UNIVERSITÀ DEGLI STUDI
DI NAPOLI FEDERICO II

TESI DI DOTTORATO DI RICERCA

Dipartimento di Scienze Sociali
Dottorato di ricerca in Scienze Sociali e Statistiche XXXIV Ciclo

HateViz: Parole d'odio contro le donne

Candidato: dott. ROCCO MAZZA

Tutor: prof.ssa MARINA MARINO

Co-Tutor: prof. ADAM E. ARVIDSSON

Sommario

1. Introduzione

Quando si discute di violenza di genere, in particolare mi riferisco a quella perpetrata contro le donne, i media broadcasting così come il singolo individuo coinvolto in una discussione fanno spesso riferimento ai cosiddetti *episodi di violenza*. Mai espressione è stata più fuorviante. La natura insita nelle forme di *hatespeech* può aiutare a comprendere tale errore. Se un fatto ritenuto episodico si ripete a diverse latitudini nel corso del tempo, la sua natura da episodica diventa sistematica e rivela un contesto necessario di attenzione. Organizzazioni, istituzioni e oramai molte comunità scientifiche ritengono l'espressione d'odio come una declinazione della violenza. Cambiano le modalità in cui questa viene eseguita ma i danni che arreca all'individuo bersaglio sono tristemente gli stessi e ben noti. Pensiamo quindi alle (oramai ben rinomate) potenzialità comunicative delle contemporanee tecnologie presenti sul web e alla facilità con cui i messaggi, attraverso queste tecnologie, possono amplificarsi. È facile ora pensare a questi messaggi caricati di un odio sistemico, culturale. La reazione di chi viene raggiunto da quei messaggi e si sente vittima del loro odio non può che generare una profonda lacerazione all'interno della comunità e dell'individuo stesso. Questo dà vita ad ansia, paura, misconoscimento nella persona bersagliata. Alla luce di un'estensione concettuale del fenomeno così importante, come possiamo ancora ritenere la violenza di genere come un *evento episodico*? Come possiamo non avvertire un sostrato culturale che permea questi atti divenuti oramai sistemici? In estrema sintesi potrebbe essere così descritta la trasformazione accaduta alla violenza di genere: trascurata da diverso tempo, relegata a realtà ritenute isolate e specchio, si credeva -erroneamente- un disagio unicamente individuale.

Il framework culturale della tesi delimita questo spazio. Se l'agire sistemico si fa carico della degenerazione valoriale e insieme arrivano a caricare le espressioni di violenza

allora esistono dei punti in comune tra queste dimensioni, ossia l'odio. Si è sempre pensato alla violenza contro le donne come un fenomeno definito nell'intimità della coppia. Compreso tra mura, relazioni strette e private. E se le matrici valoriali alla base delle pratiche violente fossero comuni? E se le nuove tecnologie avessero lacerato il muro che divide il privato dal pubblico? Le pratiche e gli esercizi della violenza sono molti, io faccio riferimento agli universi valoriali della violenza nel privato, quella che in letteratura viene definita come *Intimate Partner Violence* (IPV), e la violenza nel pubblico, esercitata tramite discorsi di odio, anche detti *hatespeech*. Per comprendere la violenza privata abbiamo l'esigenza di creare una mappa valoriale che fornisca un quadro delle pratiche attraverso cui questa si esercita. Per isolare l'hatespeech serve invece intercettare i messaggi d'odio dai flussi comunicativi della rete. Con queste considerazioni siamo arrivati a quello che è l'obiettivo della tesi. Proporre una strategia che permetta di individuare questi elementi e collocarli all'interno delle comunità di riferimento. Una strategia che possa essere versatile alle molte sfaccettature dei codici presenti sul web. La mia proposta metodologica lavora sulle dimensioni latenti presenti all'interno delle distribuzioni lessicali che descrivono i testi individuati online. L'intento è intercettare un contenuto non espresso esplicitamente.

La domanda metodologica che segue il mio percorso si riferisce proprio a questi aspetti, come posso individuare e studiare i contenuti di odio? Questa domanda non è di certo nuova. Molti sono i contributi presenti in letteratura che tentano di darle risposta. L'innovazione nel mio contributo sta in una possibile soluzione che integri tra loro vari approcci. La direzione presa dal mio lavoro cerca di individuare gli elementi estratti dal flusso comunicativo esplorando quelle che ho già presentato come latenze di significato. In particolare si tratta di dimensioni semantiche non direttamente espresse dai creatori dei contenuti messi online. Queste aiutano a classificare gli stessi contenuti in base alla loro carica emotiva. Le emozioni fanno da cornice ai gruppi semantici emergenti. La mia

proposta metodologica consiste in una concatenazione di tecniche statistiche che costituiscono una filiera di analisi. Si può facilmente immaginare il processo come una sequenza in cui ogni risultato prodotto viene processato nello step di analisi successivo. L'approccio prevalente nella tesi fa sicuramente riferimento allo studio dei dati multidimensionali. Il problema delle dimensioni latenti sottostanti alle tabelle in cui per ogni riga sono raccolte ed esplorate contemporaneamente *p* dimensioni è uno dei cardini conoscitivi della statistica contemporanea. Lo sfruttamento da parte del ricercatore del metodo statistico per esplorare ed estrarre informazioni da queste collezioni di dati è una sfida che si avverte nella letteratura scientifica

Questa tesi rappresenta il punto di arrivo di un progetto iniziato circa tre anni fa. Ho deciso di dedicare il mio lavoro di ricerca e di studio ad un tema che avesse un forte impatto sociale. Il progetto prende il nome di *HateViz* e mi ha permesso non solo di approfondire e lavorare su questi temi ma anche di incontrare molte realtà scientifiche diverse dove ho avuto modo di confrontarmi e migliorare. L'intento è stato offrire il mio contributo ad un dibattito acceso sulle degenerazioni di una frattura sociale alimentata dalla cultura dell'odio. Quella di dedicarmi alle forme di violenza di genere, per la precisione alle espressioni di odio online, mi ha permesso di crescere come ricercatore e umanamente sviluppare una nuova sensibilità verso questi temi. Questo sentimento mi ha fatto in più momenti riflettere sulle responsabilità della ricerca sociale e statistica di offrire contributi socialmente sostenibili.

Con il dottorato ho deciso di sviluppare un percorso multidisciplinare, l'obiettivo è stato attingere a più conoscenze per generare un prodotto che le integrasse. Con questo mi riferisco alle due anime che caratterizzano il lavoro presentato. Il contributo delle scienze sociali, ovvero della sociologia, è stato fornire un quadro concettuale di riferimento. La disciplina sociale mi ha permesso di delimitare il campo teorico di applicazione della mia tesi e da questo le dimensioni concettuali in cui muovermi. La

stessa mi ha fornito gli strumenti logico-metodologici che mi hanno permesso di rendere operativi tali concetti individuandone le pratiche ad essi associati. Con l'elaborazione di queste informazioni ho individuato il problema sociologico di base, operativizzandolo nelle domande di ricerca poste nella fase iniziale della pianificazione della mia proposta metodologica. Queste domande hanno guidato lo sviluppo della proposta statistica fornendo delle linee guida da seguire. L'obiettivo è stato quello di fornire uno strumento in grado di rispondere a queste domande nei vari contesti di applicazione. La statistica ha offerto un approccio in grado di aiutare a costruire questo strumento. La concatenazione tra metodi da me proposta rappresenta il risultato di uno studio molto ampio che ha coinvolto molti saperi statistici. La prospettiva adottata è definita in letteratura *applicata*, ho avuto modo studiare la letteratura metodologica e cercare il metodo che meglio sposasse i miei dati e i miei obiettivi. Queste hanno rappresentato le basi dell'integrazione tra le due discipline che caratterizzano non solo questo lavoro ma anche il mio percorso dottorale.

La tesi è organizzata secondo il seguente schema. Nel primo capitolo viene presentato il fenomeno oggetto di studio. A partire dalla rassegna della letteratura ho operato una selezione teorica e sviluppato una mappa dei concetti che si sviluppa fino a isolare le pratiche della violenza. Ho svolto la medesima operazione per le forme di violenza operate tramite i media, nello specifico tramite le tecnologie web. In questo capitolo è presente il processo di individuazione e operativizzazione dei concetti che descrivono il fenomeno oggetto del mio studio. Nel secondo capitolo è presente la letteratura statistica di riferimento. Qui ho inserito una rassegna dei numerosi lavori proposti dalla comunità scientifica dedicati ai temi e ai metodi da me trattati. Approfondisco metodi e approcci allo studio dei testi e all'identificazione delle espressioni di odio. Ripercorro i fondamenti statistici del mio progetto individuando i lavori che hanno contribuito a formare la conoscenza che ne hanno permesso la realizzazione. Nel terzo capitolo presento la mia

strategia di analisi. In queste pagine le basi poste nei primi due capitoli si incontrano per sposare esigenze e conoscenza nella formulazione di interrogativi, obiettivi e proposte. Questo capitolo rappresenta la convergenza tra le anime che compongono la tesi nella strategia di analisi. Il capitolo segue le fasi di analisi che compongono la mia proposta, per ognuna ho specificato i metodi adottati e le motivazioni scientifiche e razionali dietro tali scelte. Nel terzo capitolo è presente il primo caso studio condotto nell'ambito del progetto *HateViz*. Questo rappresenta un momento di analisi preliminare, in cui ho voluto effettuare un esercizio di raccolta di opinioni e percezioni degli utenti online riguardo il mio oggetto di studio, la violenza di genere. Questo momento preliminare ha permesso di porre le basi per lo sviluppo di un'architettura per una web app. Nel quarto capitolo è presente un caso studio condotto applicando la strategia di analisi proposta nella tesi. In queste pagine viene analizzato un set di dati estratti da un popolare social network, *Reddit*. Mediante l'applicazione della strategia si cerca di rispondere alle domande di ricerca poste in partenza.

2. Principali riflessioni sulla violenza di genere e le espressioni d'odio

1. Espressioni di odio e violenza di genere

La violenza di genere quando si è imposta con tragiche statistiche all'attenzione mondiale è finita con l'essere sovra rappresentata nei media ma comunque scarsamente presidiata in termini di prevenzione. Gli organismi internazionali, dopo anni di politiche non abbastanza incisive, sono riusciti a decriptare con nuove lenti gli squilibri insiti nelle società moderne e hanno messo in atto alcune iniziative nella speranza di innescare un circolo virtuoso. Questo succedeva alla fine del secolo precedente ma le statistiche confermano ancora una volta che il percorso è ben lungi dal ritenersi concluso (Istat, 2014, 2019).

Il mancato adattamento alla trasformazione della figura femminile, lo stridore tra patriarcato e matriarcato, il dualismo tra vittima e carnefice generano una condizione di inspiegabile mortalità nei casi più gravi e di maltrattamenti nei casi più lievi. Organizzazioni nazionali e sovranazionali cercano di sensibilizzare al tema attraverso politiche culturali e iniziative mirate, un esempio è l'istituzione nel 1999 da parte dell'ONU della giornata internazionale contro la violenza sulle donne. A questo si aggiungono nei decenni successivi una serie di definizioni ufficiali, presenti in atti internazionali. Nella Convenzione di Istanbul (2011) all'espressione violenza nei confronti delle donne si associa «una violazione dei diritti umani e una forma di discriminazione contro le donne, comprendente tutti gli atti di violenza fondati sul genere che provocano danni o sofferenze di natura fisica, sessuale, psicologica o economica».

La violenza di genere rappresenta un ombrello più ampio che raccoglie un insieme di pratiche agite contro le donne in quanto appartenenti a tale genere, un atto di prevaricazione fisica e morale, considerato come tale anche se potenziale e non realizzato. Obiettivo di questo capitolo è quello di fornire un quadro generale sulle forme di violenza

contro le donne. L'intero percorso della mia tesi mira ad arricchire lo stato dell'arte degli studi sulla violenza di genere con una prospettiva certamente innovativa, incentrata sulle dimensioni concettuali prevalenti all'interno delle comunità virtuali, ambienti sempre più in crescita e ricchi di contenuti.

1. Violenza di genere

Il percorso concettuale che ci porterà a definire le chiavi di lettura del nostro lavoro di ricerca parte dalla definizione di violenza di genere nel quadro della nostra letteratura di riferimento. È quindi necessario individuare le diverse forme in cui questa si declina, comprendendone le differenze ed i punti in comune.

Per violenza intendiamo «ogni costrizione o espressione di natura fisica o psicologica che provochi danno o sofferenza di un essere umano» (Giomi Magaraggia, 2018).

Nella definizione di Shepherd (2013) della violenza di genere «sia genere che violenza costituiscono pratiche che funzionano da principi di ordinamento concettuale, e che danno forma alla nostra realtà sociale», poiché il genere definisce quali siano i comportamenti appropriati, quali desideri avere, quali modelli seguire. Secondo l'autore «la violenza ha sempre una dimensione di genere e allo stesso tempo è costitutiva del genere»; in questa definizione il termine "violenza" determina «ogni costrizione di natura fisica o psicologica che provochi danno, sofferenza o morte di un essere umano» (Giomi Magaraggia, 2018).

Chiamare la violenza operata dagli uomini sulle donne come "violenza di genere" è stata una grande conquista dei movimenti femministi, ma il binomio "uomini – autori" e "donne – vittime" è la chiave per l'interpretazione della connessione tra la violenza ed il genere. La violenza si può esercitare in spazi sia pubblici sia privati (Giomi Magaraggia, 2018). Una particolare tipologia di violenza viene definita attraverso l'acronimo IPV (*Intimate Partner Violence*), che indica specifiche violenze di genere subite dalle donne all'interno di relazioni personali, in particolare dal proprio partner. Questo tipo di

violenza viene collegato ad uno sbilanciamento di potere all'interno della coppia, in cui l'uomo agisce con violenza verbale, fisica e/o psicologica da una parte per «preservare il proprio potere o perché ci si sente vulnerabili e si ha la sensazione di dover difendere la propria identità» (Giomi Magaraggia, 2018). La spiegazione degli autori si adatta alla definizione di violenza di genere, in quanto diventa una modalità strutturale della relazione. La violenza di genere è quindi un problema legato non solamente al mondo esterno, ma anche alla sfera relazionale intima, e più in generale rappresenta un problema culturale relativo alla costituzione dei ruoli tra uomo e donna. In molti studi, un'altra declinazione della violenza di genere è quella domestica, pertinente a ciò che avviene all'interno della propria casa, è indicata come la principale motivazione di morte o lesioni verso le donne tra 16 e 44 anni, violenze che vengono effettuate nella maggior parte dei casi da membri della famiglia o all'interno della relazione di coppia (partner, ex partner, marito, padri, zii) con una maggiore esposizione al rischio da parte delle donne più giovani (Istat, 2014, 2019). La violenza esercitata negli spazi privati aiuta a fornirci una chiave di lettura generale del fenomeno, è infatti possibile isolarne la cornice culturale per ricavarne un framework concettuale in grado di porre le basi per avere una più chiara comprensione dei meccanismi non solo relazionali ma anche sociali alla base. Questo viene poi declinato in una struttura costituita dai principali concetti che ordinano e specificano la nostra conoscenza. Con questo obiettivo i principali riferimenti in letteratura sono stati raccolti e ordinati nell'immagine in figura 1, ho isolato i concetti chiave che permetteranno una comprensione ad ampio spettro della base teorica di riferimento e un tentativo di fare chiarezza sul fenomeno studiato. Il grafico consiste in un'organizzazione concettuale ad albero in cui da un primo livello generale (lo si potrebbe considerare come il *concetto madre*) si diramano sottolivelli in cui tale dimensione si esprime, fino ad arrivare alla base, la quale consiste nella manifestazione pratica (da

intendere come pratiche culturali) del fenomeno analizzato. La colorazione dei vari livelli segue questo ordine.

La violenza nella sua forma IPV è il concetto chiave da cui parte l'intera organizzazione, a questa fa riferimento la *coercive controlling violence*, questa trova nel controllo e nella coercizione la sua matrice valoriale: «We will use the term Coercive Controlling Violence for such a pattern of emotionally abusive intimidation, coercion, and control coupled with physical violence against partners» (Pence Paymar, 1993). Sul medesimo livello ho inserito la *situational couple violence*, tale fenomeno presenta caratteristiche più subdole, le quali rappresentano la degenerazione di specifiche dinamiche situazionali, «Situational Couple Violence results from situations or arguments between partners that escalate on occasion into physical Violence. One or both partners appear to have poor ability to manage their conflicts and/or poor control of anger» (Johnson, 1995, 2006). Per quanto riguarda la prima delle due forme, solitamente i ricercatori identificano la Coercive Controlling Violence utilizzando esclusivamente un modello basato sul potere e controllo (Pence Paymar, 1993). Nel mio caso, basandomi su una serie di lavori presenti in letteratura in cui il concetto subisce un'ulteriore estensione verso altre dimensioni semantiche sulla base di evoluzioni nella cultura e nella società contemponea (Amnesty International, 2020; Falloppa, 2015), ho deciso di aggiungere due ulteriori dimensioni: benaltrismo e buonismo. Il benaltrimo rappresenta la tendenza a minimizzare l'evento, sottolinenado l'esistenza di eventi più gravi, mentre il buonismo indica un rovesciamento retorico in cui la vittima diviene un soggetto che lamenta eccessivamente una situazione inesistente (Falloppa, 2015). Sempre sullo stesso livello, ma in riferimento a una dimensione situazionale abbiamo dinamiche basate sulla degenerazione della gelosia (Kelly, 2008) e dei rapporti di coppia(Ellis Stuckless, 1996), queste forme di violenza si esprimono nel momento in cui il partner non riesce più a gestire specifiche dinamiche e gli equilibri che la caratterizzano (Babcock et al. 2004),

sentendosi minacciato, oppure con espressioni di gelosia degenerate (Leone et al., 2004). I concetti presenti nell'ultimo livello dell'immagine descrivono le pratiche più comuni associate alle precedenti dimensioni emerse. Come pratiche non necessitano della medesima attenzione alla definizione, trattandosi di espressioni in uso anche lessico giornalistico, se non in quello quotidiano, tuttavia necessitano di una contestualizzazione in letteratura operata nella tabella 1.

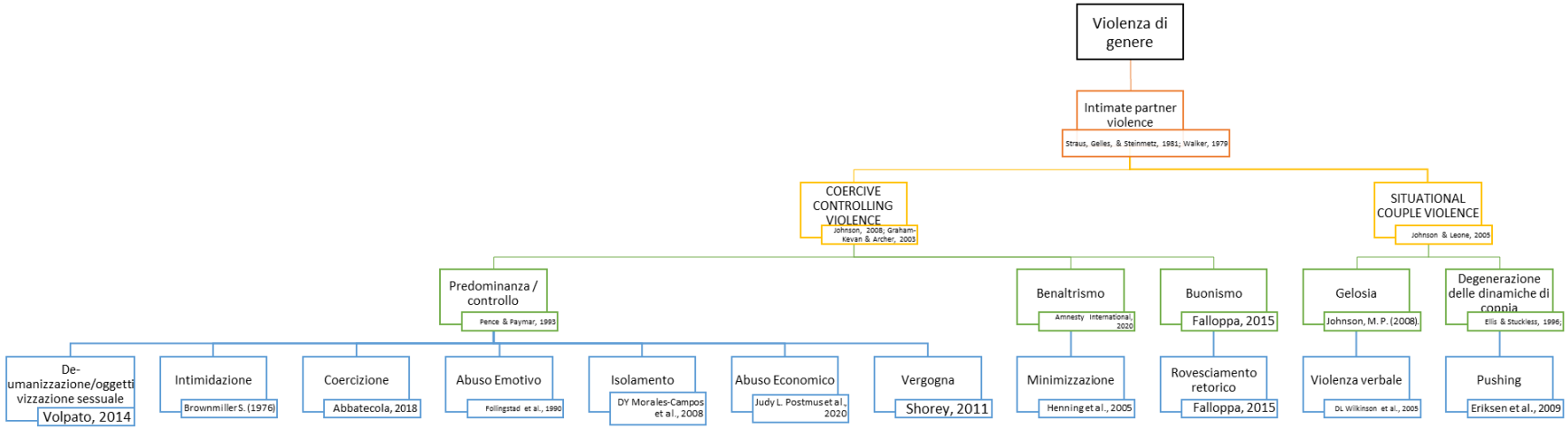

Figura 1 – mappa concettuale

Pratica	*Riferimento*	*Concetto*
Deumanizzazione/oggettivizzazione	Volpato, 2014	Predominanza/ controllo
Intimidazione	Brownmiller S., 1976	Predominanza/ controllo
Coercizione	Abbatecola, 2018	Predominanza/ controllo
Abuso Emotivo	Follingstad et al.,1990	Predominanza/ controllo
Isolamento	DY Morales-Campos et al., 2008	Predominanza/ controllo
Abuso Economico	Judy L. Postmus et al., 2020	Predominanza/ controllo
Vergogna	Shorey, 2011	Predominanza/ controllo
Minimizzazione	Henning et al., 2005	Benaltrismo
Rovesciamento teorico	Falloppa, 2015	Buonismo
Violenza verbale	DL Wilkinson et al., 2005	Gelosia
Pushing	Eriksen et al., 2009	Degenerazione dinamiche di coppia

Tabella 1- Riferimenti in letteratura

2. Violenza di genere nei media

I media svolgono un ruolo centrale nella descrizione del fenomeno della violenza di genere. In quanto amplificatori di messaggi questi hanno un importante ruolo nella costruzione sociale del fenomeno, fungendo da mediatori rispetto a delle specifiche pratiche culturali. In base alla chiave di lettura con cui i media si approcciano al tema emergono due tipologie di frame, conferendo all'argomento due prospettive: il "frame episodico" si concentra su aspetti soggettivi e peculiari del singolo caso trattato, suggerisce episodi isolati; il "frame tematico" tratta il fatto di cronaca con una visione più

estesa del fenomeno, esponendo strumenti e consigli utili a chi vive esperienze simili (Giomi Magaraggia, 2018).

Il pericolo di amplificare l'impatto della violenza di genere si trova quando i media trattano reati a sfondo sessuale per mano di una persona sconosciuta (fenomeno definito per questo "stranger danger"), poiché si corre il rischio di colpevolizzare la vittima, attraverso fenomeni noti come "victim blaming" (Randall, 2010), rivittimizzazione o vittimizzazione secondaria, «attribuendole la responsabilità dell'aggressione subita poiché "la sopravvissuta aveva attivamente provocato i suoi abusatori"» (Giomi Magaraggia, 2018).

Un esempio di minimizzazione mediale del problema consiste nel motivare la IPV maschile come la perdita di raziocinio da parte dell'aggressore, andando quindi a definire l'atto come "un attimo di follia" o "un gesto in preda alla rabbia" causato da situazioni problematiche su diversi ambiti, economici o emotivi o per difficoltà fisiche o mentali. Queste giustificazioni possono portare ad una "deresponsabilizzazione del violento" (Giomi Magaraggia, 2018), sollecitando una visione meno colpevolizzante verso l'uomo poiché anch'esso vittima della difficile situazione; alcune di queste dinamiche conflittuali portano a trasformare la violenza del singolo come un problema di coppia, con l'effetto di estendere la colpa anche verso la sopravvissuta alla violenza (che è quindi vittima non solo della violenza ma anche delle accuse del pubblico).

Questa chiave di lettura comprende anche i delitti passionali con il movente della gelosia; tramite questo frame la gelosia viene considerata una componente della relazione, suggerendo non solo una attenuante verso l'aggressore, ma attribuendo anche una corresponsabilità alla vittima: "è stata lei, in fondo, a lasciare o tradire". In alcune di queste narrazioni le protagoniste minimizzano i comportamenti violenti dei partner, incolpandosi per aver ignorato segnali e avvertimenti, fino ad accettare la violenza come normale prezzo da pagare per amore della relazione e del proprio uomo. Questo processo

di romanticizzazione delle relazioni non sane, se decontestualizzata, può finire con l'alimentare l'accettabilità sociale delle diverse forme di violenza. Queste pratiche legittimano i fenomeni persecutori, considerati come normale parte della relazione o del corteggiamento. Una pratica simile è la responsabilizzazione della vittima; questo frame si concentra esclusivamente su ha subito la violenza, considerandola come unico soggetto da colpevolizzare per via del comportamento che essa avrebbe dovuto tenere per evitare o fermare l'abuso, sminuendo così le responsabilità dell'autore. Questa chiave di lettura minimizza il problema sociale della violenza di genere.

L'idealizzazione dell'amore romantico è raffigurazione caratteristica dell'ordine patriarcale della società, in quanto il sentimento prende una nota di smarrimento e sacrificio di sé (quasi sempre da parte della donna) invece di raccontare un senso di rispetto reciproco e armonia nella coppia. Le donne sono quindi figure che si immolano per amore, su cui gli uomini possono sfogare le proprie frustrazioni. La peggiore conseguenza di questa visione distorta della «tensione tra armonia e smarrimento di sé, è che i confini tra amore e violenza divengono non nettamente definiti» (Giomi Magaraggia, 2018), finendo per legittimare ed autorizzare gli atteggiamenti violenti del partner.

La violenza di genere più diffusa nell'immaginario mediale e sociale è la violenza sessuale. Secondo Projansky (2001) queste rappresentazioni sono radicate nel nostro immaginario da naturalizzare lo stupro, non solo in quanto evento fisico reale, ma anche come parte delle nostre quotidiane fantasie, paure, e pratiche di consumo. La critica ha quindi proposto il concetto di *rape culture*, in cui lo stupro è la forma più comune di violenza, a cui le donne sono esposte nel quotidiano.

Rispetto al passato, in ambito giornalistico si registra un aumento di consapevolezza del fenomeno, trattando la violenza di genere con una dovuta e necessaria sensibilità. La IPV viene presentata come un problema sociale e viene utilizzata una chiave di lettura di

denuncia sociale alla sottomissione femminile propria degli autori di femminicidio, spostando l'attenzione del pubblico dal movente sentimentale (come la gelosia), «individuando il reale movente nella loro volontà di possesso e nella difficoltà ad accettare il diritto all'autodeterminazione della compagna» (Giomi Magaraggia, 2018). Importante è l'impegno delle testate giornalistiche di evitare e stigmatizzare la pratica del *victim-blaming* (Randall, 2010) o dello *slut shaming* (Ringrose Renold, 2011), la condanna di condotte e desideri sessuali femminili non convenzionali o socialmente accettati.

Quanto detto finora aiuta ad inquadrare la violenza di genere come un valore radicato nelle nostre società già molto prima delle tecnologie on-line. Queste tecnologie costituiscono la base materiale su cui nascono nuove pratiche di violenza. Non solo, gli spazi digitali riescono ad estendersi al globale, istituendo nuove pratiche culturali dai confini indefiniti.

3. Violenza di genere all'interno degli spazi online

Esiste una ricca letteratura di riferimento che pone l'attenzione sulle dimensioni con cui il fenomeno della violenza di genere si espleta online, facciamo riferimento agli studi sulle comunità virtuali sia come luoghi in cui si verificano casi di violenza e sia come osservatori privilegiati in cui è possibile intercettare i discorsi e le percezioni degli utenti riguardo il fenomeno. L'utilizzo massivo delle cosiddette *information and comunication tecnologies* ha contribuito a creare un nuovo spazio su cui si sono estese le pratiche della violenza e le sue percezioni. Questa specificazione è necessaria per inquadrare una letteratura che copre un arco disciplinare molto esteso, orientandosi concettualmente sugli aspetti legali del rapporto tra violenza e libertà di espressione, ricordiamo che comunque ci riferiamo a tecnologie di comunicazione (EC, 1997, 2003; ECRI, 2016), e su quelli culturali e sociali che si interrogano sul sorgere di pratiche e comunità in cui il fenomeno è esteso.

In ambito europeo, i principali riferimenti in materia di violenza espressa ed esercitata mediante internet e il web hanno origine da due note emesse dal Consiglio d'Europa, ci riferiamo in particolare al Consiglio d'Europa del 30 ottobre 1997 e alla Convenzione di Budapest sulla criminalità informatica del 2003. In entrambi gli atti il fenomeno oggetto di attenzione sono le forme di odio e di violenza condivise online e vengono associate ad ogni forma di espressione che promuove, incita, giustifica o diffonde odio razziale, xenofobia, antisemitismo o altre forme di ostilità e disprezzo verso individui o gruppi di individui basate sull'intolleranza. Nel 2015 la Commissione Europea nello sviluppo di politiche mirate al contrastare le forme di intolleranza e razzismo include per la prima volta (insieme agli aspetti precedentemente delineati) l'accezione sessuale e di genere all'interno della definizione di espressioni di incitamento e promozione alla violenza. Quindi nelle successive iniziative di contrasto al fenomeno online portate avanti dalle istituzioni europee viene inserito come obiettivo quello di monitorare le azioni intraprese per tutelare le donne dagli abusi online, come *cyber-stalking*, sessismo, misoginia e minacce di violenza sessuale. Tutte queste espresse tramite contenuti condivisi tramite le principali tecnologie presenti sul web.

