumours*, è stato arricchito con il valore "assente" per poter classificare tutti quei contenuti che non possiedono un *parent* (ovvero non sono un'interazione ad un altro contenuto sulla piattaforma). Al seguito dell'annotazione del *pilot set*, le etichette di Pamungkas et al. [2019] mantenute insieme ad "assente" sono tre: pro, contro e neutrale. Questa scelta è stata effettuata poiché nel caso in analisi non è stato ritenuto utile differenziare le domande dai commenti e inoltre è stata notata una particolare confusione fra gli annotatori nel distinguere le *labels*.

L'intersezionalità rilevata in *Il barometro dell'odio* [Amnesty International, 2020] è stata considerata per la realizzazione dello schema sull'incrocio tra misoginia, xenofobia/razzismo e islamofobia. Il barometro dell'odio ha valutato le reazioni intorno all'argomento "donne e diritti di genere" concludendo che le donne sono state la categoria più colpita e al seguito apparivano musulmani, migranti e rifugiati. Il tema "donne e diritti di genere" risulta essere sensibile e pone al centro del discorso le donne stesse rendendole maggiormente esposte. Le quattro categorie rilevate sono associabili a quelle scelte per l'annotazione dei due *target* che sono esposti ma non alla difesa in prima linea dell'emancipazione femminile, in questo modo si cerca di valutare la misoginia riducendo il *focus* sulle disuguaglianze di genere.

Considerati gli studi di Crenshaw [1991] sull'intersezionalità tra misoginia e razzismo, le ricerche di Vox [2019] e Amnesty International [2020] che hanno visto misoginia, xenofobia e islamofobia come le categorie più colpire, è stata assunta la decisione di costruire lo schema di annotazione su misoginia, xenofobia/razzismo e islamofobia. Le tre categorie coincidono con le tipologie di linguaggio riconosciute da una prima visione del *corpus* collezionato sui bersagli.

A causa delle difficoltà nell'analisi automatica sul piano della semantica, l'annotazione è stata svolta in maniera manuale da dodici annotatori di genere e età variabile che comprendono i tre annotatori esperti e altri nove codificatori fra cui otto studenti universitari. Al fine di ridurre la polarizzazione all'interno del gruppo, si è cercato di renderlo il più possibile eterogeneo riducendo le influenze legate alle opinioni e al genere.

L'annotazione è stata divisa in due fasi. La prima fase di *training* è servita per valutare il *corpus* e lo schema di annotazione sfruttando un'estrazione di 100 elementi (50 su Elodie e 50 riguardanti Silvia Romano). Nella seconda fase è stata annotata l'intera raccolta, arricchendo le linee guida e modificando il *corpus* e lo schema multilivello di annotazione in funzione dei risultati della prima fase. Il *pilot set* è stato annotato utilizzando un foglio di calcolo Microsoft Excel e la funzione "convalida dati" mentre per la seconda parte è stato realizzato un sito web per semplificare l'operazione.

4.2. Schema multilivello

Questo paragrafo comprende la descrizione delle etichette, a grana fine e grossa, all'interno dello schema di annotazione affiancando i relativi esempi contenenti alcune *emoji* che potrebbero non essere completamente fedeli alle icone classiche, pertanto è possibile trovarne la legenda di seguito. I *tweet* sono riportati seguendo le *policy*[16] per la segnalazione di esempi offensivi avendo cura della privacy degli utenti riformulando i testi mantenendo il significato e censurando le volgarità in modo da ridurre l'impatto durante la lettura.

Legenda *emoji*:

- 🇮🇹 : bandiera italiana
- 🥒 : cetriolo
- 🤮 : rigurgito
- 👎 : pollice in giù

[16] *https://www.workshopononlineabuse.com/resources-and-policies/reporting-examples*

Misoginia

[SI/NO]. Comprende qualunque forma, esplicita e implicita, di misoginia intesa come odio nei confronti delle donne, espressa in modi differenti, inclusi avversione o repulsione. Per definire meglio i contorni del comportamento misogino che si esprime nei testi, facciamo riferimento alla definizione di Kate Manne [2017]: "*Most misogynistic behavior is about hostility towards women who violate patriarchal norms and expectations, who aren't serving male interests in the ways they're expected to. So there's this sense that women are doing something wrong: that they're morally objectionable or have a bad attitude or they're abrasive or shrill or too pushy*". In questo senso, nel contesto di questo *corpus*, è considerato misogino anche un testo che zittisce il *target* e strumentalizza la gravidanza allo scopo di incolpare la vittima.

SE MISOGINIA = SI, possono essere completate anche le seguenti due categorie che permettono una caratterizzazione più a grana fine del comportamento misogino. Questa caratterizzazione a grana fina è ispirata a quella proposta nell'ambito dello *shared task* AMI 2018@EVALITA18 [Fersini et al., 2018].

Esempio di misoginia in merito alla gravidanza: *"È anche gravida #SilviaRomano"*.

Altri esempi sono presenti nell'etichetta "Prevalenza" di seguito.

Molestie sessuali e *derailing*

[Se misoginia = SI, SI/NO]. Include *avance*, richieste di favori sessuali e qualunque forma di molestia che riguarda il sesso o discorsi in cui si giustifica l'abuso sulla donna sminuendo o eludendo la responsabilità maschile.

Esempio di *derailing* che elude la responsabilità maschile e nega la possibilità di un abuso accusando Silvia Romano di esser stata consenziente e di essersi inventata tutto (*victim blaming*) rendendola corresponsabile del suo stesso rapimento e ignorando la responsabilità dei rapinatori (uomini): *"Questa ragazza ci ha preso in giro. Era compiacente, si è sposata con il suo rapitore, è incinta, ha detto di non esser stata violentata quindi era consenziente. È un'islamica* 👎 *datemi i miei soldi* IT*"*.

Esempio di misoginia sulla sfera sessuale e xenofobia: *"La z****la, appassionata ai c***i talebani, ha orchestrato una messinscena con il tipo che se la s**pa e si è sistemata a vita con il riscatto"*.

Esempio di misoginia sulla sfera sessuale assumendo con malizia il voler tornare in Kenya per un soddisfacimento carnale consigliato dalle *emoji* del cetriolo: *"Mi gioco le p**le che tornerà dai suoi 'rapitori'* 🥒 *"*.

Esempio di misoginia sulla sfera sessuale e islamofobia: *"Ha ricevuto il servizio completo, c***o e Corano"*.

Discredito e *dominance*

[Se misoginia = SI, SI/NO]. Si parla di discredito quando un individuo *S*, tramite un atto comunicativo, danneggia l'immagine di un altro individuo *T* davanti a terzi (individuo o gruppo *A*), facendo riferimento ad azioni o caratteristiche di *T* considerate negative da *A*. Esistono più forme di discredito che possono essere espresse tramite l'insulto e sono legate alle caratteristiche fisiche, alle competenze (anche affettive), e nell'ambito del discorso misogino spesso anche alla *dominance*, che nel caso specifico si esprime tipicamente come affermazione la superiorità degli uomini sulle donne evidenziando la disuguaglianza di genere.

Esempio di *dominance* e discredito misogini nei confronti del *target* che non può essersi convertita di sua spontanea volontà coesistente con xenofobia e islamofobia: *"Conte dacci le prove del riscatto pagato dagli italiani per questa odiosa nullità e vergogna nazionale! È una bambina indottrinata, senza cervello e stupida. È andata in terre cesso per seguire le sue idiozie apparentemente umanitarie"*.

Xenofobia e razzismo

[SI/NO]. Include le forme, esplicite e implicite, di razzismo e xenofobia, ossia espressioni di razzismo, "*fondate sull'arbitrario presupposto dell'esistenza di razze umane biologicamente e storicamente «superiori», destinate al comando, e di altre*

«inferiori»"[17], dirette contro qualcuno di una etnia o razza diversa basate sulla convinzione che la propria razza sia superiore; espressioni di avversione per gli stranieri e per ciò che è straniero, che si manifesta in atteggiamenti e azioni di insofferenza e ostilità verso le usanze, la cultura e gli abitanti stessi di altri paesi. Si noti che, in considerazione delle dinamiche di *ingroup* e *outgroup* che sottendono questi fenomeni, consideriamo espressione di xenofobia anche testi dove il soggetto target e altri italiani vengono insultati o rifiutati come membri dell'*ingroup* o stigmatizzati come anti-italiani, a causa della propria vicinanza (o appoggio) a popolazioni straniere o immigrati.

Esempio di razzismo, islamofobia e discredito del genere femminile: *"Una donna bianca convertita all'Islam, esce indenne dai n*****i, belve inferocite, andate tutti a fare in c**o"*.

Esempio di xenofobia secondo il concetto di *outgroup*: *"Non prenderci per c**o, torna da dove sei venuta"*.

Altri esempi sono presenti nell'etichetta "Prevalenza" di seguito.

Islamofobia

[SI/NO]. Comprende qualunque forma, esplicita e implicita, di "*forte avversione, dettata da ragioni pregiudiziali, verso la cultura e la religione islamica*"[18]. Le principali manifestazioni includono la criminalizzazione dei *target* definendoli minacciosi, violenti e terroristi, viene spesso affiancata dalla xenofobia e include anche la deumanizzazione dei bersagli.

Esempio di islamofobia e misoginia, emerge anche la scarsa preparazione in geografia riconoscendo l'Islam come uno Stato: *"È venuta qui per fare attentati, è una terrorista, è anche incinta di un musulmano. Se stava bene in Islam rimpatriatela #convertita"*.

Altri esempi sono presenti nell'etichetta "Prevalenza" di seguito.

[17] *https://www.treccani.it/vocabolario/razzismo/*
[18] *https://www.treccani.it/vocabolario/islamofobia*

Prevalenza

[Se almeno due delle tre principali dimensioni = SI, Misoginia / Xenofobia e razzismo / Islamofobia]. Nel caso di intersezione fra le tre discriminazioni sociali da analizzare, indica quale prevale sulle altre all'interno del tweet. Con il termine intersezionalità si intende la coesistenza di molteplici forme di disuguaglianze quali etnia, genere, orientamento sessuale, classe, disabilità, età ecc. che possono generare ostacoli nel quotidiano di coloro che vengono marginalizzati perché diversi [Crenshaw, 1991].

Esempio di misoginia, xenofobia e islamofobia: *"Il governo ruba 4 milioni di euro agli italioti per pagare uno specie di riscatto al marito islamico che la mette incinta e la converte. Arriva in Italia contenta, ingrassata e viene accolta come una santa. Popolo idiota!"*.

Stereotipo

[SI/NO]. Include gli stereotipi negativi sessisti, xenofobi, razzisti e islamofobi che riguardano categorie vulnerabili a cui si rivolgono la discriminazione e i discorsi di incitamento all'odio presi in considerazione in questo studio sull'odio intersezionale (donne, immigrati, stranieri, persone di fede islamica). Lo stereotipo è una generalizzazione condotta su un gruppo di persone, in cui caratteristiche identiche vengono attribuite a tutti i membri del gruppo [Aronson et al., 2013]. Lo stereotipo si basa su una serie di credenze, non basate sull'esperienza, che le persone mettono in atto per interpretare l'ambiente che le circonda e muoversi in esso. Lo stereotipo non si basa su conoscenze di tipo scientifico, ma si basa sulla classificazione in categorie molto rigide e generali, senza tenere conto di possibili eccezioni. Molti stereotipi derivano dalla cultura in cui si vive e danno luogo a pregiudizi. Lo stereotipo è infatti il nucleo cognitivo del pregiudizio [Brown, 2010], che assume spesso il volto di comportamenti discriminatori, razzisti e di odio nelle interazioni sociali.

Esempio di stereotipi xenofobi e di genere che includono il discredito: *"Vanno a fare le splendide in Africa ma quando si accorgono che ci sono i n*****i cattivi chiedono aiuto a mamma Italia"*.

Esempio di stereotipi xenofobi e islamofobi: *"Non consideriamo la sindrome di Stoccolma? Se tornasse plagiata dai suoi aguzzini per fare danni? Io non festeggerei"*.

Difesa del target

[SI/NO]. Indica se l'utente che ha pubblicato il *tweet* assume le difese del *target* all'interno della narrazione avendo come effetto una contro narrazione nei confronti dei flussi negativi verso il bersaglio. Comprende sia il sostegno privo di discriminazioni che quello che lo integra in modo da classificare anche la difesa che non punta a contrastare le avversioni, ma solo a dirottarle su altri individui.

Esempio di difesa del *target* che dà origine a una contro narrazione positiva: *"Il privato di Silvia Romano non dovrebbe essere nel dibattito pubblico. È stata liberata da una prigione fisica ma intrappolata in una di violenze psicologiche e pregiudizi inutili e ingiusti"*.

Esempio di difesa del *target* che dà origine a una contro narrazione negativa radicata sulla *pre-mediazione* [Grusin, 2004], la diffusione di opinioni prima delle notizie: *"Siamo contenti che Silvia Romano sia salva ma è la presa per il c**o dell'islamizzazione volontaria che non buttiamo giù. Capisco che possa aver aderito al credo dei carcerieri ma suona male, secondo me nasconde qualcosa, è stata plagiata ed è ancora spaventata"*.

Esempio di difesa del *target* con *hate speech*: *"A tutti quelli che hanno insultato Silvia Romano auguro il cancro!"*.

