Lingue e Linguaggi
Lingue Linguaggi 59 (2023), 413-425
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v59p413
http://siba-ese.unisalento.it, © 2023 Università del Salento

DISCORSO D'ODIO E LESSICO CONNOTATO
Un'applicazione del modello VAD al corpus HaSpeeDe

MARIA PAOLA TENCHINI, ALDO FRIGERIO
UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Abstract – The lexicon of natural languages includes both connoted and neutral terms. Connoted terms express the speaker's attitude towards the referent of the term. By contrast, neutral terms do not express any such attitude. Connotation can be positive or negative. Hate speech (HS) is understood as any message that expresses contempt or hatred towards an individual or a target group. Hence, a quite natural hypothesis would be that HS contains a high number of negatively connoted terms. Our work aims at verifying this hypothesis. To do this, we use the model developed by Montefinese *et al.* (2014), which classifies the affective connotation of 1121 Italian words based on three different parameters: valence, arousal, and dominance. We calculated the mean value of these three dimensions in an already annotated Italian HS corpus (HaSpeeDe 2020). The result is quite unexpected as there seems not to exist any meaningful correlation between HS and negatively connoted terms. Not only negatively connoted terms are not necessary to classify a message as HS, but they are not sufficient either. Consequently, HS detection software must take other dimensions into account.

Keywords: hate speech; connoted terms; valence; arousal; dominance.

1. Introduzione

Discorso d'odio, una delle possibili traduzioni dell'espressione inglese *hate speech* (HS), è una locuzione che identifica diverse forme di enunciazioni ostili che veicolano disprezzo, denigrazione o odio nei confronti di un soggetto o di un gruppo target e che possono servire a perpetrare atti di aggressione, propaganda, subordinazione (Bianchi 2017, 2021). Non esiste a tutt'oggi una definizione univoca condivisa, nemmeno in ambito giuridico (Faloppa 2020; Galli 2020). L'HS è stato primariamente oggetto di studi giuridici e linguistici (linguistica, pragmatica e filosofia del linguaggio)[1]. In tempi recenti, in seguito alla diffusione sempre maggiore dei social network, sta ricevendo una particolare attenzione l'aspetto della manifestazione digitale dell'hate speech (cfr., per esempio, Orrù 2020; Palermo 2020; Ziccardi 2016); in quest'ambito, la linguistica computazionale, con l'ausilio della linguistica dei corpora e della psicolinguistica, si occupa del fenomeno con l'intento specifico di sviluppare strumenti atti all'*hate speech detection*[2].

[1] La bibliografia è molto ampia e in continua crescita. Per ragioni di spazio ci limitiamo a citare solo alcuni lavori recenti che offrono una panoramica generale, anche in prospettiva pluridisciplinare, sull'argomento: Meibauer 2013, Finkbeiner *et al.* 2016, Bonacchi 2017, Faloppa 2020, Petrilli 2020 e Bianchi 2021. A livello internazionale la ricerca si è occupata principalmente di slur (cfr. nota 10) e anche in questo caso la bibliografia è vastissima: in Sosa 2018 si possono trovare contributi rappresentativi di diversi approcci teorici circa lo statuto di questi termini. In Italia è fiorente la ricerca sugli insulti, tema al quale è, per esempio, dedicato il dossier di *Quaderns d'Italià* 2020 (vol. 25: https://revistes.uab.cat/quadernsitalia/issue/view/v25).

[2] Ad alcuni lavori in questo settore si fa riferimento nei prossimi paragrafi.

In questo contributo, non ci proponiamo di definire che cosa sia l'HS, né di analizzarne le motivazioni o le modalità di diffusione, né, tanto meno, di avanzare rimedi a questo fenomeno. Il nostro obiettivo è molto più circoscritto: vogliamo dimostrare che, contrariamente a quanto ci si potrebbe aspettare, l'HS non è correlato positivamente con la presenza di termini connotati negativamente. Ciò significa che *in media* l'HS non contiene un numero maggiore di termini connotati negativamente rispetto al discorso non-HS. Mentre è ampiamente noto che un discorso d'odio può non contenere alcun termine connotato negativamente, ci si aspetterebbe che, almeno in media, i discorsi HS contengano più termini connotati rispetto ai discorsi non-HS. Non sembra tuttavia che le cose stiano in questi termini, almeno nei limiti dello studio che qui proponiamo. Questo ha delle importanti conseguenze per lo sviluppo di software volti alla rilevazione dell'HS. Se i risultati di questo studio dovessero essere confermati, allora la rilevazione di parole connotate negativamente non solo non dovrebbe essere l'unico fattore su cui questi software devono fare leva (cosa già riconosciuta), ma probabilmente neppure uno dei fattori rilevanti.

Questo articolo è strutturato nel modo seguente. Nel secondo paragrafo verrà illustrata la differenza fra termini connotati e termini neutri. Per "misurare" il grado di connotazione dei termini ci riferiremo al modello *VAD* applicato al lessico dell'italiano da Montefinese *et al.* (2014). Nel par. 3 illustreremo quella che sembra essere l'ipotesi di base più ragionevole: l'HS contiene in media un maggior numero di termini connotati negativamente rispetto al discorso non-HS. Nel par. 4 presenteremo la metodologia e i risultati del nostro studio. Prenderemo in considerazione il dataset *HaSpeeDe* che contiene discorsi già annotati come HS e non-HS e mediante l'uso del modello *VAD* mostreremo che i discorsi HS non sono mediamente maggiormente connotati quanto al lessico rispetto ai discorsi non-HS. Nel par. 5 discuteremo i risultati ottenuti e trarremo qualche conclusione riguardo ai software di riconoscimento automatico del discorso HS. Il par. 6 contiene le conclusioni.