Negli ambienti digitali circolano immagini e video di persone che subiscono abusi e violenze, intensificando il problema dello slut shaming (Ringrose Renold, 2011), il fenomeno caratterizzato da commenti online discriminatori e fatti di insulti e disprezzo, cui le donne sono sempre più spesso sottoposte. Con lo sviluppo del mondo online e digitale, si è diffusa una particolare forma di violenza definita *Technology-Facilitated Sexual Violence* -TFSV- (Henry Powell, 2018), traducibile come "violenza sessuale facilitata dalla tecnologia" che include le molestie sessuali subite dalle donne nel mondo online, il *cyberstalking* (Henry Powell, 2018) e lo sfruttamento o diffusione di immagini sessuali non richieste, come nei fenomeni del *sexting coercion* (Albury Crawford, 2012) e del *revenge porn* (McGlynn, Rackley Houghton,2017). È in questo quadro che si

inseriscono anche le espressioni d'odio (da ora anche nella forma inglese hate speech) come forma di violenza di genere.

Per comprendere la dimensione culturale che alimenta l'espandersi delle pratiche di violenza bisogna definire lo spazio sociale in cui queste vengono condivise. Internet è una piattaforma globale che ospita al suo interno una moltitudine di tecnologie che permettono la creazione e la condivisione di contenuti di varia natura. Sulla base della struttura tecnica e dell'organizzazione e che queste piattaforme si danno è possibile suddividerle in varie tipologie, i social media sono una di queste. La caratteristica principale di tali tecnologie sono i cosiddetti User Generated Content, contenuti generati dagli stessi utenti che utilizzano queste piattaforme e condivisi tra i membri (Obar, Wildman, 2015). Gli UGC consistono nella produzione culturale di queste community e riescono a veicolare messaggi con una elevata potenza di trasmissione. Pratiche di odio e violenza mediate da tali messaggi possono istituirsi e prendere piede molto facilmente all'interno di queste comunità. La violenza di genere perpetrata tramite tali canali si caratterizza per essere il messaggio stesso, quindi un atto molto rapido da compiere, specialmente se si tratta di contenuti testuali o di immagini.

Lo spazio online è anche dove gli utenti confrontano le proprie visioni del mondo condividendo opinioni e considerazioni sottoforma di contenuti di varia natura. Negli ultimi anni si è sviluppata una nutrita letteratura sull'utilizzo del web come fonte per la ricerca sociale (Rogers, 2016; Corposanto et al., 2015; Molinari, 2014). Gli ambienti web divenuti aggregatori di dati disponibili allo scienziato sociale sono molti, tra questi sono presenti i social media: tecnologie computer-mediated che facilitano la creazione e la condivisione di informazioni, idee, interessi e altre forme di espressione tramite comunità virtuali e network (Obar, Wildman, 2015).

2. La violenza di genere tramite le espressioni d’odio online

1. Hatespeech

È possibile riscontrare all’interno della categoria di hatespeech una sostanziale difficoltà di definizione dei corretti confini concettuali (Besussi, 2019). Quali sono i contesti, le circostanze e le espressioni lessicali che dovrebbero essere limitate, se non addirittura censurate? Quali sono i target da ritenere sensibili a tali attacchi? Risulta evidente anche all’osservatore poco esperto sul tema una difficoltà a definire un contesto univoco in cui è possibile individuare tali espressioni. A complicare il quadro contribuisce anche una ricchezza di possibili destinatari e modalità con cui esercitare questo odio.

Tale indeterminatezza è di fatto una sfida superabile tramite l’applicazione di una categorizzazione pragmatica e facendo riferimento ad una adeguata letteratura: la prima distinzione da operare riguarda i tipi di discorso e di ambiti. Strutturando un ragionamento teorico alla base, sarà possibile definire l’oggetto della ricerca e gli ambiti di applicazione dei concetti proposti.

La ricchezza dei contributi presente in letteratura ha richiesto una sistematizzazione concettuale svolta tramite una disamina che ho deciso di organizzare all’interno di uno schema riassuntivo presente nella figura 2.

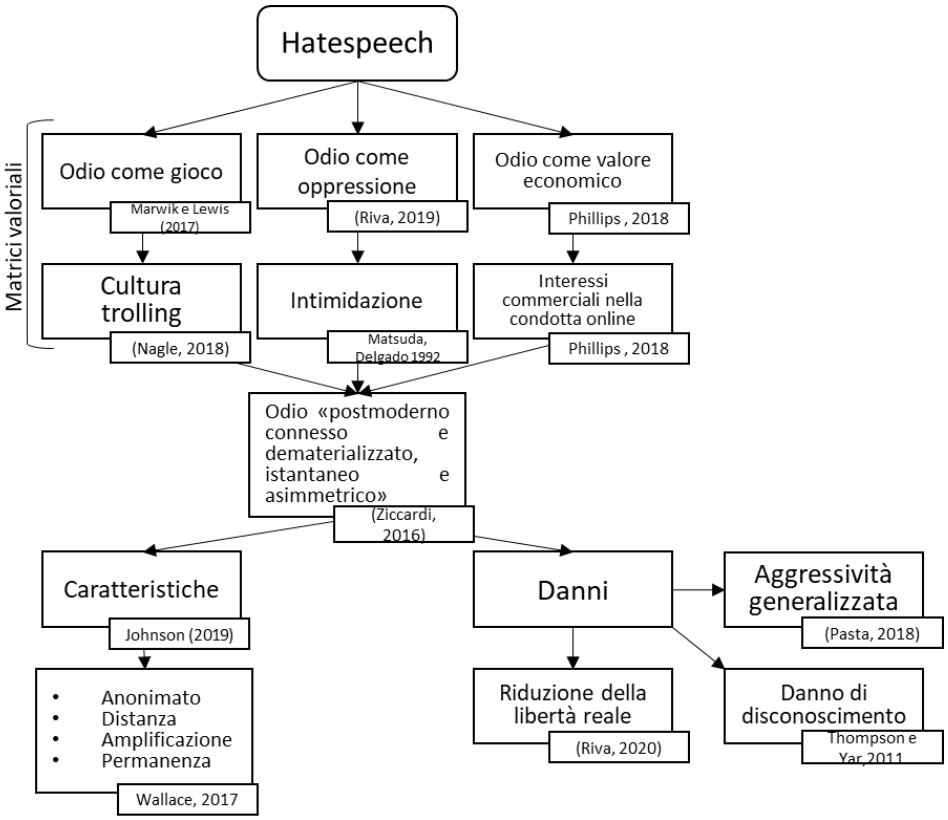

Figura 2 – *Hate speech*, schema riassuntivo

Gli interventi d'odio possono avere origine da una moltitudine di soggetti: espressioni di stampo discriminatorio, razzista o sessista da parte di gruppi organizzati; interventi spontanei e occasionali mirati a specifici obietti da parte di singoli; l'odio per gioco espresso da utenti singoli o in gruppi in rete. Marwick e Lewis (2017) ci forniscono un elenco di individui classificati in base ai comportamenti assunti nelle relazioni sociali in rete: i cosiddetti troll rappresentano i provocatori e solitamente agiscono per divertimento o per sfidare i loro interlocutori, per loro queste dinamiche rappresentano un gioco; i *non state violent actors*, seguono invece un'ideologia della violenza, che si esplica con terrore e paura, a questi possono essere associati i comportamenti di molti influencer, teorici della cospirazione ma anche alcune personalità politiche fanno proprio il linguaggio dell'odio e della paura per finalizzarlo a specifici obiettivi.

Nel primo quadro si inserisce una vera e propria cultura trolling, in cui il gruppo agisce con violenza sull'individuo. Azione esercitata contro ogni tipo di morale e vissuta come

gioco perverso che rifiuta ogni codice etico, a partire da quello della solidarietà umana (Nagle, 2018).

È possibile individuare all'interno della generica comunicazione d'odio delle specifiche caratteristiche che possono costituire le basi per la genesi di una cultura dell'odio. L'appartenenza, da parte dei destinatari di questi messaggi di odio, a gruppi sociali o a culture con una storia di oppressione alle spalle può risultare essere un'aggravante rispetto alle offese rivolte tra gli individui nelle quotidiane interazioni sociali. L'ingiuria rivolta in quanto appartenente ad uno specifico gruppo, genere o cultura rappresenta l'espressione di oppressione verso una minoranza socialmente non rispettata e considerata meno che uguale alla maggioranza della società. Il danno che sussiste nell'offesa non costituisce un danno per l'individuo preso di mira, piuttosto rappresenta un attacco simbolico all'intero gruppo, Thompson e Yar definiscono questo danno come danno di disconoscimento (2011). Secondo Besussi (2019), unilateralismo normativo e moralizzazione del discorso pubblico sono i due temi che accompagnano ogni possibile riflessione sull'hatespeech. Secondo l'autore, non ci riferiamo qui ad uno specifico universo linguistico e\o semantico, ma a pratiche linguistiche e simboliche che esprimono il loro disprezzo a gruppi sociali storicamente oppressi o tenuti in segregazione. Quindi è il fatto di essere un'aggressione perpetrata da parte del oppressore verso l'oppresso a caratterizzare il discorso d'odio e non una specifica categoria di atti linguistici.

Anche se il discorso d'odio e i comportamenti ostili sono presenti nelle società umane fin dalle origini, oggi l'hatespeech assume caratteristiche inedite a causa delle trasformazioni nella comunicazione globale online (Ziccardi, 2016). L'aggressività, la rabbia e la violenza rientrano nelle dinamiche intrinseche biologiche e psichiche degli esseri umani, e non nascono certo con Internet. Tuttavia, il discorso d'odio, che ha una lunga storia, ha assunto nuove caratteristiche con l'espansione rapidissima delle

tecnologie connettive e in particolare con i social network. Come vedremo, questo odio «postmoderno connesso e dematerializzato, istantaneo e asimmetrico» permette di colpire in modo mirato, e viene amplificato dalla rete come palcoscenico davanti a un pubblico immenso (Ziccardi, 2016).

È importante monitorare ed eliminare la diffusione sul web di messaggi basati sull'hate speech perché, data la caratteristica trasversalità dello strumento (peraltro strumento insostituibile di informazione), questo può alimentare tribalismi, aggressività generalizzata, polarizzazioni e razzismi "originari" (Pasta, 2018). Sopra è stato argomentato come gli abusi verbali costituiscano un danno simbolico, a questo può essere associato anche un danno concreto e materiale perché tale espressione costituisce anche un incitamento implicito all'abuso fisico, a intimidazioni e attacchi personali (Matsuda, Delgado 1992, Mackinnon 1993).

L'espressione d'odio è un'intimidazione che mette in ostaggio la libertà individuale dei gruppi target (Riva, 2019), andando a ledere la dignità dell'altro si mette in dubbio il suo essere un attore sociale alla pari di chi rivolge l'insulto. In questa accezione il danno consiste nella riduzione della libertà reale di chi subisce l' attacco, senza soffermarsi sulla natura del gruppo di appartenenza dell'individuo designato come vittima.

Non si può negare che le grandi società del web abbiano sfruttato questo tragico panorama per costruire vantaggi economici attraverso l'alimentazione di un traffico comunicativo molto spesso popolato da questo tipo di interazioni: l'odio urlato, le accuse, le emozioni calcate, il linguaggio offensivo rappresentano non solo valori ma anche valore aggiunto per le piattaforme online. Il volume degli interessi commerciali ed economici riesce ad esercitare molte pressioni sulla condotta delle aziende del web. Esiste una tendenza da parte dell'industria mainstream che ricicla e amplifica sistematicamente i messaggi d'odio; quindi è spesso l'infosistema tradizionale a potenziarlo, come ha dimostrato Whitney Phillips nei suoi rapporti (2018).

Il discorso d'odio, nelle accezioni qui descritte, è distinto dai crimini di odio, questi sono penalmente rilevanti e motivati da pregiudizi e intolleranza. Con il linguaggio abusivo, i danni e le offese sono intrinseche al contenuto stesso del discorso. Ciò non toglie che i modi di esprimersi possono essere perseguibili penalmente come crimini. La necessità di una risposta pubblica alle offese verso gruppi minoritari o oppressi si è tradotta in una richiesta di una serie di restrizioni a determinate pratiche che fanno riferimento esplicito a tale linguaggio, perché lesive della dignità di tali gruppi. Questa dinamica impatta con modalità problematiche l'impianto teorico delle democrazie liberali, andando a porre un freno ad una serie di libertà in una lettura critica dei principi che sussistono alla base di tali impianti politici. Tuttavia, come tutta una parte dei sostenitori delle restrizioni pensa, se le parole d'odio contro chi appartiene a una storia di oppressione e discriminazione inducano direttamente condotte persecutive, a molestie e attacchi fisici diviene necessario porre un limite (Waldron 2012).

Come precedentemente accennato, la legislazione europea intreccia il discorso d'odio alla discriminazione e all'incitamento e in tali fattispecie costituisce reato; questo però esclude una buona parte di espressioni dilaganti che si presentano come forme di aggressività e ostilità liquide, banalizzate e popolari, ugualmente pericolose e incentivanti determinati comportamenti pericolosi. Sulla base della Carta dei Diritti dell'Unione, il consiglio e la commissione europea hanno promosso una serie di interventi legislativi e azioni di sensibilizzazione in materia di hatespeech, l'obiettivo principale è rendere reato i messaggi sessisti, razzisti e omofobi (Wongher 2015).

Qual è il ruolo dei social media in questo panorama oscuro del web contemporaneo? Alla classica idea di un progressivo declino delle relazioni sociali, gli antropologi sostengono la teoria secondo cui «i social media segnano un ritorno al significato e alla vitalità di gruppi come la famiglia, la casta e la tribù, e un ripudio della precedente tendenza verso i network individuali» (Miller et al., 2018). Henry Jenkins si oppone alla

fruizione passiva del mezzo, proponendo la sua idea di una cultura che pur sempre resta ancorata a principi di ordine partecipativo attraverso un modello di cultura con «persone che plasmano, condividono e remixano contenuti in modi che in precedenza non si sarebbero neppure potuti immaginare» (Jenkins, Ford Green, 2013).

I social media, con la loro particolare struttura comunicativa orizzontale, sono diventati una parte imprescindibile del nostro sistema di comunicazione. Con ciò ci riferiamo alla oramai totale immersione delle nostre individualità (per questo sempre più mediali) in un flusso continuo di dati e informazioni, strutturate in messaggi di varia natura. Tale è la portata della grande ideologia di internet di cui ci parla Lovink (2019), rispetto alla quale, anziché esprimere una generica condanna morale o atteggiamenti di negazione, vanno elaborate strategie alternative. Queste tecnologie riproducono costantemente il nostro sé all'esterno: l'obiettivo è affermare la propria identità o ricercare una socialità (ricorrendo anche a valori aggressivi e offensivi)? Sono in gioco delle meccaniche da branco di ancestrale memoria? Resta irrisolta la contraddizione tra «il soggetto iper individualizzato e i comportamenti imitativi tipici dei social» (Lovink, 2019). Sherry Turkle (2012) scrive che con i social e il web siamo «insieme ma soli», con una nuova socialità che si esprime nei confini di nuovi mezzi per definire il noi e l'altro. Sicuramente le caratteristiche dello spazio mediatico online influenzano in qualche modo anche le dinamiche interpersonali e la possibilità di esprimersi, anche quando gli utenti si scambiano l'odio fin qui descritto. Riguardo la possibilità dell'esistenza di serie di fattori che facilitano l'aggressività in modi diversi negli ambienti online abbiamo il contributo di Patricia Wallace (2017). L'anonimato e la distanza fisica sono tra i fattori che possono avere effetti di disinibizione nell'individuo che attacca un'altra persona, azione che risulterebbe più semplice da compiere rispetto all'offendere qualcuno di persona; anche l'amplificazione è un ulteriore elemento critico, per cui un insulto o una diffamazione può raggiungere una vasta platea; ultima tra quelle ritenute più significative ai fini di questo

lavoro è la permanenza, dato che, come si è detto, la rimozione rapida è stata finora molto difficile. È Johnson (2019) ad individuare nei social una struttura favorevole per il valore dell'odio, un habitat in cui proliferare ad una velocità e una pervasività molto elevata. In quest'ottica l'orizzonte su cui agire si espande, superando la dimensione socio-culturale e politica, ci estendiamo sul piano delle emozioni e dei valori.

All'interno del flusso informativo che attraversa il web sono presenti una quantità molto elevata di contenuti moralmente rilevanti, che vanno a costituire una parte molto importante della nostra vita quotidiana (Crockett, 2017). Gli utenti dei social media e, nello specifico, dei social network, consumano questi contenuti morali nei loro pellegrinaggi quotidiani all'interno dell'offerta mediale digitale (Duggan e Smith, 2016).

Il dualismo che vede contrapporsi la sostanziale contraddizione tra la razionalità dei mercati economici e la ricerca del piacere e delle emozioni nasconde un'ipocrisia di fondo. È il mercato che trasforma in valore quei valori che popolano la sfera del se e dell'emotività, in cui si vende la costruzione del se, l'autenticità e la realizzazione di un proprio essere. L'emozione svolge un ruolo centrale, è merce posta su un piedistallo e mostrata al pubblico passaggio, ciò rende essenziale avere cura anche di questo portato emotivo, diventando quindi importanti per quello che significano e per cosa dicono di noi (Illouz, 2017). La carica simbolica dei prodotti di consumo, ben nota alla teoria sociologica, e in particolare della loro dimensione emotiva si riflette nell'uso dei social media (Turkle, 2012).

Secondo Brady e Crockett (2019) i social media, tramite la loro architettura di sistema, potrebbero essere la causa di un arresto dell'azione collettiva contro le disuguaglianze e il progresso sociale, sfruttando l'indignazione come un fine in sé. È chiaro che bisogna operare una valutazione delle ripercussioni offline di questo odio diffuso. È importante chiedersi il ruolo che svolgono i social media e tutte le tecnologie della comunicazione ospitate sul web nella formazione e nell'esercizio dei sentimenti empatici. Come

valutiamo le nostre azioni? Come ci comportiamo nei confronti dell'aggressività altrui? È importante dunque chiedersi il significato che le persone attribuiscono ai loro comportamenti online, di sicuro la risposta risiede nella nostra mente, individuale e collettiva, e nel modo siamo socializzati allo strumento e al significato dell'altro.

Venendo meno il contatto fisico, i protocolli comunicativi delle relazioni interpersonali online sono più rapidi e depersonalizzati. Il sentire l'altro, alla base della sensazione empatica, viene meno. Negli studi di Konrath, O'Brie e Hsing (2011) è chiaro come le generazioni dei giovani americani siano meno empatiche.

In vari lavori sull'argomento (Turkle, 2016; Twenge, 2018) si evidenzia come l'utilizzo sempre più comune di tecnologie del web per la comunicazione possa indurre a un'inibizione della capacità empatica, a causa di un distanziamento che non permette il dialogo in presenza e quindi viene meno la possibilità di osservare direttamente gli effetti sull'altro dei nostri messaggi.

Si può cercare di circoscrivere una dimensione concettuale all'odio espresso con tali modalità anche spostando l'attenzione da chi esprime il messaggio a chi lo riceve, in tal senso Fumagalli (2019) sposta le sue considerazioni dal parlante agli uditori. Sostiene, in pratica, che il mittente di uno specifico messaggio non fa altro che sintonizzarsi su chi ha di fronte, quindi la platea riconosce e condivide le sue parole d'odio. C'è sempre il rischio che tale sincronizzazione tra l'emittente e la sua platea non riesca con successo, tuttavia l'autore ritiene che le parole d'odio trovino, nella maggior parte dei casi, una platea già orientata in tal senso.

In quest'ottica il discorso d'odio diviene una conferma, un codice comune che permette al parlante e all'uditore di riconoscersi e condividere queste malsane idee. Tali espressioni diventano pratiche comuni e condivise tra questi attori e così facendo gli si dà la possibilità di essere diffuso, in questo caso l'espansione a macchia d'olio di tale pratica rende l'idea della facilità di propagazione e dell'urgenza di arresto. In base a queste

considerazioni per arrestare tali pratiche non bisogna ricorrere esclusivamente ad atti normativi e legali ma agire sulle pratiche dell'odio e indebolirle, «in altri termini, non è attraverso restrizioni legali o sociali dell'hate speech che si combattono razzismo, sessismo, omofobia ecc., ma è piuttosto confrontandosi con le diffuse pratiche sociali che incorporano atteggiamenti e convinzioni razziste, sessiste ecc. che si cureranno i discorsi d'odio che sono un sintomo di quelle pratiche e non produttrici di odio» (Galeotti, 2019).

Un ruolo importane nella ridefinizione dei comportamenti aggressivi deve essere rimandata a un processo di educazione che svolga un ruolo di contro peso a queste pratiche d'odio costruite su codici linguistici e forme di espressione razziste, sessiste, xenofobe ecc. (Dworkin 1993, Ross 2015).

L'odio rappresenta un rischio per la coesistenza pacifica tra gruppi di individui, le espressioni d'odio rappresentano un attacco pratico a questa convivenza. Non le si può tollerare perché vanno a ledere la dignità delle persone, ponendo delle distanze tra chi è prima e chi è dopo. E poiché la stessa censura di tali elementi rappresenta una questione di esercizio delle libertà individuali si dovrà ricorrere a un riaggiustamento della civility, la virtù sociale che consente un'interazione rispettosa e fluida fra agenti sociali (Galeotti, 2019).

3. La misoginia nei social networks

Esistono molte espressioni che servono a descrivere le dinamiche di odio e di violenza di genere online: *gendered cyberhate*, *technology-facilitated violence*, *tech-related violence*, *online abuse*, *hatespeech online*, *digital violence*, *networked harassment*, *cyberbullying*, *cyberharassment*, *online violence against women* e *online misogyny*.

La misoginia, che possiamo definire come l'assunzione di comportamenti di odio o di stampo pregiudizievole contro le donne, può essere verbalmente manifestata in numerosi modi, che vanno dall'esclusione sociale e la discriminazione a espressioni più pericolose legate a minacce di violenza e oggettivazione sessuale (Anzovino et al., 2018).

Un recente studio del Pew Research Center (2014) ha rilevato che molte donne percepiscono come ostili molte web communities. Inoltre, le donne riferiscono un maggiore disagio emotivo a causa delle molestie online, indicando che le loro esperienze sono particolarmente insidiose e probabilmente qualitativamente diverse da quelle degli uomini (Pew Research Center, 2014).

Sebbene il nostro focus sia sui fenomeni online, siamo consapevoli che spesso esiste un continuum tra le manifestazioni online e offline di violenza facilitata dalla tecnologia, ad esempio nel caso di abuso IPV. Amnesty International nel 2017 ha rilevato che il 41% delle donne che hanno subito abusi o molestie online ha affermato che, in almeno un'occasione, queste esperienze online le hanno fatte sentire che la loro sicurezza fisica era minacciata.

È importante sottolineare, tuttavia, che le tecnologie digitali non si limitano a facilitare o aggregare forme esistenti di misoginia, ma ne creano anche di nuove che sono inestricabilmente connesse con le offerte tecnologiche dei nuovi media, le politiche algoritmiche di alcune piattaforme e le culture che si esprimono su di esse alimentando gli individui e le comunità che le utilizzano.

Il lavoro di D.Ging (2017) sulla manosphere si concentra su specifici ambienti online che permetterebbero ai cosiddetti maschi beta di usare come arma la misoginia e il razzismo nel tentativo di rivendicare questi spazi come bianchi e maschi, sebbene questi uomini possano avere un basso capitale culturale ed economico offline e una vita sociale poco coltivata.

Il fenomeno della misoginia online non è nuovo. I lavori accademici più rilevanti partono dagli anni 2000 ma è stato solo successivamente agli eventi del cosiddetto *gamergate* nell'agosto del 2014 che i media mainstream hanno iniziato a coglierla. Con essi si è mobilitata anche una parte sostanziosa di accademia.

Il lavoro preveggente di Jill Filipovic (2007) sulla relazione tra misoginia di Internet e molestie nel mondo reale è stato pubblicato sullo Yale Journal of Law and Feminism, mentre P. Turton-Turner (2013) ha pubblicato sulla semiotica visiva della misoginia e della libertà di parola online in Forum on Public Policy, e il saggio di Karla Mantilla (2013) sul *gendertrolling* è apparso in Feminist Studies. J. Jenson e S. De Castell (2013) e S. Chess e A. Shaw (2015) hanno risposto specificamente alla questione della misoginia nei giochi e nella cultura del gioco, mentre due dei contributori di quel numero speciale, Emma Jane e Adrienne Massanari, erano in primo piano nella ricerca su questo argomento da una prospettiva specificamente femminista di studi sui media o studi su Internet. Il lavoro di E.A. Jane (2014) su e-*bile* è stato un intervento importante in termini di tematizzazione specifica dell'odio misogino online come argomento di studio, mentre l'analisi di A. Massanari (2017) riguardo il gamergate and *the fappening* su Reddit ha sottolineato in modo cruciale il ruolo di algoritmi nel determinare la politica della piattaforma. Anche lo studio di Angela Nagle sulla misoginia su *4/chan/b* (2015) e il lavoro di S. Banet-Weiser e K. M. Miltner (2016) sulla mascolinità tossica hanno costituito nuovi importanti modi di affrontare il problema.

Un altro concetto molto importante per comprendere le dinamiche misogine sulle piattaforme online è il cosiddetto *ambient sexism*, se gli atteggiamenti e il comportamento sessisti sono prevalenti in uno spazio social, questa atmosfera può avere effetti negativi indipendentemente dal fatto che un utente sia direttamente preso di mira da molestie sessuali o da atteggiamenti offensivi in generale (Glomb et al., 1997).

Un’ulteriore elemento su cui porre attenzione è che queste espressioni sessiste non sono sempre caratterizzate da un’ostilità e un’aggressività verbale; molto spesso, in difesa di commenti sessisti, si dice che sono da intendersi in modo scherzoso (Ford e Ferguson, 2004). Indipendentemente dal tono, il sessismo online è pernicioso; come nota Marwick

(2013), l'umorismo sessista in contesti online "rafforza il diritto maschile e gli stereotipi di genere convenzionali normalizzando un comportamento incredibilmente sessista".

In effetti, il "sessismo quotidiano", compreso l'umorismo sessista, può essere dannoso quanto altre forme di sessismo, indipendentemente dalle sue presunte intenzioni (Swim, Hyers, Cohen e Ferguson, 2001). Le donne sperimentano affetti negativi dall'umorismo sessista (LaFrance e Woodzicka, 1998) e in molti contesti, le battute sessiste sono vissute allo stesso modo delle molestie sessuali (Boxer e Ford, 2010). Gli uomini esposti all'umorismo sessista successivamente credono che il sessismo sia più normativo, si incolpano meno per il proprio comportamento sessista e si impegnano in comportamenti più discriminatori (Ford, Boxer, Armstrong, Edel, 2008; Ford, Wentzel e Lorion, 2001).

3. Le espressioni d'odio nella letteratura statistica

L'hatespeech è, in prima istanza, un problema relativo al linguaggio e al suo rapporto con le emozioni. Le pratiche e i valori culturali, espressi tramite concetti, delineati nel precedente capitolo vengono veicolati, codificati e trasmessi tramite specifiche forme linguistiche. La ricerca statistica aiuta ad automatizzare la fase di individuazione di queste forme linguistiche che necessitano di censura. Mi riferisco a contenuti trasmessi in forma testuale, considerando la parola l'unità minima di senso e la *feature* principale che caratterizza il contenuto testuale preso in esame. È possibile leggere questo tipo di ricerca come ad un processo. Questo consiste in una moltitudine di step in cui il contenuto viene acquisito dal repository online e trasformato in un dato trattabile statisticamente. Nei paragrafi successivi mi soffermerò sulle fasi che costituiscono l'intera filiera di analisi al fine di delineare le basi metodologiche allo studio statistico dei testi, in letteratura definito *text mining*. In particolare approfondirò le fasi metodologiche che curano la trasformazione dei testi, da una forma non strutturata fino alla creazione della matrice di analisi. Successivamente mi soffermerò in particolare sugli strumenti utili ad estrarre un'informazione sul contenuto emotivo dei testi inseriti all'interno del processo di analisi. Questi due importanti filoni teorici costituiscono la base per la comprensione e contestualizzazione del lavoro svolto per la progettazione della strategia di analisi da me proposta.