Presa di posizione (o *stance*)

[Se non esiste un *parent tweet* ASSENTE, altrimenti PRO / CONTRO / NEUTRALE]. Identifica la posizione (o *stance*) dell'utente che ha pubblicato il tweet nei confronti del *parent tweet* a cui ha reagito. L'etichetta assume il valore "*pro*" se è d'accordo con il parent tweet, "*contro*" se lo contesta e "*neutrale*" se dal testo non è possibile asserire alcuna *stance*. Se invece il tipo del contenuto pubblicato dall'utente è diverso da *quote* o *reply*, non esiste alcun *parent tweet* pertanto il *tweet* non è una reazione e il valore di default è "*assente*".

Esempio di presa di posizione "*pro*":

- Parent tweet: *"Quindi di quale liberazione parliamo?"*.

- Child tweet: *"Silvia non prenderci per c**o, torna da dove sei venuta"*.

Esempio di presa di posizione "*contro*" e "*neutrale"*:

- Parent tweet: *"La liberazione di Silvia Romano è una bella notizia. L'aspettiamo in Italia, ringraziamo i nostri servizi di Intelligence e coloro che hanno contribuito a questo importante obiettivo"*.
- Child tweet "*contro*": *"Assolutamente no! Vanno a fare le splendide in Africa ma quando si accorgono che ci sono i n*****i cattivi chiedono aiuto a mamma Italia"*.
- Child tweet "*neutrale*": *"È tornata e sono contento per la sua famiglia. Ora non trasformiamola in un'arma di distrazione di massa e non tiriamola fra un programma televisivo e l'altro"*.

5.Analisi corpus-based del discorso d'odio intorno ai target

In questo capitolo sono esposte le considerazioni in merito all'annotazione sul piano dei *target* descrivendo separatamente la prima e la seconda fase, vengono introdotti i *background* dei bersagli annotati per comprenderne meglio il contesto e quindi come gli utenti discutono gli avvenimenti descrivendo le dinamiche. Inoltre, il capitolo è arricchito con l'analisi dell'*agreement*/*disagreement* fra gli annotatori in merito alle dimensioni singole.

5.1.Considerazioni

L'annotazione è stata effettuata in due fasi: nella prima, che ha avuto la funzione di *training*, è stato classificato un *pilot set*, su Elodie e Silvia Romano, che ha permesso di formarci una prima immagine delle raccolte sui *target*. La seconda fase ha visto l'annotazione dei due *dataset* ottenuti dal filtraggio per parole chiave su Silvia Romano e dalle reazioni ai *tweet* scelti pubblicati da giornali, partiti e politici su Twitter.

Al seguito della fase di *training* è emerso che gli annotatori avevano più consapevolezza sulle vicende intorno a Silvia Romano piuttosto che su quelle di Elodie, pertanto la fase successiva ha incluso esclusivamente la classificazione di contenuti riguardanti Silvia Romano. Il *training* ha permesso anche di effettuare una prima valutazione dello schema di annotazione multilivello che ha visto alcuni cambiamenti tra cui la revisione dell'etichetta "presa di posizione (o *stance*)" (descritta nel §4.1) e delle linee guida (descritte nel §4.2) rendendole maggiormente accurate.

Essendo entrambi i bersagli connessi alla scena politica, Elodie per la discussione con Matteo Salvini e Silvia Romano per la liberazione e il ritorno in Italia, molti contenuti pubblicati contengono riferimenti, opinioni e attacchi a partiti e politici di ambe le parti mostrando come coesistono sia gli attacchi che le difese dei bersagli intorno alle affermazioni e interviste dei rappresentanti dei cittadini. Allo stesso modo, la raccolta contiene anche *tweet* rivolti ai politici col solo scopo di attaccare e contraddire

indipendentemente dal tema trattato all'interno del *thread* rendendo il contenuto *off-topic*. Risultano essere molto comuni anche le manifestazioni di dissenso sul piano economico sostenendo che non fosse necessario il presunto riscatto considerate le necessità del popolo italiano. In questo caso, sono apparsi *tweet* non classificabili con le dimensioni dello schema di annotazione poiché non era completamente chiaro se il motivo della disapprovazione fosse di natura discriminatoria. Entrambi i *target* sono stati giudicati molto sulla base dell'apparenza estetica vedendo l'aspetto fisico e l'abbigliamento come oggetto di dialogo, soprattutto su Silvia Romano al seguito dell'arrivo all'aeroporto di Ciampino con l'*hijab* e un orologio riconosciuto subito come un Rolex nonostante non fosse possibile distinguerlo dai contenuti diffusi dai media.

5.2. Silvia Romano

Silvia Romano è una volontaria italiana di Africa Milele Onlus, rapita il 20 novembre 2018 in Kenya. È stata liberata il 9 maggio 2020 ed è arrivata il giorno seguente in Italia animando il dibattito sulle testate giornalistiche, i media e i social network in merito a diversi temi: l'eventuale riscatto, l'attività di volontariato, la gravidanza, il presunto matrimonio, l'*hijab* e la conversione all'Islam. Quando il 9 maggio 2020 la notizia della liberazione di Silvia Romano si è diffusa sui social, la giovane è diventata bersaglio di una campagna d'odio molto violenta, con attacchi verbali riconducibili a comportamenti misogini, dove si intersecano anche l'odio razzista e islamofobo. Valeria Scandurra [2020] ha scritto un saggio per l'Università Jagellonica di Cracovia in cui descrive gli attacchi subiti da Silvia Romano, influenzati dalle dinamiche dei *social media*, identificabili come violenze psicologiche spesso legittimate e ignorate rispetto a quelle fisiche. Scandurra [2020] spiega come *"a violent reaction has occurred from the Italian media that described Silvia Romano as an ungrateful, selfish and a traitor of her country for converting to the Islamic religion"* che ha contribuito all'atmosfera di intolleranza e islamofobia nel Paese. Espone anche gli attacchi ricevuti da giornali e personaggi noti, tra cui politici, che oltre a discriminarla rispetto alle sue varie dimensioni, non le hanno risparmiato minacce o auguri di morte [ANSA, 2020].

Gli utenti, i giornali, i media e la scena politica hanno contribuito al montare di un acceso dibattito sul *target*, caratterizzato da toni provocatori e discriminatori, connotati da una polarizzazione principalmente negativa. Per completezza, è necessario tener

presente che la liberazione di Silvia Romano è avvenuta a maggio in corrispondenza della fine del *lockdown* volto al contrasto della pandemia da *Covid-19*. Si potrebbe quindi ipotizzare che il periodo di tensione possa avere amplificato l'islamofobia, la xenofobia e la misoginia presenti nel Paese. D'altro canto, questa ipotesi sembra in contrasto con quello che si osserva su Luca Tocchetto[19]. Quest'ultimo è stato liberato e ha raggiunto l'Italia solo due mesi prima di Silvia Romano e quindi in un periodo di maggior confusione dovuto all'inizio dell'epidemia ma, nonostante questo, la sua vicenda non ha ricevuto la stessa attenzione mediatica e reazione sui social media (vedere *figura 31*).

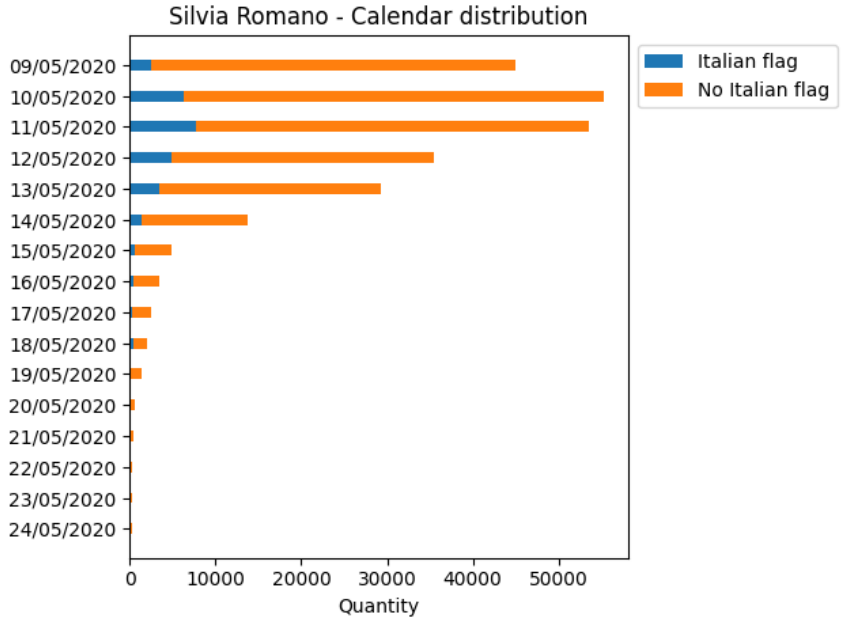

Figura 15: distribuzione giornaliera dei tweet raccolti su Silvia Romano confrontando i contenuti con il tricolore nel name, screen_name, description *o* nel tweet

Silvia Romano è il *target* maggiormente discusso, spesso in maniera negativa. Lo stesso trattamento non è stato riservato a Luca Tacchetto liberato a marzo 2020, Alessandro Sandrini e Sergio Zanotti arrivati in Italia rispettivamente a maggio e aprile 2019. Sono stati collezionati 248.240 *tweet* su Silvia Romano, di cui 237.031 collocati nei primi sette giorni (vedere *figura 15*), mentre gli altri tre sono arrivati massimo a poche migliaia. La loro principale differenza con Silvia Romano risulta essere il genere poiché

[19] Luca Tacchetto è un architetto padovano rapito in Burkina Faso nel dicembre 2018 assieme alla compagna canadese Edith Blais [Il Sole 24 Ore, 2020].

anche due di loro sono arrivati in Italia convertiti all'Islam[20]. Questo sottolinea come il sessismo possa apparire in forme implicite all'interno di un Paese dove la cultura ha radici patriarcali che definiscono i diritti e i doveri dei generi secondo consolidati stereotipi.

5.3.Elodie

La cantante italiana Elodie ha preso parte al dibattito con Lega Salvini in merito alla sua affermazione "Salvini? Piccolo uomo. Offende gratuitamente scatenando odio. Non mi piace come la Lega accalappia i voti. In Italia c'è troppa ignoranza". L'11 agosto 2020, Lega Salvini cita Elodie con una fotografia evidenziando le parti più delicate e scrive in maiuscolo "Non la pensi come lei? Sei ignorante!", questa discussione ha scatenato molte reazioni negative nei confronti della cantante che il 14 agosto 2020 ha collezionato alcuni commenti e li ha citati in un *tweet* scrivendo "E noi donne vi mettiamo pure al mondo"[21].

Considerato il tema del dibattito e i soggetti coinvolti, la sezione della raccolta annotata nel *pilot set* contiene molti *tweet* politici, anche *off-topic*, rivolti al contrasto del politico piuttosto che al dibattito. È una considerazione aspettata in quanto la collezione è stata ottenuta dalla selezione di *thread*, anche di partiti e politici, direttamente coinvolti nella vicenda.

La figura 16 mostra la quantità di *tweet* pubblicati con e senza il tricolore nel name, screen_name, description o nel tweet. La distribuzione dei contenuti è giornaliera e comprende la discussione di agosto e quella precedente di giugno 2020.

[20] *https://www.globalist.it/news/2020/05/12/tre-storie-di-italiani-rapiti-convertiti-all-islam-su-cui-nessuno -ha-aperto-bocca-non-erano-donne-2058060.html*

[21] *https://music.fanpage.it/elodie-e-gli-attacchi-sessisti-dopo-la-polemica-con-la-lega-e-noi-donne-vi-mettiamo-pure-al-mondo/*

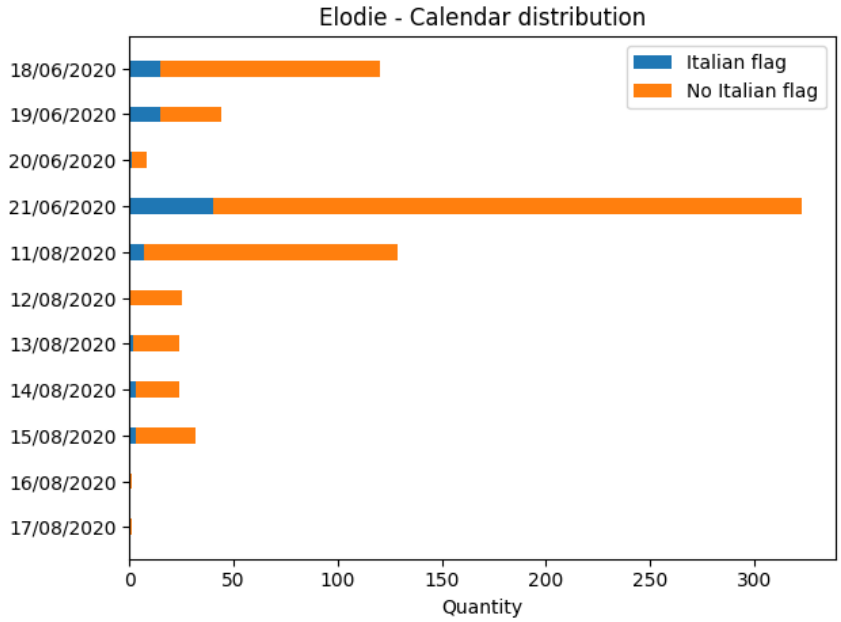

Figura 16: distribuzione giornaliera dei tweet raccolti su Elodie confrontando i contenuti con il tricolore nel name, screen_name, description o nel tweet

5.4.Prima fase: training

Il *pilot set*, composto da cento elementi, include:

- 50 reazioni in data 11 agosto 2020 selezionate dalla raccolta su Elodie;
- 25 risposte pubblicate il 10 maggio 2020 prelevate dalla collezione su Silvia Romano ottenute dai *thread* conversazionali di 32 *tweet* selezionati;
- 25 contenuti condivisi l'11 maggio 2020 estratti dalla raccolta su Silvia Romano filtrata per nome e cognome nel periodo 9-24 maggio 2020.