2. Termini neutri e termini connotati

Esistono nel lessico delle lingue termini che possiamo chiamare *connotati*[3]. Un termine connotato è un termine che esprime un atteggiamento da parte di chi lo usa nei confronti del referente del termine stesso. I connotati si contrappongono ai termini neutri che invece non esprimono alcun atteggiamento nei confronti di tale referente. Ad esempio:

(1) Maria è uscita con la sua barca
(2) Maria è uscita con la sua bagnarola

"Barca" e "bagnarola" (in alcuni dei loro sensi) sono rispettivamente un termine neutro e un termine connotato: "bagnarola", al contrario di "barca", esprime un atteggiamento di scherzoso disprezzo nei confronti della barca di Maria.

[3] Per una introduzione alla connotazione, cfr. Kerbrat-Orecchioni (1977).

La connotazione spesso (ma non necessariamente[4]) è una componente di significato che si *aggiunge* alla componente neutra. Per esempio "la sua barca" e "la sua bagnarola" si riferiscono in (1) e (2) entrambi alla barca di Maria. Tuttavia, "bagnarola" fa qualcosa che "barca" non fa, cioè esprime l'atteggiamento del parlante. Possiamo quindi dire che molti termini connotati hanno due componenti di significato: una componente neutra, che serve a individuare il referente o i referenti, e una componente connotativa.

Mediante un termine connotato è possibile esprimere vari tipi di atteggiamento nei confronti dei referenti di un termine. Alcuni termini esprimono un atteggiamento positivo o di rispetto del parlante nei confronti del referente ("egregio", "onorevole"). Altri termini esprimono affetto ("amore", "tesoro", ecc.). I peggiorativi invece sono quei termini connotati mediante i quali il parlante esprime disprezzo o addirittura odio nei confronti del referente. Per esempio:

(3) Wolfgang è tedesco
(4) Wolfgang è un crucco

"Tedesco" e "crucco" hanno una componente neutra in comune: entrambi si riferiscono a una classe di persone che viene individuata per il fatto di avere una certa nazionalità. Tuttavia, "crucco" è un termine connotato negativamente perché esprime un atteggiamento di lieve disprezzo nei confronti dei tedeschi (e quindi anche nei confronti di Wolfgang). La forza dell'atteggiamento varia da termine a termine (si confronti, per esempio, "crucco" con "negro") e anche da epoca storica a epoca storica. La forza dell'atteggiamento che un termine esprime può infatti variare nel tempo e al limite un termine che era connotato negativamente può diventare neutro (è quello che è successo ad esempio al lessema "gay"). Per contro, un termine neutro può diventare peggiorativo: per esempio, l'espressione "handicappato", per evitare la connotazione moderatamente negativa che si è nel tempo associata a tale termine, è stata sostituita prima da "disabile", poi da "diversamente abile" e infine da "persona con disabilità"[5].

Ci sono diverse possibilità di classificare e "misurare" la componente connotata delle parole. Qui ci rifaremo al modello di Montefinese *et al*. (2014), che a sua volta si rifà a Bradley e Lang (1999). Questo modello si basa sulle *emozioni* che le parole suscitano nei parlanti. Avendo infatti a che fare con vari tipi di atteggiamento, la componente connotata è naturalmente collegata alla dimensione emozionale: tramite la componente connotata, infatti, un parlante esprime le emozioni che qualcosa o qualcuno suscita in lui/lei e inoltre l'espressione di tali emozioni ne suscita altre nei suoi ascoltatori. Il modello di Montefinese *et al*. (2014) classifica i termini connotati secondo tre dimensioni (che costituiscono il cosiddetto *VAD model*)[6].

[4] Esistono infatti termini connotati nei quali l'unica componente di significato è quella espressiva e che quindi sono privi di referenti: ad esempio, "accidenti!", "mamma mia!", alcuni termini volgari, ecc. In questa sede non ne tratteremo.

[5] Ci sono ovviamente delle differenze individuali: alcuni parlanti possono percepire come molto peggiorativo un termine che altri parlanti percepiscono come poco peggiorativo. Tuttavia, ci sono dei limiti alla variazione individuale. Se qualcuno usa un termine come "negro", giustificandosi mediante il fatto che non "sente" quel temine come negativo, lo si può accusare di non conoscere (o di far finta di non conoscere) il significato della parola che sta utilizzando. In altri termini, nei peggiorativi la componente connotativa è convenzionalizzata e sottoposta in modo limitato alla variazione individuale. Essa, infatti, è indipendente dai possibili stereotipi associati a quel termine (su questo punto si vedano per esempio Jeshion (2013, p. 322) e Vallée (2014, p. 86)).

[6] Per un apprendimento del modello VAD e delle scale utilizzate, rimandiamo al par. 4.

La prima dimensione è quella della *valenza* (*valence*, in inglese). Questa dimensione può variare tra le dimensioni del piacere e del dispiacere. Come abbiamo detto, i termini connotati possono esprimere atteggiamenti sia positivi sia negativi. La valenza indica la polarità del termine, cioè quanto le emozioni che esso esprime e suscita siano emozioni positive o negative. Per esempio, nell'indagine di Montefinese "amato" viene indicato come avente una valenza molto positiva (8,37 in una scala da 1 a 9), mentre "disgustato" una valenza negativa (2,09).

La seconda dimensione è quella della *carica emozionale* (*arousal,* in inglese). Essa codifica la forza dell'emozione che il termine esprime e suscita. Quindi, mentre i termini neutri producono il grado più basso di carica emozionale, quanto più un termine esprime un atteggiamento forte e marcato, quanto più produce eccitazione. È chiaro che questa dimensione è collegata alla forza della componente connotata del termine: quanto più tale forza è marcata, tanto più il termine è atto a suscitare emozioni e tanto più forte sarà probabilmente la reazione emotiva negli ascoltatori. Per esempio, "divano" viene giudicato avere un basso grado di carica emozionale (1,9 su una scala da 1 a 9), mentre "litigio" un alto grado (7,19).