1. Text mining: il processo di strutturazione del dato

Il testo e il linguaggio sono gli oggetti di ricerca su cui la letteratura in materia costruisce il suo impianto metodologico. L'analisi automatica dei testi e la statistica testuale sono approcci che nascono con l'evoluzione tecnologica contemporanea, questa ha permesso il proliferare di contenuti testuali e di conseguenza è cresciuto l'interesse da parte di scienziati sociali e sviluppatori informatici per il cosiddetto *natural language*

process e l'analisi automatica dei testi (Lebart, et al., 1988). La letteratura propone una ricca varietà di approcci a tali studi, a tal proposito la prima contrapposizione è tra una logica di tipo linguistico, orientata alla ricerca di regolarità nella lingua, ad una logica di tipo lessicale, la quale mira a comprendere gli usi dei lemmi nei loro significati e accezioni (Guiraud, 1954; Muller, 1977). La produzione scientifica si è ulteriormente estesa negli ultimi anni grazie alla grande disponibilità di dati testuali prodotti negli ambienti digitali. Gli obiettivi della ricerca e l'intento di individuare delle dimensioni emotive all'interno delle unità testuali analizzate orientano la letteratura di riferimento verso un approccio lessicale piuttosto che linguistico. Le prime statistiche elementari presenti in letteratura hanno lo scopo di confrontare tra loro testi di autori differenti, queste si sono evolute fino alla formalizzazione di analisi statistiche di dati testuali basate su tecniche fattoriali. Questo passaggio ha permesso il diffondersi di una nuova logica lessico-testuale che permette di superare gli ostacoli posti dall'insita ambiguità della lingua e permette un'estrazione di informazione più efficiente (Lebart, et al., 1988). La letteratura più recente si rivolge ad una crescente mole di fonti disponibili in formato elettronico. Ciò ha reso possibile il ricorso a strategie più complesse per l'estrazione, l'analisi e l'organizzazione della conoscenza, con lo scopo di soddisfare i diversi bisogni informativi (Balbi, Misuraca, 2005). Attualmente, con la sempre crescente generazione di dati online da parte dell'IoT (Internet of Things), sta crescendo anche la necessità di un'adeguata elaborazione. Una percentuale considerevole dei dati disponibili online è costituita da siti web, blog, riviste, social network e simili, che includono una grande quantità di testo. Questa enorme quantità di dati non può nella sua forma attuale essere elaborata dagli esseri umani o dall'elaborazione convenzionale dei dati, portando il mondo verso il miglioramento e la ricerca nella scienza dei dati. Obiettivo del Text Mining (TM) è proprio l'estrazione di informazione rilevante da dati non strutturati che risiedono in file di testo. Obiettivo dell'*hatespeech detenction* è quello di estrarre informazioni relative al

contenuto emotivo dei testi, in particolare io mi riferisco allo spettro *negativo*, relativo alle espressioni di odio e violenza.

1. Text mining

Il Text Mining (TM) è un campo di ricerca di analisi testuale che si occupa di estrarre informazioni da documenti o testi, servendosi di tecniche statistiche, linguistiche e informatiche. La classificazione delle unità testuali e la valutazione dell'appartenenza di un testo ad una certa categoria è una delle task principali di questa disciplina. Questa operazione consiste in una ricerca nei testi di informazioni utili a produrre nuove conoscenze. La caratteristica fondamentale risiede nel fatto che questo tipo di analisi lavora con dati non strutturati (per esempio file di testo narrativo, oggetti multimediali, ecc.). Questi sono privi di un'organizzazione e di una struttura tale da essere processati statisticamente quando vengono raccolti, in pratica è necessario definire un processo di codifica che trasformi le parole in numeri. Secondo Hotho et al. (2005) possiamo distinguere tre diverse prospettive di text mining: estrazione di informazioni, data mining, e il processo KDD (Knowledge Discovery in Databases[1]). Questo schema ci permette di individuare tre fasi di analisi:

- la definizione degli obiettivi e l'acquisizione dei documenti: prevede la definizione dell'oggetto di studio da analizzare e la fase di raccolta del materiale testuale da sottoporre alle fasi successive
- la codifica dei dati: il processo che parte dall'acquisizione dei documenti fino alla definizione della matrice termini-documenti, fulcro dell'analisi testuale
- l'estrazione dell'informazione, la quale prevede la scelta della tecnica statistica più idonea per il raggiungimento degli obiettivi prefissati.

[1] Il processo globale di analisi di grossi database, finalizzata ad estrarre della conoscenza nascosta, è noto come "**Knowledge Discovery in Databases**" (KDD). Il **Data mining** (DM) è il cuore del processo di KDD e comprende gli algoritmi e le tecniche per esplorare ed apprendere dai dati, implementando il processo di conoscenza.

È importante sottolineare che il processo di codifica può comporsi, a sua volta, di due momenti:

- la scelta delle unità di analisi;
- il sistema di pesi da adottare.

Tali scelte però non sono assimilabili alla cosiddetta fase di pulizia dei dati tipica di ogni analisi statistica su dati di tipo numerico. La definizione di unità e pesi nell'ambito del TM è da considerarsi parte integrante dell'analisi stessa. Da esse infatti deriveranno i risultati, la loro interpretabilità e soprattutto la loro inerenza con gli obiettivi prefissati. È chiaro quindi che "la determinazione delle unità di analisi e dei pesi deve avvenire immediatamente dopo la definizione degli obiettivi della ricerca ed insieme alla scelta delle tecniche statistiche da utilizzare" (Infante, 2007).

2. I fondamenti statistici all'analisi automatica dei testi

La statistica testuale dispone di un ricco vocabolario tecnico di riferimento sul quale è importante soffermarsi. Il *corpus* è «una collezione di unità di contesto, o *frammenti*, che si ritengono tra loro coerenti e pertinenti per essere studiate sotto un qualche punto di vista o proprietà» (Bolasco 2013). Questa definizione di corpus può essere applicata a fonti testuali più disparate: dalle analisi sulle domande aperte dei primi questionari politici, ai messaggi pubblicitari fino al linguaggio utilizzato attualmente sul Web. Secondo Salem (1994) un corpus deve essere omogeneo, ovvero costituito da testi con caratteristiche lessico-metriche confrontabili e in condizioni di enunciazione simili. I *frammenti* sono «testi, documenti, loro sezioni o semplici frasi, generati sia da testi scritti sia dalla trascrizione di discorsi orali» (Bolasco 2013). Ogni frammento costituisce il livello elementare del corpus testuale di riferimento, il quale è composto da unità elementari definite, arbitrariamente, parole

Il *dato statistico* che viene rilevato è il numero di volte in cui una unità lessicale si è presentata nella raccolta d'esame: nell'analisi testuale viene definito *occorrenza*. Vien da

sé che l'occorrenza rimanda al concetto statistico di *frequenza;* nel momento in cui è necessario effettuare dei confronti tra corpora di ampiezza differente è necessario utilizzare le *occorrenze normalizzate*, ovvero delle frequenze relative ottenute dividendo le occorrenze di ogni parola per una quantità data, definita dalla grandezza del corpus.

L'insieme delle parole individua il vocabolario del corpus. Tale insieme è rappresentato da una lista, in cui a ciascuna parola è associato il numero di occorrenze. L'ampiezza del vocabolario V è definita dal numero di parole presenti nel testo:

$$V = V_1 + \ldots + V_k + \ldots + V_{f(max)}$$

dove V_1 rappresenta il numero di parole che appaiono una sola volta nel testo, ossia l'insieme degli hapax, V*k* è il numero di parole che si presenta k volte, e $V_{f(max)}$ esprime il valore delle occorrenze della parola con il maggior numero di occorrenze del vocabolario.

La scelta dell'unità di analisi determina le modalità di trattamento dell'informazione testuale. L'unità elementare del linguaggio è la parola, nell'ambito della statistica testuale è definita come una forma grafica, una «catena di caratteri di un alfabeto delimitata da due separatori» (Lebart, Salem 1988 p.28). L'operazione di riconoscimento nel corpus delle forme grafiche determina una perdita di informazione sul significato, il contesto e lo stile comunicativo. Il problema della scelta dell'unità di testo si verifica quindi nel decidere quale tipo di riconoscimento adottare. Una scelta opportuna è quella di far riferimento alle unità minimali di senso, in modo da limitare le ambiguità. Le unità di senso possono essere tanto delle forme grafiche, quanto dei segmenti di testo che esprimono contenuto autonomo. I segmenti ripetuti sono disposizioni di *p* forme che si ripetono più volte nel corpus, e possono essere vuote, cioè solo formate da parole grammaticali, oppure caratteristiche se costituiscono unità di senso indipendenti. Nell'ottica della statistica testuale è opportuno considerare come unità elementare

d’analisi la forma testuale, ovvero una componente significativa minima del discorso non soggetta ad ulteriore scomposizione (Bolasco, 2013).

Una volta selezionata l’unità d’analisi, è necessario ricorrere alla fase di pre-trattamento dei documenti, in modo tale da poter effettuare delle successive analisi statistiche.

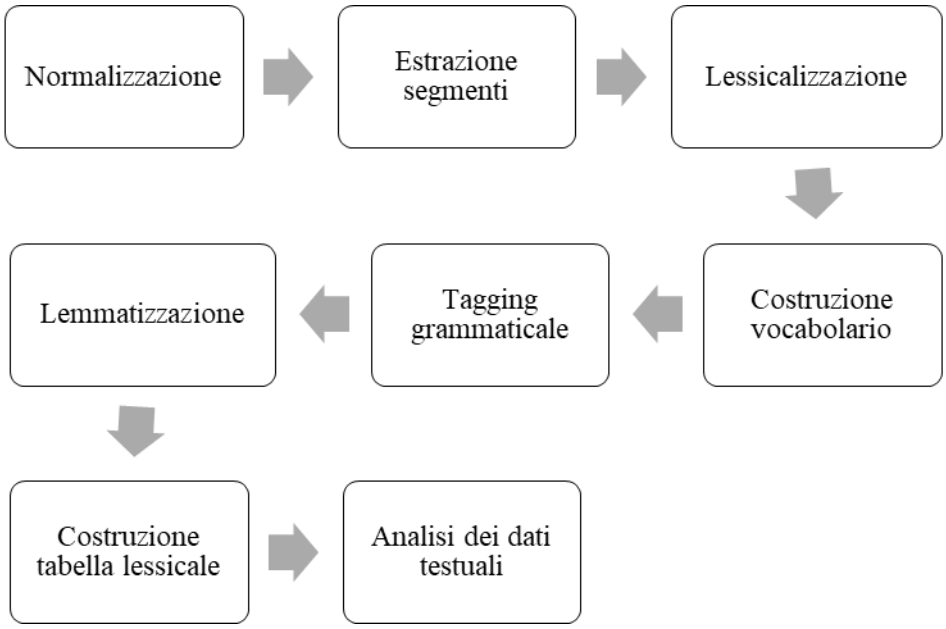

Figura 3 – Processo di pretrattamento del dato testuale

Le operazioni presentate nella figura 2 sono le fasi di pretrattamento del corpus. Queste sono fondamentali perché strutturano il contenuto testuale e permettono il processamento statistico. Il primo passo è individuare le successioni di caratteri dell’alfabeto, comprese tra i separatori attraverso una procedura di *parsing*, la quale consente di selezionare i testi e di ridurli ad un elenco di forme grafiche. Un documento appartenente a una data raccolta può essere visto come una sequenza di caratteri. Questi ultimi, infatti, vengono analizzati e *tokenizzati*, ottenendo un insieme di stringhe distinte (*token*) separate da spazi, segni di punteggiatura o, a seconda del particolare fenomeno analizzato, altri tipi di caratteri (es: hashtag). Successivamente è possibile ridurre i caratteri di tutti i termini in minuscolo, questa operazione è detta di *normalizzazione*. Per tenere traccia della varietà del linguaggio e procedere in modo uniforme nell'intera raccolta, è possibile eseguire anche

altre operazioni di normalizzazione come eliminare caratteri alfanumerici. È anche possibile eliminare le *stopwords*, parole che sono utili per comporre una frase di senso compiuto ma che, prese da sole, non danno alcuna informazione (es. preposizioni e articoli). Questi token vengono detti *parole vuote* o *stopwords* poiché non danno alcuna informazione aggiuntiva.

Qui inizia la fase di strutturazione del dato e conversione in valore numerico. Le scelte del ricercatore in questa fase determinano la natura del dato e di conseguenza la scelta di tecniche e la validità dei risultati ottenuti. Gli approcci più comuni per la codifica numerica delle parole sono il Bag of Words (BOW) e il Part of Speech (POS).

L'approccio BOW rappresenta la modalità più rapida e computazionalmente più semplice per effettuare la codifica numerica del dato testuale. Prevede la fase di lessicalizzazione, questa opera sulle unità minime di senso definite all'interno del disegno di ricerca. Questa fase permette la costruzione del vocabolario delle forme testuali. Le singole parole presenti all'interno di un documento sono raggruppate in un vocabolario senza impostare l'ordine di ogni parola (come volessimo conservarle in un sacco), le occorrenze di ogni singola forma\lemma sono calcolate per ogni segmento, questa operazione genera un vettore numerico il quale contiene questi conteggi. Ogni vettore ha la stessa lunghezza, che equivale alla dimensione del vocabolario. Di seguito è presente uno schema che illustra il processo (figura 4), nell'esempio ho utilizzato un breve estratto da *A Tale of Two Cities* di Charles Dickens. Ho diviso il periodo in 4 testi, tagliando ad ogni virgola.

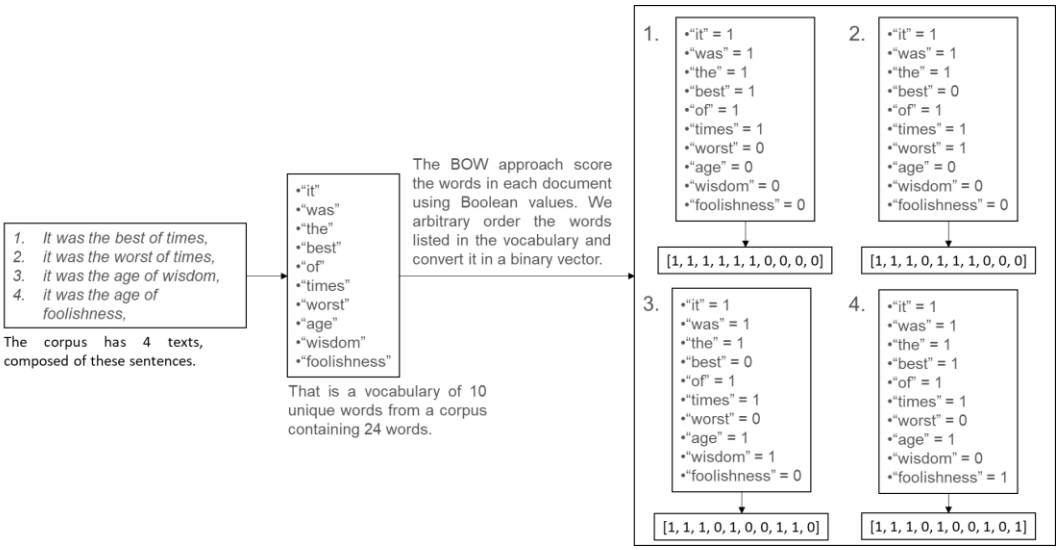

Figura 4 – *BOW*, schema riassuntivo

Il BOW ha una serie di debolezze che risulta importante elencare:

- Il vocabolario: il vocabolario richiede una fase di progettazione molto accorta e delicata, che ovviamente prende molto tempo all'analisi e risente dell'intervento della soggettività del ricercatore. Le caratteristiche di questo *oggetto*, come la dimensione e le occorrenze dei lemmi dei documenti, possono avere un grande impatto sulla rappresentazione e analisi del corpus, faccio riferimento ai casi di elevata sparsità e presenza di documenti vuoti alla fine del processo
- La sparsità: come precedente detto, la rappresentazione sparsa della matrice documenti per lemmi impatta negativamente sui nostri modelli sia per motvi computazionali (legati ad un dispendio maggiore di potenza e tempo) sia per motivazioni legate alla realizzazione di un'informazione che apporti una conoscenza solida e adeguata.
- Significato: molto spesso, agendo direttamente sui lemmi, il ricercatore è portato ad effettuare drastici tagli di frequenza sulle parole, questo, se non accompagnato da uno studio attento può determinare un'importante perdita di significato e di informazione all'interno del nostro corpus.

L’approccio POS prevedere di operare un *tagging grammaticale.* Un processo di marcatura di ogni parola in un testo una particolare parte del discorso. Questa fase risulta essere centrale perché permette il riconoscimento del POS (Part of speech) funzionali all’individuazione delle categorie delle parole. È possibile riconoscere, a tal proposito, due gruppi:

- Le categorie lessicali, che rappresentano la classe più numerosa e sono in costante formazione ed aggiornamento in quanto, al loro interno, vi è un processo continuo di acquisizione di nuove parole;
- Le categorie funzionali, le quali presentano un numero di elementi limitato ma caratterizzati da un utilizzo inedito all’interno di una grammatica;

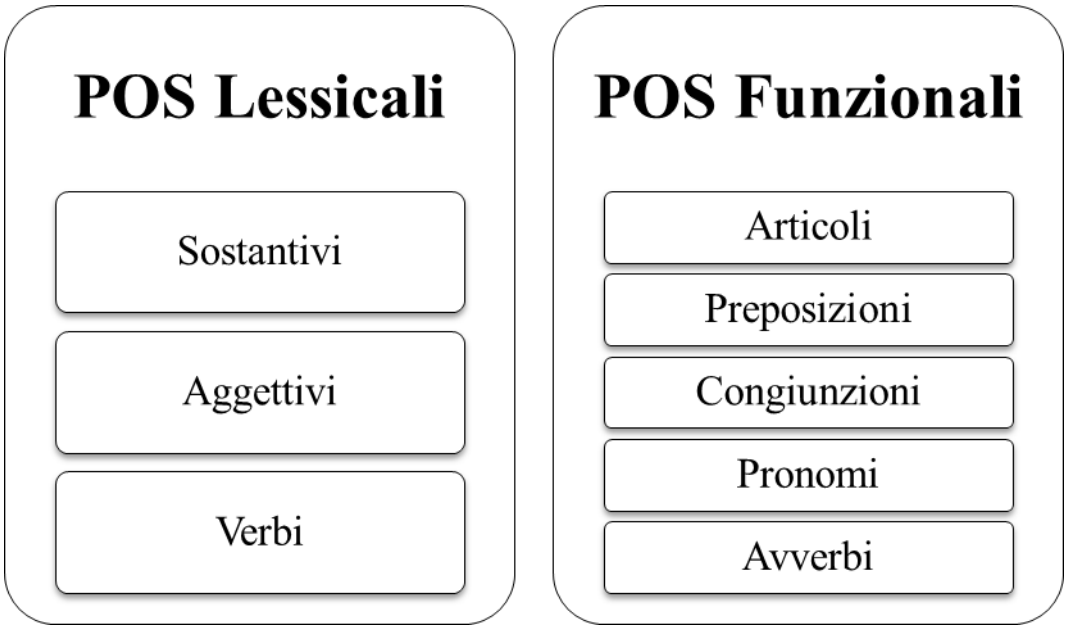

Figura 5 -: Classificazione del part of speech

La fig. 5 mostra 8 differenti categorie di POS, le quali dipendono dalla lingua a cui afferiscono e sono relative ai processi morfologici, quali la formazione del plurale da una parola singolare o del passaggio dal femminile al maschile.

Quando si definisce la POS di appartenenza, si procede con la fase di lemmatizzazione, la quale prevede di ricondurre ogni forma flessa al lemma di appartenenza. Per lemma si

intende la «forma base o forma canonica con cui una parola è citata in un dizionario della lingua (come entrata della voce o articolo della lista)» (Bolasco 2013). Nell'analisi testuale ciò indica che si considera l'infinito per le forme verbali, il singolare per i sostantivi e il singolare maschile per gli aggettivi. La lemmatizzazione dipende dall'identificazione corretta della parola all'interno del contesto in cui è inserita. Un'ulteriore step che può essere svolto se non si vuole ricorrere alla lemmatizzazione è lo stemming, la procedura di riduzione delle parole flesse (o talvolta derivate) alla loro parola radice, base o radice, generalmente una forma verbale scritta. La radice non deve essere identica alla radice morfologica della parola; di solito è sufficiente che parole correlate mappino alla stessa radice, anche se questa radice non è di per sé una radice valida. Gli stemmer sono in genere più facili da implementare ed eseguire più velocemente. La caratteristica principale di questa procedura è che opera su una singola parola senza la conoscenza del contesto, e quindi non può discriminare tra parole che hanno significati diversi a seconda della parte del discorso. La "precisione" ridotta potrebbe non avere importanza per alcune applicazioni. Tuttavia, «Anche se gli stemmer sono più facili e veloci da implementare, è preferibile usare i lemmatizzatori per conservare il ruolo sintattico di ogni termine e migliorare la leggibilità dei risultati» (Misuraca, Spano 2020).

È possibile procedere alla costruzione del vocabolario lessicale anche utilizzando la procedura POS, grazie ad una serie di trasformazioni sugli oggetti creati nell'ambiente di analisi. La procedura prevede la sovrapposizione dei token/termini identici e il conteggio del numero di occorrenze di ogni voce del vocabolario nella collezione dei documenti.

Concluse queste fasi è possibile calcolare un indice che permette di definire in maniera veloce ed efficace la variabilità lessicale: il *type/token ratio (TTR*). Queto è definito come il numero totale di parole uniche (type) diviso per il numero totale di parole (token) in un dato segmento di linguaggio. I valori di questo rapporto oscillano tra 0 e 1: valori vicini

allo 0 indicano che il vocabolario del testo non è molto grande, e quindi non molto vario; il valore massimo 1 si ottiene quando la grandezza del vocabolario è pari alla lunghezza del testo, ovvero quando il testo è interamente formato da hapax, cioè da parole che occorrono una volta sola all'interno del corpus. Alla fine di questo processo, il contenuto dei documenti è pronto per essere processato sottoforma di dato strutturato.

A questo punto è possibile far congiungere i due approcci, utilizzando il modello dello spazio vettoriale (Salton et al., 1975) per rappresentare un documento da un punto di vista algebrico. La successiva fase di codifica pone all'analista delle scelte circa il peso da attribuire alle singole parole all'interno delle parti del testo. Dopo aver operato una trasformazione dei testi tramite uno degli approcci presentati, è possibile organizzare i documenti in vettori in uno spazio multidimensionale. Consideriamo una collezione D contenente n differenti documenti. Ogni documento di può essere visto come un vettore (wi1, wi2…wij…wip) in uno spazio vettoriale Rp formato dai termini appartenenti al vocabolario. Quando un termine j compare nel documento, il suo valore nel vettore wij non è zero. Quest'ultima quantità può essere vista come un peso che rappresenta l'importanza del termine j in dij, cioè quanto il termine contribuisce a spiegare il contenuto del documento. Secondo Salton e Buckley (1988), tre componenti devono essere presi in considerazione in uno schema di ponderazione:

- Term frequency, utilizzata per esprimere l'importanza relativa dei termini in ogni documento (peso locale);

$$TF = tf_{ij}$$
$$LogTF = log_2(tf_{ij} + 1)$$

- Collection frequency, usata per catturare il potere di discriminazione dei termini rispetto a tutti i documenti (peso globale);

$$IDF = log(\frac{|D|}{df_i})$$

- Normalisation, utilizzata per evitare distorsioni introdotte da lunghezze disuguali dei documenti, dove la lunghezza è rappresentata come il numero totale di token utilizzati in un documento

Una distinzione principale può essere fatta tra schemi di ponderazione dei termini non supervisionati e supervisionati, a seconda dell'uso delle informazioni disponibili sulla composizione dei documenti. Gli schemi di ponderazione non supervisionati includono:

- binary weights: lo schema di ponderazione Booleano è basato sulla presenza o l'assenza di una determinata forma testuale all'interno di un documento: questo sistema assegna valore wij=1 qualora la forma j-esima è presente nel documento i-esimo, altrimenti la forma avrà importanza wij=0

$$Boolean = \begin{cases} 1, & if\ tf_{ij} > 0 \\ 0, & if\ tf_{ij} < 0 \end{cases}$$

- *raw frequency* weights (tf): sono calcolati come il numero di occorrenze di un termine in un documento e corrispondono alla frequenza assoluta

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

- *normalised frequency* weights: definiscono 1/ (max tf_{ij}) come fattore di normalizzazione delle frequenze grezze, dove max tf_{ij} è il più alto numero di occorrenze osservato nel documento d_i. (Misuraca, Spano 2020).

$$w_{ij} = \frac{tf_{ij} \cdot log(\frac{n}{n_i})}{\sqrt{\sum_{d_i} (tf_{ij} \cdot log(\frac{n}{n_i}))^2}}$$

Gli schemi supervisionati di ponderazione dei termini seguono la stessa logica del tf-idf, utilizzando come pesi globali altre misure sviluppate per la selezione delle caratteristiche, come $\chi 2$, gain ratio o information gain (Debole e Sebastiani, 2003; Deng et al., 2004). Una volta che i documenti sono stati pre-processati e convertiti in vettori, è possibile incorporare i dati strutturati in matrici di vario tipo (Misuraca, Spano 2020).

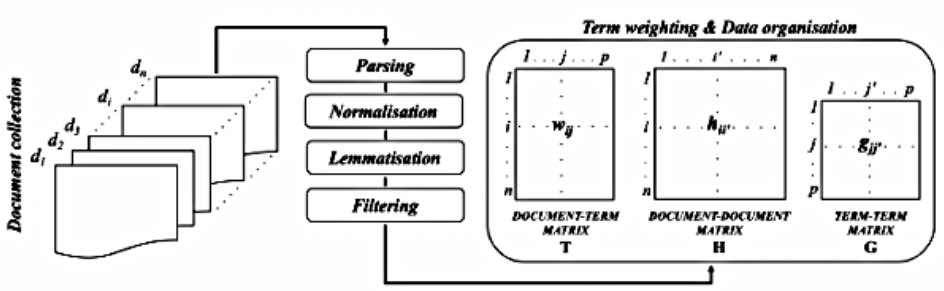

Figura 6 – Processo di creazione del dato strutturato (Misuraca, Spano 2020).

I vettori dei documenti possono essere giustapposti per formare una matrice **T** documento-termine con *n* righe e *p* colonne. Questa matrice è una particolare tabella di contingenza le cui distribuzioni marginali forniscono informazioni differenti a seconda dello schema di ponderazione scelto. La trasposizione della matrice in una matrice termine-documento non cambia la lettura dei dati. In una matrice binaria documento-documento $\mathbf{T_b}$, i totali di riga marginali indicano quanti termini del vocabolario sono usati in ogni documento, mentre i totali di colonna marginali indicano il numero di documenti in cui ogni termine compare. In una matrice di frequenza grezza $\mathbf{T_{tf}}$, i totali marginali di riga indicano le lunghezze dei documenti differenti, mentre i totali marginali di colonna rappresentano la distribuzione dei termini nella collezione complessiva, cioè il vocabolario. Quando si usano altri schemi di ponderazione, i totali marginali non sono direttamente interpretabili. Per esempio, le due distribuzioni marginali di una matrice tf-idf documento-termine rappresentano rispettivamente il tf-idf totale per documento e il tf-idf totale per termine (Misuraca, Spano 2020).

Quando l'interesse della ricerca riguarda l'associazione tra termini, è possibile costruire una matrice quadrata termine-termine **G:**

$$g_{jj'} = \sum_{i=1}^{n} w_{ij} \cdot w_{ij'}$$

L'elemento diagonale gjj′ corrisponde ai totali marginali di colonna di Tb. La matrice di co-occorrenza può anche essere ottenuta dividendo in frasi il documento della collezione e costruendo una matrice binaria frase-termine Sb. Il generico elemento gjj′ di $\hat{G} = S_b T S_b$ può essere letto come il numero di frasi in cui il termine *j* e il termine *j′* co-occorrono e più propriamente può essere interpretato come la co-occorrenza dei due termini nella collezione. È possibile ottenere una granularità differente delle co-occorrenze dei termini cambiando il modo in cui i documenti vengono divisi, ad esempio spezzando i documenti in corrispondenza di segni di punteggiatura più forti come punti, o dividendo ogni documento in frasi. Seguendo la stessa logica della matrice termine-termine, una matrice quadrata documento-documento H può essere derivata per rappresentare l'associazione di documenti a coppie nella collezione (Misuraca, Spano 2020). La matrice documento-documento si ottiene dal prodotto della matrice TbTbT, dove hii′ è il numero di termini condivisi dai documenti di e di′:

$$h_{ii'} = \sum_{j=1}^{p} w_{ij} \cdot w_{i'j}$$

Gli elementi diagonali $h_{ii'}$ corrispondono ai totali marginali di riga di $\mathbf{T_b}$. Come sopra per la matrice termine-termine G, misure differenti possono essere usate per esprimere l'associazione a coppie in una matrice documento-documento **H** (Sternitzke, Bergmann 2009).