Visionando le annotazioni effettuate dai codificatori, è stato possibile notare che, come ci si aspettava, nel caso di Elodie l'etichetta "islamofobia" non risulta utile a differenza delle altre categorie. Le etichette "misoginia" e "xenofobia/razzismo" però difficilmente coesistono, questo rende complicata la misurazione dell'intersezionalità fra le due diseguaglianze. Inoltre, non sono stati riconosciuti alcuni contenuti negativi nei confronti della cantante poiché non è noto l'intero *background* e le *reference* coinvolte. Questo potrebbe ridurre il numero di attacchi al *target* classificati rendendo l'annotazione non completa. Lo stesso non succede con Silvia Romano che è stata un soggetto

particolarmente discusso nei programmi televisivi, nei giornali ma soprattutto sui social network. Rispetto a Elodie, i codificatori sono molto più consapevoli della narrativa che la coinvolge. Sono state rilevate molte dimensioni e la loro coesistenza, questo è utile per poter valutare la prevalenza delle varie etichette e per effettuare un paragone lessicale con le collezioni raccolte intorno a Luca Tacchetto, Alessandro Sandrini e Sergio Zanotti.

La misoginia è stata riconosciuta prevalentemente dai codificatori femminili vedendo solo 1/3 di quelli maschili avvicinarsi alle classificazioni delle prime. È un risultato aspettato in quanto è comprensibile che le donne, essendo colpite dalle discriminazione di genere, risultano più sensibili al tema rispetto agli uomini che potrebbero non riconoscere sufficientemente le forme implicite.

Sono apparse molte divergenze sull'etichetta "stereotipi", alcuni annotatori ne hanno riconosciuti tanti e altri molti meno. La valorizzazione del campo è risultata abbastanza soggettiva e alcune volte ripetitiva poiché la presenze di misoginia, xenofobia/razzismo e islamofobia, di solito, indicano rispettivamente stereotipi di genere, provenienza/etnia e culto religioso. Per ridurre la divergenza fra gli annotatori sono state affinate le linee guida in modo da chiarire alcuni casi che potrebbero porre dubbi.

Inoltre, è apparsa poco chiara la distinzione fra i valori "domanda" e "commento" all'interno dell'etichetta "presa di posizione (o *stance*)" ispirata a Pamungkas et al. [2019] poiché spesso le domande sono apparse retoriche e il confine può essere sottile. I due valori sono stati utilizzati da Pamungkas et al. [2019] per l'analisi dei *rumours*, la loro distinzione risulta meno significativa per questo studio pertanto sono stati sostituiti con l'etichetta "neutrale" che va a indicare la mancanza di schieramento dell'utente e le etichette "pro" e "contro" hanno preso il posto di "supporto" e "contestazione" per semplificare la comprensione.

Le divergenze apparse fra le annotazioni del *pilot set* sottolineano le difficoltà del dominio e la presenza di *bias* nei codificatori in funzione del genere e dell'età, per questo motivo sono state affinate le linee guida in modo da chiarire le forme implicite che si vogliono integrare all'interno della classificazione della seconda fase di annotazione.

5.5. Seconda fase

La seconda fase, a differenza della prima, è stata effettuata su un sito web adattato allo schema di annotazione stabilito. L'interfaccia grafica ha aiutato a velocizzare e migliorare la qualità dell'annotazione dei codificatori, mentre le linee guida affinate hanno sanato i dubbi presentati nel *pilot set*. I contenuti sono stati annotati da due persone e, per quelli in cui non si è ottenuto l'*agreement*, è stato richiesto un terzo annotatore. Alla fine di questo processo, 1680 sono risultati in *agreement* tra almeno due annotatori (vedere §5.6). La collezione annotata consiste in 1680 *tweet* (1500 estratti per *keys* e 180 dalle reazioni con il tricolore ai contenuti selezionati), ognuno è stato annotato da due annotatori differenti generando coppie con genere vario e una terza classificazione è subentrata in caso di *disagreement*.

Entrambe le collezioni contengono *tweet* che riconducono al fenomeno della "pre-mediazione" [Grusin, 2004] vedendo la diffusione della propria opinione, o quella di qualcun altro, come una reazione preliminare alla conoscenza e comprensione della notizia favorendo l'immediatezza e l'emotività nella comunicazione. Succede con Silvia Romano che viene attaccata fin da subito nonostante non fosse chiaro il quadro completo della vicenda (i contenuti annotati sono stati estratti in maniera randomica dalla raccolta dei primi tre giorni dal suo arrivo in Italia). Le informazioni più discusse sono state in merito alla conversione religiosa, la gravidanza, il presunto matrimonio e il sostenimento di non esser stata maltrattata né costretta a aderire all'Islam. Ancor prima di aver chiaro il quadro psico-fisico della ragazza, l'opinione pubblica ha elaborato, diffuso e condiviso affermazioni riguardanti il bersaglio fino a portare i suoi *hashtags* nei *trend topics* su Twitter.

Il piano politico ed economico è molto comune nei testi all'interno della collezione vedendo reazioni positive e negative nei confronti di politici che, a seconda del loro schieramento, vengono incolpati di non pensare agli italiani in periodo di Covid-19 e usare i quattro milioni di euro (in alcuni *tweet* è stato aggiunto anche uno zero) per liberare "un'ingrata e un irresponsabile che se l'è cercata andando in posti dove non doveva andare che il volontariato lo poteva fare in Italia dove c'è tanta gente bisognosa, in più è tornata convertita". Allo stesso modo, altri personaggi noti vengono affrontati per l'insensibilità mostrata. Sul piano economico non è sempre stato possibile riconoscere la presenza delle dimensioni poiché non esplicitamente espresse, quindi non sempre è stato

possibile assumere le motivazioni che potessero aver portato l'utente a condividere la propria disapprovazione nei confronti del riscatto.

Collezione filtrata per le keys "Silvia" e "Romano"

La presenza del nome e del cognome del bersaglio all'interno degli *hashtags* riguardanti Silvia Romano ha favorito l'estrazione dei contenuti sia positivi che negativi sulla vicenda. Sono stati compresi i *tweet* informativi, neutrali, aggressivi e che assumono le difese del *target* rendendo la collezione mista e riducendo il rumore bianco.

Le discriminazioni di stampo islamofobo sono state le più diffuse all'interno del *corpus* riconoscendo l'islamofobia nel 21.1% dei casi, seguita da misoginia e xenofobia comparse rispettivamente nel 19.2% e 10.2% dei *tweet* annotati (vedere *figura 17*).

L'islamofobia è apparsa principalmente sotto forma di disprezzo nei confronti del suo nuovo nome "Aisha" e della conversione spontanea della ragazza in contrasto con l'idea diffusa dell'Islam definita come una religione fatta di "sottomissione, violenze, stupri e attacchi terroristici". Spesso è stata usata come presupposto per assumere che la liberazione, e soprattutto il riscatto, non fossero necessari alla ragazza in quanto "stava bene lì dov'era". Silvia Romano è stata spesso chiamata "la musulmana", "l'islamica" e "la terrorista" per screditare la sua nuova fede e definendola membro dell'*outgroup* prendendo le distanze dalla sua persona.

La xenofobia e il razzismo si sono presentati comprendendo, oltre gli attacchi verso i nativi, l'avversione nei confronti della ragazza. Nonostante sia italiana, ha deciso di aderire alle usanze altrui e a occuparsi, attraverso il volontariato, degli abitanti di un altro paese generando i presupposti per poterla definire quasi "non più italiana" e quindi distante dall'italiano che scrive e la posiziona nell'*outgroup*.

La misoginia è la seconda categoria di discriminazione sociale rilevata. Include, fra le principali forme, il rifiuto che Silvia Romano possa aver preso la decisione di convertirsi, in quanto donna, ignorando il suo diritto nel farlo (discredito e *dominance*), la strumentalizzazione della gravidanza allo scopo di incolpare la vittima (*victim blaming*), l'affermazione dei presunti istinti carnali, non controllati, che avrebbero dovuto essere responsabili del matrimonio e della gravidanza spostando il *focus* su un argomento

più confortevole agli occhi di chi scrive (molestie sessuali e *derailing*). In alcuni casi la misoginia è apparsa anche in forma di "*benevolence*" screditando la ragazza definendola "poveretta" in quanto appartenente al genere femminile come se non potesse aver preso una decisione e fosse impotente, ad esempio "quella lì a cui hanno fatto subire/che l'hanno fatta convertire". Fra le etichette della misoginia a grana fine (vedere *figura 18*) emergono il discredito e la *dominance* nel 85.8% dei casi di misoginia, mentre le molestie sessuali e il *derailing* appaiono nel 12.5%. Il *derailing* risulta essere più complicato da classificare e meno evidente anche a causa del dominio del *corpus* diverso rispetto a quello di altre situazione con al centro episodi di stupri e *revenge porn*.

I codificatori femminili hanno annotato il 32.6% del *corpus* rispetto al 67.4% dei soggetti maschili (vedere *figura 21*). Le prime hanno riconosciuto la misoginia nel 22.1% dei tweet, le molestie sessuali e il *derailing* nel 24.1% e il discredito e la *dominance* nel 92.1% dei casi di misoginia, a differenza dei secondi che hanno riconosciuto le dimensioni rispettivamente nel 20.3%, 18.6% e 84.4% del totale annotato.

Gli stereotipi e la difesa del *target* sono comparsi rispettivamente nel 26.3% e 33.4% dei *tweet*. I primi hanno spesso coesistito con le altre dimensioni evidenziando stereotipi di genere, etnia e culto. La difesa del target invece è l'etichetta maggiormente contrassegnata che mostra come una buona percentuale di utenti abbia cercato di contrastare gli attacchi subiti da Silvia Romano generando delle contro-narrazioni con connotazioni sia positive che negative. Le prime cercando di contrastare l'abuso della parola mentre le seconde sostituiscono esclusivamente il destinatario alterando eventualmente l'intensità e l'argomento. Alcuni utenti cercano di far ragionare gli odiatori, altri puntano sui sensi di colpa e infine altri ancora "difendono il bersaglio ma…" dando origine a contraddizioni che vedono una intersezione fra la difesa del *target* e altre dimensioni d'odio. La presa di posizione (o *stance*) è stata invece difficilmente misurabile all'interno del *dataset* filtrato per *keys* poiché dipende dal contesto del *parent tweet* che può variare all'interno del *corpus* assumendo caratteristiche pro/contro ma anche *off-topic*.

La maggior parte dei contenuti sono stati pubblicati il 10 e l'11 maggio 2020 (vedere *figura 15*), le discriminazioni sono principalmente concentrate nel giorno del suo arrivo in Italia vendendo un aumento nella fascia pomeridiana e notturna (vedere *figura 19*).

La divisione equa del *corpus* fra i contenuti con o senza il tricolore nel *name*, *screen_name*, *description* o *tweet*, ha permesso un confronto fra i contenuti discriminatori pubblicati dagli utenti appartenenti alle due diverse sezioni. I contenuti con la bandiera italiana hanno caratterizzato il 78.5% della misoginia, il 73.2% della xenofobia e del razzismo e il 74.4% dell'islamofobia nel *corpus* (vedere *figura 20*). Pertanto, le percentuali sono elevate ma è chiaro come il tricolore non possa essere fonte di pregiudizi per classificare i contenuti in quanto ce ne sono altri che hanno assunto forme positive al sostenimento dello Stato, ad esempio "Non importano le dinamiche della liberazione di Silvia Romano, è importante che sia libera. Per una volta lo Stato italiano trionfa! 🇮🇹", dove 🇮🇹 è l'*emoji* della bandiera italiana.