La terza dimensione è quella della *dominanza* (*dominance*, in inglese). Tale dimensione varia dal "fuori controllo" al "completo controllo". Per esempio, "affogare" viene giudicato avere una bassa dominanza o basso controllo (3,15 su una scala che va da 1 a 9), mentre "vittoria" un alto grado di dominanza o controllo (8,15). Sulla nozione di dominanza esistono in letteratura due posizioni diverse, di fatto tra loro speculari: l'una interpreta la dominanza come il grado di controllo che un individuo sente di avere su uno stimolo o una situazione[7], l'altra la interpreta come controllo esercitato dalla parola-stimolo sull'individuo[8]. Nel nostro lavoro viene intesa come controllo sullo stimolo/situazione adottando l'interpretazione del modello qui applicato. Ci si potrebbe chiedere se, o fino a che punto, questa dimensione sia davvero distinta dalla seconda, dato che da alcuni studiosi viene esclusa dall'analisi, per cui il modello passa da *VAD* a *VA* (Buechel, Hahn 2017, p. 578)[9]. Si potrebbe infatti pensare che quanto più sia alta la carica emozionale quanto meno il soggetto si senta in controllo. Tuttavia, si è potuto appurare che i termini che sono connotati positivamente, anche se hanno una carica emozionale molto alta, fanno sentire il soggetto meno fuori controllo di quanto non facciano i termini negativi. Inoltre, anche tra i termini negativi, non c'è una uniformità nella dominanza: i termini che tendono a suscitare emozioni di ansia, tristezza, imbarazzo tendono a far sentire i soggetti esposti a tali termini più fuori controllo rispetto a quelli che suscitano emozioni di odio, disgusto e disprezzo (Montefinese *et al*. 2014, pp. 896-898). Pertanto, Montefinese *et al*. (2014) concludono evidenziando che "the exploration of emotional meaning of affective words by means of simple two-dimensional models, such as the valence-versus-arousal model, may fail to account for important sources of variation in the emotion domain. In other words, we suggest that all of the three affective dimensions are

[7] Si vedano ad esempio, Bradley e Lang (1994, 1999), Montefinese *et al.* (2014), Buechel e Hahn (2017), Scott *et al.* (2019). Tuttavia, Bradley e Lang (1994: 57), oltre a ricordare che questa dimensione non gode di una denominazione univoca, riconoscono la possibilità di una doppia lettura di questo fattore, in quanto relazionale: "[s]ince this rating is inherently relational, dominance judgments will clearly need to specify which member of the interaction [the perceiver and the perceived] is being judged…".

[8] Cfr. Warriner *et al.* (2013). Essi la definiscono come "[...] the dominance/power of the word - the extent to which the word denotes something that is weak/submissive or strong/dominant" (p. 1192).

[9] In una tabella comparativa in prospettiva interlinguistica, Montefinese *et al*. (2014, p. 895) su dodici modelli presi in considerazione ne annoverano sette che non presentano la dimensione della dominanza.

needed to adequately represent the subtle similarities and differences in the affective information of emotion words" (pp. 897-898)[10].

Il modello VAD presuppone che la connotazione dei termini non sia una questione di sì e di no, ma sia una dimensione scalare dipendente da tre dimensioni anch'esse scalari. Da un certo punto di vista anche i termini cosiddetti neutri possiedono una dimensione emozionale, ancorché molto bassa.

3. Discorso d'odio e termini connotati

Nell'HS l'individuo o la classe target vengono sminuiti, aggrediti, insultati. È quindi abbastanza naturale ipotizzare che l'HS contenga un alto numero di termini connotati negativamente. In particolare, sarebbe naturale aspettarsi un lessico che abbia mediamente bassi valori di *valence*. La classe target dovrebbe essere indicata tramite slur[11] o comunque tramite termini peggiorativi e, in generale, il discorso d'odio dovrebbe contenere termini indicanti un atteggiamento negativo da parte del parlante. Ci si aspetterebbe, inoltre, che i termini utilizzati abbiano mediamente una *arousal* alta perché chi produce discorsi d'odio dovrebbe utilizzare termini dalla carica emozionale alta. Infine, ci si aspetterebbe che il lessico utilizzato da chi produce discorsi d'odio abbia un valore medio di *dominance* alto derivante dal senso di superiorità o di controllo che chi odia ha sui suoi target.

Il discorso non-HS dovrebbe presentare invece valori medi più alti di *valence*, dato che esso contiene molti termini e testi "neutri", cioè né emozionalmente positivi né emozionalmente negativi. Esso inoltre dovrebbe avere una *arousal* mediamente più bassa perché, di nuovo, esso contiene molti termini e discorsi non carichi dal punto di vista emozionale. Infine, ci si aspetterebbe un valore di *dominance* mediamente più basso per i discorsi non-HS perché esso dovrebbe contenere molti termini e discorsi che presentano un livello di controllo basso del soggetto nei confronti dello stimolo[12].

È noto che l'HS non può essere individuato solo tramite parametri lessicali (sull'italiano cfr., per esempio, Femia 2020; Ferrini, Paris 2019; Palermo 2020; Petrilli 2020; Vedovelli 2020). In altri termini, il fatto che un discorso non contenga peggiorativi non può essere considerato sufficiente per classificare quel discorso come non-HS.

[10] Questa posizione era già stata sostenuta con forza da Russell e Mehrabian (1977, p. 279) che giudicano "necessario" l'apporto della dimensione della dominanza in quanto "[o]nly dominance makes it possible to distinguish angry from anxious, alert from surprised, relaxed from protected, and disdainful from impotent [the first word in each pair involves dominance; the second involves submissiveness]" (Russell, Mehrabian 1977, p. 292).

[11] Gli slur, in italiano reso con 'epiteti denigratori', sono termini fortemente connotati che denotano una classe di individui accomunati da razza, provenienza geografica, nazionalità, sesso, orientamento sessuale, stato psico-fisico, professione, ecc. La particolarità di questi termini è che essi esprimono odio, denigrazione o disprezzo nei confronti sia del denotato sia dell'intero gruppo target di cui il denotato fa parte *proprio* in virtù delle caratteristiche condivise. Per esempio, il termine "negro" denota gli individui aventi un certo colore della pelle ed esprime disprezzo e/o odio nei loro confronti proprio perché hanno quella caratteristica fisica.