2. Metodi per identificare i discorsi d’odio

Il numero di ricerche sviluppate negli ultimi anni riguardo la detenzione automatica e l’esplorazione del contenuto del materiale *hatespeech* è elevato. La letteratura presentata in questa sezione riguarda una rassegna e un’organizzazione dei metodi statistici per rilevare tali i contenuti. Gli approcci sono vari e si basano su diverse logiche e strategie, io propongo un'organizzazione di questi lavori basati su diversi metodi. Una gran parte della letteratura concorda sul fatto che il problema dell'hate speech è un compito di classificazione dei documenti (Schmidt, et al 2017). Zhang e Luo (2018) hanno diviso i metodi in due categorie:

- I metodi classici, questi dipendono dall'elaborazione manuale delle caratteristiche che vengono poi computate da algoritmi basati su metodi statistici come la SVM, la Naive Bayes e la Logistic Regression (Burnap, Williams, 2015; Davidson et al., 2017; Djuric et al. 2015; Greevy, Smeaton., 2004; Kwok, Wang, 2013; Mehdad, Tetreault, 2016; Warner, Hirschberg, 2012; Waseem, 2016; Waseem, Hovy, 2016; Xiang et al, Yuan et al. 2016).
- Metodi basati sul *deep learning*, molti studi recenti si basano su questa tipologia di modelli i quali rappresentano uno dei più recenti paradigmi metodologici. Questi impiegano reti neurali per apprendere automaticamente caratteristiche astratte da dati grezzi con processi computazionali multistrato (Björn, Sikdar, 2017; Nobata et al., 2016; Park e Fung, 2017; Del Vigna et al., 2017).

Il primo gruppo necessita di una fase di progettazione e codifica che trasforma i dati in vettori riconoscibili non solo dalla macchina ma anche dal riceratore, questi vettori sono poi direttamente utilizzati dai classificatori. Le *features* utilizzate nello stato dell'arte sono solitamente varie (Schmidt et al., 2017). Queste includono bag of words, *n-gramms*

di parole e caratteri speciali selezionali appositamente per la rilevazione dell'*hatespeech* (Burnap, Williams, 2015, 2016; Davidson et al., 2017; Greevy, Smeaton, 2004; Kwok, Wang, 2013; Del Vigna et al., 2017, Warner, Hirschberg, 2012; Waseem, 2016; Waseem, Hovy, 2016).

Djuric et al. (2015) hanno proposto un lavoro molto interessante utilizzando i commenti estratti dal portale Yahoo Finance in cui dimostrano che le rappresentazioni delle distribuzioni dei commenti inferite utilizzando il modello *paragraph2vec* (Le e Mikolov, 2014) superano in performance le rappresentazioni *bag-of-words* (BOW) più semplici in un contesto di classificazione supervisionata per il rilevamento dell'hate speech. Nobata et al. (2016) partendo da questo lavoro, hanno migliorato i risultati di Djuric et al. addestrando il loro classificatore su una combinazione di caratteristiche tratte da quattro diverse categorie: linguistiche (ad esempio, conteggio delle parole offensive e aggressive), sintattiche (utilizzando un tagging grammaticale basato sul procedimento POS), *distributional semantic* (ad esempio, *embedding* di parole e commenti) e basate su BOW (n-grammi di parole e caratteri). Tra queste categorie i risultati migliori sono stati ottenuti utilizzando un approccio misto, con tutte le features combinate. Invece, tra le *features* considerate singolarmente, gli n-grammi di caratteri hanno contribuito più alle prestazioni di tutte le altre.

Sono presenti in letteratura anche degli schemi basati su rapporti dicotomici. Alfina et al. (2017), Ross et al. (2017) e Gao et al. (2017) per individuare contenuti di *hatespeech*, da Nithyanand et al. (2017) per contenuti offensivi e da Hammer (2016) per le minacce violente. *Frameworks* che si basano su dimensioni lessicali e semantiche più sfumate cercano di evidenziare caratteristiche particolari, questi in pratica lavorano sulle aree del linguaggio più ambigue e di difficile codificazione. In questi lavori vengono utilizzate delle etichette per la classificazione dei testi inseriti all’interno del modello, in Mathur et al. (2018) usano l'etichetta abusivo, in Mubarak et al. (2017) usano le etichette osceno,

offensivo e pulito. Altri autori, come Olteanu et al. (2018) e Fiser et al. ˇ (2017), usano schemi più complessi ed elaborati. I primi, in particolare, hanno sperimentato uno schema di annotazione basato sul rating, riportando un basso indicatore di accordo. Anche Sanguinetti et al. (2018) utilizza uno schema complesso in cui i contenuti di odio vengono annotati sia per la presenza (valore binario: assente/presente) sia per la sua intensità (scala di valutazione 1-4). Queste strategie forniscono dei preziosi contributi al problema, tuttavia, basandosi su questa procedura che prevede l'aggiunta di un'etichetta ai singoli testi, rendono l'intero processo di molto dispendioso in termini di tempo.

Esistono anche studi che seguono processi di individuazione diversi basati sempre sul rintracciamento di contenuti offensivi e di abuso (Chen et al., 2012; Mehdad, Tetreault, 2016; Nobata et al., 2016), discriminazione (Yuan et al. 2016), a questi ci aggiungo anche contributi interessanti in termini metodologici come quello sul cyberbullismo di Zhong et al., 2016. Oltre allo studio della variabile testuale, altri studi di interesse si concentrano su altre componenti del contenuto pubblicato. Faccio riferimento agli studi dedicati alle menzioni, URL, hashtag, punteggiature, lunghezza delle parole e dei documenti, capitalizzazione (Chen et al., 2012; Davidson et al., 2017; Nobata et al., 2016).

Esiste inoltre una vasta letteratura che lavora su testi non etichettati, andando a rintracciare l'informazione offensiva direttamente all'interno dei vettori testuali che compongono il database. In questa letteratura gli autori ricorrono a rappresentazioni vettoriali di parole a bassa dimensione e dense (non sparse), solitamente analizzate tramite clustering (Warner, Hirschberg, 2012), topic modelling (Xiang et al. 2012; Zhong et al., 2016) , word embeddings (Djuric et al. 2015; Nobata et al., 2016; Del Vigna et al., 2017; Yuan et al., 2016). Gli autori che contribuiscono ad alimentare questo approccio costruiscono vettori di caratteristiche dei messaggi con diverse rappresentazioni di parole, un esempio è la sentiment analysis (Davidson et al., 2017), oppure con informazioni linguistiche con approccio PoS per studiare le stesse relazioni di dipendenza alle *feature*

grammaticali (Zhong et al., 2016). A tal proposito, anche le meta informazioni sono utilizzate, si riferiscono a dati sui messaggi, come l'identità di genere di un utente associato a un messaggio (Waseem, Hovy, 2016) o l'alta frequenza di quelle che nella letteratura anglosasone vengono definite *profanities* nella *storyline* dei post di un utente.

Le variabili associate ai profili utente (ad esempio, numero di amici, follower, sesso, posizione geografica, stato di anonimato, stato attivo vs. non attivo, tra gli altri) hanno dimostrato di essere utili per identificare comportamenti aggressivi e antisociali (Cheng, Danescu-Niculescu-Mizil, e Leskovec 2015; Waseem 2016; Chatzakou et al. 2017; Wulczyn, Thain, e Dixon 2017). Tuttavia, Fehn Unsvåg e Gambäck (2018) hanno dimostrato che queste variabili migliorano solo leggermente la capacità della strategia di classificazione o dell'analisi in generale di rilevare l'hate speech, testando la loro ipotesi su tre set di dati Twitter con un modello di Regressione Logistica. Un altro svantaggio è che queste caratteristiche utente sono spesso limitate o non disponibili.

I metodi basati sul *deep learning* (DNN) utilizzano algoritmi di intelligenza artificiale che si sviluppano su diversi livelli di rappresentazione. Questi corrispondono a gerarchie di caratteristiche o concetti, dove i concetti di alto livello sono definiti sulla base di quelli di basso livello. Possiamo immaginare diverse tecniche basate su reti neurali artificiali organizzate in diversi strati, dove ogni strato calcola i valori per il successivo in modo che l'informazione venga elaborata sempre più completamente. L'input di solito sono i dati grezzi dei testi o una delle codifiche delle caratteristiche utilizzate anche nei metodi classici e l'output è la classificazione dell'hate speech. Gli algoritmi basati su questi metodi possono imparare dai dati in ingresso, grazie alla struttura multistrato. Per questo motivo «questi metodi tipicamente spostano la loro attenzione dall'ingegneria manuale delle caratteristiche alla struttura della rete, che è accuratamente progettata per estrarre automaticamente le caratteristiche utili da una semplice rappresentazione delle caratteristiche di input» (Zhang, Luo, 2018). I metodi DNN sono frequenti in letteratura

(Badjatiya et al., 2017; Björn, Sikdar, 2017; Park, Fung, 2017). Le architetture prevalenti sono la *Convolutional Neural Network* (CNN) e la *Reccurrent Neural Network* (RNN). La CNN è una rete utilizzata per lavorare come estrattore di *features*, la RNN modella sequenze di problemi di apprendimento (Ordóñez, Roggen, 2016). Riguardo al rilevamento o alla classificazione dell'hate speech, le CNN identificano combinazioni di parole o caratteri (Badjatiya et al., 2017; Björn, Sikdar, 2017; Park, Fung, 2017) (ad esempio, frasi, n-grammi), le RNN imparano le dipendenze di parole o caratteri (informazioni ordinate).

Tra i contributi presenti in letteratura che sfruttano questi modelli sono prenti Zhang, Robinson e Tepper (2018), i quali hanno applicato embeddings di parole pre-addestrate e CNN con Gated Recurrent Units (GRU) alla modellazione di cosiddette *long dependencies* tra le *features*. Founta et al. (2018) hanno costruito due classificatori neuronali: uno per il contenuto testuale e un altro per le caratteristiche utente. Con risultati notevoli perché riescono a dimostrare che quando le reti sono addestrate congiuntamente, le prestazioni complessive aumentano. Inoltre lo stesso lavoro riesce ad approfondire la letteratura già presente proponendo un modello neurale molto originale. In primo luogo, utilizza un nuovo modello (biLSTM) per adattare l'incorporamento preaddestrato al dominio degli studi in materia di *hatespeech detection*. In secondo luogo, impiega una configurazione di apprendimento basata sullo sfruttamento di più set di dati per costruire un singolo set di *hatespeech.*

3. Le radici emotive del discorso: sentiment ed emozioni

La revisione che segue nel seguente paragrafo utilizza gli schemi, le logiche e i metodi della letteratura presentata finora inserendo nel framework due fattori fino a questo punto trascurati: il sentiment e l’emozione. Performare una sentiment o una emotional analysis vuol dire effettuare una classificazione dei nostri testi rispetto a determinate categorie emotive. Pertanto è necessario soffermarci su quei metodi che permettono di elaborare i

dati testuali allo scopo di estrarre informazione rispetto alle tipologie indicate. Entrando nel merito della letteratura, la comunicazione online viene analizzata tramite queste procedure per molti scopi, esempi molto calzanti possono essere il marketing (Alalwan , Rana, Dwivedi, Algharabat, 2017; Kapoor et al., 2018), la previsione del comportamento di voto (es, Greco, Maschietti, Polli, 2017; Grover, Kar, Dwivedi, Janssen, 2018), la gestione dei disastri (Singh et al, 2017), le indagini sulle campagne (Afful-Dadzie Afful-Dad zie, 2017), e nella valutazione dei siti web, dell'efficacia delle recensioni dei clienti e delle percezioni dei clienti del digitai marketing (Antonacci , Fronzetti Colladon, Stefanini, Gloor, 2017; Aswani, et al. 2018; Gloor, Fronzetti Colladon, Giacomelli, Saran, Grippa, 2017; Rekik, Kallel, Casillas, Alimi , 2018; Singh, Irani et al. , 2017). In particolare le analisi che hanno come oggetto di indagine la percezione di sentimenti ed emozioni sono sempre più utilizzate nel cosiddetto campo *dell'opinion* mining (ad esempio, Aswani et al., 2018; Ceron, Curini, Iacus, 2016; Gloor, 2017; Hopkins King, 2010; Liu, 2012).

In Greco et al. (2017) l'emotional text mining viene definite come una particolare tioplogia di text mining basata «su un approccio socio-costruttivista e su un modello psicodinamico, che permette di identificare gli elementi che determinano le interazioni, il comportamento, gli atteggiamenti, le aspettative e la comunicazione delle persone». Così, secondo un approccio semiotico all'analisi dei dati testuali, l'ETM permette di realizzare una profilazione sociale. Questo è già stato applicato in diversi campi che vanno dal dibattito politico, al fine di profilare gli utenti dei social media e anticipare le loro scelte politiche (Greco, Alaimo, Celardo, 2018; Greco, Celardo, Alaimo, 2018; Greco et al, 2017; Greco Polli, 2019), all'efficacia della formazione professionale alla Sapienza di Roma (Cordella, Greco, Meoli, Palermo , Grasso, 2018), alla struttura del cervello (Laricchiuta et al., 2018) e all'impatto del diritto sulla società (ad esempio, Greco, 2016; Cordella, Greco, Carlini, Greco, Tambelli, 2018).

Le tecniche e le tipologie di sentiment analysis che è possibile trovare in letteratura sono molteplici. In questo caso il problema di effettuare una classificazione si associa con l'utilizzo di dimensioni emotive a cui associare le nostre unità.

Per Yu et al (2012) gli approcci alla sentiment analysis esistenti in letteratura sono basati su insiemi di dati da sfruttare come risorse esterne al database di partenza (lexicon e dizionari ad esempio) o sul machine learning. La prima tipologia di analisi è incentrata su liste predeterminate di parole positive e negative, tale polarità del documento in esame è determinata dalla frequenza delle parole che appaiono nel documento. Mentre da un lato questo approccio è molto comune e agevole, presenta limitazioni perché richiede un elevato sforzo computazionale (Short et al., 2008). L'approccio basato sul machine learning prevede solo in training l'utilizzo di un oggetto esterno al corpus testuale analizzato, è poi il modello a operare una classificazione sul automatica del totale delle unità. Nel caso della sentiment I dati in training sono annotati manualmente, a partire da questi vengono successivamente etichettati i restanti. Un'ulteriore classificazione proposta in letteratura è tra lavori che si concentrano su metodi per polarizzare il contenuto delle unità testuali e ricerche che invece si concentrano sulla cosiddetta *feature extraction* (Gruhl et al.2005; Joshi et al.2010). A questo si aggiungono i lavori in materia di *sentiment tracking* che rappresentano uno sviluppo sempre più significativo dell'argomento di studio. A tal proposito sono stati fatti progressi significativi nelle tecniche *public mood detection* direttamente dai contenuti dei social media, come i contenuti dei blog e in particolare i feed di Twitter su larga scala (Gilbert, 2010; Liu et al. 2010; Mishne et al., 2006; Pak et al. 2010).

In Lì et al. (2010) viene approfondito e problematizzato il ruolo del *sentiment score* all'interno dei lavori in letteratura. Gli autori sostengono che obiettivo dell'analisi è associare uno score ai testi o alle parole estratte da questi relativo ad una scala che si muove tra le polarità negativo\ positivo oppure, come nel caso della *emotional analisys*,

tra varie emozioni. Questo indicatore può riferirsi sia ai testi sia alle parole, i metodi adottati possono estrarre il livello emotivo o di sentiment ed esprimerlo tramite un valore numerico. La maggior parte dei processi analitici presenti in letteratura procede con l'assegnazione di uno score alle parole estratte dalla fase di pretrattamento e l'agglomerazione di tale misura tramite una specifica funzione. Di solito, la funzione è la media o la somma. Queste procedure assegnano questo valore con differenti modalità. Cesarano et al (2006) introduce questa strategia chiedendo a un certo numero di soggetti umani di assegnare un punteggio a documenti che esprimono opinioni, poi applica la *pseudo-expected value word* che si ispira al concetto dei valori attesi. Un'altra strategia chiamata *Pseudo Standard-Deviation Adjective Scoring strategy* è stata introdotta nello stesso articolo. Queste strategie risentono fortemente della soggettività dell'individuo che compila la valutazione. Un'ulteriore via consiste nello sfruttare strumenti messi a disposizione online e che sfruttano database lessicali, un esempio è wordnet per la lingua inglese (Miller, 1990). Kamps e Marx (2004) sono stati i primi a utilizzare questo strumento-. Il loro lavoro si concentra solo sul punteggio degli aggettivi. Due parole di riferimento "buono" e "cattivo" sono state selezionate per indicare la direzione positiva/negativa. Per ogni aggettivo w, le distanze d (w, buono) e d (w, cattivo) possono essere misurate. Si ritiene che gli aggettivi con una distanza più breve da "buono" siano più positivi e quelli che sono più vicini a "cattivo" siano più negativi. L'approccio chiamato PMI-IR (Turney, 2021), *Pointwise Mutual Information-Information Retrieval*, funziona sul presupposto che i termini che co-occorrono frequentemente tendono ad avere un significato simile. Sulla base di questo presupposto, la distanza tra due parole può essere misurata dalla statistica del tasso di co-occorrenza di queste due parole. Per qunato riguarda il machine learning, la ricerca più nota che utilizza l'apprendimento automatico supervisionato è stata condotta da Pang et al. nel 2002. Questi hanno introdotto gli approcci di base di classificazione del sentimento sulle recensioni di film (Turney, 2002)

e hanno presentato un metodo per estrarre le frasi oggettive per migliorare gli esperimenti precedenti nel 2004 (Pang e lee, 2004). Nel lavoro sono stati adottati tre classificatori, Naïve Bayes, la classificazione di massima entropia (ME) e la support vector machine.

Per la sentiment analysis vengono inoltre utilizzate tecniche di clustering. Il processo di clustering mira a scoprire raggruppamenti naturali, e quindi presentare una panoramica delle classi in una collezione di documenti (Salton e Bukley, 1988). L'algoritmo più usato è il k-means (Hartigan, 1985). Questi algoritmi non hanno bisogno né di un processo di training di base né di conoscere in anticipo la classe di un documento. A differenza dei tagging manuali non risente della soggettività del ricercatore che assegna uno score sul sentiment. Tecniche di clustering sono già state utilizzate in studi simili da Hatzivassiloglou e McKeown (1997). Queste erano in grado di determinare l'orientamento semantico degli aggettivi. La clusterizzazione dei documenti consiste nel raggrumento delle unità in due o più cluster, dipende se l'obiettivo della classificazione è una bipartizione in diversi poli o in uno spettro emotivo.

Il topic model è un metodo raramente applicato. Esistono lavori che incorporano il sentiment in modelli come il *probabilistic latent semantic indexing* (PLSI) o la *latent Dirichlet allocation* (LDA). Questi effettuano una modellizzazione sul processo generativo delle opinioni per i testi analizzati (Cesarano, 2010). Punto debole di questa strategia è solitamente l'impostazione su base bipolare del sentiment basato su negativo\positivo. Per superare questo ostacolo sarebbe necessario utilizzare una tecnica che permetta di etichettare i topic sulla base di informazioni esterne a quelle estratte dal corpus.

Negli ultimi anni si è sviluppata una nutrita letteratura che applica la sentiment analysis seguendo l'approccio basato sull'utilizzo di lexicon. Usando questi dizionari, la polarità delle parole può essere considerata nelle fasi di *word embedding* (Naderalvojoud Sezer, 2020). La scelta di un lexicon in riferimento ad un campo preciso è guidata da una serie

di scelte ragionate orientate ad ottimizzare il processo di analisi. I dizionari più basilari difficilmente riescono a cogliere molte delle sfumature lessicali presenti all'interno dei testi, specialmente quando si lavora con elevata ambiguità lessicale e di senso, per questo motivo molti di questi algoritmi includono l'estensione automatica dei lexicon utilizzati (Wang et al., 2021). Questo procedimento viene solitamente integrato con alcuni metodi specifici tra cui il Naive Bayes, l' *attention model* e lo sfruttamento di simboli presenti nel testo che aiutano a calcolare lo score emotivo delle nuove parole incluse (Kiichi et al., 2019; Li, Li, Jin 2020). I contributi che sfruttano questi ultimi elementi sono molti in letteratura, considerando come elementi dell'analisi anche emoji e interi segmenti della frase che fanno riferimento a modi di dire o a *slang* (Huang et al., 2018) si conferisce una maggiore profondità emotiva al lexicon utilizzato e si possono sfruttare più categorie emotive, superando la polarità tra negativo e positivo (Tago Jin, 2018), in questo modo vengono riconosciute espressioni di emozioni più fini (Cao et al., 2021; Wang Guo, 2020). Ricapitolando, questi metodi utilizzano lexicon per estrarre specifiche features su cui andare a modellare l'analisi. Il vantaggio di questo approccio sta nell'essere facilmente interpretabile e gestibile, molti di questi processi non hanno bisogno di conoscenze statistiche molto avanzate e generano risultati facilmente comprensibili. Il limite sta nell'utilizzo di un insieme definito di parole come riferimento principale per l'analisi, questo implica una forte dipendenza dei risultati dal dizionario utilizzato.

In questa categoria di metodi la *support-vector machine* (SVM) rappresenta in molti casi il fulcro metodologico dei modelli proposti. Che si tratti di un modello basato su SVM multi-core (Peng et al., 2017; Peng et al., 2019), o di addestrare un modello SVM con un set di dati multidimensionale (Aurangzeb et al., 2021), o di usare SVM per fondere campioni di diversi tipi di voce e testo (Atmaja Akagi, 2021), esiste un approccio comune di base. Esistono studi che valutano positivamente questo approccio alla senitment analysis, questi lavori sviluppano benchmark delle performance computazionali (Asghar

et al. 2020). I metodi di ricerca basati sul machine learning si fondano sulla selezione di features emotive come le parole del testo (Halim et al., 2020) ma anche simboli o le emoticon (Ullah et al., 2020). Il machine learning può eseguire in successione funzioni di apprendimento automatico basate su specifici indici estratti dai testi (TF, TF-IDF) o basate su vocabolari dati in training (parole positive e negative, connotazione) per migliorare la precisione della sentiment analysis (Keerthi et al., 2019). La combinazione di più classificatori di base (Ghanbari-Adivi Mosleh, 2019) può migliorare la classificazione in uscita dal modello. Il vantaggio di questi metodi è che riescono a lavorare agevolmente su più dimensioni emotive associate ai testi e filtrare le parole più centrali garantendo un'interpretabilità maggiore. La loro limitazione si basa sull'estrazione del sentiment score su queste parole, avendo comunque bisogno di un lexicon o comunque di un set di dati da cui partire. Gran parte del loro lavoro si basa sulla conoscenza a priori delle caratteristiche emotive di una parte dei dati.

Una classe di metodi utilizzata per analizzare dataset di grandi dimensione sono le reti neurali. Queste consistono infatti non solo in un metodo ma anche in una vera e propria tecnologia, la quale ha dimostrato avere una certa compatibilità anche con la sentiment analysis (Liu Shen, 2020; Ronran et al., 2020). Mentre nei capitoli precedenti ho presentato una rassegna che vede l'utilizzo di questi metodi nell'analisi testuale automatica, ora mi concentrerò nello specifico su questo tipo di analisi. Uno dei modelli più ricorrenti in letteratura è la *long short-term memory* (LSTM). Molti studiosi utilizzano la LSTM come nucleo operativo per effettuare questo tipo di studi, lavorano attraverso una *dynamic decomposition* in cloud e su sistemi distribuiti per ottenere un processamento più rapido (You et al., 2020), esistono anche proposte di modelli ibridi in cui LSTM o BiLSTM vengono combinati con CNN per costruire un modelli più efficienti (Jang et al., 2020; Wu et al., 2021). Questa tipologia permette di creare profilazioni e analisi sulla personalità che mirano ad esplorare anche le caratteristiche motivazionali dietro ai

comportamenti online. Questi ibridi presentano molte applicazioni. In Wang, Yu et al., (2020) il *region based CNN* prevede il livello di emotività espresso all'interno di un testo, in Li et al. (2018) il *conditional random fields* (CRFs) aiuta a discriminare i componenti di entità regolari ed entità irregolari. Il vantaggio di questi metodi è che non si basano sulla conoscenza preliminare delle caratteristiche emotive del testo, e possono cogliere la semantica emotiva nel testo mentre addestrano i dati attraverso la rete neurale. Lo svantaggio è che sono tutti modelli a scatola nera, ed è difficile spiegare le caratteristiche emotive del testo apprese.

4. HateViz, una dashboard per studiare il fenomeno dell'odio online e della violenza di genere

Nella fase di pianificazione del progetto di ricerca oggetto della mia tesi ho deciso di inserire uno studio preliminare dedicato alla percezione della violenza di genere sulle community online. L'obiettivo è arricchire la rassegna teorica con le opinioni degli utenti di una community riguardo il fenomeno oggetto di studio. In questa occasione ho avviato il progetto *HateViz.* Il cui obiettivo è quello di crescere e riuscire ad implementare la strategia sopra presentata in un'unica applicazione web.

Per lo studio preliminare è stato selezionato il social network Twitter, la scelta è stata orientata dal facile accesso alle API e alla natura della piattaforma, luogo virtuale privilegiato per lo studio delle percezioni e delle opinioni degli utenti su molti fenomeni.

Twitter è un sito di microblogging in cui gli utenti registrati possono trasmettere e leggere aggiornamenti di stato di lunghezza fino a 280 caratteri, noti come tweet. Un considerevole volume di dati generato dagli utenti attualmente attivi rende Twitter una ricca fonte di dati. Sul social network sono molto comuni espressioni sessiste, espresse con diversi livelli di aggressività, molte donne sono state oggetto di abusi e minacce misogine facilmente amplificabili, considerato il volume del flusso di contenuti generato sul social. Di questi molti sono status condivisi con intenti ironici o addirittura con l'esplicita volontà di marginalizzare il pubblico femminile della community. Molti degli hashtag sessisti sono anche divenuti virali all'interno di Twitter, promuovendo stereotipi o ostilità nei confronti delle donne ma anche controllo maschile sulle loro vite e di sottomissione all'autorità maschile. Si arriva fino ad aggregatori più spaventosi che promuovono e mitizzano lo stupro e disumanizzano le donne come oggetti la cui unica funzione è il sesso. Alcuni esempi sono: #LiesToldByFemales, #IHateFemalesWho, #RulesForGirls, #MyGirlfriendNotAllowedTo, #ThatsWhatSlutsDo, #ItsNotRapeIf. La

presenza di questi sentimenti misogini è un ulteriore motivazione della decisione di utilizzare il social network come campo per la ricerca preliminare.

La ricerca è stata condotta mediante l'utilizzo di uno strumento finalizzato a monitorare la percezione del fenomeno della violenza di genere e della discriminazione rivolta alle donne. In particolare, gli obiettivi consistono nell'analizzare i flussi di contenuti condivisi sul social network e mostrare i principali temi di discussione interni inerenti ad uno specifico fenomeno. Individuare una metodologia di analisi quantitativa utile a cercare nuove e migliori interpretazioni della struttura delle relazioni sociali che si instaurano tra i soggetti di una community come quella del social network di microblogging rappresenta un punto di partenza per individuare meccanismi, anche latenti, da parte degli individui, che possono portare ad innescare fenomeni discriminatori nei confronti della figura femminile. La piattaforma proposta con il nome di HateViz può quindi fornire un contributo quantitativo ed esplorativo per l'interpretazione e l'individuazione delle strutture sociali di quelle che sono le diverse modalità di interazione che si creano tra gli attori della community. La dashboard è stata sviluppata cercando di fornire un layout intuitivo per una semplice user experience.

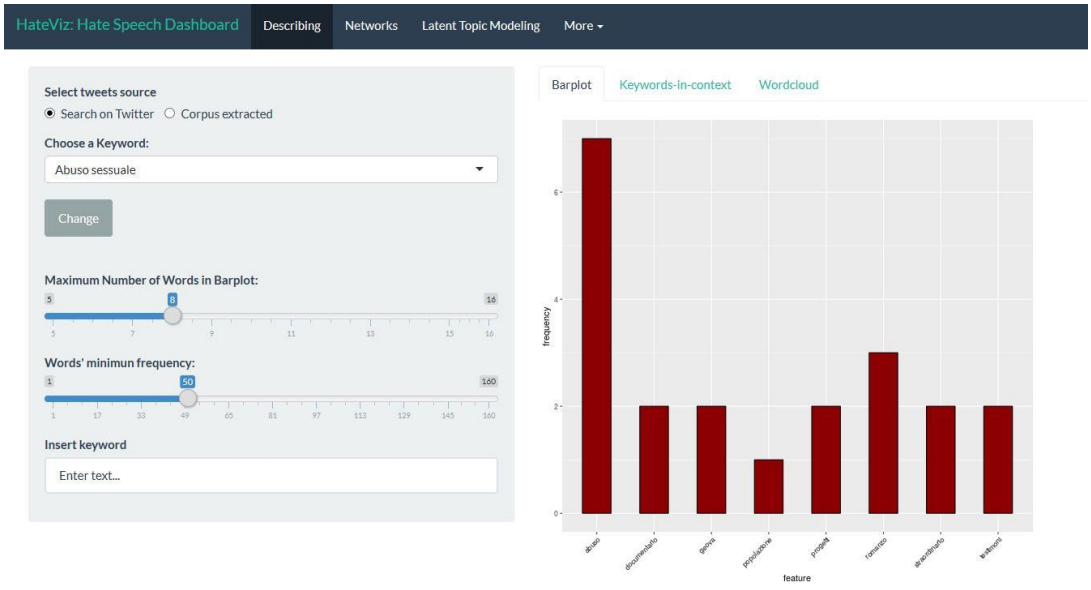

Figura 7 – *User inteface* della dashboard

L’applicazione è stata sviluppata in Java nell’ambiente R, utilizzando il pacchetto shiny è possibile sviluppare e pubblicare delle applicazioni dinamiche per la visualizzazione di grafici e dati. Questo strumento è stato progettato con un’interfaccia chiara e intuitiva, con una divisione tra la schermata di controllo a sinistra e quella di output grafico, nella parte destra dello schermo (figura 7).