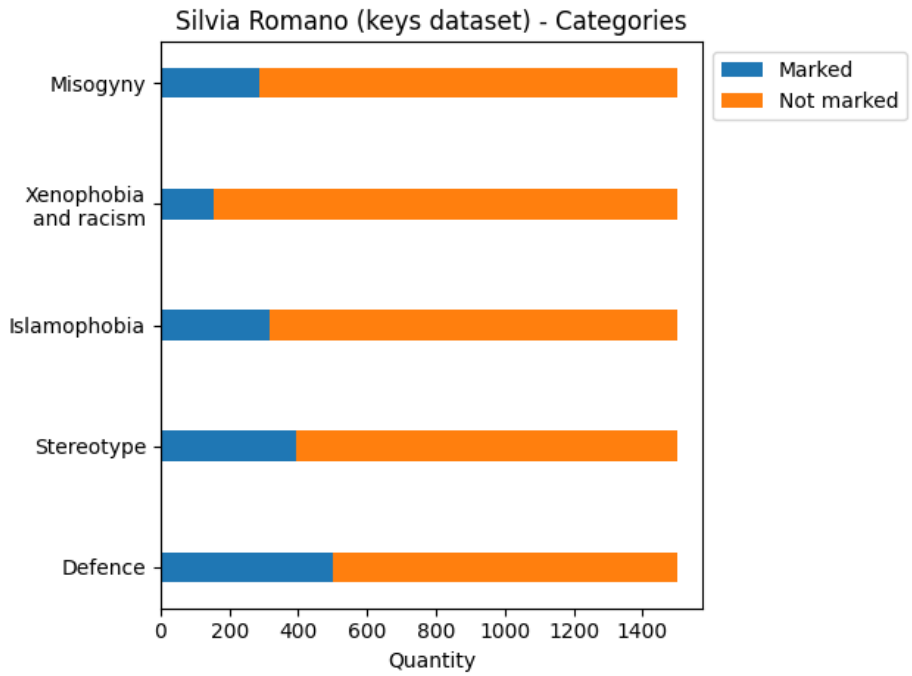

Figura 17: principali dimensioni discriminatorie nei tweet filtrati per keys

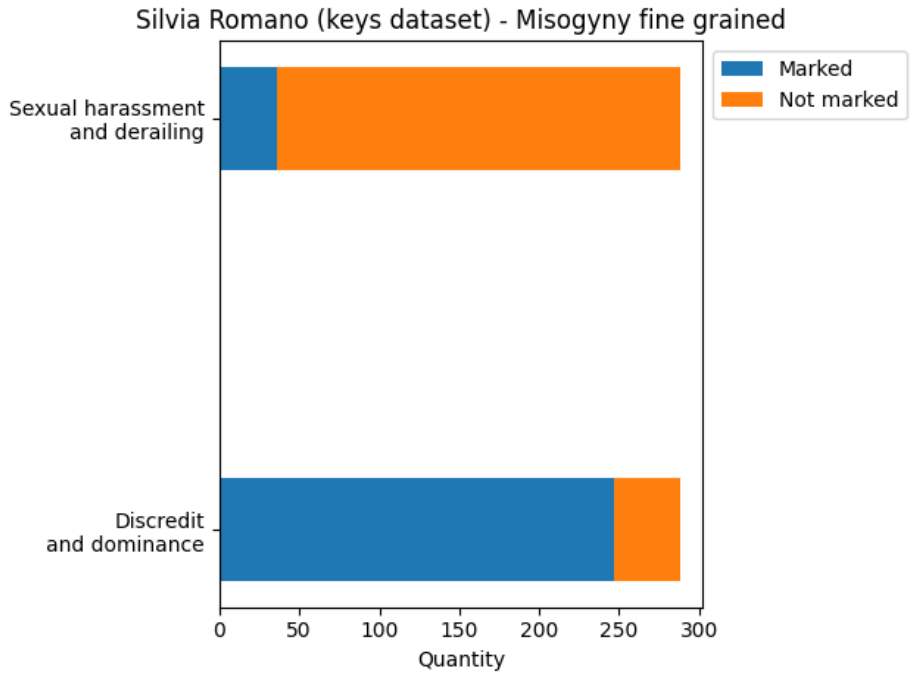

Figura 18: dimensioni a grana fine della misoginia nei tweet filtrati per keys

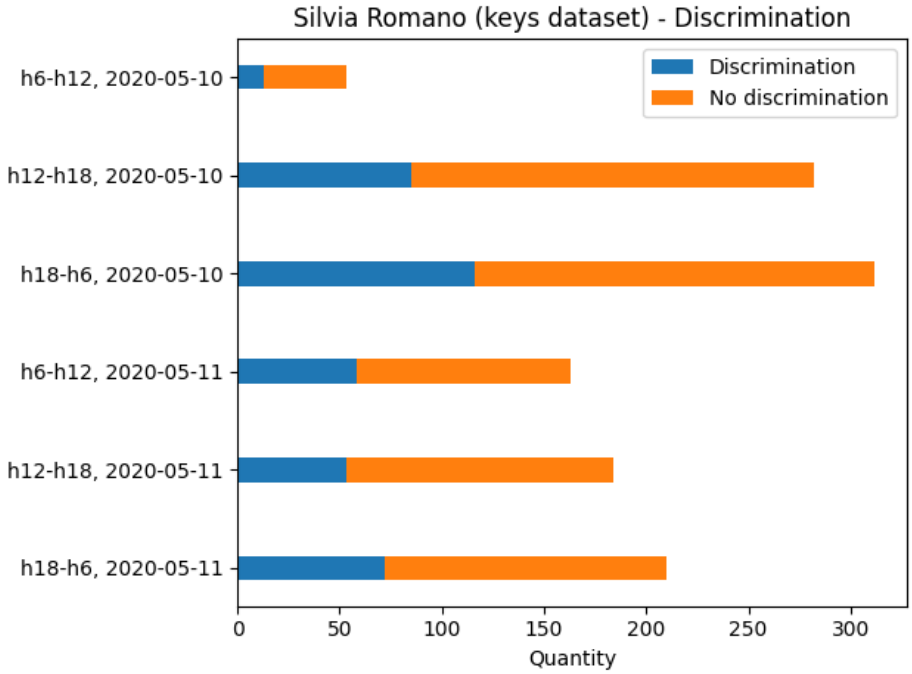

Figura 19: distribuzione oraria delle discriminazioni nei primi due giorni in Italia tra i tweet filtrati per keys

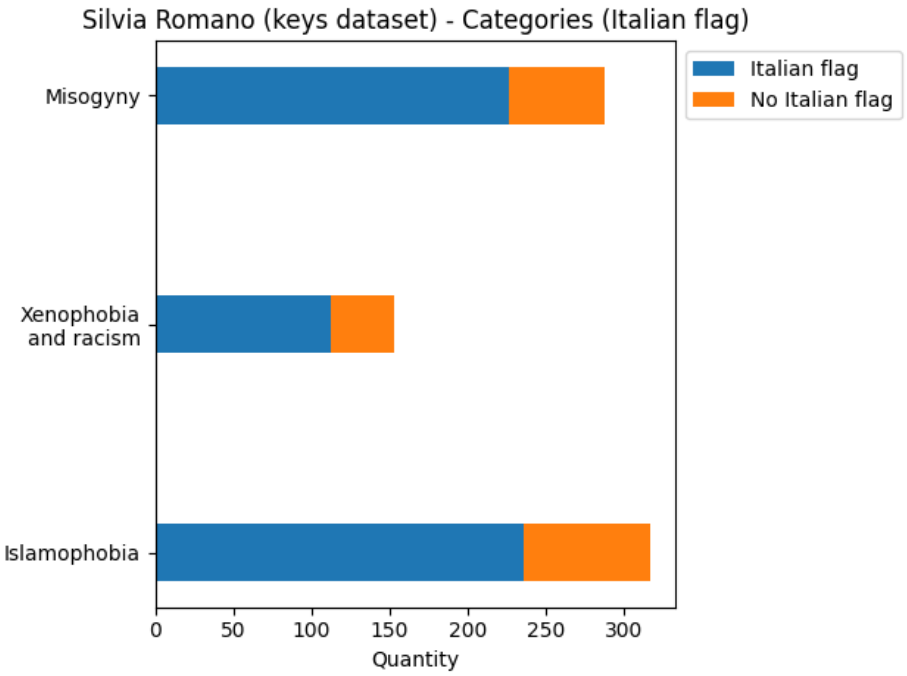

Figura 20: confronto tra le discriminazioni riconosciute nei tweet con e senza il tricolore tra i tweet filtrati per keys

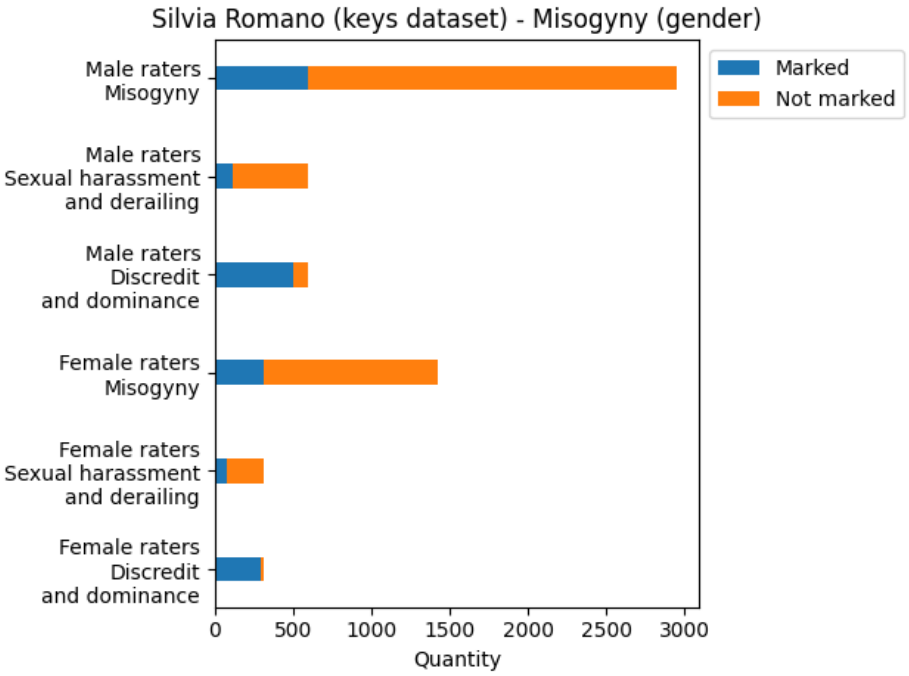

Figura 21: confronto annotazione della misoginia a grana fine e larga effettuata dai codificatori maschili e femminili sui tweet filtrati per keys

Collezione con le reazioni ai tweet scelti

La raccolta contiene le reazioni, con il tricolore in *name, screen_name, description* o *tweet,* ai contenuti selezionati fra quelli provocatori oppure apparentemente positivi o informativi ma che all'interno della *filter bubble* hanno fomentato contrasti nei *thread*. La categoria discriminatoria maggiormente classificata è l'islamofobia nel 48.9% dei *tweet,* a seguire la xenofobia e il razzismo nel 44.4% e la misoginia nel 43.3% dei contenuti (vedere *figura 22*). Nelle etichette a grana fine della misoginia (vedere *figura 23*) compaiono il discredito e la *dominance* nel 82.1% delle volte e le molestie sessuali e il *derailing* nel 23.1%. A parte il discredito e la *dominance*, le percentuali sono più che raddoppiate rispetto a quelle ottenute dall'annotazione della raccolta filtrata per *keys*, appare una leggera variazione nella sequenza vedendo le discriminazioni di stampo etnico leggermente più frequenti rispetto alla misoginia. Questi valori sono influenzati dalle scelte fatte nella selezione dei dati. Le tipologie di discriminazione rilevate sono le medesime descritte nel paragrafo precedente sulla raccolta per *keys* mostrando con maggior frequenza il tema politico a causa del contenuti selezionati. Sono aumentati anche gli stereotipi arrivando al 71.7%, mentre l'unica etichetta diminuita risulta essere la difesa del target che appare nel 5.6% mostrando come le reazioni siano arrivate maggiormente da utenti indifferenti, o che disapprovano la liberazione e il ritorno di Silvia Romano in Italia.

Molti dei contenuti selezionati sono stati pubblicati da giornali, partiti e politici. La responsabilità nella diffusione delle notizie dovrebbe garantire un'informazione il più possibile neutrale rispetto alle parti. Alcuni media, giornali, partiti e politici hanno rivestito un ruolo particolarmente rilevante nella diffusione della notizia manifestando, in maniera riconoscibile, la propria polarizzazione che ha influenzato le dimensioni delle reazioni ai loro contenuti. Nei casi di Tacchetto, Sandrini e Zanotti, è piuttosto visibile come i titoli degli articoli giornalistici si siano espressi in modo più neutrale e soprattutto quanto poco sono stati discusse le loro liberazioni rispetto a quella di Silvia Romano (vedere *figura 31*). Alcuni *tweet* con testi positivi hanno evocato discussioni influenzate da chi pubblica il *parent tweet* e della *filter bubble* in cui si trovano i *followers*. Alcuni esempi sono il *tweet* pubblicato da Giorgia Meloni:

- *"La liberazione di #SilviaRomano è una gran bella notizia. In attesa di vederla tornare in Italia il ringraziamento di Fratelli d'Italia va ai nostri*

servizi di Intelligence e a tutti coloro che hanno contribuito a raggiungere questo importante obiettivo. #silvialibera"

e i due di Matteo Salvini:

- *"Bentornata a casa #SilviaRomano! Un abbraccio a lei, alla sua famiglia e ai suoi amici. E un ringraziamento agli straordinari operatori dei Servizi Segreti italiani!";*
- *"#Salvini: bentornata a Silvia, complimenti agli uomini e alle donne dei Servizi italiani che rischiano la vita per salvare altre vite. #mezzorainpiù".*

In *figura 26*, è evidente come le reazioni ai tre contenuti appena riportati abbiano ridotto il supporto al 1.9% e la neutralità al 7.7% mentre è aumentata la contestazione arrivando al 90.4% a differenza della *stance* misurata sull'intera collezione di 180 elementi che ha visto le dimensioni rispettivamente al 7.8%, 8.3% e 83.9% (vedere *figura 25*). La contestazione si è manifestata in varie forme partendo dalla genuina disapprovazione, fino ad attacchi più pesanti e alla negazione del proprio voto alle prossime elezioni in quanto le parole del *tweet* non rispecchiano la politica e le opinioni usuali del politico.