[12] O quanto meno questa ci sembra la previsione più plausibile. Alternativamente, si potrebbe pensare che le emozioni di odio e di disprezzo provate dai soggetti produttori di discorsi HS determinino un valore di dominanza basso perché tali emozioni fanno sentire il soggetto fuori controllo. Le diverse interpretazioni date alla dimensione della dominanza in letteratura non ci permettono discernere quale sia la previsione più appropriata. Comunque stiano le cose, come vedremo, nessuna di queste ipotesi trova conferma nei nostri risultati.

Possiamo infatti convogliare disprezzo o odio nei confronti di un soggetto o di un gruppo target anche senza utilizzare termini connotati negativamente. Ad esempio:

> *C'è qualche #immigrato che mi può fare da prestanome per avere un alloggio e aiutare un amico che purtroppo è italiano come me? #lagabbia*
>
> *E basta con sta storia che questi migranti scappano dalle guerre! La Spagna è distante solo km ma stranamente arrivano i* [sic] *italia #Naufragio*
>
> *Questi immigrati sono il futuro per l'Italia!!! Bella roba.....*

Di per sé questi tweet tratti dal dataset di *HaSpeeDe* non contengono parole connotate ma solo lessico neutro, eppure sono classificabili come HS.

Tuttavia, si potrebbe affermare che, se l'assenza di termini connotati negativamente non è una condizione sufficiente per classificare il discorso come non-HS, almeno la loro presenza dovrebbe essere una condizione sufficiente per classificare il discorso come HS, o comunque tale presenza dovrebbe essere un indizio che si tratti di HS[13].

In questo articolo vogliamo dunque verificare questa ipotesi, ossia se la presenza di termini a valenza negativa, con un grado alto di eccitazione e alto di dominanza sia correlata positivamente con l'HS. Per verificare questa ipotesi ci avvarremo di un corpus in cui i discorsi sono già classificati come HS e non-HS e accerteremo nei due tipi di discorso il grado di presenza dei termini classificati come peggiorativi secondo le tre dimensioni che abbiamo citato. Come anticipato, il risultato sarà molto interessante: non sembra esserci *alcuna* correlazione significativa tra HS e termini connotati in modo negativo. Non solo i termini connotati negativamente non sono necessari per la classificazione di un discorso come HS, essi non sono neppure sufficienti.

4. Metodologia e risultati

Per verificare la nostra ipotesi abbiamo utilizzato il dataset di *HaSpeeDe* 2020 (Sanguinetti *et al*. 2020)[14]. Tale corpus contiene 8102 Tweet che fanno riferimento a minoranze, in particolare a immigrati, Musulmani e Rom.

Ogni tweet è stato annotato secondo le seguenti dimensioni:

[13] Chiaramente le due ipotesi sono indipendenti. Potrebbe infatti darsi che esistano discorsi HS privi di termini connotati negativamente e che, ciononostante, *in media* i discorsi HS presentino un maggior numero di termini connotati rispetto ai discorsi non-HS. Potrebbe darsi anche, sebbene si tratti di una ipotesi alquanto bizzarra, che non esistano discorsi HS privi di termini connotati e che ciononostante i discorsi HS non presentino *in media* termini maggiormente connotati dei discorsi non-HS. Mentre in letteratura la tesi secondo cui esistono discorsi HS privi di termini connotati negativamente è stata ampiamente sostenuta, a nostra conoscenza la tesi che i discorsi HS non contengono in media più termini connotati negativamente dei discorsi non-HS è originale. Ringraziamo un referee anonimo per averci sollecitato a chiarire questo punto.

[14] https://github.com/msang/haspeede. Questo corpus costituisce una versione aggiornata di IHSC 2018 (https://github.com/msang/hate-speech-corpus, cfr. Bosco *et al*. 2018; Sanguinetti *et al*. 2018), integrata da un sottoinsieme di tweet tratti dal progetto "Contro l'odio" (Capozzi *et al*. 2019, 2020).

i. hate speech[15]: no – sì
ii. aggressività: no - debole - forte
iii. offensività: no - debole - forte
iv. ironia: no - sì
v. stereotipo: no - sì
vi. intensità: 0 - 1 - 2 - 3 - 4, dove 0 sta per intensità nulla e 4 per intensità massima

Ogni tweet è stato annotato da due annotatori. Quando i giudizi dei due annotatori erano discordanti, un terzo annotatore indipendente forniva un parere.

Dal corpus sono state filtrate le *stopwords*[16] utilizzando la lista di *stopwords* per l'italiano del *Natural Language Toolkit*[17] ed è stato estratto il vocabolario per le categorie di testo HS e non-HS. Si è deciso di escludere i token meno frequenti e di lemmatizzare solo quelli che presentano almeno 10 occorrenze. A questo punto, i due sotto-corpora – i tweet classificati come HS (3388 tweet) e quelli classificati come non-HS (4714 tweet) – sono stati integrati con l'adattamento del corpus *Affective Norms for English Words (ANEW*, cfr. Bradley, Lang 1999*)* per l'italiano (Montefinese *et al*. 2014). Il corpus fornisce per 1121 parole italiane i valori delle tre dimensioni che abbiamo introdotto nel paragrafo precedente (*valence*, *arousal*, *dominance*), più altri indici soggettivi (familiarità, immaginabilità e concretezza) e oggettivi (lunghezza e frequenza di parola, grandezza del vicinato ortografico, frequenza media del vicinato ortografico). Ai fini di questo lavoro abbiamo preso in considerazione solo i valori di *valence*, *arousal* e *dominance*, perché strettamente connessi alla dimensione connotata ed emozionale del lessico.