Il pannello di controllo della dashboard presenta una serie di funzioni cui l’utente può ricorrere per modificare a suo gradimento l’output grafico. Un primo comando che l’utente può selezionare è la fonte di dati (figura 8), può scegliere se estrarre i dati direttamente dallo stream del social network tramite una ricerca o utilizzare dei dataset precedentemente caricati. In basso è presente il menù a finestra in cui è possibile selezionare il dataset di riferimento, potendo scegliere tra le diverse keyword analizzate durante il progetto: Abuso sessuale, Femminicidio, Molestia, Stalking, Stupro, Violenza domestica, Violenza (di) genere, Violenza sessuale.

Figura 8 – Selezione dei dataset precompilati

Oltre a queste keywords, all’interno del menù è presente anche la voce “all data” che consente di analizzare l’intero dataset di oltre 400mila tweet; la pesantezza del dataset non permette però di poterne effettuare l’analisi poiché, al momento, il server cui si appoggia la piattaforma non sostiene la computazione di una quantità così elevata di dati.

HateViz unisce tre metodi di analisi: il textual mining, la textual network analysis, la declinazione testuale delle reti sociali, e il latent topic modelling. Per ognuno di questi metodi sono previste delle diverse tecniche per l'analisi e la rappresentazione dei dati. A partire dai corpus estratti la dashboard effettua delle operazioni di pretrattamento. Da questo processo si genera una matrice Documenti-Termini (DT, tweets in riga e termini in colonna), a cui sono state rimosse sia le parole "sparse" ossia quelle parole che apparivano isolate sia i documenti vuoti. La matrice DT è il core della dashboard: permette di rappresentare il corpus dei tweet grazie all'utilizzo delle più comuni misure descrittive e grafiche. Queste rappresentazioni, forniscono una lettura immediata ed intuitiva di quelli che sono i termini più utilizzati nella community tra quelli presenti nel corpus. Considerata la sua struttura a due dimensioni, la DT può essere trattata come una matrice di affiliazione. In questo modo è possibile generare i textual network, individuando sia le relazioni tra documenti e parole (occorrenze) ma anche quelle tra le parole stesse (cooccorrenze). Questa parte dell'analisi attraverso la visualizzazione consente di cogliere ulteriori connessioni semantiche che in una lettura preliminare non possono essere individuate.

La possibilità di personalizzare una serie di parametri tra cui il numero di parole da plottare all'interno dei grafici e la soglia di occorrenza minima permette all'utente di ottenere output in grado di rappresentare le dimensioni semantiche più rilevanti.

La prima analisi disponibile è la misurazione delle frequenze attraverso l'uso dei diagrammi a barre ("barplot"). Il grafico a barre fornisce una indicazione della frequenza con cui un termine ricorre all'interno del corpus; una frequenza maggiore indica una maggiore occorrenza di un termine rispetto agli altri. Un secondo output fornito dalla piattaforma è la Wordcloud. Questa rappresentazione, molto utilizzata da utenti con una bassa literacy statistica, fornisce una lettura immediata ed intuitiva di quelli che sono i termini più utilizzati nel set di dati tra quelli presenti nel corpus. Le dimensioni delle

parole vengono definite dal numero delle occorrenze di tale termine; le parole di dimensioni maggiori indicano una ricorrenza maggiore.

La terza analisi disponibile all'utente è la Social Network Analysis, per ottimizzare la lettura del grafo, questo è accompagnato da una serie di misure di centralità che ne permettono una più intuitiva rappresentazione. In più è possibile visualizzare l'ego-network di specifici lemmi tramite una barra di ricerca in cui è possibile inserire la parola da plottare.

Uno degli output disponibili è il grafico delle associazioni tra i termini. L'output mostra i termini selezionati con il corrispettivo grado di associazione ad ogni altra parola presente nella matrice termini-documenti. Il grado di associazione ha un range di variazione tra 0 ed 1; un valore vicino allo zero sta a significare che i termini non compaiono mai nello stesso documento, mentre un valore vicino ad 1 indica che i termini appaiono quasi sempre in coppia all'interno dei documenti. Il calcolo dell'indice di associazione viene effettuato a partire dai documenti; quindi per ogni parola presa in considerazione, che appare in un documento, vengono associati gli altri termini presenti in quello specifico documento. Se il termine ricercato non compare in un documento, questo non viene preso in considerazione.

Un'ulteriore analisi che è possibile selezionare è quella dei Latent Topics. L'ultima analisi disponibile per gli utenti è la Words Network Topic Model, definita anche WNTM (Jiang, Liu Wang, 2018). A differenza di altri metodi di analisi, questa tecnica permette di trattare in contemporanea sia la sparsità sia la densità del reticolo, in quanto la WNTM modella la distribuzione dei topic da ogni termine, invece che estrarre la tematica da ogni documento, in modo da aumentare la densità semantica del dataset garantendo al tempo stesso una discreta sensibilità nel riconoscere anche i topic meno frequenti. Questa tecnica pone le basi sul considerare le co-occorrenze dei termini nel network piuttosto che dai documenti (Zuo, Zhao Xu, 2015).

1. Come leggere gli output della dashboard, un esempio con la keyword "*femminicidio*"

Come precedentemente descritto, la dashboard è stata sviluppata con l'implementazione di gruppo di dataset precompilati già caricati all'interno. Questi sono associati ad un set di keyword, concetti chiave compresi all'interno dello spettro della violenza di genere.

A tal proposito presento un breve caso studio dedicato alla keyword *femminicidio,* tra tutte considerata quella che ha smosso di più lo spirito di cronaca della community. La presente keyword raccoglie tutti i contenuti che indicano una propensione a discutere dell'argomento, quindi anche l'interesse che la comunità in oggetto ha nei confronti del fenomeno.

Questa rappresenta un quadro di come la community vede e interpreta il fenomeno oggetto di studio e permette di mostrare le funzionalità dello strumento utilizzato. La costruzione del dataset analizzato è stata effettuata estraendo i tweets tramite API (Charu et al. 2015; Puschmann, 2017) selezionando la keyword "femminicidio". I testi sono stati pubblicati nel periodo tra luglio 2018 e maggio 2019, il numero è pari a 78501.

La figura 9 mostra la wordcloud ed il barplot, in cui sono presenti i termini con una maggiore ricorrenza. Nella wordcloud, le dimensioni delle parole vengono definite dal numero delle occorrenze di tale termine, dunque una maggiore dimensione indica un termine con un uso più frequente.

Le principali tematiche sottese alle parole messe in risalto dalla wordcloud riguardano l'ambiente politico e l'aspetto legislativo in cui si inserisce il contesto della violenza di genere. Un altro aspetto importante che viene evidenziato riguarda la relazione tra vittima e carnefice, sottolineato dai diversi termini che riconducono alla violenza in ambito famigliare.

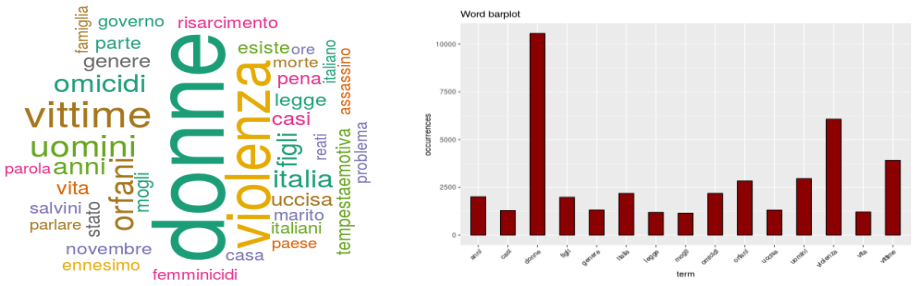

Figura 9 – Parole più frequenti

Dalla rappresentazione del network e dall'analisi delle misure di centralità è possibile effettuare alcune osservazioni. Nel grafo sono utilizzate alcune funzioni che permettono di identificare la forza del legame che unisce due nodi della rete, rappresentato attraverso lo spessore dell'arco di collegamento. Una coppia di nodi connessa attraverso un legame debole avrà quindi un arco più sottile rispetto ad una coppia di termini che condividono un legame più forte, che presenterà un collegamento dallo spessore maggiore.

Guardando la disposizione dei termini all'interno del network (figura 10) è possibile fare alcune osservazioni. La posizione particolarmente centrale all'interno del grafico di termini come "donne", "uomini", "omicidi" e "vittime", indica un forte legame e una importante connessione relazionale di questi lemmi con la maggior parte degli altri termini; ciò significa che queste parole sono il cuore dei testi analizzati. Guardando il network più in generale è possibile osservare tre aree in cui le parole si dispongono con diverse densità. È presente il gruppo centrale, descritto in precedenza, un gruppo di termini che si pone all'esterno del reticolo e un gruppo intermedio di parole. Questi tre livelli si differenziano non solo per la posizione più o meno centrale all'interno del network ma anche per il numero di archi che li connettono. Il gruppo esterno presenta un numero minore di legami rispetto al gruppo centrale; gli stessi archi che collegano i termini più marginali della rete sono caratterizzati da legami relazionali più deboli rispetto alle relazioni che si instaurano tra i termini più centrali del network.

Anche da questa analisi emerge la dimensione relazionale tra l'individuo che commette violenza e la vittima. Termini come "marito", "ex", "moglie" legano la violenza alla sfera domestica e sentimentale, confermando la letteratura precedentemente osservata che vede molti dei casi di violenza di genere legati a crimini passionali.

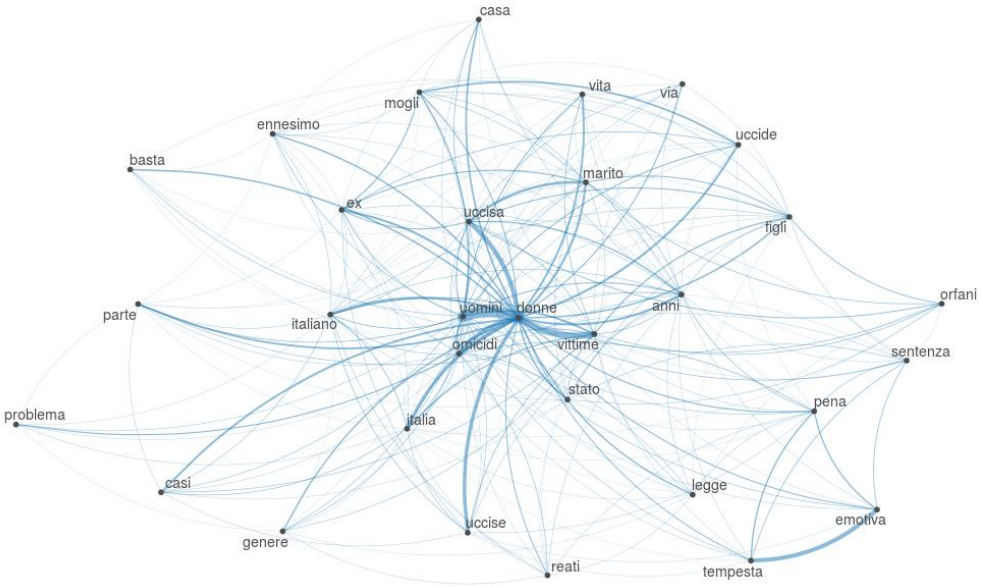

Figura 10 – Network delle cooccorrenze

Dopo aver analizzato la rappresentazione del network, è fondamentale trovare conferme attraverso le principali misure di centralità che definiscono la posizione dei nodi del network in base all'importanza relazionale che la parola assume rispetto agli altri termini. All'interno della dashboard è presente una sezione in cui vengono fornite alcune delle principali misure di centralità.

	DEGREE	***BETWEENNESS***
ABUSO	11	0.1260
STATO	11	0.1260
MINORE	11	0.1260
ANNI	11	0.1260
PERSONE	11	0.1260
UOMINI	11	0.1260
DONNE	11	0.1260
VITA	10	0.0625

VITTIME	10	0.0278
ACCUSE	10	0.0278

Tabella 2 – Principali misure di centralità

Con un modello di Latent Topic Modeling sono state estratte quattro tematiche, caratterizzate da termini differenti al loro interno. Le tematiche che emergono mettono in evidenza alcune dimensioni semantiche:

- Topic 1: mostra aspetti di ambito legale e giudiziario, a cui fanno riferimento termini come “reato”, “omicidi” e “pena”;
- Topic 2: si lega alla dimensione politica, a cui fanno riferimento i lemmi “salvini”, “italia”, “italiani”, “casi”;
- Topic 3: con un riferimento alla dimensione culturale del fenomeno per via di parole come “genere”, “violenza”, “donne”, “casa”;
- Topic 4: riferito alla dimensione istituzionale, con i termini “stato” e “governo”.

TOPIC 1	***TOPIC 2***	***TOPIC 3***	***TOPIC 4***
DONNE	vittime	donne	mogli
OMICIDI	italia	violenza	uccisa
CASA	legge	genere	stato
VITA	casi	omicidi	governo
PENA	parte	italiani	orfani
MORTE	salvini	casa	marito
UOMINI	italiani	italia	anni
REATI	assassino	novembre	figli
ESISTE	problema	parlare	risarcimento
TEMPESTA	italiano	ore	ennesimo

Tabella 3 – Parole maggiormente associate ai topics

L’uso congiunto del topic modeling e della textual network analysis permette di costruire un modello che meglio definisce il contenuto e le relazioni tra ogni topic latente. Un topic che prende una posizione centrale all’interno del network rappresenta la principale area semantica identificata nel dataset. Un termine che rappresenta un nodo di

connessione tra differenti topic indica non solo la sua presenza in entrambi i gruppi, ma rappresenta anche una connessione che mette in relazione le diverse aree semantiche associate ad ogni topic (Zuo, Zhao and Xu, 2015).

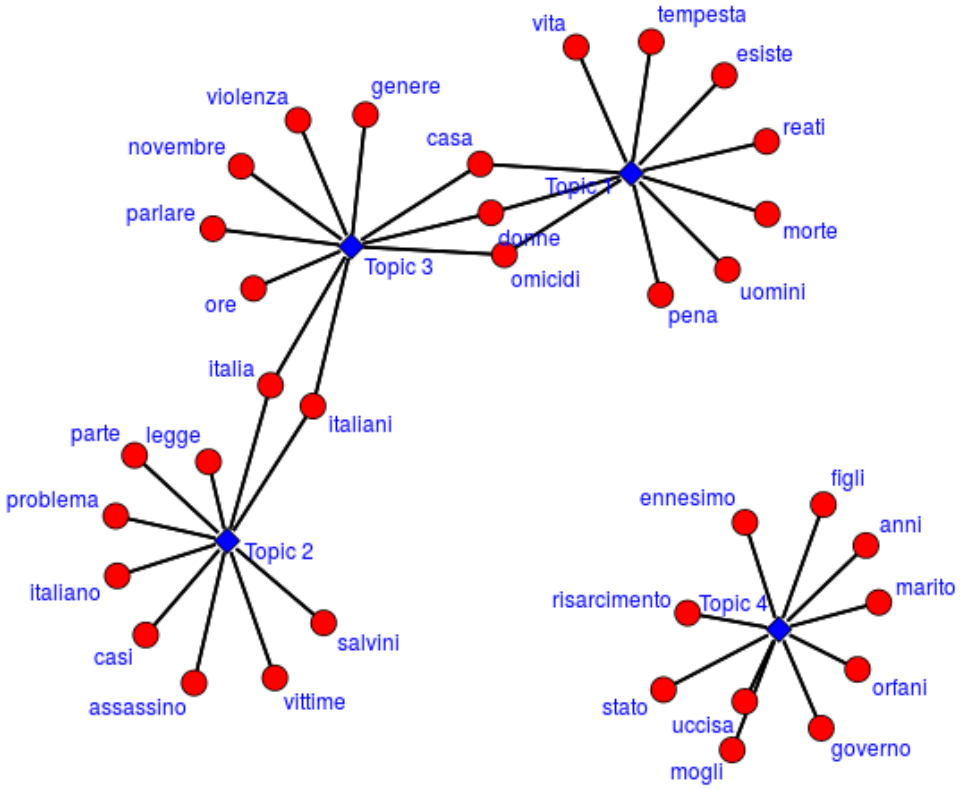

Figura 11 – Network topic e parole

Il network termini-topic (figura 11) analizzato mostra come non ci sia un topic più centrale rispetto agli altri, dunque nei testi analizzati non si trova una tematica più rilevante rispetto alle altre aree semantiche. La posizione del topic 3 (culturale), legata alla dimensione culturale, si pone come "ponte" e collegamento tra i topic 1 (giudiziario) e 2 (politico). In particolare, il topic 1 (giudiziario) e quello 3 (culturale) sono collegati dai termini "casa", "donne" e "omicidi", mentre termini di contatto tra i topic 2 (politico) e 3 (culturale) sono rappresentati dalle parole "italia" e "italiani". Il topic 4 non presenta termini in comune con gli altri, ciò significa che non presenta connessioni semantiche con le altre aree.

2. Conclusioni dello studio preliminare

Dallo studio emerge la forte componente relazionale tra l'individuo violento e la vittima. In particolare, gli utenti rivolgono la loro attenzione alla violenza che avviene all'interno della sfera sentimentale, dove rapporti non sereni all'interno della coppia portano l'uomo ad avere atteggiamenti violenti nei confronti della propria compagna, fino ad ucciderla. Secondo gli studi effettuati gran parte dell'attenzione del pubblico segue le notizie relative alla IPV, confermando la letteratura precedentemente osservata, che vede molti dei casi di violenza di genere essere legati prevalentemente a moventi passionali.

Lo studio mette in evidenza users che commentano gli squilibri della società e che al contempo vanno alla ricerca di una forma di giustizia per riequilibrare una situazione che genera sempre più stupore, stupore dovuto dalla mancanza di una adeguata formazione su come potersi difendere in queste situazioni. La forma di giustizia che viene richiesta dal pubblico riguarda non solo il punto di vista giuridico, dove un violento viene punito per i propri crimini. Soprattutto emerge l'aspetto istituzionale e legislativo; gli user chiedono un intervento profondo da parte di Stati e governi, promuovendo leggi più severe da applicare per i casi di violenza di genere.

Concludo il paragrafo sottolineando che questo ha rappresentato uno studio preliminare, quindi la strategia adottata e descritta in queste pagine differisce leggermente da quella presentata nel paragrafo precedente. Le motivazioni sono relative allo sviluppo del progetto e nel cambiamento di alcuni orientamenti all'interno della strutturazione della ricerca. La scelta di inserire comunque il paragrafo è dovuta alla volontà di dare spazio ad uno strumento che ha rivestito un ruolo molto importante nello sviluppo di uno scenario di riferimento.

5. Una strategia di analisi per identificare le espressioni di odio online

In questo capitolo presento la strategia di ricerca oggetto della tesi. Riprenderò brevemente le fila del discorso sostenuto finora orientandomi alla definizione delle domande di ricerca e delle caratteristiche alla base della mia proposta metodologica. La letteratura di riferimento approfondita nel capitolo primo evidenzia quanto sia importante soffermarsi sugli abusi verbali, in quanto questi costituiscono un danno simbolico al quale può essere associato anche un danno concreto e materiale perché tale espressione costituisce anche un incitamento implicito all'abuso fisico, a intimidazioni e attacchi personali (Matsuda e al., 1992; Mackinnon 1993). Amnesty International (A. I., 2017, 2020) e il Pew Research Center (2014) hanno rilevato che molte utenti donne riferiscono un maggiore disagio emotivo a causa delle molestie online, indicando che le loro esperienze sono particolarmente insidiose. In più molte delle donne (il 41% secondo A.I.) che hanno subito abusi o molestie online ha affermato che, in almeno un'occasione, queste esperienze online hanno fatto sentire la loro sicurezza fisica sotto minaccia. Le esperienze online influenzano il vissuto offline, andando a definire percezioni e visioni del sé all'interno delle società di appartenenza. Ciò che succede è che una parte della popolazione è emarginata, prova disagio emotivo a causa di queste molestie, come sostengono le ricerche citate. In sostanza, l'esperienza online provoca dei risentimenti sulla sicurezza percepita offline. Ciò ha portato ad interrogarmi sulla possibilità di riscontrare all'interno delle dimensioni semantiche emergenti dall'odio online le pratiche che caratterizzano la violenza esercitata nel privato. Nella fattispecie mi riferisco alle dimensioni di coercizione, dominio e aggressività. Alla luce di queste considerazioni, inserite nel più ampio spazio dedicato al tema in termini di valori e contesti sociali, ho sviluppato le domande di ricerca che hanno guidato lo studio condotto:

- QR1: È possibile ritrovare all'interno delle dimensioni semantiche emergenti dall'odio on-line le pratiche che caratterizzano la violenza esercitata nel privato?
- QR2: È possibile risalire alle sub-communities degli haters presenti online tramite lo studio delle dimensioni semantiche che li caratterizzano?
- QR3: In che modo le categorizzazioni delle pratiche della violenza privata si rispecchiano e si differenziano dalle dimensioni semantiche che si trovano online?

Le domande sopra delineate rappresentano la struttura conoscitiva dell'intero progetto, rappresentano i quesiti alla base a cui la strategia messa in cantiere deve rispondere. Per raggiungere tale scopo c'è bisogno in un primo momento di un processo automatico che identifichi le principali tematiche che caratterizzano le discussioni online, classificarle tramite una procedura automatica e identificare le espressioni in cui le dimensioni associate ai contenuti di odi sono prelevanti. È necessaria quindi una strategia di analisi che permetta di rispondere contemporaneamente alle domande qui presentate, fornendo le risposte come frutto di un unico processo analitico. La complessità dei quesiti posti alla base trova la sua origine dalla natura esplorativa degli stessi, tale logica non lascia spazio a semplificazioni concettuali, anzi mira ad individuare dimensioni latenti di significato all'interno dei contenuti posti in esame. Considerate tali premesse, al fine di sviluppare uno strumento in grado di fornire risposte soddisfacenti ai quesiti di ricerca, mi sono posto il seguente obiettivo:

- Sviluppare uno strumento per esplorare le espressioni di odio rispondendo alle tre domande di ricerca.

È stato necessario impostare a priori delle caratteristiche che avrebbe dovuto avere questo strumento per costruire qualcosa che potesse risultare utile e innovativo:

- Versatilità, ossia la capacità di analizzare contenuti estratti da diverse web communities e con vari framework culturali. Questo vuol dire che dovrà poter rispondere in modo ottimale non solo alle diverse caratteristiche assunte dai contenuti pubblicati sulle piattaforme ma dovrà avere la capacità di classificare testi prodotti in diversi contesti culturali.
- Visualization friendly, produrre output grafici accessibili e sintetici. La comprensibilità e la facile accessibilità dei grafici da parte di utenti con un diversificato livello di literacy statistica non può non essere una priorità per uno strumento destinato all'utilizzo da parte di un pubblico diversificato.
- A basso costo computazionale, questa rappresenta una sfida in comune tra tutti gli strumenti che lavoreranno sul web e che quindi avranno a disposizione un server remoto il quale dovrà processare una variabile quantità di dati.

Inserite le domande di ricerca nel framework teorico relativo alle scienze sociali, mi collego tramite queste note riguardo il mio obiettivo di sviluppo al quadro teorico statistico-metodologico. Nel secondo capitolo presento una rassegna della letteratura molto ricca riguardo gli strumenti di text mining, natural language processing (NLP) e hatespeech detection. Trovo superfluo soffermarmi ulteriormente su questi aspetti se non per inserire le componenti e le fasi di analisi da me proposte all'interno di una cornice di riferimento. Volendo trovare una classificazione sintetica della letteratura esplorata possiamo fare riferimento al grafico qui proposto (fig. 12).

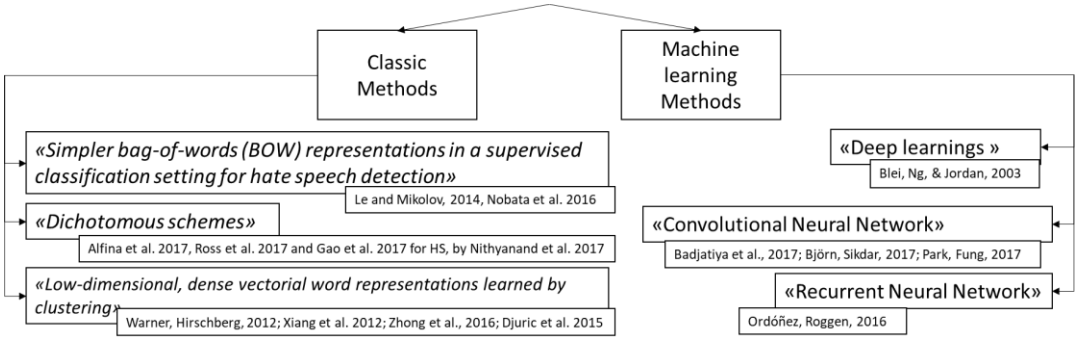

Figura 12 – Hate speech detenction, schema letteratura

La differenza fondamentale sta nella automazione del processo di classificazione, il quale da un lato è effettuato da una macchina che apprende automaticamente, mentre dall'altro risente dell'intervento del ricercatore il quale effettua delle scelte su una base di criteri ragionati e razionali. A questo proposito l'apporto innovativo sta nel proporre una soluzione semi automatica, che integri in un unico processo fasi automatiche e altre che necessitano dell'intervento dello studioso. La mia strategia si compone di tre fasi:1) pretrattamento automatico eseguito tramite una rete con approccio POS; 2) topic model; 3) emotional analysis tramite un'analisi in componenti principali in un sottospazio di riferimento (ACPR). L'innovazione sta nell'integrare all'interno di un'unica strategia - basata su un pretrattamento automatico eseguito tramite rete neurale - tecniche machine learning per l'estrazione di topic con una tecnica di analisi multidimensionale per effettuare la classificazione emotiva. Vengono sfruttati lessici esterni e le relazioni tra le varie dimensioni sono rappresentate all'interno di un piano ortogonale rispetto ai topic emergenti. Recapitolando questi 3 step:

1. Pretrattamento del dato testuale non strutturato eseguito tramite un modello di annotazione
2. Topic model per generare gruppi semantici
3. Emotional analysis tramite PCAR labeling emotivo dei topic estratti e visualizzazione dei risultati

Ogni fase è stata progettata con riferimento alle caratteristiche in obiettivo:

1. Il modello di annotazione è disponibile in più lingue (Versatilità)
2. Il Topic model selezionato non risente della lunghezza dei testi analizzati (Versatilità, basso costo comuputazionale)
3. Con la PCAR è possibile visualizzare agevolmente i risultati del topic model e dell'emotional analysis

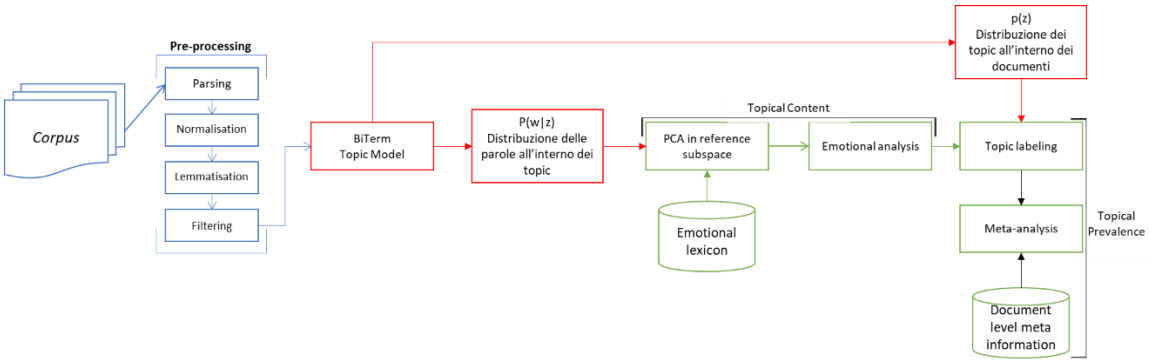

Figura 13 – *Flowchart* strategia

In figura 13 è rappresentato il flowchart che descrive i singoli passaggi della strategia. Lo schema è stato pensato per rispondere positivamente alle caratteristiche poste in obiettivo. Questo è ovviamente un ulteriore apporto dato dal processo rappresentato. Inoltre la capacità di automatizzare alcune fasi e di modellare il content dei gruppi semantici tramite dei dati esterni aiuta a rendere lo strumento più versatile ed efficiente.

1. I Fase: Pretrattamento del dato testuale

La prima fase del lavoro consiste nel pretrattamento dei testi al fine di ottenere un dato strutturato e processabile dagli algoritmi implementati nella mia strategia di analisi. Nello specifico mi serve uno strumento che effettui automaticamente le fasi di normalizzazione, pulizia e annotazione del dataset testuale processato. L'oggetto finale dello step sarà un corpus strutturato in token lessicali.