I 180 elementi sono stati annotati da tre codificatori, due uomini e una donna, l'ultima ha riconosciuto di solito una percentuale più alta di misoginia a grana fine rispetto ai primi (vedere *figura 24*) ma non è possibile trarre conclusioni sui *bias* di genere in quanto ci sono stati pochi annotatori in questa collezione.

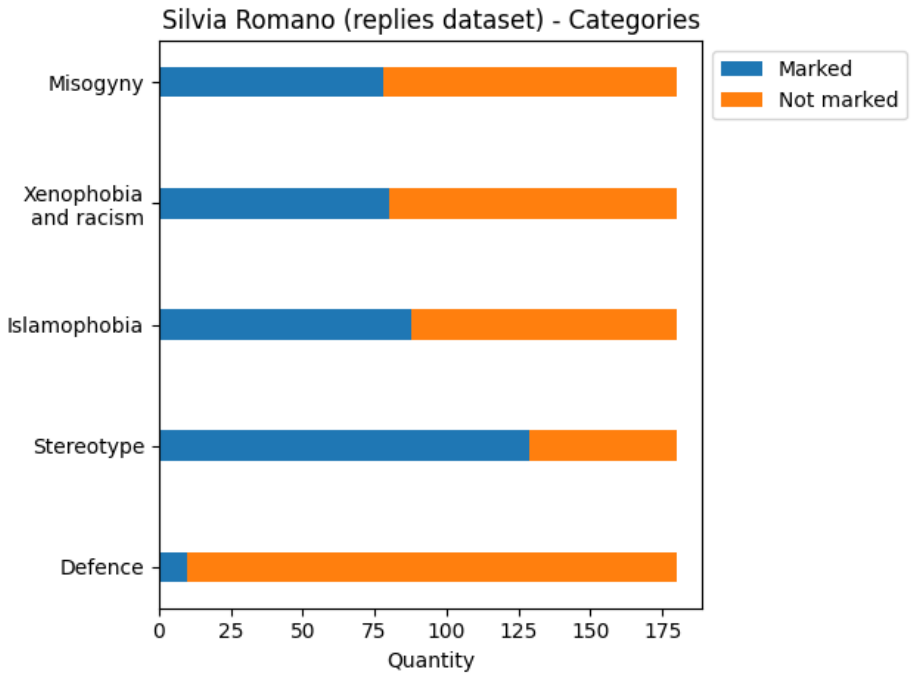

Figura 22: principali dimensioni discriminatorie nelle reazioni ai tweet scelti

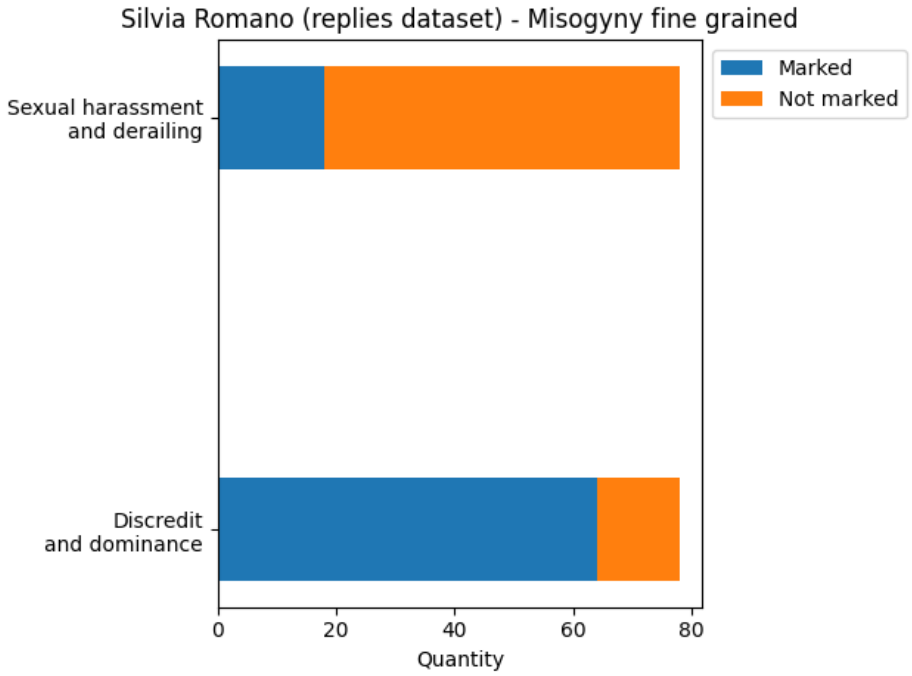

Figura 23: dimensioni a grana fine della misoginia nelle reazioni ai tweet scelti

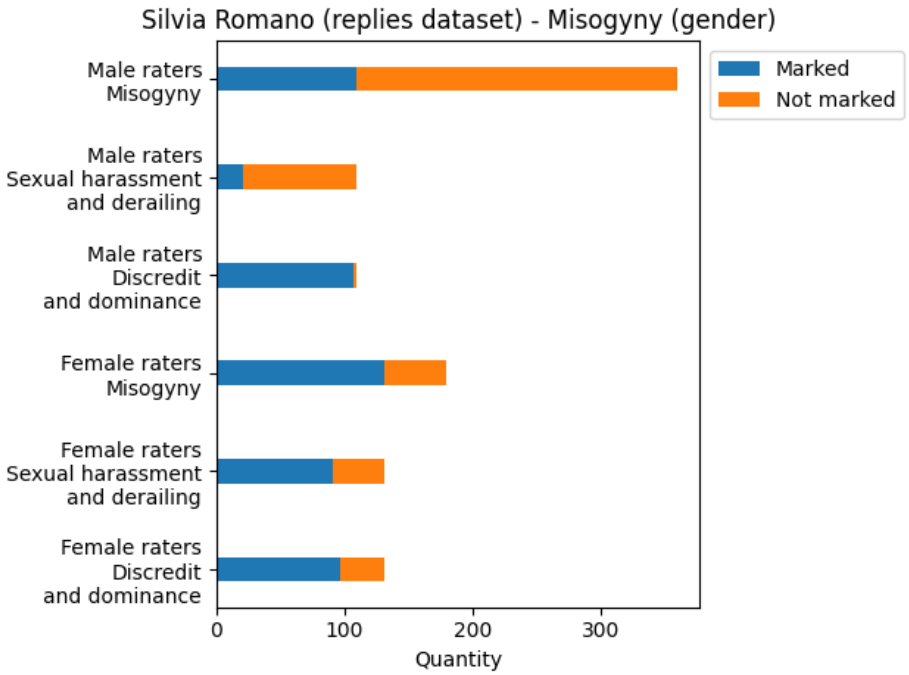

Figura 24: confronto annotazione della misoginia a grana fine e larga effettuata dai codificatori maschili e femminili sulle reazioni ai tweet scelti

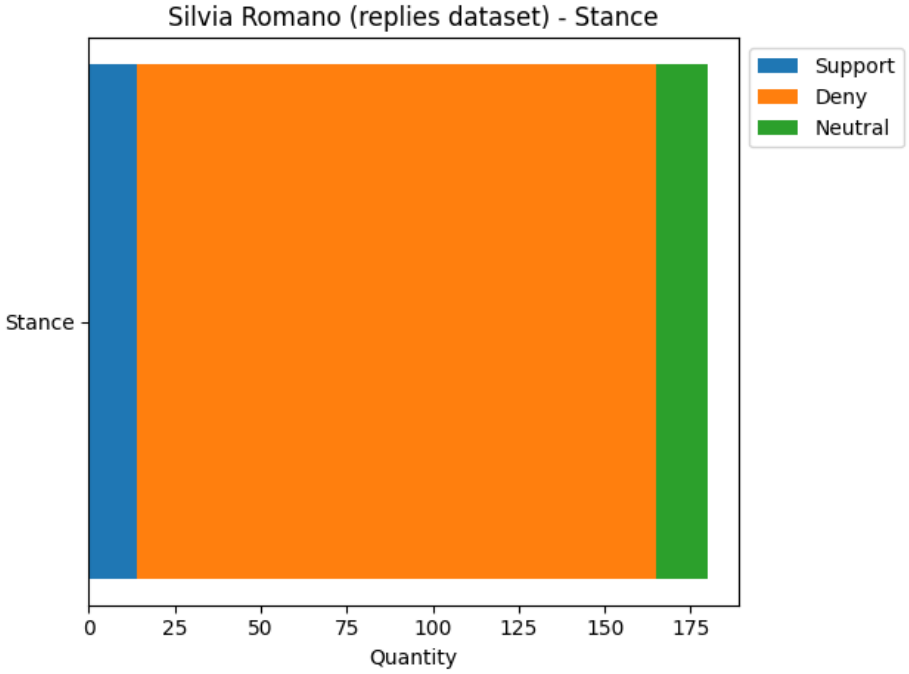

Figura 25: presa di posizione (o stance) nei confronti dei tweet scelti

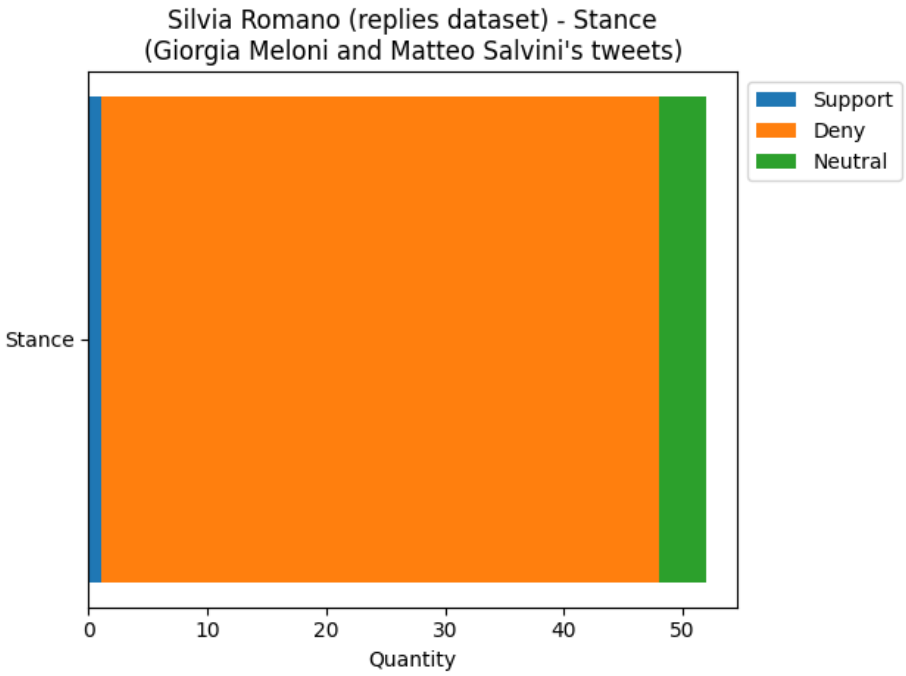

Figura 26: presa di posizione (o stance) nei confronti dei tre tweet scelti di Giorgia Meloni e Matteo Salvini

5.6. Agreement tra gli annotatori

Questo paragrafo contiene la misurazione dell'*agreement* tra i codificatori nelle due fasi dell'annotazione.

Prima fase

Il *pilot set* (descritto nel §5.4) è stato annotato da dieci codificatori, sei di genere maschile e quattro femminile. Ogni *tweet* è stato classificato da tutti i valutatori.

In *tabella 3* è visibile il calcolo dell'*agreement*, calcolato attraverso la Kappa di Fleiss[22], effettuato sull'insieme e filtrando per genere in modo da riconoscere un eventuale polarizzazione dovuta ai *bias*. Assumendo che la percezione soggettiva conti molto per la rilevazione dei fenomeni compresi nello schema multilivello, i *background* degli annotatori e il contesto sociale in cui si trovano, le opinioni e le credenze possono influenzare il modo in cui i soggetti si approcciano al problema. In *tabella 3* si nota come risulti esserci più *agreement* fra le donne nel riconoscimento della presenza di misoginia

[22] https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

a grana fine e larga nei post, piuttosto che fra gli uomini, un segnale di come ci sia una diversa percezione del fenomeno in base al genere. Come suggerito in alcuni studi recenti [Basile, 2020] i sentimenti e l'abuso della parola sono percepiti diversamente da un individuo all'altro, pertanto non sono sempre riconoscibili in maniera oggettiva e bisogna tenere in considerazione i vari punti di vista e le opinioni a riguardo, il che mostra quanto il *task* di riconoscimento di questi fenomeni sia difficile da effettuare anche per gli umani. Sono state quindi affinate le linee guida in modo da rendere più chiare le forme implicite che si vogliono comprendere nella seconda frase espressa di seguito.