I valori *valence*, *arousal* e *dominance* sono stati ricavati da Montefinese *et al*. (2014) chiedendo a 1084 studenti di classificare le 1121 parole secondo queste tre dimensioni. Le scale vanno da 1 a 9. Per quanto concerne la *valence,* 1 indica una emozione molto spiacevole e 9 molto piacevole. In riferimento all'*arousal*, cioè il grado di carica emozionale o di attivazione energetica che un individuo prova in presenza di uno stimolo, 1 corrisponde a molto calmo e 9 a molto eccitato. Infine, nel caso della *dominance,* il punteggio 1 indica un basso controllo del soggetto nei confronti dello stimolo, il punteggio 9 un alto controllo.

Abbiamo quindi assegnato a ogni occorrenza di un lemma appartenente al corpus *HaSpeeDe* un punteggio di *valence*, *arousal* e *dominance*, quando tale lemma fa parte del database dell'adattamento italiano di *ANEW* di Montefinese *et al*. (2014). I lemmi non appartenenti a tale database sono stati esclusi. Abbiamo quindi calcolato le medie di tali valori.

I risultati ottenuti sono i seguenti (Tabella 1):

[15] Come riportato dagli autori, affinché un tweet possa essere annotato come HS devono sussistere contemporaneamente due aspetti: (a) il target deve coincidere con uno dei tre gruppi minoritari oggetto della ricerca (immigrati, Musulmani, Rom), o comunque con un individuo discriminato per caratteristiche riconducibili al proprio gruppo di appartenenza (sono quindi escluse le caratteristiche individuali); (b) la forza illocutoria dell'enunciato contenuto del tweet deve incitare, promuovere, diffondere o giustificare condotte violente nei confronti del gruppo target, eventualmente deumanizzando, intimidendo o delegittimando gli individui che ne fanno parte.

[16] Ovvero, parole molto comuni nel lessico di una lingua e ritenute poco significative per gli obiettivi di questo lavoro, per esempio gli articoli, le preposizioni e gli ausiliari.

[17] https://www.nltk.org/.

	***HaSpeeDe,* HS**	***HaSpeeDe, non*-HS**
Valenza	5,27	5,48
Arousal	5,93	5,95
Dominance	5,27	5,34

Tabella 1.

Come ulteriore evidenza, di seguito sono presentati i 10 sostantivi più frequenti per ogni sotto-corpus (Tabella 2):

***HaSpeeDe,* HS**	***HaSpeeDe,* non-HS**
immigrato	rom
migrante	migrante
Italia/italia	Roma
rom	immigrato
italiano	Italia
islam	italiano
casa	campo
clandestino	immigrazione
straniero	straniero
terrorista	nomade

Tabella 2.

Di fatto, le parole più frequenti nel lessico HS e non-HS tendono a coincidere. In particolare, 6 dei 10 lemmi più frequenti coincidono.

5. Discussione

La nostra ricerca smentisce le ipotesi formulate nel par. 2 circa le differenze fra i discorsi HS e non-HS. Una delle ipotesi di partenza è che i discorsi HS abbiano una *valence* media più bassa di quelli non-HS. Nel corpus analizzato la *valence* media delle occorrenze dei termini utilizzati nei discorsi HS è di 5,27, mentre quella dei discorsi non-HS è di 5,48. Sebbene il valore dei discorsi HS sia in effetti più basso, i due valori sono molto vicini. Per calcolare se la differenza tra le due medie sia significativa dal punto di vista statistico, abbiamo applicato il test t di Student, pensato proprio a questo scopo. Assumendo un intervallo di confidenza del 95% (cioè $p<0,05$), il test ha dato esito negativo: la differenza tra le due medie non è statisticamente significativa.

Una seconda ipotesi di partenza è che la *arousal* media dei discorsi HS sia più alta

rispetto a quella dei discorsi non-HS. Nel corpus analizzato, la media della *arousal* delle occorrenze dei termini utilizzati nei discorsi HS è 5,95, quella dei discorsi non-HS 5,93. Sebbene i discorsi HS presentino una *arousal* più alta, la differenza tra le due medie non è ancora una volta statisticamente significativa secondo il test t di Student ($p<0,05$).

Infine, per quanto riguarda la *dominance*, la media delle occorrenze dei termini utilizzati nei discorsi non-HS è addirittura superiore a quella dei discorsi HS (5,34 contro 5,27), smentendo l'ipotesi iniziale secondo cui i discorsi HS dovrebbero dimostrare una *dominance* mediamente più alta[18].

Questi risultati hanno conseguenze importanti circa la definizione di un discorso come HS e circa le metodologie per individuare tali discorsi tramite strumenti di apprendimento automatico. Non solo i discorsi HS non sono individuabili tramite il lessico, ma esso sembra avere un'importanza del tutto marginale nella discriminazione tra discorsi HS e discorsi non-HS.

A riprova di ciò abbiamo calcolato il numero di occorrenze di slur espliciti nei discorsi HS. Nel dataset di *HaSpeeDe*, "negro" appare 44 volte in forma plurale, incluso un interessante caso di modificazione in "negroidi", e 1 al singolare; la parola "zingaro", compare 57 volte al plurale e 5 al singolare (di cui 3 volte al femminile: 2 come sostantivo e 1 come aggettivo); la parola "giudeo" compare 2 volte al singolare. Infine, ricorre 8 volte "puttana", tra singolare e plurale, 2 volte "troia" e 1 volta "baldracca". Poiché tra i 10 sostantivi più frequenti non rientrano slur e dato che, talvolta, può succedere che uno stesso tweet contenga la ripetizione di uno slur[19] o l'occorrenza di più di uno degli slur menzionati[20], ciò conferma che la grande maggioranza dei tweet HS non contiene slur espliciti.