Lo strumento utilizzato per questo lavoro è UDPipe (Straka et al., 2016), questa libreria permette la creazione di modelli di annotatori automatici. Di seguito illustrerò nel dettaglio il modello alla base di questo algoritmo e le logiche che segue, facendo riferimento ai lavori dell'autore. UDPipe è un'abbreviazione di Universal Dependencies Pipeline, rappresenta un processo automatico di pulizia e normalizzazione multilingua dei testi. L'espressione pipeline si riferisce alla sua struttura a catena, in cui i dati vengono processati secondo un algoritmo organizzato per step consecutivi e concatenati. La prima operazione svolta consiste nella separazione delle singole parole e nella creazione di

stringhe di caratteri che il processo automatizzato potrà riconoscere e processare. Il testo viene oranizzato su diversi livelli, secondo la seguente struttura: documento –> paragrafo –> frase –> token. Una rete bidirezionale GRU a sinolo strato si occupa della segmentazione delle frasi e dello split delle parole in sinoli token, questa predice la posizione e il ruolo di ogni sinolo carattere. Per quanto riguarda i suffissi lessicali, questi vengono riconosciuti come delle regole fisse generate automaticamente che il modello ricorda. Il processo generativo delle regole fa riferimento ai dati in training, il sistema riesce a mantenere questo processo fixando meno di molti altri modelli. Il segmentatore utilizza una strategia per ottimizzare la fase dedicata. Apprende automaticamente le interruzioni dei sinoli segmenti e omette le chiusure di pararafo o frase, perché queste molto spesso sono parte del layout, ciò può portare il sistema in errore portandolo a dividere delle espressioni regolari.

La procedura di tagging grammaticale sfrutta un doppio processo computazionale. Vengono generate delle triplette semantiche per ogni token: UPOS (), XPOS () e FEATS (), la procedura si basa considerando contemporaneamente sia la radice del lemma sia gli ultimi caratteri di ogni parola. In questo caso viene utilizzato per l'esecuzione un guesser che si basa su pattern predeterminati. Mentre un modello di rete neurale, un percettrone, attraverso un set di regole fissate disambigua i risultati della fase precedente. Nell'ambito del machine learning un percettrone è un modello di classificatore binario che mappa i suoi ingressi *x* (un vettore di tipo reale) in un valore di output *f(x)* (uno scalare di tipo reale) calcolato con:

$$f(x) = \chi(\langle w, x \rangle + b)$$

dove *w* è un vettore di pesi con valori reali, l'operatore ‹.,.› è il prodotto scalare (che calcola una somma pesata degli input), *b* è il bias, un termine costante che non dipende da alcun valore in input e *χ(y)* è la funzione di output.

Il processo di lemmatizzazione funziona con una procedura analoga, il modello produce pattern semantici che vanno a definire delle regole fisse e gli UPOS sopra definiti. I primi li ottiene eliminando prefissi e suffissi da una parola, e aggiungendone di nuove. Per generare pattern semantici corretti vengono quindi utilizzati gli ultimi caratteri di una parola ma l'intero prefisso. Come per la fase di tagging, la disambiguazione dei lemmi è eseguita da una rete *percetron*. I processi sopra descritti di tagging grammaticale e lemmatizzazione sono consecutivi e non contemporanei per migliorarne la performance del sistema.

Il modello fin qui descritto è stato selezionato perché risponde meglio alle esigenze della ricerca, risultando più idoneo a soddisfare le caratteristiche in obbiettivo. L'approccio alternativo a questo presentato è il cosiddetto BOW, il quale non è stato selezionato perché non compatibile con le esigenze messe in campo dal progetto. Il ricercatore può infatti selezionare più linguaggi su cui far operare il pretrattamento (lo strumento sviluppato sarà predisposto alle lingue inglese e italiano). Questa caratteristica, unita alla possibilità di automatizzare un processo lungo e dispendioso per l'utente, compensa l'impiego di una forza computazionale maggiore rispetto ad altri possibili scelte.

2. II Fase: BiTerm topic model

La seconda fase della strategia prevede l'integrazione di due diversi approcci di analisi. Il primo step comprende l'estrazione dei topic semantici presenti all'interno del corpora processato, successivamente questi vengono proiettati all'interno di un sottospazio di riferimento definito dalle dimensioni semantiche inserite all'interno della catena di analisi. Per quanto riguarda il topic model, questo ha obiettivo di estrarre informazione sotto forma di gruppi semantici latenti all'interno del corpus. L'approccio all'information extraction è di tipo probabilistico, in questa famiglia vengono inclusi modelli statistici generativi che operano su pattern non osservabili direttamente da una lettura umana dei

testi. Viene allo stesso tempo assegnata una probabilità ad ogni documento rispetto allo spazio generato da questi topic. Questi cluster sono inoltre facilmente interpretabili perché composti dagli stessi termini dei testi, allo stesso tempo questa tecnica risente di una difficolta di rappresentazione grafica degli oggetti in outcome. La logica alla base di questi modelli nasce nel 1999 con un lavoro di Hoffmann con l'introduzione della probabilistic-Latent semantic analisys. A differenza delle tecniche basate sulla scomposizione a valori singolari, la pLSA effettua una decomposizione della matrice lessicale basandosi su distribuzioni probabilistiche. in questo modello la probabilità è definita da una mistura di distribuzioni multinomiali condizionalmente indipendenti. A differenza del modello basato sulla scomposizione a valori semplici, i valori che assegnano le parole ai topic e i topic ai documenti sono tutti non-negativi e a somma 1. Inoltre tutti i topic hanno una certa probabilità di generare un documento, quindi hanno tutti valore superiore a 0. In pratica i documenti contengono tutti i topic con proporzioni differenti. Dal lavoro di Haffman sono stati sviluppati altri modelli che superano alcuni limiti del lavoro iniziale. Un esempio molto famoso è la Latent dirichlet allocation (LDA: Blei et al., 2003), si tratta di un modello bayesiano più efficace che supera alcuni limiti della pLSA. La logica alla base dei due modelli è piuttosto simile, tuttavia nel secondo la distribuzione dei topic è disegnata da una Dirichelet. I topic si distribuiscono sul vocabolario, ogni topic contiene l'intera collezione dei termini del corpus testuale, con diverse probabilità. Il BiTerm topic model (BTM) proposto da Yan et al. (2013) agisce direttamente sui pattern di cooccorrenza dei termini per ottimizzare la computazione dei topic all'interno della collezione di documenti. Rispondere all'esigenza di poter analizzare e modellare documenti composti da poche parole è una task molto richiesta in varie applicazioni pratiche come per sintetizzare QA all'interno di websites, messaggi di status nei social media e review su siti di e-commerce. Questi testi sono difficili da processare per molti modelli in uso, perché applicandoli questi risentiranno della sparsità

delle matrici prodotte in fase di processamento. Nello specifico, le occorrenze calcolate sui vocabolari di termini estratti da documenti brevi hanno un ruolo meno importante rispetto ai documenti più lunghi dove il modello ha a disposizione abbastanza conteggi da sapere come le parole sono tra loro relazionate. Inoltre un'unità di contesto più breve rende più difficile per il modello identificare le parole ambigue all'interno dei documenti. In letteratura sono presenti svariate proposte che cercano di ridurre i problemi generati dalle caratteristiche dei testi brevi. L'aggregazione dei testi in un'unità di contesto più lunga è una di queste, Weng et al. (2010) hanno aggregato i tweet pubblicati da un singolo utente in un unico documento prima di addestrare LDA. Un ulteriore proposta si basa sull'aggregazione dei documenti contenenti la medesima parola, in questa occasione Hong et al. (2011) hanno anche dimostrato che questi modelli funzionano meglio della classica LDA. Questi modelli, pur avendo un riscontro positivo in termini computazionali e statistici, sono altamente dipendenti dai dati. Se non disponiamo delle informazioni sugli utenti che hanno pubblicato quello specifico contenuto o se disponiamo di una collezione in cui sono presente un'alta varietà di utenti, difficilmente si potranno aggregare le informazioni in nostro possesso. Il BTM computa i topics a partire dai testi modellando direttamente i biterms all'interno dell'intero corpus. Per biterm si intende una coppia di parole non ordinate che co-occorrono all'interno di un cosiddetto contesto, nel caso di questo modello questa finestra di parole è ovviamente molto breve. Il processo logico-generativo alla base del BTM è che il corpus consiste in una miscela di topics, dove ogni biterm è tratto da uno specifico topic. Rispetto ad altri topic models, il BTM modella direttamente i biterm, piuttosto che i documenti, per ottimizzare l'apprendimento del modello. Questo vuol dire che genera automaticamente i biterm e successivamente avvia l'analisi a partire da una matrice con questi elementi come unità. Successivamente sarà possibile inferire l'allocazione dei documenti all'interno dei singoli topics. Inoltre per risolvere il problema della sparsità a livello dei singoli documenti il modello opera su

pattern aggregati sull'intero corpus. Yan et al. (2013) dimostrano inoltre che il modello riesce a superare il rendimento della LDA anche sui testi più lunghi.

Per riassumere, il modello qui presentato invece che lavorare sulla distribuzione delle parole all'interno dei documenti, modella le co-occorrenze basate sui biterms. È importante chiarire come vengono generate questi elementi e cosa rappresentano. Un biterm consiste in una coppia di parole non ordinate che co-occorrono in un breve contesto. Qui il contesto breve si riferisce a una vera e propria finestra di testo contenente co-occorrenze di parole significative. Nei testi brevi, poiché i documenti sono di solito brevi e specifici, prendiamo ogni documento come una singola unità di contesto. Estraiamo qualsiasi due parole distinte in un documento di testo breve come un biterm. Il processo generativo del modello può essere descritto come segue:

1. Per ogni topic viene definita una distribuzione $\varphi z \sim Dir(\beta)$
2. Definisce una distribuzione $\theta \sim Dir(\alpha)$ per l'intera distribuzione
3. Per ogni biterm b all'interno del set costituito dall'insieme di tutti i biterm B
    a. definisci $z \sim Multi(\theta)$
    b. definisci wi, wj $\sim$ Mulit(φz)

Seguendo questa procedura, la proprietà congiunta di un biterm $b = (w_i, w_j)$ può essere scritta come segue:

$$\begin{aligned} P(b) &= \sum_z P(z)P(w_i|z)P(w_j|z). \\ &= \sum_z \theta_z \phi_{i|z} \phi_{j|z} \end{aligned}$$

Mentre la probabilità dell'intero corpus è:

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z}$$

In questo modo il metodo ci permette di modellare direttamente il pattern costituito dalle co-occorrenze delle parole.

3. III Fase: Proiezione dei topic in un sottospazio definito dalle emozioni

La terza fase del processo di analisi consiste nell'emotional analisys condotta mediante un approccio lexicon based e l'applicazione di un'analisi in componenti principali in un sottospazio di riferimento (PCAR). I risultati del topic model verranno modellati all'interno di una PCAR, la quale permetterà di visualizzare le relazioni che intercorrono tra i gruppi lessicali estratti e le dimensioni emotive inserite nell'analisi.

Quando parliamo di acp in un sottospazio di riferimento da un punto di vista algebrico intendiamo la definizione – tramite trasformazioni delle matrici di partenza – di un operatore tale che ci sarà possibile arricchire le informazioni contenute in una matrice di dati con ulteriori informazioni esterne ad essa.

Da un punto di vista statistico analizziamo le relazioni che intercorrono tra vari elementi di una matrice all'interno di una cornice informativa esterna. Nel nostro caso, ci sarà possibile studiare le relazioni tra topic e parole all'interno di una corniche definita dalle emozioni che compongono il lessico esterno.

Quindi la cornice informativa esterna è rappresentata dai gradi emotivi contenuti nel lexicon sfruttato dalla emotional analysis. La quale ci permetterà di dara un'etichetta emotiva ai topic. L'approccio utilizzato supera la classica dimensione binaria, la quale si estende tra i due poli negativo e positivo, e considera la multidimensionalità dell'emozione, per questo si è preferito inserirla in una cornice relativa all'analisi multidimensionale dei dati.

La necessità di riuscire ad esplorare un set di dati all'interno di una cornice informativa esterna alla matrice di partenza è un argomento particolarmente rilevante per chi si occupa di analisi testuale, questo è dovuto alla tipologia di dati con cui lavora chi solitamente è abituato a lavorare in questo ambito. Le informazioni contestuali ai testi e quelle che solitamente vengono perse durante le fasi di strutturazione del dato sono molte, la possibilità di includerle all'interno delle operazioni di analisi è per questo uno sviluppo

interessante all'interno delle questioni aperte in letteratura. Le informazione a disposizione possono essere relative al processo di categorizzazione dei documenti oppure possono riferirsi al livello dei termini utilizzati, andando ad arricchire la conoscenze contestuale dei nostri oggetti di studio per compensare l'ambiguità intrinseca di questo dato. Questi metadati possono essere utilizzati per ottimizzare l'interpretazione personale del ricercatore, quindi con lo scopo di arricchire la conoscenza estraibile dai dataset costruiti, oppure possono essere incorporate direttamente nei dati studiati. Balbi e Misuraca (2010) individuano due diversi tipi di informazioni in un'analisi testuale:

- Informazione intratestuale: viene definita quantitativa e corpus driven, relativa alle relazioni che intercorrono tra termini e documenti.
- Informazione extra-testuale: di natura qualitativa, riferita al contesto di appartenenza dei documenti estratti e quindi metadati non direttamente leggibili dal dataset su cui viene applicata l'analisi testuale.

Di seguito viene fornito un approfondimento metodologico dedicato alla PCAR, al fine di chiarirne le fondamenta statistiche alla base. C. R. Rao in un lavoro del 1964 introduce nel dibattito accademico la necessità di inserire informazioni aggiuntive all'interno dell'analisi esplorativa di strutture multivariate. Il suo lavoro presenta un set di q variabili esplicative numeriche strumentali all'interno della PCA. Nel 1991 Takane e Shibayama propongono un metodo di analisi che, combinando regressione e PCA, include informazioni aggiuntive all'interno dell'analisi sia sugli individui sia sulle variabili. Takane continua i suoi studi in questa direzione e nel 1997 sviluppala cosiddetta Constrained Principal Component Analysis (CPCA) (Takane, 1997). Diversamente, sempre sulla scia del medesimo dibattito accademico, D'Ambra e Lauro nel 1982 presentano un lavoro sulla Principal Component Analysis onto a Reference Subspace (PCAR), il metodo selezionato all'interno della mia strategia di analisi. La strategia da

me proposta si avvicina a quella di Balbi e Misuraca (2010), i quali propongono una doppia strategia di proiezione, utilizzando contemporaneamente proiettori ortogonali sugli spazi attraversati dalle variabili aggiuntive relative ai documenti e ai termini.

Da un punto di vista geometrico, una proiezione ortogonale di un oggetto k-dimensionale su un sottospaizo d-dimensionale definito dalle d colonne linearmente indipendenti di una matrice P(n,d) si , si ottiene considerando un operatore di proiezione P(P P)-1P , simmetrico e idempotente. Da un punto di vista statistico proiettare una struttura di dati su un sottospazio di riferimento significa analizzare le relazioni tra le righe e le colonne nel quadro delle informazioni elencate in P.

Entrando nello specifico dei metodi sopra presentati risulta essenziale definire i processi statistico algebrici che differenziano la CPCA e la PCAR, per poi chiarire le motivazioni che mi hanno portato a preferire il secondo metodo. La CPCA parte da una matrice Z individui per variabili, due matrici G (per gli individui) e H (per le variabili) rappresentano le informazioni esterne. Non esistono per questa tecnica assunzioni in termini di ipotesi distributive, di analisi preliminare dei dati o di particolari scelte sulla metrica da adottare. La CPCA consiste in due fasi analitiche principali. Nel primo, la cosiddetta analisi esterna, Z viene proiettato ortogonalmente sugli spazi attraversati da G e H, in modo da scomporre l'influenza delle variabili "esterne" nella somma di quattro termini: il primo riguarda ciò che può essere spiegato congiuntamente da G e H, il secondo e il terzo, rispettivamente, riguardano ciò che può essere spiegato da G e H, mentre il quarto è una matrice di residui. Questa soluzione si ottiene in un quadro di stima dei minimi quadrati minimizzando la matrice dei residui. Nel secondo passo si esegue l'analisi interna sulle matrici di decomposizione per mezzo di una o più PCA.

La struttura dei dati PCAR è data da due matrici Z e X. PCAR mira a visualizzare, in un quadro geometrico adeguato, la dipendenza di Z da X. Vale a dire che PCAR cerca le componenti principali della proiezione ortogonale di Z sullo spazio attraversato dalle

colonne di X. Può essere visto come un caso speciale di un'analisi interna CPCA, quando si considera solo il primo termine della decomposizione e si vuole introdurre informazioni esterne solo sulle variabili. Inoltre le variabili in Z sono centrate e spesso standardizzate. In questo senso è una vera e propria PCA. I vantaggi della PCAR sono strettamente legati agli aspetti grafici e all'interpretazione. Infatti le mappe fattoriali mostrano sia le correlazioni all'interno dello stesso insieme di variabili che le correlazioni tra i due insiemi.

Per quanto il topic model è uno strumento molto utile per estrarre dimensioni semantiche latenti all'interno di corpus di documenti, questo metodo manca di una resa grafica dei risultati che permetta di presentare correttamente le informazioni ottenute. Inoltre, considerato la strategia in cui è stato inserito all'interno di questo lavoro, si ha la necessita di far emergere delle relazioni che intercorrono rispetto ad una fonte di dati esterna ai documenti raccolti. La PCAR ci permette di ottenere una rappresentazione su un piano fattoriale dei topic estratti all'interno di una cornice definita da un'informazione esterna. Nel nostro caso questa cornice è data dalle dimensioni emotive che costituiscono il lexicon esterno utilizzato per etichettare i topic. Per formalizzare il processo, partiamo con il definire la matrice ottenuta successivamente alla fase di pretrattamento. La matrice T(n,q) composta in riga dai documenti e in colonna dai termini. A questa matrice applico il topic model BiTerm, il quale estrae la distribuzione dei termini all'interno dei topic. Questa è definita dalla matrice Φ(q,k). Per costruire il proiettore a partire dalla matrice C(q,h) che contiene l'informazione sui termini applichiamo la formula C'(CC')-1C, in questo modo otterremo un operatore simmetrico ed idempotente. La proiezione ortogonale all'interno del sottospazio di riferimento sarà ottenuta tramite la matrice ΦC ottenuta tramite il prodotto Φ*x*C.

6. Un caso studio dedicato alla community di Reddit

La strategia descritta nel capitolo 3 è stata applicata all'interno di un caso studio riportato nel presente capitolo. Ho analizzato 6567 post contenuti in un repository a libero accesso ed estratti dal social network Reddit da febbraio a maggio 2020, scritti in lingua inglese. Come chiavi per l'estrazione sono stati utilizzati 104 *subreddit* raccolti consultando una letteratura di riferimento, lo strumento per l'estrazione è stato un *wrapper* costruito in Python (Boe,2020). Di seguito ho approfondito tutte le informazioni riguardo il social network Reddit e i dati utilizzati. L'approccio metodologico seguito all'interno del caso studio è di tipo misto, questo prevede una prima fase *bottom-up* per le fasi di organizzazione, pretrattamento e analisi dei dati e una *theory-driven* per la lettura dei risultati e la ricerca delle categorie emergenti dall'analisi.

1. Il dataset

Il dataset utilizzato per il caso studio è stato realizzato da (Guest et al., 2021) e depositato in un archivio online pubblico a disposizione di future ricerche. Il dataset è stato selezionato perché consiste in un corpus testuale di recente acquisizione da una web community attiva e con contenuti misogini, in modo da poter applicare la strategia sviluppata. Il database impiegato è già stato utilizzato per una ricerca dedicata allo studio dei contenuti misogini online ed è stato costruito seguendo una procedura mirata che assicurasse la raccolta di testi con hatespeech contro le donne. Questa procedura mira con il costruire un campione ragionato di subreddit seguendo diversi criteri di selezione, dai *thread* individuati sono stati successivamente scaricati i testi da includere nel corpus. La letteratura di riferimento ha permesso di invididuare i primi 12 subreddit già classificati come misogini da ricerche passate (es.: *r/MensRights, r/seduction, r/TheRedPill*). Gli autori hanno identificato ulteriori 22 discussioni a partire dai gruppi raccomandati dai moderatori e proprietari dei 12 iniziali. Gli autori ci tengono a specificare che anche se

non tutti i contenuti sono dichiaratamente contro le donne, sono comunque gruppi di discussione dedicati all'argomento di genere, in cui è comunque presente una certa tensione tra le parti che intervengono. Nella tabella 4 è presente la selezione effettuata dagli autori del dataset sulla base delle ricerche consultate. Vengono qui elencati tutti i gruppi selezionati e la modalità con cui sono stati inclusi nel campione, in tutto sono 34 subreddit. I dati sono stati raccolti in un range di 11 settimane, per ogni subreddit, gli autori hanno raccolto le intere discussioni dei 20 post più popolari di quella settimana.

Il campione così progettato ha reso il dataset ancora più interessante da utilizzare, poiché gli autori hanno scelto di usare i gruppi di discussione piuttosto che le parole chiave per la pianificazione campionaria. Tale scelta ha garantito una maggiore varietà linguistica all'interno dei documenti scaricati. Minimizzando la quantità di bias, dato che parole chiave come "*slut*" sono associate a forme di abuso più esplicite e meno sottili. Ovviamente questa strategia porta con se delle limitazioni, campionare solo dalle comunità sospettate di misoginia secondo alcuni studi potrebbe generare un errore tale che lo strumento di classificazione identificherebbe solo le forme di misoginia trovate in quei contesti mirati (Davidson et al., 2017; Wiegand et al., 2019). La strategia adottata dagli autori per arginare questa possibile distorsione consiste nel campionare i contenuti di 71 subreddit scelti a caso, ciò ha inoltre consentito una maggiore generalizzabilità. Essi rappresentano il 18% delle discussioni e il 16% delle voci nel nostro set di dati. Per ogni subreddit selezionato a caso, hanno raccolto il thread del post più popolare. Tutti i thread sono in inglese. I post e i commenti sono stati raccolti da febbraio a maggio 2020 utilizzando il pacchetto python PRAW, un wrapper per le API di Reddit (Boe, 2020). I post su Reddit hanno un titolo di testo e un corpo che può essere testo, un'immagine o un link. Per i post con un corpo di testo hanno combinato questo con il titolo del post per creare una singola unità di testo. Per il 29% dei post in cui il corpo era un'immagine hanno anche raccolto l'immagine.

Reddit è un social network che ricorda molto i forum testuali del web 2.0 dove gli utenti registrati possono condividere contenuti in forma di testo o *link*, gli utenti sono anche chiamati *redditors*. Precedentemente conosciuto come "la prima pagina di internet", Reddit è una piattaforma che ospita oltre 130.000 sotto-forum e comunità. I singoli post sono aperti all'intera comunità di Reddit, gli utenti possono esprimere un proprio giudizio sui contenuti pubblicati votando con un click su un'icona a forma di pollice verso l'alto o il basso. I post più popolari sono presentati direttamente in prima pagina. I subreddit sono disponibili per categoria e i Redditors possono seguire subreddit selezionati rilevanti per i loro interessi e anche controllare quali contenuti vedono sulla loro prima pagina personalizzata. Alcuni dei subreddit più popolari sono *r/AskReddit* o *r/AMA* - il formato "*Ask Me Anything*". Secondo la società, Reddit ha ospitato 1.800 AMA nel 2018, con una vasta gamma di argomenti e *host*.

SUBREDDIT	NUMERO	TIPO DI SELEZIONE
ALTTRP	2	Snowball
ASKFEMINISTS	263	Snowball
ASKSEDDIT	142	Snowball
BADWOMENSANATOMY	430	Farrell et al. (2019)
BECOMEAMAN	2	Snowball
EGALITARIANISM	115	Snowball
EXREDPILL	113	Snowball
FEMRADEBATES	195	Snowball
GEOTRP	11	Snowball
INCELSINACTION	110	Farrell et al. (2019)
INCELSWITHOUTHATE	325	Farrell et al. (2019)
KOTAKUINACTION	373	Qian et al. (2019)

MARRIEDREDPILL	87	Snowball
MASCULISM	34	Snowball
MENSRANTS	4	Ging (2017)
MENSRIGHTS	364	Ging (2017); Qian et al. (2019); Zuckerberg (2018)
MENSRIGHTSLAW	2	Snowball
MENSRIGHTSMETA	4	Snowball
MGTOW	601	Farrell et al. (2019); Ging (2017); Qian et al. (2019)
MGTOWBOOKS	2	Snowball
MRACTIVISM	8	Snowball
NOMAAM	2	Snowball
PUA	10	Snowball
PURPLEPILLDEBATE	221	Snowball
PUSSYPASS	344	Qian et al. (2019)
PUSSYPASSDENIED	262	Qian et al. (2019)
REDPILLPARENTING	12	Snowball
REDPILLWIVES	61	Snowball
REDPILLWOMEN	217	Snowball
THANKTRP	8	Snowball
THEREDPILL	338	Ging (2017)
THEREDPILLRIGHT	10	Snowball
TRUFEMCELS	434	Farrell et al. (2019)

Tabella 4 – Selezione Subreddit (Guest et al., 2021)

2. Il lexicon

Una strategia di emotional analysis che si basa su un lexicon di riferimento risente molto, ovviamente, della scelta del dizionario. Il processo analitico presentato nella tesi sarà la base computazionale e operativa di una applicazione liberamente fruibile sul web, all'interno di questa sarà possibile inserire il dizionario che più si addice agli interessi del ricercatore. Nel presente caso il lessico su cui è modellata la emotional analysis è composto da 20,000 parole inglesi codificate su tre dimensioni emotive, realizzato facendo riferimento agli strumenti della linguistica computazionale e alla letteratura neuro-psicologica sugli stimoli neurologici alle emozioni. Per comprendere i risultati dell'analisi è importante presentare il dizionario, procederò quindi con un approfondimento sulle dimensioni emotive che lo costituiscono e successivamente sulla metodologia adottata per costruirlo.

Le dimensioni emotive su cui è stato costruito il lexicon sono tre, *valence, arousal* e *dominance* (da qui in avanti mi riferirò alle dimensioni utilizzando anche l'abbreviazione VAD). Ho deciso di mantenere in inglese i nomi delle dimensioni al fine di garantirne l'autenticità semantica rispetto alla letteratura di riferimento, anch'essa in lingua inglese. Il riferimento principale a queste dimensioni è lo studio del 1987 condotto da James A. Russel sul cosiddetto *core affect*, un nucleo emotivo prototipale che gli esseri umani utilizzano per orientarsi nel mondo. Consiste negli elementi primitivi che possono guidare la risposta del singolo ai vari livelli della scala emotiva e affettiva. Secondo l'autore è lo stato neurofisiologico di base, cosciente ma non riflessivo. Quando il core viene indirizzato verso un oggetto prende forma un'emozione, per questo motivo comprende le emozioni ma non consiste di per se in un insieme ben definito e conforme. A seconda della risposta a specifici oggetti vengono quindi a delinearsi le emozioni su cui si estende il lessico utilizzato per l'analisi, la lingua con la sua espressione orale o scritta difatti rappresenta un importantissimo catalizzatore di questi sentimenti. Per questo motivo

risulta tanto importante studiarla e comprenderla nelle sue dimensioni semantiche ed emotive. Per conferire un senso e un significato alla risposta emotiva agli stimoli che provengono dall'esterno, dall'esperienza quotidiana. Il core affect può aiutare a comprendere e giustificare comportamenti umani, in letteratura sono presenti molti riferimenti relativi alla spiegazione di diversi comportamenti, anche senza uno specifico riferimento al core è comunque presente la matrice emotiva che lo definisce. Possiamo individuare comportamenti relativi ad attitudini sessuali (Abramson Pinkerton, 1995), alimentari (Pinel, Assanand, Lehman, 2000), atteggiamenti aggressivi (Berkowitz, 1993), dipendenze e abusi di droga (R. L. Solomon, 1977), mantenimento dell'autostima (Tesser, 2000). Una grande varietà di effetti cognitivi possono essere spiegati da variazioni nel core affect (Ashby, Isen, Turken, 1999; Bower, 1992; Eich, 1995; Forgas, 1995). Per questo motivo non è stato difficile collegare questa matrice neurofisiologica alla violenza verbale, considerata anche l'affinità che questa ha con molti comportamenti sopra delineati.