	***Agreement* tra tutti gli annotatori**	***Agreement* tra le donne**	***Agreement* tra gli uomini**
Misoginia	0.28	0.30	0.23
Molestie sessuali	0.34	0.38	0.25
Discredito, *dominance* e *derailing*	0.25	0.26	0.19
Xenofobia e razzismo	0.33	0.26	0.37
Islamofobia	0.37	0.25	0.44
Prevalenza	0.12	0.0	0.13
Stereotipi	0.14	0.18	0.06
Difesa del *target*	0.64	0.60	0.67
Presa di posizione (o *stance*)	0.36	0.17	0.52

Tabella 3: agreement nel pilot set utilizzando da kappa di Fleiss[23]

Seconda fase

Per la raccolta con le *reply* ai contenuti scelti (180 elementi) abbiamo raccolto tre annotazioni indipendenti da tre codificatori differenti, due di genere maschile e uno femminile, mentre per quella filtrata per *keys* (1500 elementi) abbiamo raccolto tre annotazioni indipendenti per 1373 *tweet*. Per i restanti 127 *tweet* sono state raccolte due

[23] https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

annotazioni indipendenti e un terzo annotatore è intervenuto per risolvere i casi di disagreement.

Gli annotatori coinvolti in questa annotazione sono stati in tutto dodici, equamente divisi per genere, a cui sono state assegnate diverse porzioni del dataset da annotare (vedere *figura 21* e *figura 24* per i dettagli).

Calcolando l'Inter-Annotator Agreement (IAA), definita come la percentuale media di volte in cui due annotatori sono d'accordo tra loro, sono emersi 1006 accoppiamenti in *agreement* e 2784 in *disagreement* ottenendo un IAA medio del 26.5%. Il Kappa di Cohen[24] medio tra gli annotatori e gli attributi è 0.40, mentre il Kappa di Fleiss[25] per ogni etichetta all'interno delle due raccolte è visibile in *tabella 4*. Il calcolo del Kappa di Fleiss per i tweet con tre annotazioni indipendenti per *tweet* ha evidenziato come sia stato complicato classificare il dominio attraverso le dimensioni dello schema multilivello stabilito.

Nel caso della raccolta per *keys*, la presa di posizione o *stance* (0.81) ha ottenuto un *agreement* quasi perfetto, a seguire la difesa del *target* (0.59), l'islamofobia (0.53), la misoginia (0.49) e il discredito e *dominance* (0.44) con un *agreement* moderato, le restanti dimensioni sono state caratterizzate da un *agreement* equo (vedere *tabella 4*).

All'interno della collezione contenente le reazioni ai *tweet* scelti, la categoria maggiormente in *agreement* è la difesa del *target* (0.52) con un *agreement* moderato, successivamente appaiono la xenofobia e il razzismo (0.36), il discredito e la *dominance* (0.34), l'islamofobia (0.32), gli stereotipi (0.30) e la misoginia (0.24) equamente in *agreement*. I rimanenti hanno ottenuto un *agreement* leggero tra cui la *stance* che in questa collezione assume un valore completamente diverso rispetto al filtraggio per *keys* (vedere *tabella 4*). Il *disagreement* dipende molto anche dalla percezione soggettiva del fenomeno che varia tra gli annotatori e dalle caratteristiche degli stessi. È facile pensare come i soggetti femminili siano più propensi a riconoscere atteggiamenti misogini rispetto a quelli maschili in funzione dei *bias* di genere e della vita quotidiana dei soggetti che vedono le donne nel mirino.

[24] *https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html*
[25] *https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html*

	Raccolta per keys (tre annotazioni)	**Reply (tre annotazioni)**
Misoginia	0.49	0.24
Molestie sessuali e *derailing*	0.29	-0.04
Discredito e *dominance*	0.44	0.34
Xenofobia e razzismo	0.32	0.36
Islamofobia	0.53	0.32
Prevalenza	0.33	0.15
Stereotipi	0.33	0.30
Difesa del *target*	0.59	0.52
Presa di posizione (o *stance*)	0.81	0.04

Tabella 4: agreement nella seconda fase di annotazione utilizzando da kappa di Fleiss[26]

[26] *https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html*

6.Analisi corpus-based dell'intersezionalità e dei segnali lessicali attraverso Hurtlex

Il sesto capitolo comprende la descrizione dell'intersezionalità classificata nella seconda fase dell'annotazione, la valutazione del modello, l'analisi e il confronto lessicale mediante Hurtlex [Bassignana et al., 2018] paragonando i contenuti su Silvia Romano a quelli riguardanti altri tre soggetti liberati tra il 2019 e 2020.

6.1.Modello per identificare le intersezionalità

Lo schema di annotazione multilivello definito ha permesso la classificazione dell'intersezionalità nei casi in cui, all'interno dei contenuti, fossero coesistenti più di una dimensione fra misoginia, xenofobia e razzismo e islamofobia. Nelle annotazioni che riconoscono una o nessuna categoria di discriminazione sociale, l'etichetta ha assunto il valore di *default* "assente", pertanto vengono esclusi quei contenuti dai grafici a seguire.

Non è sempre risultato semplice validare questa *label* in quanto la presenza di più valori e i *background* dei codificatori, e delle persone che le circondano, potrebbero aver influenzato la scelta riconoscendo in prevalenza una categoria rispetto ad un'altra. Inoltre, essendo possibile validare il campo solo nel caso in cui fossero state assegnate almeno due dimensioni, aumentano i casi di disaccordo fra gli annotatori che hanno annotato diversamente le etichette propedeutiche. L'etichetta non è in grado di rilevare alcune forme sottili come il sessismo implicito riconosciuto visionando dall'esterno i flussi di discorso intorno al *target* e altri soggetti con caratteristiche simili e divergenti. In questo modo è possibile ipotizzare le eventuali motivazioni a monte delle reazioni degli utenti (vedere §6.2).

Nel caso della collezione filtrata per *keys*, è stata osservata la coesistenza di più categorie nel 12.9% delle situazioni vedendo l'islamofobia, la misoginia e la xenofobia/razzismo rispettivamente con le seguenti percentuali: 50.8%, 32.6% e 16.6%

(vedere *figura 27*). All'interno della raccolta con le reazioni ai tre *tweet* scelti di Giorgia Meloni e Matteo Salvini, è stata classificata la prevalenza nel 41.1% dei casi con l'islamofobia, la misoginia e la xenofobia/razzismo presenti nel 48.7%, 37.8% e 13.5% dei *tweet* (vedere *figura 28*). In entrambi i casi risulta essere l'islamofobia la dimensione prevalente all'interno del *corpus* dimostrando come la conversione spontanea alla religione islamica sia stata rilevante all'interno del flusso di discorso estratto. A seguire appare la misoginia spesso riconosciuta nei contenuti poiché è stata screditata come se non avesse il diritto, la possibilità e la sanità mentale per prendere le decisioni. È stata di frequente attaccata sul piano sessuale strumentalizzando la gravidanza e alludendo al piacere carnale della ragazza dietro alle vicende narrate in relazione ai pregiudizi sulla religione islamica. Le due dimensioni sono apparse quindi fortemente correlate ed entrambe con grande rilevanza. La xenofobia e il razzismo sono comunque presenti ma nell'annotazione sono apparsi come una categoria meno frequente rispetto alle altre due risultate il fulcro della narrazione.

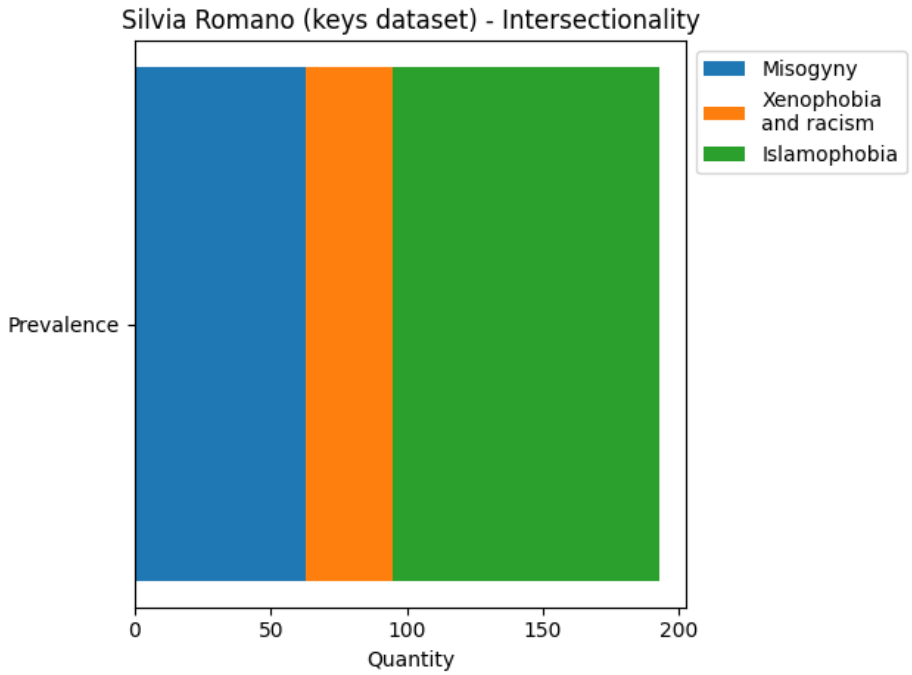

Figura 27: prevalenza delle principali dimensioni d'odio tra i tweet filtrati per keys

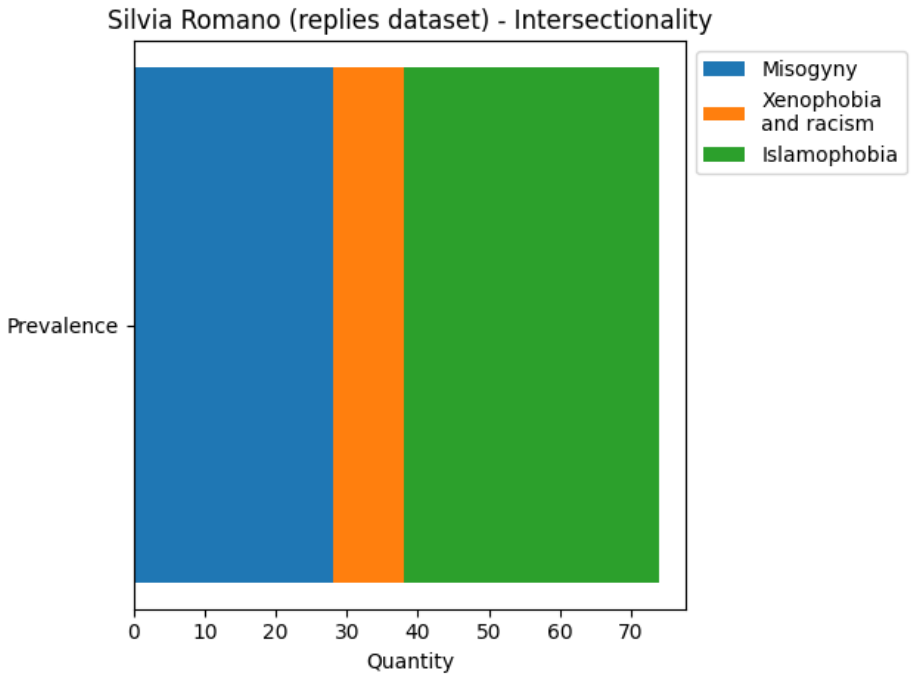

Figura 28: prevalenza delle principali dimensioni d'odio tra le reazioni ai tweet scelti

6.2. Analisi dei segnali lessicali attraverso l'uso del lessico computazionale HurtLex

L'analisi lessicale tramite Hurtlex [Bassignana et al., 2018] è stata utile per analizzare l'intero *corpus* di quasi 250.000 *tweet* dal punto di vista dei termini utilizzati e quindi del genere di lessico scelto per discutere Silvia Romano. Al seguito è stato anche possibile confrontare il *target* con Luca Tacchetto, Alessandro Sandrini e Sergio Zanotti rilevando il differente approccio ai bersagli in termini di quantità e forma dei *tweet*.

Su Silvia Romano sono stati riconosciuti 18.403 *tweet* con stereotipi rappresentando il 7.4% del *corpus*. Le categorie lessicali (descritte in *figura 11*) apparse maggiormente (vedere *figura 29 e figura 30*) sono le *parole dispregiative* (CDS, 28611), le parole relative a *crimini* e *comportamenti immorali* (RE, 25079), le potenziali *connotazioni negative* (QAS, 14501), i *difetti morali e comportamentali* (DMC, 8299) e le parole relative ai *sette peccati capitali della tradizione cristiana* (SVP, 5569). A seguire compaiono i termini sui genitali femminili (ASF, 4593) e maschili (ASM, 3082), del mondo animale (AN, 3970), le disabilità cognitive e le diversità (DDP, 3393) e gli stereotipi negativi e *insulti etnici* (PS, 2144). Le dimensioni apparse, anche se in maniera

minore a causa della mancanza dell'analisi semantica, sono riconducibili alle forme degli attacchi trovati nell'annotazione, tra questi l'utilizzo della cultura e della religione per definire cos'è immorale, il discredito del *target*, la criminalizzazione della vittima, la deumanizzazione del diverso, il sostenimento dell'incapacità mentale a prendere decisioni e i discorsi sulla sfera sessuale.