Contrariamente a quanto ci si potrebbe aspettare, quindi, in media i termini utilizzati nei discorsi d'odio non hanno una connotazione maggiormente negativa di quelli utilizzati nei discorsi classificati come non-HS. Ne consegue che nello sviluppo di algoritmi che cercano di individuare i discorsi HS, il peso del lessico dovrebbe essere limitato e ci si dovrebbe basare invece su altri fattori[21]. In effetti, si stanno tentando strade alternative. Alcune sono basate su informazioni metatestuali (informazioni sugli utenti, sulla loro provenienza geografica, sul genere, sulla loro frequentazione di certi social network, sulla sequenza temporale dei loro post, ecc.); altre si basano invece su caratteristiche dei post individuando *features* differenti e più complessi rispetto alla semplice occorrenza di certe parole: ad esempio, la presenza di certi digrammi o trigrammi con una certa struttura sintattica, l'assegnazione di un vettore a ogni parola in base alla sua

[18] Anche l'ipotesi alternativa, cioè quella secondo cui i discorsi HS dovrebbero avere un grado di *dominance* mediamente più basso perché prodotti da soggetti che si sentono fuori controllo, è smentita dal test t di Student ($p<0,05$).

[19] Per esempio: *Anche gli zingari li devi chiamare Rom ma io ho sempre continuato a chiamarli zingari Hai capito Fico? sono zingari, zingari, zingari ZINGARIII !!*

[20] Per esempio: *Devi obbligarli ad andare a vivere solo una settimana nelle periferie,in mezzo hai cannibali negri,hai rom,agli zingari,hai delinquenti italiani.Insomma fargli provare con mano cosa voglia dire vivere nel mondo democratico del cazzo.Vedresti come cambierebbero idea e partito*; oppure: *Troia baldracca puttana. Potrei essere anche io favorevole all'immigrazione non controllata con tutti i suoi soldi. Troia*. Si veda l'uso predicativo in: *Padroni in casa nostra padroni a casa nostra via la merda negra islamica e zingara dall'Italia Matteo premier.*

[21] Questo non vuole dire che il riconoscimento dei discorsi HS debba basarsi esclusivamente su fattori pragmatici. Quello che qui sosteniamo è che il ruolo del lessico è marginale. Ciò non esclude che ci siano frasi *semanticamente* classificabili come discorsi HS. Tuttavia, ciò dipenderebbe dalla semantica dell'intera frase e non dei singoli termini contenuti in essa. Ringraziamo un referee anonimo per averci sollecitato a chiarire questo punto.

vicinanza o meno con altre parole, la leggibilità del testo misurata tramite indicatori numerici (come la formula di Flesch o simili), altri indicatori di *sentiment* del post. I programmi vengono addestrati su dataset specifici di post HS e non-HS nei quali questi *features* sono annotati e sulla base di tale addestramento valutano la probabilità che un nuovo post sia HS o non-HS[22]. Sebbene non abbiano una precisione estrema, questi sistemi sembrano abbastanza promettenti.

6. Conclusioni

Questo articolo presenta i risultati di un'indagine volta a verificare l'ipotesi che la presenza di termini a valenza negativa, con un grado alto di carica emozionale e un grado alto di dominanza sia correlata positivamente con l'HS. Queste tre dimensioni sono state ampiamente applicate all'analisi delle emozioni suscitate dal lessico per quanto concerne la lingua inglese, meno per altri idiomi. Per l'italiano si conosce finora lo studio di Montefinese *et al*. 2014. La nostra indagine ha applicato il loro modello a un corpus di testi già classificati come HS e non-HS e ha verificato, nei due tipi di discorso, il grado di presenza dei termini classificati come peggiorativi secondo le tre suddette dimensioni. Il risultato conseguito ha dimostrato che non solo la presenza di termini connotati negativamente non è necessaria per classificazione di un discorso come HS, ma probabilmente neppure sufficiente.

Ciò che la ricerca lascia aperto è se la presenza di *alcuni* termini connotati negativamente (slur, termini insultanti, ecc.) sia correlata positivamente con l'HS. Per verificare questa ipotesi si dovrebbe effettuare una ricerca per certi versi opposta a quella che abbiamo condotto qui: si dovrebbero individuare le frasi in cui tali termini sono presenti e controllare quanto spesso esse siano marcate come HS o come non-HS. Come è noto, infatti, la presenza di tali termini non è un indice automatico di HS poiché potrebbero essere utilizzati in modo ironico, scherzoso o appropriativo[23].

Il nostro lavoro presenta alcuni limiti che ci proponiamo di superare in ricerche future. Anzitutto, sebbene il corpus utilizzato sia uno dei più estesi dell'italiano attualmente disponibili, si tratta tuttavia di un corpus ancora circoscritto, sia per tipologia di classi target (immigrati, Musulmani, ROM), sia per dimensioni. Ricerche future potrebbero prendere in considerazione un numero maggiore di corpora, e poiché i corpora dell'italiano che classificano i discorsi come HS/non-HS non sono al momento numerosi, si potrebbero prendere in considerazione lingue per le quali sono presenti corpora di maggiori dimensioni[24].

Un secondo limite, comune a tutti i lavori che si basano su corpora annotati, è rappresentato dal fatto che le classificazioni dei messaggi come HS e non-HS sono state effettuate da coloro che hanno raccolto il corpus. Ora è noto che esistono diverse

[22] Per alcuni tentativi in questa direzione, cfr. Davidson *et al.* (2017), Del Vigna *et al.* (2017), Joulin *et al.* (2017), Kovács *et al.* (2021), MacAvaney *et al.* (2019), Zimmerman *et al.* (2018).

[23] Ringraziamo un referee anonimo di *Lingue e Linguaggi* per averci fatto riflettere su questo punto.

[24] Va tuttavia segnalato che le dimensioni della *arousal* e della *dominance* sono maggiormente dipendenti da specificità endolinguistiche e culturali rispetto alla *valence* (Montefinese *et al.* 2014, p. 894), per cui l'utilizzo di corpora diversi dall'italiano porterebbe risultati significativi solo per la specifica lingua presa in esame.

definizioni di hate speech, non tutte sovrapponibili tra loro[25]. Di conseguenza, uno stesso post può essere classificato come HS o non-HS sulla base di definizioni differenti. Inoltre, anche assumendo una definizione unitaria, è noto che l'accordo tra i diversi annotatori non è purtroppo alto quando si tratta di hate speech perché la linea di confine tra HS e non-HS è spesso sfumata e sfuggente[26]. Inoltre, i corpora di HS attualmente disponibili si basano soprattutto su Twitter perché le politiche di uso e di distribuzione di questo social network sono le più permissive. Benché i tweet possano fornire ottimi esempi di hate speech, si tratta di un genere testuale particolare dato il numero limitato di caratteri. Altri social network permettono messaggi più lunghi, che possono entrare a far parte di discussioni più ampie su argomenti specifici. Ciò può fornire, tra l'altro, ulteriore contesto che può influenzare il significato dei testi.