Per quanto riguarda la *valence*, Davidson (2000) sostiene che è "presente in virtualmente tutti i sistemi che sono stati sviluppati per classificare l'emozione e la motivazione, dai conti comparativi che affrontano le origini filogenetiche (Schneirla, 1959) agli studi sulla struttura semantica (Osgood, Suci, Tannenbaum, 1957)". La valenza agisce anche su una dimensione inconscia, questa infatti dipende dal successo o dall'insuccesso in obiettivi non coscientemente innescati (Bargh Chartrand, 1999). La dimensione della valenza del core affect è inoltre legata all'azione. Sappiamo che la chiave di lettura per il core affect è la necessita che si materializzi un oggetto, nel caso di questa emozione, serve un obiettivo su cui far convergere l'azione. Quindi, una volta attribuito ad un oggetto, l'azione del core affect diventa la valutazione dell'oggetto e quindi un pesare, per esempio, i costi minacciati o sofferti. C'è un obiettivo edonico generale per porre fine al dispiacere e cercare il piacere. L'estremizzazione della valenza

potrebbe quindi contribuire al vigore e all'intensità dell'azione (Russel, 2003). Davidson individua anche ragioni empiriche e concettuali per collegare questa tensione all'azione a un'ampia dimensione comportamentale di approccio o ritiro (Davidson, 1992). Tuttavia, la natura dell'azione in quanto tale, ossia come sentimento all'agire, come valore strumentale intrapresa dipende dall'oggetto specifico. In alcune occasioni, il core affect piacevole è associato alla sazietà e quindi alla cessazione dell'azione, e il core affect negativo può essere associato all'avvicinamento, come quando un bambino angosciato cerca il suo *caregiver* (Bowlby, 1969).

Per quanto riguarda il sentimento definito *arousal*, che in italiano si potrebbe tradurre come eccitazione, in letteratura spazia su una dimensione che procede attraverso vari stadi di allerta che partono da una bassa condizione di eccitamento, che potrebbe essere associata alla sonnolenza, fino all'eccitazione frenetica (Russel, 2003). La sensazione collegata a questa sensazione è il senso di mobilitazione e di energia (Russel, 1991). Per comprendere meglio questo stato possiamo pensarlo in combinazione con altri sentimenti, un alto livello in *arousal* con il piacere può essere qualificato come euforia, mentre in unione al dispiacere abbiamo l'ansia. Ovviamente non può essere considerato un sentimento esaustivo dell'intera sfera emotiva, tuttavia possiamo considerarlo tra i principali stimoli emotivi che si possono esperire. Anche per questo motivo esiste una ricca letteratura riguardo questo sentimento, gli psicologi lo hanno infatti interpretato in molti modi (Cannon, 1927; Duffy, 1941; Hebb, 1955; Schachter Singer, 1962). Alcuni autori presentano il piacere come stimolo mentale e l'eccitazione come componente fisiologica. Alcuni pensano all'eccitazione come a una componente fisiologica che definisce l'intensità di un'emozione, con la cognizione che fornisce la sua direzione positiva o negativa. Nella prospettiva adottata dai riferimenti principali utilizzati per la costruzione del lexicon, l'approccio è meno determinista rispetto alla letteratura di riferimento. Le dimensioni delineate sono stati del sistema nervoso centrale, queste hanno

correlati fisiologici periferici ed entrambi sono sperimentati soggettivamente come eventi mentali. Ciò è molto più evidente grazie all'esperienza linguistica, le lingue umane conosciute forniscono le frasi "mi sento bene" e "mi sento male". Infatti, i concetti di sentire, buono e cattivo sono universali e semanticamente primitivi (Wierzbicka, 1999). Uno studio del modo in cui in cui le emozioni sono descritte in varie lingue ha suggerito che non solo il bene e il male, ma anche l'eccitazione sono probabilmente dimensioni semantiche universali dell'emozione (Russell, 1991). Gli stati di alto eccitamento sono preparativi per l'azione, sia stimolati da eventi esterni che creati da un adeguato riposo. Gli stati di bassa eccitazione sono momenti di inazione, creati da sazietà, bisogno di riposo o abbandono degli obiettivi.

La terza dimensione si riferisce al dominio/potere della parola: la misura in cui la parola denota qualcosa che è debole/sottomesso o forte/dominante. Russell e Mehrabian in una ricerca del 1977 hanno sviluppato un modello affidabile per individuare e collocare le emozioni all'interno di spettri semantici ben definiti, ai sopra citati stadi emotivi ne aggiungono un ulteriore, la dominanza. Questo si sviluppa in un asse emotivo che va dalla dominanza alla sottomissione. In pratica dalla capacità di sopraffare alla sua negazione, la tendenza a sottomettersi all'altro, all'accettare la visione del mondo imposta dall'altro. La presenta dimensione è legata alla potenza e all'esercizio della stessa, ponendosi come tentativo di sopraffazione dell'altro. La dominanza va da sentimenti di totale mancanza di controllo o influenza sugli eventi e sull'ambiente circostante all'estremo opposto, ossia al sentirsi come capaci di imporre il proprio sè e la propria visione del mondo su chiunque, sentendosi influenti e capaci di controllare la situazione (Russell, Mehrabian, 1977). La parola in questa eccezione è potere, la componente linguistica assume qui un importante significato. Si tratta di un potere non subdolo, nascosto, ma espresso con forza e convinzione. È sostanzialmente l'attuazione di un volere emotivo forte e deciso.

Secondo queste definizioni, "emozione" non include semplicemente stati passionali occasionali. Piuttosto, una persona è vista come uno stato emotivo in ogni momento, uno stato che può essere descritto come una regione all'interno di uno spazio tridimensionale. Definite le dimensioni emotive su cui si estende il lessico selezionato, è importante presentare la metodologia adottata dagli autori. Questa si basa su un metodo chiamato Best-Worst Scaling (BWS), con il quale il team ha ottenuto un valore rispetto alle emozioni *valence*, *arousal*, and *dominance* per circa 20000 parole inglesi tramite *crowdsourcing*. Con questo termine si intende uno sviluppo collettivo di un progetto, operato interpellando un collettivo di individui esterni al team di sviluppo. La BWS è una tecnica di annotazione comparativa che riesce a superare le limitazioni delle scale tradizionali (Louviere, 1991; Louviere et al., 2015). Gli scores sono valori reali e continui compresi in intervallo che va da 0 (basso livello dell'emozione specifica) a 1 (elevato livello dell'emozione specifica). Per valutare la bontà di questi scores viene utilizzato l'indice di correlazione calcolato tra le annotazioni rispetto ad ogni item (in questo caso rispetto ad ogni parola), la metrica utilizzata è la split-half reliability (SHR). In riferimento al lexicon utilizzato i valori di SHR sono r = 0.95 per la valence, r = 0.90 per la arousal, and r = 0.91 per la dominance (Mohammad, 2018). Stando a quanto riporta l'autore, questi sono valori ben superiori rispetto ad altri sviluppi di dizionari basati sullo spettro VAD, come il lavoro di Warriner et al. (2013). Per controllare la selezione di individui che ha risposto al questionario dedicato agli stimoli VAD è stata aggiunta una scheda in cui gli intervistati potevano rispondere a domande socio-demografiche. In questo modo è stato possibile fornire informazioni sulla loro età, sul loro genere e altre informazioni riguardo la personalità e i rapporti sociali. In questo modo il lexicon sviluppato è accompagnato da una scheda descrittiva riguardo il collettivo con cui è stato possibile svilupparlo.

Il lexicon è stato quindi sviluppato a partire da un insieme di lemmi a cui un collettivo di individui ha assegnato uno score sulle dimensioni emotive VAD. Questo set di parole è stato composto riunendo parole dell'inglese comune. In particolare gli autori affermano di aver preferito parole che connotassero emozioni. La selezione è stata orientata anche verso un lessico tipico da social media, in particolare twitter. Le risorse utilizzate sono state (Mohammad, 2018):

- 14000 parole dal lessico delle amozioni NRC (Mohammad e Turney, 2013). I termini sono etichettati e indicano a una delle otto emozioni di base: rabbia, anticipazione, disgusto, paura, gioia, tristezza, sorpresa e fiducia (Plutchik, 1980).
- Tutti i 4.206 termini classificati nel dizionario positive/negative del General Inquirer (Stone et al., 1966).
- 1.061 termini elencati in ANEW (Bradley e Lang, 1999).
- 13.915 termini presenti nel lessico di Warriner et al. (2013).
- 520 parole dalle categorie del Roget's Thesaurus corrispondenti alle otto emozioni di base di Plutchik.
- 1000 termini di contenuto ad alta frequenza, comprese le emoticon, dal Hashtag Emotion Corpus (HEC) (Mohammad, 2012).

3. I Risultati

In questo paragrafo sono presentati i risultati dell'analisi condotta sui post estratti da Reddit inclusi all'interno del dataset. I risultati posso essere letti seguendo una doppia linea interpretativa. Da un lato viene modellata la topical prevalance, ossia la prevalenza, in termini di distribuzione, delle proporzioni di topic all'interno dei documenti (a partire dalle coppie di parole, i biterm). Questa, come mostrerò, permette la classificazione dei documenti e la possibilità di rappresentare le distribuzioni rispetto alle variabili associate

alle nostre unità. Dall'altro è presente il topical content, lo studio delle proporzioni rispetto al contenuto semantico dei nostri topics, mi riferisco alle distribuzioni di parole all'interno dei temi e di conseguenza alle proporzioni semantiche che ci aiutano ad etichettare i topic con le emozioni inserite nel modello e ad esplorare il contenuto lessicale latente che l'analisi è riuscita a far emergere.

In questo caso ho organizzato i risultati seguendo l'ordine consequenziale di lavoro della strategia adottata, prima opero l'etichettatura dei topic, successivamente sfrutto le etichette per rappresentare la topical prevalance e le distribuzioni rispetto alle variabili, e concludo con la rappresentazione delle dimensioni latenti rispetto ai termini.

La strategia di analisi ha permesso di leggere le relazioni tra i topics emergenti all'interno dello spazio definito dalle emozioni. Il modello ha dato in outcome le correlazioni tra topic ed emozioni, tramite la lettura di questo indice è stato possibile effettuare un'etichettatura automatica dei topic. Le relazioni all'interno del piano sono rappresentate in figura 14, il grafico va letto seguendo i criteri relativi al cerchio delle correlazioni, le proiezioni delle dimensioni emotive hanno un ruolo puramente illustrativo e determinano il sottospazio di riferimento delineato per la proiezione dei topics.

Sempre in figura 14 vengono quindi rappresentate le relazioni tra topics ed emozioni. Come si può notare, le emozioni sono distribuite uniformemente tra i topics, alcuni di questi sono poco correlati rispetto agli assi estratti dal modello. Si nota tuttavia come si definiscono dei gruppi visibili di topics in prossimità dei segmenti rappresentati dalle emozioni.

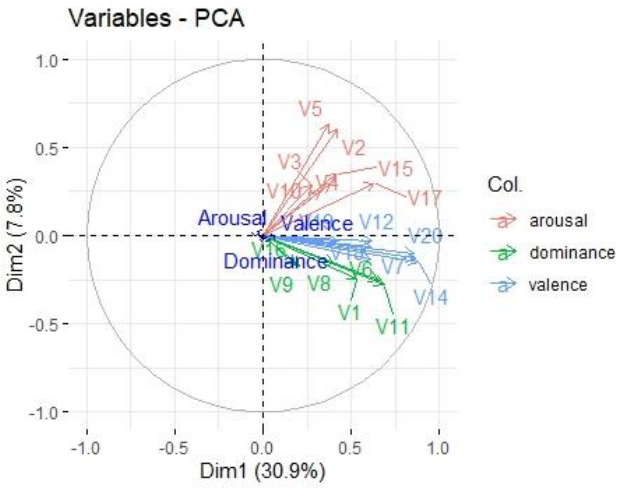

Figura 14 – Spazio delle variabili

Nel modello selezionato, dalle cooccorrenze tra i termini sono stati computati i topic. A partire dalle distribuzioni ottenute è possibile effettuare una classificazione dei documenti sfruttando le probabilità assegnate. Per infierire la distribuzione dei topics all'interno di un documento, viene utilizzato un algoritmo che a partire dai singoli biterms (coppie di parole) generati da uno specifico documento genera delle nuove distribuzioni per documento. La logica alla base è molto semplice, parte dall'assunto che le proporzioni dei topic per documento (topical prevalence) siano inferibili a partire dalle proporzioni rispetto ai biterms generati da quello stesso documento. In questo modo è possible assegnare i documenti ai vari topic, modellando la *topical prevalence* (fig 15). Utilizzando le etichette dei topic è stato possibile esplorare l'estensione emotiva del corpus rispetto alle dimensioni definite quando è stato selezionato il lexicon. Si nota come la valenza sia la dimensione emotiva prevalente all'interno del corpus.

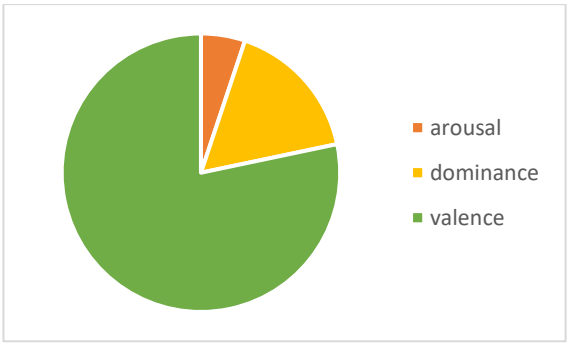

Figura 15 – Emozioni prevalenti

Sfruttando le informazioni associate ai testi relative ai subreddit si è potuto effettuare una *mappatura emotiva* per sotto-comunità. Con questa espressione mi riferisco alla distribuzione delle dimensioni emotive all'interno dei sottogruppi in cui sono divisi i nostri testi. Ovviamente è possibile applicare questa rappresentazione utilizzando qualsiasi variabile categoriale associata ai testi. Quella qui rappresentata è una mappatura emotiva del mio corpus, realizzata mediante una meta-analisi sviluppata sfruttando la classificazione dei documenti come informazione associata ai testi. In questo caso le variabili sono relative all'appartenenza di un testo ad uno specifico subreddit. Data la numerosità dei subreddit selezionati la rappresentazione risulta difficoltosa.

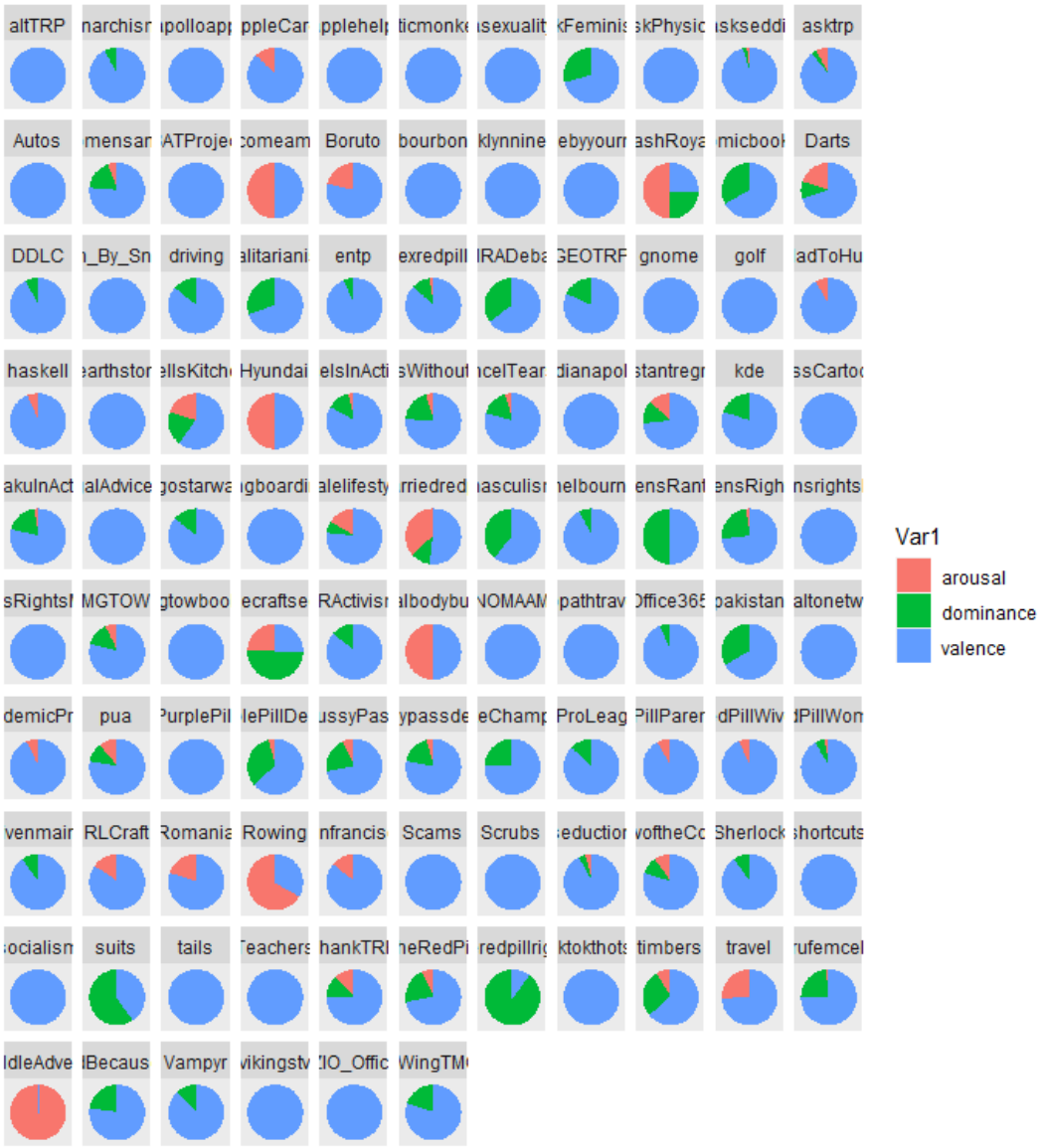

Figura 16 – Mappa emotiva

Per questa difficoltà nell'ottenere una chiara rappresentazione, ho selezionato 9 subreddit a partire dalla revisione in letteratura operata per costruire il corpus e presentata nella figura 17.

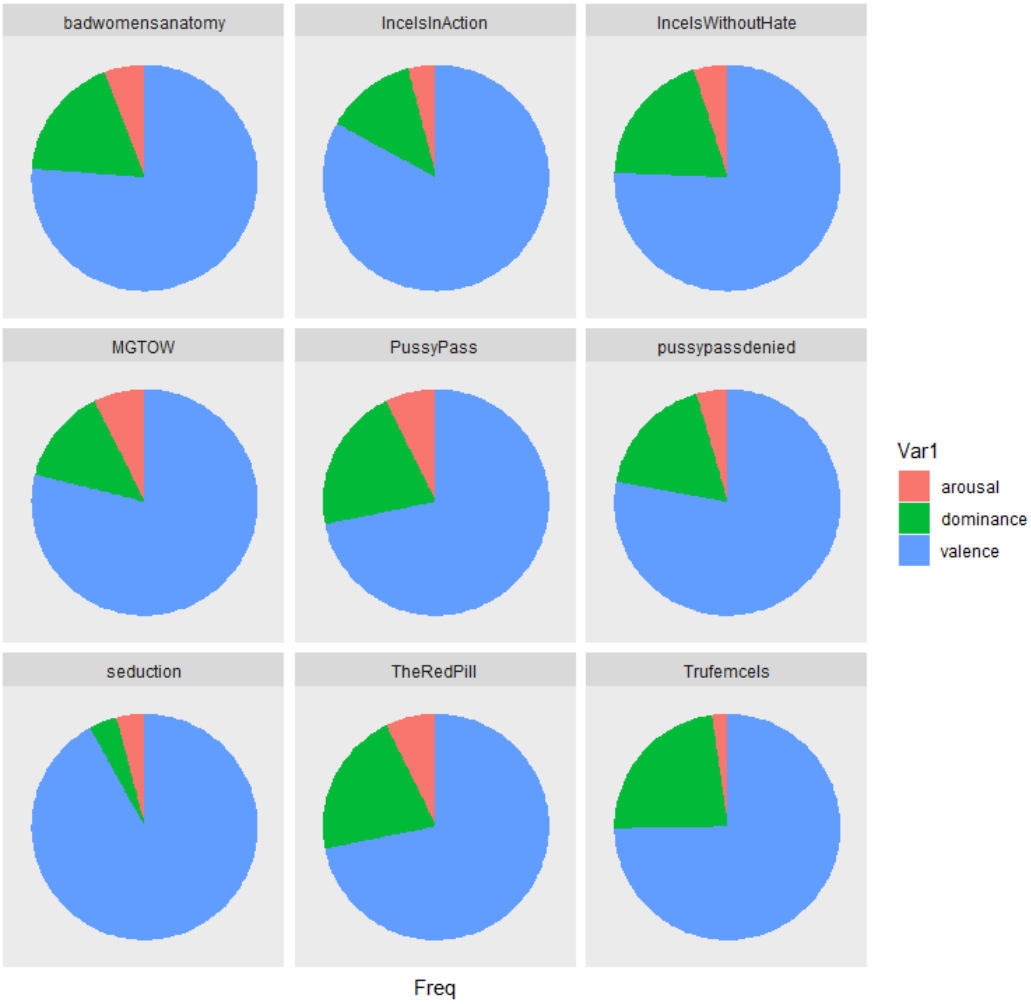

Figura 17 – Selezione di sub-reddit

Queste sono le sub-community con una varietà emotiva maggiore, in cui persiste tuttavia una prevalente attitudine al linguaggio aggressivo e dominante.

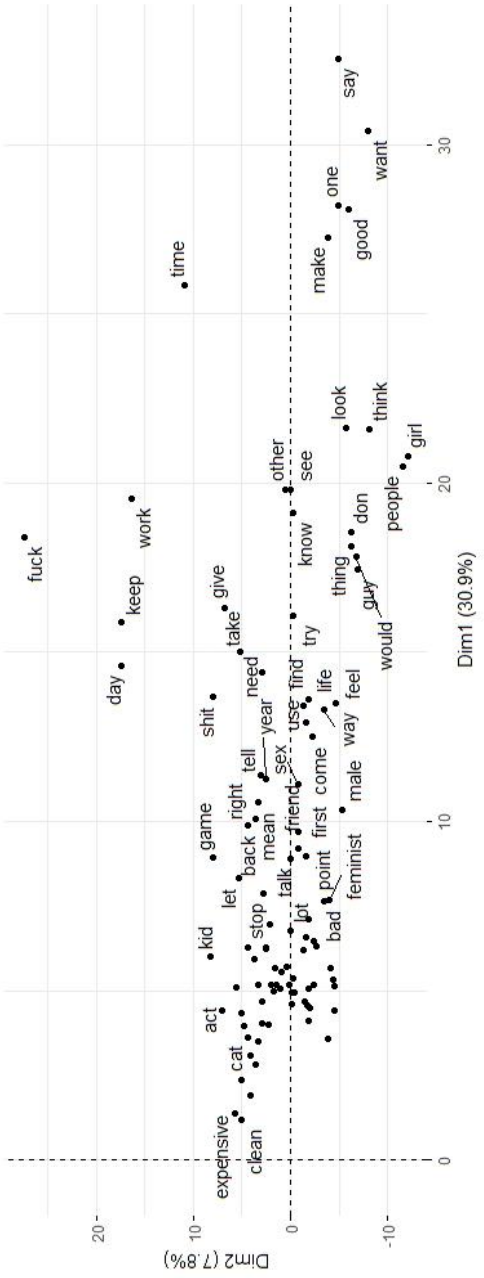

Figura 18 – Dimensioni semantiche

Ritornando al contenuto, è stato infine mappato lo spettro semantico relativo ai topic all'interno del sottospazio definito dalle emozioni (Figura 18). In modo da poterne esplorare il contenuto in relazione alla dimensione emotiva. I contenuti più carichi rispetto alla dimensione aggressiva-dominante sono quelli che più si avvicinano alla sfera della femminilità e al genere.

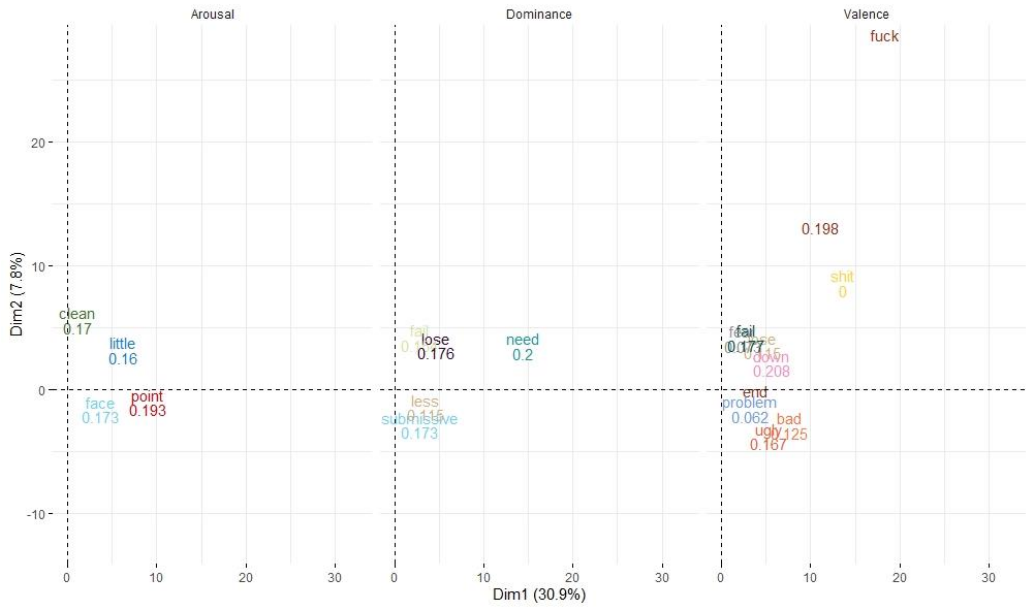

Figura 19 – Piani per dimensione emotiva

In figura 19 sono rappresentati tre piani relativi ognuno ad una diversa dimensione emotiva. Sono proiettate le parole che presentano la maggior correlazione rispetto alle emozioni e che hanno contribuito maggiormente alla creazione degli assi. In questo modo si evidenziano le specificità lessicali rispetto alle singole direzioni emotive, ben chiara è ad esempio la differenza tra la dimensione della valenza rispetto alla dominanza. Si nota come la valenza abbia una componente aggressiva più accentuata e marcata.

7. Conclusioni

In conclusione, la strategia di analisi presentata nelle precedenti pagine ha permesso di realizzare una serie di output grafici che sintetizzano facilmente le informazioni emerse dallo studio sui testi presenti nel corpus. In base ai dati raccolti e alle informazioni esterne al corpus utilizzato (lexicon), la dimensione emotiva più forte è quella aggressiva. La presenza di un marcato atteggiamento aggressivo nell'ambiente giustifica la bassa percezione di sicurezza percepita dalle donne e denunciata nelle ricerche precedentemente studiate. L'ambiente caratterizzato da tali atteggiamenti risulta ostile e pregiudica la percezione di sicurezza e accettazione degli individui target. In questo habitat è possibile rintracciare il filo conduttore che lega le pratiche misogine online e offline (QR1). L'emotional analysis ha permesso di mappare la qualità dell'emozione espressa dagli utenti delle comunità e individuare quei sottogruppi in cui sono prevalenti specifici stati d'animo (QR2), per questo è stato possibile risalire facilmente a quelle sub-comunità in cui sentimenti vicini alle dimensioni dell'odio sono prevalenti. Il processo di labeling condotto modellizzando il topical content ha permesso di dare automaticamente un'identità ai singoli topic e risalire all'informazione relativa all'emozione espressa ai documenti e studiarla in relazione alle variabili associate. L'esplorazione del contenuto ha evidenziato che le dimensioni semantiche che più si avvicinano ai topic etichettati come negativi riguardano la sfera della femminilità (QR3). Non esistono riferimenti diretti alla sfera sessuale ne linguaggio profano. Questi risultati mostrano la necessità di utilizzare un metodo che operi sulle dimensioni latenti di significato, in modo da andare ad individuare e mettere in evidenzia delle sfere semantiche che rimandino a pratiche e culture altrimenti sottese nel testo. Le quali p possibile raggiungere solo attraverso lo studio delle relazioni che intercorrono tra essere. La strategia proposta riesce a proiettare

agilmente queste relazioni all’interno di uno spazio emotivo, facendo risaltare le dimensioni concettuali inerenti alla sfera culturale.

8. Bibliografia

1. . J. Allaire, Francois Chollet, Yuan Tang, Daniel Falbel, Wouter Van Der Bijl, and Martin Studer. 2018. R interface to 'keras'. Computer software manual](R package version 2.1.6). Retrieved from https://CRAN. R-project. org/package= keras
2. Abbatecola, E. (2018). Trans-migrazioni: Lavoro, sfruttamento e violenza di genere nei mercati globali del sesso. Trans-migrazioni, 1-174.
3. Abramson, P. R., & Pinkerton, S. D. (Eds.). (1995). Sexual nature/sexual culture. University of Chicago Press.
4. Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. Telematics and Informatics, 34(7), 1177-1190.
5. Albury K. and Crawford K. (2012). Sexting, consent and young people's ethics: Beyond Megan's Story. Continuum: Journal of Media Cultural Studies, vol. 26, n. 3.
6. Alfina, I., Mulia, R., Fanany, M. I., Ekanata, Y. (2017, October). Hate speech detection in the Indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 233-238). IEEE.
7. Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45(4):1191–1207
8. Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and

characterization of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media.