La stessa analisi lessicale è stata applicata ai dati raccolti per Tacchetto, Sandrini e Zanotti. Il confronto tra Romano, Tacchetto, Sandrini e Zanotti (vedere *figura 31* e *figura 32*) è volto a misurare come gli utenti si sono approcciati ai quattro soggetti in funzione del fatto che sono stati tutti rapiti e liberati nel periodo 2019-2020. Come Silvia Romano, Luca Tacchetto e Alessandro Sandrini sono tornati convertiti all'Islam[27] e con un aspetto riconducibile alle usanze del posto, come la barba lunga. Sandrini e Zanotti sono stati liberati grazie a un negoziato[28] e quindi la principale differenza fra i quattro risulta essere il genere biologico. Il confronto in *figura 31* è stato costruito sulla base dei contenuti filtrati per *keys* (nome e cognome coesistenti) da Twita pubblicati nei primi sette giorni dall'arrivo in Italia evidenziando con il colore blu il numero di *tweet* contenenti almeno un termine nel lessico Hurtlex sul totale delle condivisioni. È visibilmente chiaro come Silvia Romano sia stata maggiormente discussa rispetto agli altri tre soggetti (vedere *figura 31*). In sette giorni sono stati pubblicati 237.031 contenuti su Romano, 1668 su Tacchetto, 546 su Sandrini e 2206 su Zanotti di cui rispettivamente 88.233 (37.2%), 642 (38.5%), 203 (37.2%) e 1312 (59.5%) contenenti termini in Hurtlex (vedere *figura 31* e *figura 32*). L'analisi lessicale possiede i suoi limiti dovuti alla mancanza dell'analisi sul piano semantico necessaria per comprendere la natura dei contenuti, ad esempio un tweet informativo che riporta un attacco viene classificato allo stesso modo di quello che l'attacco l'ha espresso.

[27] *https://www.globalist.it/news/2020/05/12/tre-storie-di-italiani-rapiti-convertiti-all-islam-su-cui-nessuno -ha-aperto-bocca-non-erano-donne-2058060.html*

[28] *https://www.globalist.it/news/2020/05/12/tre-storie-di-italiani-rapiti-convertiti-all-islam-su-cui-nessuno -ha-aperto-bocca-non-erano-donne-2058060.html*

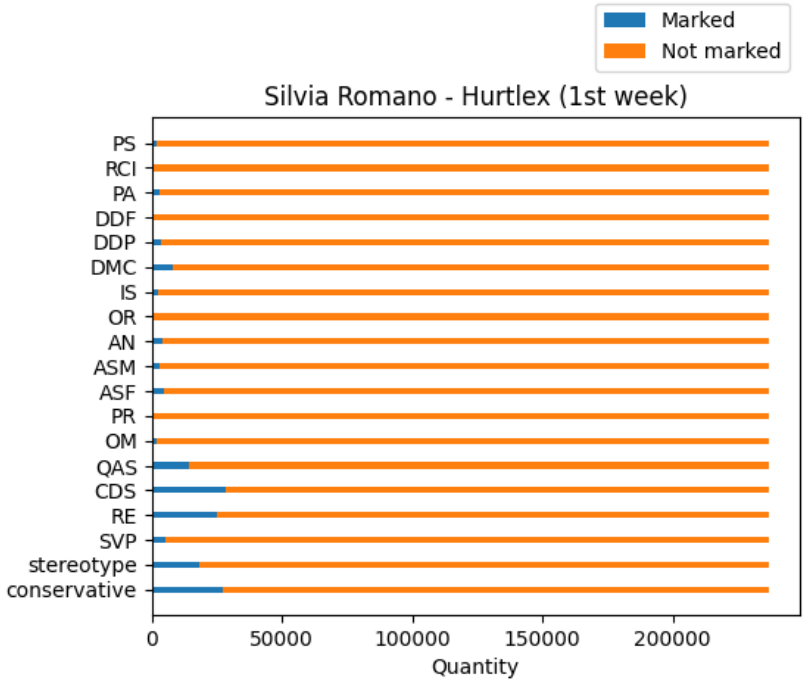

Figura 29: distribuzione delle categorie di Hurtlex nei tweet filtrati per keys

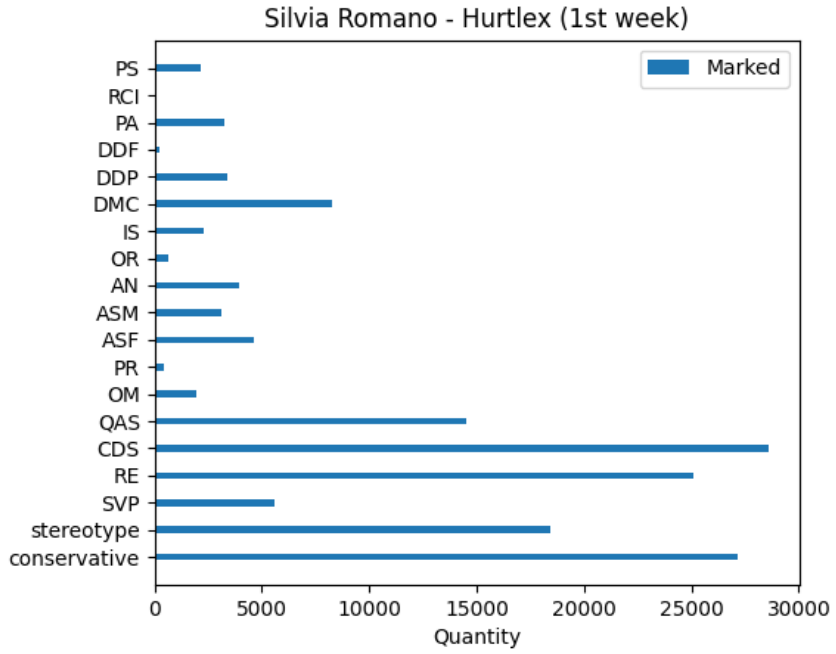

Figura 30: zoom sulla distribuzione delle categorie di Hurtlex nei tweet filtrati per keys

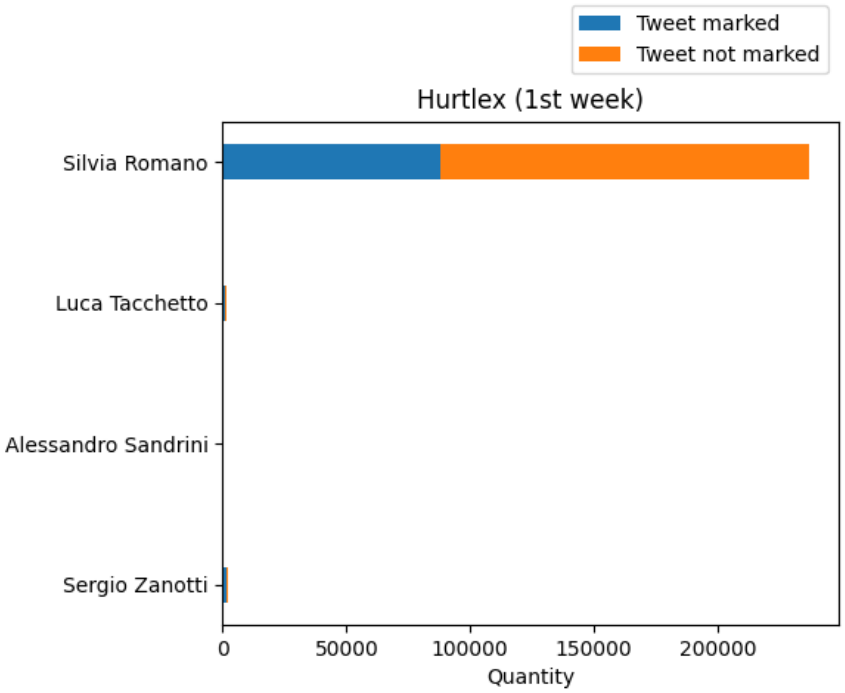

Figura 31: confronto tra i tweet, filtrati per keys dei tre soggetti liberati, contenti almeno un termine in Hurtlex

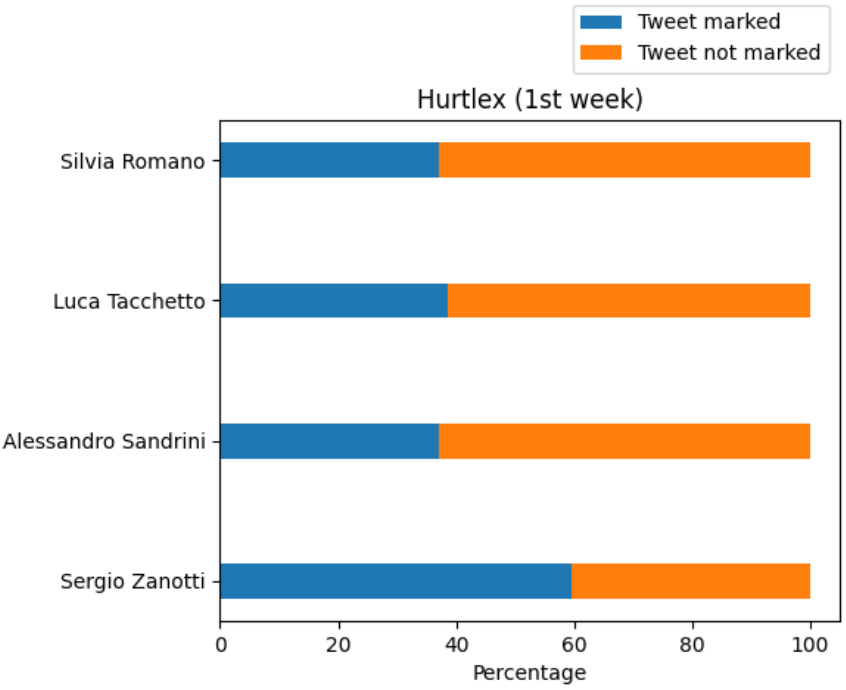

Figura 32: percentuale di tweet con termini in Hurtlex sul totale estratto per soggetto

7.Conclusioni

In questa tesi abbiamo proposto un'analisi dell'odio intersezionale in Twitter, con particolare attenzione a come la dimensione dell'odio misogino si interseca con altre dimensioni come l'islamofobia e il razzismo. In questo studio abbiamo deciso di partire dalle vittime femminili di campagne d'odio; questo approccio ha guidato la selezione dei dati e ha permesso di fare venire alla luce alcune intersezioni fra diverse discriminazioni ponendo il *focus* sulla misoginia al centro. Come mostrato dalle analisi di Vox [2019], la concentrazione dei *tweet* e dei contenuti negativi sui social cresce in corrispondenza di specifici avvenimenti nel mondo reale. Questo ha portato alla decisione di analizzare più nello specifico alcuni eventi come la discussione tra la cantante Elodie e Lega Salvini nel *pilot set*, e la liberazione e l'arrivo in Italia di Silvia Romano nello sviluppo del corpus annotato *IntersHate*. È apparsa evidente la relazione fra il mondo reale e digitale che tende a commentare e giudicare la quotidianità, e i rapporti causa-effetto fra di essi.

Parte dello studio si è focalizzata sull'analisi di post rilevanti per il dibattito in cui gli utenti scrivevano in risposta a un insieme di tweet generati da account che si ipotizza rivestire un ruolo particolare nei dibattiti online, ossia profili di testate giornalistiche e di politici attivi sui social, ipotizzando di potere in questo modo avere un canale di ingresso per identificare contenuti testuali e contesti/narrazioni potenzialmente provocanti. L'annotazione di queste reazioni all'insieme di *tweet* scelti mostra una particolare polarizzazione degli utenti, legata all'appartenenza a una certa "fazione" con la quale si identificano, resa esplicita e quasi dichiarata dagli utenti stessi attraverso l'uso di specifici simboli (specificatamente *emoji*) nei profili. Abbiamo osservato inoltre, uno scarso numero di post a difesa della ragazza e un sostanziale incremento di espressioni riconducibili a discriminazioni e degli stereotipi. Le dinamiche di funzionamento dei social network hanno sicuramente favorito la diffusioni dei contenuti di personaggi molto seguiti, *tweet* e *hashtags* diventati virali, come *#convertita* nel caso di Silvia Romano. L'uso scorretto delle piattaforme ha caratterizzato il flusso di contenuti favorevoli e contrari ai bersagli vedendo gli utenti aderire a opinioni spesso ignorando le ripercussioni delle violenze psicologiche meno evidenti rispetto a quelle fisiche.

All'interno del *corpus* si è distinto il fenomeno della *pre-mediazione* che ha visto i contenuti emotivi e impulsivi in anticipo rispetto alla diffusione e alla conferma delle notizie.