Infine, il calcolo della media delle tre dimensioni qui prese in esame si è basato sui 1121 lemmi presenti nel corpus di Montefinese *et al*. (2014). Anche questo limite può essere superato in futuro espandendo questo corpus, oppure adattando per l'italiano corpora di altre lingue.

L'estensione della ricerca potrebbe rendere più solidi i risultati qui ottenuti. Se per esempio i risultati qui ottenuti si replicassero anche per corpora più ampi, questo rafforzerebbe la tesi della marginalità del lessico nella individuazione dei discorsi HS.

Bionote: Maria Paola Tenchini è ricercatrice di Linguistica generale presso l'Università Cattolica del Sacro Cuore (sede di Brescia) dove tiene corsi di Linguistica generale, Lingua e fonologia tedesca e Apprendimento e didattica delle lingue moderne. Tra i suoi principali interessi di ricerca: semantica e pragmatica dei peggiorativi, il discorso riportato, l'ordine delle parole in tedesco, il linguaggio non verbale, temi di storia della linguistica.
Aldo Frigerio è professore associato di Filosofia del linguaggio presso l'Università Cattolica del Sacro Cuore (Milano) dove insegna Semantica e Informatica. I suoi interessi di ricerca vertono sulla semantica dei sintagmi nominali, sulla semantica e pragmatica dei peggiorativi, sulla logica e semantica degli enunciati temporali.

Recapiti autori: paola.tenchini@unicatt.it; aldo.frigerio@unicatt.it

[25] Per alcune definizioni usate nell'ambito dei social network e della rilevazione automatica del discorso HS, cfr. Davidson *et al.* (2017), Fortuna e Nunes (2018), de Gibert *et al.* (2018), nonché le definizioni date da Facebook (https://www.facebook.com/communitystandards/objectionable_content) e Twitter (https://help.twitter.com/en/rules-and-policies /hateful-conduct-policy).

[26] Sul disaccordo tra gli annotatori per quanto riguarda gli HS, cfr. Ross *et al.* (2016).

Riferimenti bibliografici

Bianchi C. 2017, Linguaggio d'odio, autorità e ingiustizia discorsiva, in "Rivista di estetica", 64, pp. 18-34.

Bianchi C. 2021, *Hate speech. Il lato oscuro del linguaggio*, Laterza, Bari-Roma.

Bonacchi S. (a cura di) 2017, *Verbale Aggression. Multidisziplinäre Zugänge zur verletzenden Macht der Sprache*, De Gruyter, Berlin-Boston.

Bradley M.M. and Lang P.J. 1994, *Measuring emotion: The Self-Assessment Manikin and the semantic differential*, in "Journal of Behavior Therapy and Experimental Psychiatry", 25, pp. 49-59.

Bradley M.M. and Lang P.J. 1999, *Affective norms for English words (ANEW): Instruction manual and affective ratings,* in *Technical Report C-1*, The Center for Research in Psychophysiology, University of Florida. https://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf (19/7/2021).

Bosco C., Dell'Orletta F., Poletto F., Sanguinetti F. and Tesconi M. 2018, *Overview of the EVALITA 2018 Hate Speech Detection Task*, in Caselli T., Novielli N., Patti V. and Rossi P. (eds.) *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, CEUR, vol. 2263, http://ceur-ws.org/Vol-2263/paper010.pdf (19/7/2021).

Buechel S. and Hahn U. 2017, *EMOBANK: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis*, in Lapata M., Blunsom P. and Koller A. (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2, Short Papers, ACL, pp. 578–585, 2017. https://aclanthology.org/E17-2092.pdf (19/7/2021).

Capozzi A.T.E., Lai M., Basile V., Poletto F., Sanguinetti M., Bosco C., Patti V., Ruffo G., Musto C., Polignano M., Semeraro G. and Stranisci M. 2019, *Computational Linguistics Against Hate: Hate Speech Detection and Visualization on Social Media in the "Contro L'Odio" Project*, in R. Bernardi, R. Navigli, and G. Semeraro (eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics, CLiC-it 2019*. http://ceur-ws.org/Vol-2481/paper14.pdf (19/7/2021).

Capozzi A.T.E., Lai M., Basile V., Poletto F., Sanguinetti M., Bosco C., Patti V., Ruffo G., Musto C., Polignano M., Semeraro G. and Stranisci M. 2020, *"Contro L'Odio": A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media*, in "Italian Journal of Computational Linguistics", 6[1], pp. 77-97. https://journals.openedition.org/ijcol/659?lang=it (19/7/2021).

Davidson T., Warmsley D., Macy M. W. and Weber I. 2017, *Automated Hate Speech Detection and the Problem of Offensive Language*, in *Proceedings of the eleventh international AAAI conference on web and social media (ICWSM)*, AAAI Press, Palo Alto, California, pp. 512-515. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843 (19/7/2021).

Del Vigna F., Cimino A., Dell'Orletta F., Petrocchi M. and Tesconi M. 2017, *Hate Me, Hate Me Not: Hate Speech Detection on Facebook*, in Armando A., Baldoni R. and Focardi R. (eds.), *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pp. 86-95. http://ceur-ws.org/Vol-1816/paper-09.pdf (19/7/2021).

de Gibert O., Perez N., García-Pablos A. and Cuadros M. 2018, *Hate Speech Dataset from a White Supremacy Forum*, in *2nd Workshop on Abusive Language Online (ALW2)*, Association for Computational Linguistics, pp. 11–20. https://aclanthology.org/W18-5102.pdf (19/7/2021).

Faloppa F. 2020, *#ODIO. Manuale di resistenza alla violenza delle parole*, Utet, Milano.

Femia D. 2020, *Discorso dell'odio e risorse per il trattamento automatico delle lingue. Metodi, ipotesi, proposte*, in Petrilli R. (a cura di), Hate Speech. *L'odio nel discorso pubblico. Politica media e società*, Round Robin, Roma, pp. 147-164.

Ferrini C. e Paris O. 2019, *I discorsi dell'odio. Razzismo e retoriche xenofobe sui social network*, Carocci, Roma.

Finkbeiner R., Meibauer J. and Wiese H. (ed.) 2016, *Pejoration*, John Benjamins, Amsterdam-Philadelphia.

Fortuna P. and Nunes S. 2018, *A Survey on Automatic Detection of Hate Speech in Text*, in "ACM Computing Surveys" 51 [4], art. n. 85, pp.1–30. https://doi.org/10.1145/3232676 (19/7/2021).

Galli M. 2020, *Soltanto parole? Discorsi d'odio e intervento penale*, in Petrilli R. (a cura di), Hate Speech. *L'odio nel discorso pubblico. Politica media e società*, Round Robin, Roma, pp. 23-40.

Jeshion, R. 2013, *Slurs and Stereotype*, in "Analytic Philosophy" 54, pp. 314-329.

Joulin A., Grave E., Bojanowski P. and Mikolov T. 2017, *Bag of Tricks for Efficient Text Classification*, in Lapata M., Blunsom P. and Koller A. (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, Short Papers, ACL, pp. 427-431. https://www.aclweb.org/anthology/E17-2068 (19/7/2021).

Kerbrat-Orecchioni C. 1977, *La connotation*, PUL, Lyon.
Kovács, G., Alonso, P. and Saini, R. 2021, *Challenges of Hate Speech Detection in Social Media. Data Scarcity, and Leveraging External Resources*, in "SN Computer Science", 2, art. n. 95, https://doi.org/10.1007/s42979-021-00457-3 (19/7/2021).
MacAvaney S., Yao H.-R., Yang E., Russell K., Goharian N. and Frieder O. 2019, *Hate speech detection: Challenges and solutions*, in "PLoS ONE", 14(8), e0221152. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152 (19/7/2021).
Meibauer J. (a cura di) 2013, *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, Gießener Elektronische Bibliothek.
Montefinese, M. Ambrosini E., Fairfield B. and Mammarella N. 2014, *The adaptation of the Affective Norms for English Words (ANEW) for Italian,* in "Behavior Research Methods" 46, pp. 887-903.
Orrù P. 2020, "Capra sarai tu e anche maleducato!!": insulto e interazione nei commenti Facebook, in "Quaderns d'Italià", 25, pp. 65-82.
Palermo M. 2020, *L'insulto ai tempi dei social media: costanti e innovazioni*, in "Lingue e Culture dei Media", 4, 2, pp. 2-15.
Petrilli R. 2020, *Il meccanismo dell'odio nel discorso pubblico*, in Petrilli R. (a cura di), Hate Speech. *L'odio nel discorso pubblico. Politica media e società*, Round Robin, Roma, pp. 41-58.
Petrilli R. (a cura di) 2020, *Hate Speech. L'odio nel discorso pubblico. Politica media e società*, Round Robin, Roma.
Ross B., Rist M., Carbonell G., Cabrera B., Kurowsky N. and Wojatzki M. 2016, *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*, in Dipper S. (ed.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (Bochumer Linguistische Arbeitsberichte 16), Ruhr-Universität Bochum, Bochum, pp. 6-9. https://arxiv.org/pdf/1701.08118.pdf (19/7/2021).
Russel J.A. and Mehrabian A. 1977, *Evidence for a Three-Factor Theory of Emotions*, in "Journal of Research in Personality" 11, pp. 273-294.
Sanguinetti M., Poletto F., Bosco C., Patti, V., and Stranisci M. 2018, *An Italian Twitter Corpus of Hate Speech against Immigrants*, in Calzolari N. *et al.* (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2798-2805. https://aclanthology.org/L18-1443.pdf (19/72021).
Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V. e Russo, I. 2020, *HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task*, in Basile V., Croce D., Di Maro M., and Passaro L.C. (eds.), *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, CEUR, vol. 2765. http://ceur-ws.org/Vol-2765/paper162.pdf (19/7/2021).
Scott G.G., Keitel A., Becirspahic M., Yao B., Sereno Sara C. 2019, *The Glasgow Norms: Ratings of 5,500 words on nine scales*, in "Behavior Research Methods" 51, pp. 1258-1270. https://doi.org/10.3758/s13428-018-1099-3 (19/7/2021).
Sosa D. (a cura di) 2018, *Bad Words. Philosophical Perspective on Slurs*, Oxford University Press, Oxford.
Vallée R. 2014, *Slurring and Common Knowledge of Ordinary Language*, in "Journal of Pragmatics" 61, pp. 78-90.
Vedovelli M. 2020, *Il caso "cambia vita": il razzismo comunicativo dall'insulto al messianismo*, in Petrilli R. (ed.), Hate Speech. *L'odio nel discorso pubblico. Politica media e società*, Round Robin Roma, pp. 129-146.
Warriner A.B., Kuperman V. and Brysbaert M. 2013, *Norms of valence, arousal, and dominance for 13,915 English lemmas*, in "Behavior Research Methods" 45, pp. 1191-1207, https://doi.org/10.3758/s13428-012-0314-x (19/7/2021).
Ziccardi, G. 2016. *L'odio online: Violenza verbale e ossessioni in rete*, Raffaello Cortina, Milano.
Zimmerman S., Kruschwitz U. and Fox C. 2018, *Improving Hate Speech Detection with Deep Learning Ensembles*, in Calzolari N. *et al., Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018*, Miyazaki, Japan), European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/pdf/292.pdf (19/7/2021).