La categoria maggiormente colpita è l'islamofobia che si posiziona al primo posto anche in termini di prevalenza fra le dimensioni all'interno della raccolta annotata. A seguire misoginia, xenofobia e razzismo che assumo percentuali simili nelle *reply*, mentre nella raccolta per *keys* la prima appare in quasi il doppio dei casi rispetto alle ultime. Il discredito e la *dominance* sono presenti in maniera analoga fra le due collezioni a differenza delle molestie sessuali e del *derailing* che sono il doppio nelle reazioni. Gli stereotipi risultano essere particolarmente diffusi sia nei due *dataset* annotati che nell'analisi lessicale effettuata su Silvia Romano, le difese del *target* invece sono molto diffuse nei contenuti filtrati per *keyword* ma poco nelle reazioni ai contenuti scelti, probabilmente a causa dell'orientamento politico degli utenti dei tweet scelti. Alcune categorie come la *stance*, gli stereotipi e il *derailing* sono risultate più difficile da identificare a causa del dominio del *corpus* e delle diverse percezioni soggettive degli annotatori dovute ai *bias* che ognuno possiede. L'annotazione delle *reply* ai contenuti scelti, pubblicati da giornali, partiti e politici, ha mostrato come la contestazione del *parent tweet* risulti esse alta soprattutto dei casi in cui i personaggi noti pubblicano contenuti positivi che però non sono in sintonia con la linea politica e le opinioni generalmente espresse; questo mostra come i loro *tweet* possano influenzare le reazioni nei *thread* social e come per gli stessi *leader* sia difficile discostarsi da una linea di pensiero definita all'interno di queste bolle. Anche le *emoji* associate ai profili sono apparse rilevanti come predittori di post con contenuti discriminatori, vedendo il tricolore italiano apparire in oltre il 70% delle discriminazioni nel *corpus*.

L'analisi lessicale applicata ai dati raccolti utilizzando il lessico computazionale HurtLex ha messo luce alcune categorie linguistiche di parole per ferire apparse nel *corpus* di quasi 250.000 *tweet,* che sono state riconosciute anche all'interno dell'annotazione. Tra queste categorie: le azioni immorali, la religione, la sfera sessuale, il discredito e le terminologie di insulti ispirata al mondo animale. Il confronto con Luca Tacchetto, Alessandro Sandrini e Sergio Zanotti ha permesso di evidenziare quanto diversamente Silvia Romano è stata discussa rispetto agli altri tre uomini nei primi sette giorni dall'arrivo in Italia. Il volume di tweet pubblicati nel dibattito su Silvia Romano è

di gran lunga superiore. Abbiamo osservato un volume di tweet da 107 a 434 volte superiore rispetto a quello osservato per i tre uomini aventi caratteristiche in comune con Silvia Romano (inclusa la conversione all'Islam di due di loro).

Amnesty International [2020] ha notato che solo un uomo è stato soggetto a una gogna pubblica, tra quelle rilevate. In maniera simile, nel nostro studio l'intersezionalità della misoginia, oltre ad apparire attraverso lo schema di annotazione, è visibile in maniera sottile e implicita sotto forma di sessismo osservando la differenza fra la mole di materiale pubblicato su Twitter intorno alla ragazza e quella sui tre uomini.

È rassicurante notare come una parte degli utenti si occupi delle contro narrazioni andando a difendere il *target* e quindi riducendo l'impatto negativo della campagna d'odio a cui è soggetto. Sarebbe quindi interessante analizzare i flussi di narrazioni alternative generati dalla posizione assunta nei confronti del bersaglio. Allo stesso modo si potrebbe valutare l'ironia e la misoginia espressa sotto forma di "*benevolence*" per arricchire il quadro complessivo sulla misoginia e sull'intersezionalità. Inoltre, è importante cercare di rendere l'annotazione il più possibile equamente distribuita tra i generi per ridurre i *bias*. Abbiamo osservato come questo porti come *side effect* a una naturale polarizzazione fra i codificatori nel riconoscimento del fenomeno della misoginia, un fenomeno interessante che intendiamo studiare a fondo in futuro attraverso gli strumenti introdotti in [Akthar et al., 2019]. Trattandosi di notizie sui *target,* l'analisi delle reazioni potrebbero essere utile per analizzare l'impatto delle fonti di informazione sul fenomeno.

Considerando l'aumento dell'odio online negli anni raccontato anche da Vox [2019] e Amnesty International [2020], è necessaria la ricerca di una visione più inclusiva della realtà. Non dimenticandoci che ognuno può scegliere le parole giuste da usare.

Ringraziamenti

I migliori ringraziamenti vanno a tutt* coloro che hanno preso parte alla realizzazione di questa tesi, tassello dopo tassello. Innanzitutto, i miei due relatori Viviana Patti e Mirko Lai che hanno reso possibile questo percorso e tutti gli annotatori che in pochissimo tempo si sono mobilitati. Un grazie speciale a Noemi Rabassi, Martina Rossi, Sara Placenti, Athena Parodi, Valeria Scandurra, Federico Torrielli, Davide Polimeni, Gabriele Scanu, Marco Stranisci, Lorenzo Sciandra e i relatori che hanno avuto la pazienza di annotare il corpus IntersHate rilasciando importanti feedback. Ringrazio nuovamente Noemi Rabassi, Federico Torrielli e Martina Rossi per i confronti avuti negli ultimi mesi che hanno permesso di mettermi in discussione e ampliare i miei punti di vista su un argomento molto delicato che non può ignorare chi viene colpito in prima persona e su come questo può accadere online. Un grazie a tutti i miei amici che mi sorreggono, i miei parenti e soprattutto i miei genitori che mi hanno concesso di percorrere questo cammino raggiungendo i più importanti checkpoint. In conclusione, voglio ringraziare infinitamente mia madre che, senza accorgersene, mi ha dato tutto ciò di cui avevo bisogno.

Fonti

Bibliografia

Akhtar, S., Basile, V. & Patti, V. (2019). A New Measure of Polarization in the Annotation of Hate Speech. AI*IA, pp. 588-603.

Aronson, E., Wilson, T. D., Akert, R. M. (2013). Social Psychology, 8th edition, Pearson Education Inc.

Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Atlanta, pp. 100–107.

Basile, V., Lai, M., & Sanguinetti, M. (2018). Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. Turin: CEUR-WS, pp. 1-6.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis: Association for Computational Linguistics, pp.54-63.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In *CEUR Workshop Proceedings*. Turin: CEUR-WS, 2253, pp. 1-6.

Brown, R. (2010). Prejudice: Its Social Psychology. Oxford: Wiley-Blackwell.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. In *Stanford Law Review*, 43(6), pp. 1241-1299.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Caselli, T., Novielli, N., Patti, V., & Rosso, P. (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*. Turin: Accademia University.

Grusin, R. (2004). Premediation. Criticism, 46(1), 17-39.

Kim, J.Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). Intersectional Bias in Hate Speech and Abusive Language Datasets. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Manne, K. (2017). Down girl: The logic of misogyny. In *Oxford University Press*.

McCombs, M., & Shaw, D. (1972). The Agenda-Setting Function of Mass Media. The Public Opinion Quarterly, 36(2), 176-187.

Pamungkas, E.W., Basile, V., & Patti, V. (2019). Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure. In *Proceedings of 2nd RDSM*, 2018.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. *LREC*.

Scandurra, V. (2020) (Gender) Violence in mass media: how Italian media describe women. Essay. Jagiellonian University in Kraków, Poland.

Sitografia e materiale multimediale

AGI (Agenzia Giornalistica Italia) (2020, May 9). Chi è Silvia Romano, la cooperante italiana liberata dopo 18 mesi in Africa. Available at: *https://www.agi.it/cronaca/news/2020-05-09/chi-e-silvia-romano-8565666/*

Amnesty International (2020). Barometro dell'odio – sessismo da tastiera. Available at: *https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2020/03/15212126/Amnesty-Barometro-odio-aprile-2020.pdf*

ANSA (Agenzia Nazionale Stampa Associata Soc. Coop.) (2020, May 12). Silvia Romano: choc su Fb, 'impiccatela'. Available at: *https://www.ansa.it/sito/notizie/topnews/2020/05/12/silvia-romano-choc-su-fb-impiccatela_fa5cafa4-4de4-4aaf-9c06-c72c0ca75d24.html*

Basile, V. (2020). On the Impact of the Pre-aggregation on the Evaluation of Highly Subjective Tasks. AIxIA DP. Paper_81. Available at: *https://vimeo.com/477462102*

Crenshaw, K. (2016). The urgency of intersectionality. TEDWomen 2016. Available at: *https://www.ted.com/ talks/kimberle_crenshaw_the_urgency_of_intersectionality*

ECRI (European Commission against Racism and Intolerance) (2007). Expert seminar: combating racism while respecting freedom of expression. Available at: *https://www.ivir.nl/publicaties/download/Proceedings_ECRI_2007.pdf*

Fanpage (2020, August 14). Elodie e gli attacchi sessisti dopo la polemica con la Lega: "E noi donne vi mettiamo pure al mondo". Available at: *https://music.fanpage.it/elodie-e-gli-attacchi-sessisti-dopo-la-polemica-con-la-lega-e-noi-donne-vi-mettiamo-pure-al-mondo/*

GitHub, Hurtlex. Available at: *https://github.com/valeriobasile/hurtlex*

GitHub, Tweepy. Available at: *https://github.com/tweepy/tweepy*

Globalist (2020, May 12). Tre storie di italiani rapiti convertiti all'Islam su cui nessuno ha aperto bocca: non erano donne. Available at: *https://www.globalist.it/news/2020/05/12/tre-storie-di-italiani-rapiti-convertiti-all-islam-su-cui-nessuno-ha-aperto-bocca-non-erano-donne-2058060.html*

Il Sole 24 Ore (2020, March 14). Mali, liberi Luca Tacchetto e la fidanzata: «Sono fuggiti dai rapitori». Available at: *https://www.ilsole24ore.com/art/liberato-luca-tacchetto-l-architetto-padovano-rapito-burkina-faso-ADFsuKD*

Istat (2018). Le molestie e i ricatti sessuali sul lavoro. Available at: *https://www.istat.it/it/archivio/209107*

La Repubblica (2014, April 18). Paragonò la Kyenge a una scimmia: due mesi all'ex assessore leghista bresciano. Available at: *https://milano.repubblica.it/cronaca/2014/04/18/news/paragon_la_kyenge_a_un_orango_due_mesi_all_ex_assessore_leghista_bresciano-83936467/*

La Repubblica (2019, November 7). Liliana Segre sotto scorta, dopo le minacce assegnata la tutela alla senatrice a vita. Il Centro Wiesenthal: "Una vergogna per l'Italia". Available at: *https://www.repubblica.it/cronaca/2019/11/07/news/scorta_segre-240463251/*

La Repubblica (2020, May 2). Giovanna Botteri, body shaming in tv: "Modelli stupidi e anacronistici che non hanno più ragione di esistere". Available at: *https://www.repubblica.it/cronaca/2020/05/02/news/giovanna_botteri_striscia_la_notizia-255505796/*

Libero (2020, August 26). Matteo Salvini contro Faraone: "La vergogna della sinistra anti-italiana". Il renziano denuncia Musumeci, spunta questa foto. Available at: *https://www.liberoquotidiano.it/news/politica/24320459/matteo-salvini-davide-faraone-denuncia-musumeci-foto-carola-rackete-vergogna-sinistra-anti-italiana.html*

Odiare ti costa. Available at: *https://www.odiareticosta.it/*

Python, random. Available at: *https://docs.python.org/3/library/random.html*

Scikit-learn, sklearn.metrics.cohen_kappa_score. Available at: *https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html*

Statsmodels, statsmodels.stats.inter_rater.fleiss_kappa. Available at: *https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html*

Towards Data Science (2018). Overview of Text Similarity Metrics in Python. Available at: *https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50*

Treccani, “hate speech” in Vocabolario. Available at: *https://www.treccani.it/vocabolario/hate-speech_res-2f344fce-89c5-11e8-a7cb-00271042e8d9_(Neologismi)*

Treccani, “islamofobìa” in Vocabolario. Available at: *https://www.treccani.it/vocabolario/islamofobia*

Treccani, “razzismo” in Vocabolario. Available at: *https://www.treccani.it/vocabolario/razzismo/*

Twitter, Standard Search API. Available at: *https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/ intro-to-tweet-json*

Vox – Osservatorio italiano sui diritti (2019). La mappa dell'intolleranza 4.0. Available at: *http://www.voxdiritti.it/wp-content/uploads//2019/06/190610_VOX-Comunicato-mappa-2019_-completo-compresso.pdf*

Wikipedia, Carola Rackete. Available at: *https://it.wikipedia.org/wiki/Carola_Rackete*

Wikipedia, Cathy La Torre. Available at: *https://it.wikipedia.org/wiki/Cathy_La_Torre*

Wikipedia, Elodie (cantante). Available at: *https://it.wikipedia.org/wiki/Elodie_(cantante)*

Wikipedia, Laura Boldrini. Available at: *https://it.wikipedia.org/wiki/Laura_Boldrini*

Wikipedia, Liliana Segre. Available at: *https://it.wikipedia.org/wiki/Liliana_Segre*

Workshop on online abuse, Policy | Reporting examples of abusive content. Available at: *https://www.workshopononlineabuse.com/resources-and-policies/reporting-examples*

Zollo, F. (2018). Where do misinformation come from? TEDxBari. Available at: *https://www.ted.com/ talks/fabiana_zollo_where_do_misinformation_come_from*

DICHIARAZIONE DI ORIGINALITÀ

Dichiaro di essere responsabile del contenuto dell'elaborato che presento al fine del conseguimento del titolo, di non avere plagiato in tutto o in parte il lavoro prodotto da altri e di aver citato le fonti originali in modo congruente alle normative vigenti in materia di plagio e di diritto d'autore. Sono inoltre consapevole che nel caso la mia dichiarazione risultasse mendace, potrei incorrere nelle sanzioni previste dalla legge e la mia ammissione alla prova finale potrebbe essere negata.

This page intentionally left blank