9. Antonacci, G., Colladon, A. F., Stefanini, A., Gloor, P. (2017). It is rotating leaders who build the swarm: Social network determinants of growth for healthcare virtual communities of practice. Journal of Knowledge Management.
10. Ashby, F. G., & Isen, A. M. (1999). A neuropsychological theory of positive affect and its influence on cognition. Psychological review, 106(3), 529.
11. Aswani, R., Kar, A. K., Ilavarasan, P. V., Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from Twitter and SEOClerks. International Journal of Information Management, 38(1), 107-116.
12. B. Pang and L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL) in Barcelona, Spain, pp.271–278, July21-26 2004.
13. Babcock, J. C., Green, C. E., & Robie, C. (2004). Does batterers' treatment work? A meta-analytic review of domestic violence treatment. Clinical psychology review, 23(8), 1023-1053.
14. Badjatiya, P., Gupta, S., Gupta, M., Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).
15. Balbi S., Misuraca M., (2005), Pesi e metriche nell'analisi dei dati testuali, Quaderni di statistica, vol.7
16. Balbi, S., & Misuraca, M. (2010). A doubly projected analysis for lexical tables. In Advances in Data Analysis (pp. 13-19). Birkhäuser Boston.

17. Banet-Weiser, S., and K. M. Miltner. 2016. “# MasculinitySoFragile: Culture, Structure, and Networked Misogyny.” Feminist Media Studies 16 (1): 171–174.
18. Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. American psychologist, 54(7), 462.
19. Batsell, R. R., & Louviere, J. J. (1991). Experimental analysis of choice. Marketing letters, 2(3), 199-214.
20. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In Proceedings of the international AAAI conference on web and social media (Vol. 14, pp. 830-839).
21. Berkowitz, L. (1993). Aggression: Its causes, consequences, and control. Mcgraw-Hill Book Company.
22. Besussi, A. (2019). Hate speech. Una categoria inattendibile1. iblioteca della libertà, 54(224), 39.
23. Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Proceedings of the Workshop on Natural Language Processing for ComputerMediated Communication (NLP4CMC), pages 6–9, Bochum, Germany
24. Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 991-1022. DOI: 10.5555/944919.944937
25. Boe, B. (2020). PRAW: The python reddit API wrapper. PRAW, 2(1), 21.
26. Bolasco S., (2013), L’analisi automatica dei testi. Fare ricerca con il text mining, Carocci, Roma

27. Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., Maurizio, T. (2018). Overview of the evalita 2018 hate speech detection task. In EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (Vol. 2263, pp. 1-9). CEUR.
28. Bower, G. H. (1992). How might emotions affect learning. The handbook of emotion and memory: Research and theory, 3, 31.
29. Bowlby, J. (1969). Attachment and loss: volume I: attachment. In Attachment and Loss: Volume I: Attachment (pp. 1-401). London: The Hogarth Press and the Institute of Psycho-Analysis.
30. Boxer, C. F., Ford, T. E. (2010). Sexist humor in the workplace: A case of subtle harassment. In J. Greenberg (Ed.), Insidious workplace behavior (pp. 175–206). Boca Raton, FL: Routledge. Brock, A. (2012).
31. Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report No. C-1). Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.
32. Brady W.J., Crockett M.J. (2019). How effective is online outrage?.Trends in Cognitive Sciences, 23(2), 79-80. Konrath S.H., O'Brien E.H., Hsing C. (2011). Changes in Dispositional Empathy in American College Students Over Time: A MetaAnalysis. Personality and Social Psychology Review, (15)2, 180-198.
33. Brownmiller, S. (1976). Against Our Will men women and rape Harmo nds Worth: Pengui.
34. Burnap, P., Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy internet, 7(2), 223-242.

35. Burnap, P., Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data science, 5, 1-15.
36. Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., ... Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. Technological Forecasting and Social Change, 95, 96-108.
37. Cannon, W. B. (1927). The James–Lange theory of emotion: A critical examination and an alternative theory. American Journal of Psychology, 39, 106–124
38. Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledgemanagement, pages 1980–1984, Maui, HI, USA.ACM.
39. Ceron, A., Curini, L., Iacus, S. M. (2016). First-and second-level agenda setting in the Twittersphere: An application to the Italian political debate. Journal of Information Technology Politics, 13(2), 159-174.
40. Cesarano et al., OASYS: An Opinion Analysis System, in AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, pp. 21–26, 2006.
41. Charu R., Amit V., Bharath Y., Rama K. and Kiran S. (2014). API-FICATION, Hcl Technologies, testo disponibile al sito: https://www.hcltech.com/sites/default/files/apis_for_dsi.pdf, 05/06/2020.
42. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A. (2017, April). Measuring# gamergate: A tale of hate, sexism, and bullying. In Proceedings of the 26th international conference on world wide web companion (pp. 1285-1290).

43. Cheng, J. (2015). Danescu-Niculescu-Mizil, C.—Leskovec, J.: Antisocial Behavior in Online Discussion Communities. In AAAI International Conference on Weblogs and Social Media (ICWSM).
44. Chess, S., and A. Shaw. 2015. "A Conspiracy of Fishes, or, How We Learned to Stop Worrying about #GamerGate and Embrace Hegemonic Masculinity." Journal of Broadcasting Electronic Media 59(1): 208–220.
45. Chopra, S., Sawhney, R., Mathur, P., Shah, R. R. (2020, April). Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 386-393).
46. Cordella, B., Greco, F., Meoli, P., Palermo, V., Grasso, M. (2018). Is the educational culture in Italian Universities effective? A case study. In JADT'18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (pp. 157-164). Universitalia Rome, IT.
47. Corposanto C., and Molinari B. (2015). Rilevare dati sul web: la cassetta degli attrezzi 2.0. In: Sannella A. e Toniolo F., a cura di, Le sfide della società italiana tra crisi strutturale e social-innovation. 33-49, Venezia: Edizioni Cà Foscari-Digital Publishing.
48. Crockett M.J. (2017). Moral outrage in the digital age. Nature Human Behaviour, 1(11), 769-771. Duggan M., Smith A. (2016). The Political Environment on Social Media. Pew Research Center. Disponibile in: http://assets.pewresearch.org/wpcontent/uploads/sites/14/2016/10/24160747/PI_2016.10.25_Politics-and-Social Media_FINAL.pdf. [12/09/2019].
49. D'AMBRA L., et LAURO C. (1982). Analisi in componenti principali in rapporto ad un sottospazio di riferimento, Rivista di Statistica Applicata, 15, 1, 51-67

50. Davidson, R. J. (2000). The functional neuroanatomy of affective style. In R. D. Lane & L. Nadel (Eds.), Cognitive neuroscience of emotion (pp. 371–388). New York: Oxford University Press.
51. Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1).
52. Debole, F., Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In Text mining and its applications (pp. 81-97). Springer, Berlin, Heidelberg.
53. Deng, Z. H., Luo, K. H., Yu, H. L. (2004). A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications, 41(7), 3506-3513.
54. Djuric, N. et al. (2015) 'Hate Speech Detection with Comment Embeddings,' pp. 29–30.
55. Duffy, E. (1941). An explanation of "emotional" phenomena without the use of the concept "emotion." Journal of General Psychology, 25, 283–293.
56. Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., Madden, M. (2015). Social media update 2014. Pew research center, 19, 1-2.
57. Dworkin R. (1997), Freedom's Law. The Moral Reading of the American Constitution, Oxford, Oxford University Press.
58. E. Gilbert, K. Karahalios, C. Sandvig, The network in the garden: designing social media for rural life, American Behavioral Scientist 53 (9) (2010) 1367–1388.
59. Eich, E. (1995). Searching for mood dependent memory. Psychological Science, 6(2), 67-75.

60. Ellis, D., & Stuckless, N. (1996). Mediating and negotiating marital conflicts. SAGE Publications, Incorporated.
61. Eriksen, C. (2013). Gender and wildfire: Landscapes of uncertainty. Routledge.
62. Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017., pages 86–95.
63. Faloppa F. (2015), Buonisti o cattivisti? Meglio realisti. Consultabile a: https://www.cartadiroma.org/news/buonisti-o-cattivisti/
64. Filipovic, J. (2007). Blogging while female: How internet misogyny parallels real-world harassment. Yale JL Feminism, 19, 295.
65. Follingstad, D. R., Rutledge, L. L., Berg, B. J., Hause, E. S., & Polek, D. S. (1990). The role of emotional abuse in physically abusive relationships. Journal of family violence, 5(2), 107-120.
66. Ford, T. E., Ferguson, M. A. (2004). Social consequences of disparagement humor:A prejudiced norm theory. Personality and Social Psychology Review, 8, 79–94. http://dx.doi.org/10.1207/S15327957PSPR0801_4.
67. Ford, T. E., Boxer, C. F., Armstrong, J., Edel, J. R. (2008). More than ''just a joke'': The prejudice-releasing function of sexist humor. Personality and Social Psychology Bulletin, 34, 159–170. http://dx.doi.org/10.1177/0146167207310022.
68. Ford, T. E., Wentzel, E. R., Lorion, J. (2001). Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. European Journal of Social Psychology, 31, 677–691. http://dx.doi.org/10.1002/ejsp.56.

69. Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). Psychological bulletin, 117(1), 39.
70. Fortuna, P., Bonavita, I., Nunes, S. (2018, December). Merging datasets for hate speech classification in Italian. In CEUR Workshop Proceedings (pp. 218-223).
71. Fumagalli, C. (2019). Discorsi d'odio come pratiche ordinarie. Biblioteca della libertà, 54(224).
72. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross and K. Miller, WordNet: An online lexical database, International Journal of Lexicography, vol. 3, iss. 4, pp. 235-244, 1990.
73. G. Mishne, N. Glance, Leave a reply: an analysis of weblog comments, Third Annu-al Workshop on the Weblogging Ecosystem, 2006.
74. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing Management, vol. 24, issue. 5: 513–523, 1988.
75. Gambäck, B., Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In Proceedings of the first workshop on abusive language online (pp. 85-90). Park, J. H., Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206.
76. Gao, L., Huang, R. (2017). Detecting online hate speech using context aware models. arXiv preprint arXiv:1710.07395.
77. Ging, D., Lynn, T., & Rosati, P. (2020). Neologising misogyny: Urban Dictionary's folksonomies of sexual abuse. new media & society, 22(5), 838-856.

78. Giomi E. and Magaraggia S.M. (2018). La questione maschile. La violenza degli uomini contro le donne nella realtà e nelle rappresentazioni mediali. Sociologia Italiana, pp. 73-94, Milano: Egea Editore, Università Milano Bicocca. DOI: 10.1485/AIS_2018/12_3435533.
79. Glomb, T. M., Richman, W. L., Hulin, C. L., Drasgow, F., Schneider, K. T., Fitzgerald, L. F. (1997). Ambient sexual harassment: An integrated model of antecedents and consequences. Organizational Behavior and Human Decision Processes, 71, 309–328. http://dx.doi.org/10.1006/obhd.1997.2728.
80. Gloor, P. A. (2017). Sociometrics and human relationships: Analyzing social networks to manage brands, predict trends, and improve organizational performance. Emerald Group Publishing.
81. Gloor, P., Colladon, A. F., Giacomelli, G., Saran, T., Grippa, F. (2017). The impact of virtual mirroring on customer satisfaction. Journal of Business Research, 75, 67-76.
82. Greco F., Maschietti D., Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration, Rivista Italiana di Economia Demografia e Statistica, Vol. 71, No. 2, pp. 125-136
83. Greco, F., Polli, A. (2020). Emotional Text Mining: Customer profiling in brand management. International Journal of Information Management, 51, 101934.
84. Greco, F., Alaimo, L., Celardo, L. (2018, June). Brexit and Twitter: The voice of people. In JADT'18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (pp. 327-334). Universitalia Rome, IT.

85. Greco, F., Maschietti, D., Polli, A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. Rivista Italiana di Economia Demografia e Statistica, 71(2), 125-136.
86. Greevy, E. P., Smeaton, A. F. (2004). Text categorisation of racist texts using a support vector machine. Proceedings of 7es Journées internationales d'Analyse statistique des Données Textuelles JADT (1), 533-544.
87. Grey, R., & Shepherd, L. J. (2013). "Stop rape now?" Masculinity, responsibility, and conflict-related sexual violence. Men and Masculinities, 16(1), 115-135.
88. Griffiths T.L. and Steyvers M. (2002). A probabilistic approach to semantic representation. In: Proceedings of the annual meeting of the cognitive science society, Vol. 24, No. 24.
89. Griffiths T.L. and Steyvers M. (2003). Prediction and semantic association. In: Advances in neural information processing systems.
90. Griffiths T.L. and Steyvers M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences. 101: 5228-5235.
91. Grover, P., Kar, A. K., Dwivedi, Y. K., Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes–Can twitter analytics predict changes in voting preferences. Technological Forecasting and Social Change, 145, 438-460.
92. Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and
93. Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021, April). An expert annotated dataset for the detection of online misogyny. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 1336-1350).

94. Guiraud, P. (1954). Les caractères statistiques du vocabulaire. Paris: PUF.
95. Hammer, H. L. (2016, October). Automatic detection of hateful comments in online discussion. In International Conference on Industrial Networks and Intelligent Systems (pp. 164-173). Springer, Cham.
96. Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In IJCAI, pages 3952– 3958, New York City, NY, USA. IJCAI/AAAI Press.
97. Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). Psychological Review, 62, 243–254.
98. Henry N. and Powell A. (2018). Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research. Trauma, Violence, Abuse, 19, 2.
99. Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).
100. Hopkins, D. J., King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229-247.
101. Hotho, A., Nürnberger, A., Paaß, G. (2005, May). A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).
102. Illouz E. (Ed.) (2017). Emotions as Commodities. Capitalism, Consumption and Authenticity. London: Routledge.
103. Infante, G., & Misuraca, M. (2007). Text Mining Strategies for Analyzing Semi-Structured Corpora. Classification and Data Analysis, 267-270.
104. ISTAT, (2014). Indagine sulla Sicurezza delle donne

105. ISTAT, (2019). Indagine sulla Sicurezza delle donne

106. J. A. Hartigan, Clustering Algorithms, Wiley, 1985.

107. J. Kamps, M. Marx, R. J. Mokken and M. D. Rijke, Using WordNet to measure semantic orientation of adjectives, in International Conference on Language Resources and Evaluation, vol. IV, pp. 1115-1118, 2004.

108. J.C. Short, T.B. Palmer, The application of DICTION to content analysis research instrategic management, Organizational Research Methods 11 (4) (2008) 727–752.D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomkins, The predictive power of onlinechatter, Proceedings of the eleventh ACM SIGKDD international conference onKnowledge discovery in data mining, ACM, 2005, pp. 78–87.

109. Jane, E. A. 2014. "'Your a Ugly, Whorish, Slut' Understanding E-Bile." Feminist Media Studies 14 (4): 531–546.

110. Jenkins H., Ford S., Green J. (2013). Spreadable media. I media tra condivisione, circolazione, partecipazione. Sant'Arcangelo di Romagna: Apogeo.

111. Jenson, J., and S. De Castell. 2013. "Tipping Points: Marginality, Misogyny and Videogames." JCT (Online) 29 (2): 72.

112. Jenson, J., Taylor, N., de Castell, S., Dilouya, B. (2015). Playing with our selves: Multiplicity and identity in online gaming. Feminist Media Studies, 15(5), 860-879.

113. Jiang, M., Liu, R., & Wang, F. (2018, March). Word network topic model based on word2vector. In 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 241-247). IEEE.

114. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, pages 261–270. ACM, 2010

115. Johnson N.F., Leahy R., Johnson Restrepo N., Velasquez N., Zheng M., Manrique P., Devkota P., Wuchty S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. Nature, 573, 261-265.

116. Kapoor, R., Kumar, Y., Rajput, K., Shah, R. R., Kumaraguru, P., Zimmermann, R. (2019, July). Mind your language: Abuse and offense detection for code-switched languages. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 9951-9952).

117. Kelly, J. B., & Johnson, M. P. (2008). Differentiation among types of intimate partner violence: Research update and implications for interventions. Family court review, 46(3), 476-499.

118. Kiran, G. C., Kumar, J. S., Mohideen, G. M., Jack, P., Saheb, M. C. P. Offensive Tweet using Bidirectional.

119. Kwok, I., Wang, Y. (2013, June). Locate the hate: Detecting tweets against blacks. In Twenty-seventh AAAI conference on artificial intelligence.

120. Kwok, I., Wang, Y. (2013, June). Locate the hate: Detecting tweets against blacks. In Twenty-seventh AAAI conference on artificial intelligence.

121. LaFrance, M., Woodzicka, J. A. (1998). No laughing matter: Women's verbal and nonverbal reactions to sexist humor. In J. Swim C. Stagnor (Eds.), Prejudice: The target's perspective (pp. 61–80). San Diego, CA: Academic Press.

122. Laricchiuta, D., Greco, F., Piras, F., Cordella, B., Cutuli, D., Picerni, E., Petrosini, L. (2018). The grief that doesn't speak": Text mining and brain

structure. In JADT'18: Proceedings of the 14th international conference on statistical analysis of textual data (pp. 419-427). Universitalia Rome, IT.

123. Lebart L., Salem A., Berry L. (1998) Correspondence Analysis of Lexical Tables., Exploring Textual Data. Text, Speech and Language Technology, vol 4. Springer, Dordrecht.

124. Lee, Y., Yoon, S., Jung, K. (2017). Comparative studies of detecting abusive language on twitter. arXiv preprint arXiv:1808.10245.

125. Leone, J. M., Johnson, M. P., Cohan, C. L., & Lloyd, S. E. (2004). Consequences of male partner violence for low-income minority women. Journal of Marriage and Family, 66(2), 472-490.

126. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

127. Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). Best-worst scaling: Theory, methods and applications. Cambridge University Press.

128. Lovink G. (2019). Nichilismo digitale. L'altra faccia delle piattaforme. Milano: EGEA.

129. Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34(8), 1381-1388.

130. M. Joshi, D. Das, K. Gimpel, N.A. Smith, Movie reviews and revenues: an experi-ment in text regression, Human Language Technologies: The 2010 Annual Con-ference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 293–296.

131. MacKinnon, C. A. (1993). Only words. Harvard University Press.

132. Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. Feminist Studies, 39(2), 563-570.

133. Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In International Conference on Applications of Natural Language to Information Systems, pages 57–64. Springer.

134. Marwick, A. (2013). Gender, sexuality, and social media. In T. Senft J. Hunsinger (Eds.), The social media handbook (pp. 59–75). New York: Routledge. <http://www.tiara.org/blog/wp-content/uploads/2014/03/Marwick_gender_sexuality_chapter_2013.pdf>.

135. Marwick, A., Lewis, R. (2017). Media manipulation and disinformation online. New York: Data Society Research Institute.

136. Massanari, A. 2017. "# Gamergate and the Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." New Media Society 19 (3): 329–346.

137. Matsuda M.J., Lawrence C.R. III, Delgado R., Crenshaw K.W. (1993) (a cura di), Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment, Boulder, Westview Press.

138. McGlynn C., Rackley E. and Houghton R., (2017). Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. Fem Leg Stud 25, 25–46. DOI: 10.1007/s10691-017-9343-2.

139. McKinnon C.A. (1987), Feminism Unmodified: Discourses on Life and Law, Cambridge (MA), Harvard University Press.

140. Miller D., Costa E., Haynes N., McDonald T., Nicolescu R., Sinanan J., Spyer J., Venkatraman S., Wang X. (2018). Come il mondo ha cambiato i social media. Milano: Ledizioni

141. Misuraca M., Spano M. (2020). Unsupervised Analytic Strategies to Explore Large Document Collections. In Text Analytics. Advances and Challenges, Springer International Publishing

142. Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 174-184).

143. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3), 436-465.

144. Morales-Campos, D. Y., Casillas, M., & McCurdy, S. A. (2009). From isolation to connection: Understanding a support group for Hispanic women living with gender-based violence in Houston, Texas. Journal of immigrant and minority health, 11(1), 57-65.

145. Mubarak, H., Rashed, A., Darwish, K., Samih, Y., Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. arXiv preprint arXiv:2004.02192.

146. Muller, C. (1992). Principes et méthodes de statistique lexicale. Paris: Larousse, 1977, réimpression Champion-Slatkine, 1992.

147. Nagle A. (2018). Contro la vostra realtà. Come l'estremismo del web è diventato mainstream. Roma: Luiss University Press.

148. Nagle, A. 2015. "An Investigation into Contemporary Online Anti-Feminist Movements." [Doctoral dissertation]. Dublin City University).

149. Nisbett, R. E., & Schachter, S. (1966). Cognitive manipulation of pain. Journal of Experimental Social Psychology, 2, 227–236

150. Nithyanand, R., Schaffner, B., Gill, P. (2017). Online political discourse in the Trump era. arXiv preprint arXiv:1711.05303.

151. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y. (2016, April). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

152. Nouri, F., Behmanesh, J., MOHAMMAD, N. B., & Rezaei, H. (2012). Evaluation of WMS/HEC-HMS model in flood forecasting of Ghorve watershed.

153. Obar J.A. and Wildman S. (2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy, 39(9): 745-750.

154. Olteanu, A., Castillo, C., Boy, J., Varshney, K. (2018, June). The effect of extremist violence on hateful speech online. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1).

155. Ordóñez, F. J., Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors, 16(1), 115.

156. Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning (No. 47). University of Illinois press.

157. P. Turney, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, in Proceedings of the Twelfth European Conference on Machine Learning in Springer Verlag, Berlin, pp. 491-502, 2001.

158. P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 417–424, 2002.

159. Pak, P. Paroubek, Twitter based system: using Twitter for disambiguating sentiment ambiguous adjectives, Proceedings of the 5th

International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2010, pp. 436–439.

160. Park, J. H., Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206.

161. Pasta S. (2018). Razzismi 2.0. Analisi socio-educativa dell'odio online. Brescia: Scholé.

162. Pence and Paymar (1993) Education Groups for Men who Batter: The Duluth Model, New York: Springer

163. Phillips W. (2018). The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online. Disponibile in: datasociety.net/wpcontent/uploads/2018/05/FULLREPORT_Oxygen_ of_Amplification_DS.pdf. [12/09/2019].

164. Phillips, R., Kelly, L., & Westmarland, N. (2013). Domestic violence perpetrator programmes: an historical overview.

165. Pinel, J. P., Assanand, S., & Lehman, D. R. (2000). Hunger, eating, and ill health. American Psychologist, 55(10), 1105.

166. Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Theories of emotion (pp. 3-33). Academic press.

167. Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In 4th Italian Conference on Computational Linguistics, CLiC-it 2017 (Vol. 2006, pp. 1-6). CEUR-WS.

168. Postmus, J. L., Hoge, G. L., Breckenridge, J., Sharp-Jeffs, N., & Chung, D. (2020). Economic abuse as an invisible form of domestic violence: A multicountry review. Trauma, Violence, & Abuse, 21(2), 261-283.

169. Projansky S. (2001). The elusive/ubiquitous representation of rape: a historical survey of rape in U.S. film, 1903-1972. Cinema Journal, vol. 41, no. 1.

170. Puschmann C., and Ausserhofer J. (2017). Social Data APIs: Origin, Types, Issues. In: Schäfer M.T and Van Es K., The Datafied Society: Studying Culture through Data. Amsterdam: Amsterdam University Press. DOI: 10.25969/mediarep/12401.

171. Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251.

172. Randall, M. (2010). Sexual Assault Law, Credibility, and "Ideal Victims": Consent, Resistance, and Victim Blaming. Canadian Journal of Women and the Law, 22: 397–433. DOI: 10.3138/cjwl.22.2.397.

173. Rekik, R., Kallel, I., Casillas, J., Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. International journal of information management, 38(1), 201-216.

174. Renauer, B., & Henning, K. (2005). Investigating intersections between gender and intimate partner violence recidivism. Journal of Offender Rehabilitation, 41(4), 99-124.

175. Ringrose J. and Renold E. (2011) Slut-shaming, girl power and 'sexualisation': thinking through the politics of the international SlutWalks with teen girls. Gender and Education, vol. 24, n. 3.

176. Riva, Nicola. "Il principio del danno e le espressioni d'avversione o d'odio." Biblioteca della libertà 54.224 (2019): 19.

177. Rogers R. (2013). Digital methods. MIT press. , testo disponibile al sito: https://mitpress.mit.edu/books/digital-methods, 06/06/2020.

178. Ross C. (2015), Lessons in Censorship, Cambridge (MA), Harvard University Press.

179. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv preprint arXiv:1701.08118.

180. Russell, J. A. (1991). Culture and the categorization of emotion. Psychological Bulletin, 110, 426–450.

181. Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological review,110(1), 145.

182. Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of research in Personality, 11(3), 273-294.

183. Salem, A. (1987). Pratique des segments répétés. Paris : Klincksieck.

184. Salem, A. (1995). La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert. In: Langages de La Révolution (1770-1815) (Actes Du 4ème Colloque International de Lexicologie Politique). Paris: Klincksieck

185. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing Management 24(5):513-523

186. Salton G, Wong A, Yang C (1975) A vector space model for automatic indexing. Communications of the ACM 18(11):613-620.

187. Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. Psychological Review, 69, 379–399.

188. Schmidt, A., Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

189. Schmidt, A., Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

190. Schneirla, T. C. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal.

191. Serra, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., Vakali, A. (2017, August). Class-based prediction errors to detect hate speech with out-of-vocabulary words. In Proceedings of the first workshop on abusive language online (pp. 36-40).

192. Shepherd, L.J. (2013). Gender, violence and popular culture: Telling stories. Oxon: Routledge.

193. Shorey, R. C., Stuart, G. L., & Cornelius, T. L. (2011). Dating violence and substance use in college students: A review of the literature. Aggression and violent behavior, 16(6), 541-550.

194. Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 283(1), 737-757.

195. Solomon, R. L. (1977). Addiction: an opponent-process theory of acquired motivation: the affective dynamics of addiction.

196. Sternitzke, C., Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. Scientometrics, 78(1), 113-130.

197. Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

198. Straka, M., Hajic, J., & Straková, J. (2016, May). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4290-4297).

199. Swim, J. K., Hyers, L. L., Cohen, L. L., Ferguson, M. J. (2001). Everyday sexism:Evidence for its incidence, nature, and psychological impact from three daily diary studies. Journal of Social Issues, 57, 31–53. http://dx.doi.org/10.1111/0022-4537.00200.

200. Takane Y., Shibayama T. (1991). Principal Component Analysis with External Information on both subjects and variables, Psychometrika, 56, 97-120.

201. Takane, Y. (1997, October). CPCA: A comprehensive theory. In 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation (Vol. 1, pp. 35-40). IEEE.

202. Tesser, A. (2000). On the confluence of self-esteem maintenance mechanisms. Personality and Social Psychology Review, 4(4), 290-299.

203. Thompson S., Yar. M. (2011), The Politics of Misrecognition, Farnham, Ashgate.

204. Turkle S. (2012). Insieme ma soli. Perché ci aspettiamo sempre più dalla tecnologia e sempre meno dagli altri. Torino: Codice.

205. Turkle S. (2016). La conversazione necessaria. La forza del dialogo nell'era digitale. Torino: Einaudi.

206. Turton-Turner, P. (2013, March). Villainous avatars: The visual semiotics of misogyny and free speech in cyberspace. In Forum on Public Policy: A Journal of the Oxford Round Table. Forum on Public Policy.

207. Twenge J.M. (2018). Iperconnessi. Perché i ragazzi oggi crescono meno ribelli, più tolleranti, meno felici e del tutto impreparati a diventare adulti. Torino: Einaudi.

208. Unsvåg, E. F., Gambäck, B. (2018, October). The effects of user features on twitter hate speech detection. In Proceedings of the 2nd workshop on abusive language online (ALW2) (pp. 75-85).

209. V. Hatzivassiloglou and K. R. McKeown, Predicting the semantic orientation of adjectives, in Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, ES, pp. 174–181, 1997.

210. Vickery, J.R., and T. Everbach, eds. 2018. Mediating Misogyny: Gender, Technology, and Harassment. Palgrave Macmillan. https://www.amazon.com/dp/3319729160/ref=rdr_ext_tmb

211. Waldron J. (2012), The Harm in Hate Speech, Cambridge (MA), Harvard University Press. Wongher V. (2015), Disciplina della libertà di espressione sull'hate speech nell'Unione Europea e negli Stati Uniti d'America: profili a confronto, https://tesi.luiss.it/15958/1/wongher-valeria-tesi-2015.pdf.

212. Wallace P. (2017). La psicologia di Internet. Milano: Raffaello Cortina

213. Warner, W., Hirschberg, J. (2012, June). Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media (pp. 19-26).

214. Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods, 45(4), 1191-1207.

215. Waseem, Z. (2016, November). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science (pp. 138-142).

216. Waseem, Z. (2016, November). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science (pp. 138-142).

217. Waseem, Z., Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

218. Wierzbicka, A. (1999). Emotions across languages and cultures. New York: Cambridge University Press.

219. Wilkinson, D. L., & Hamerschlag, S. J. (2005). Situational determinants in intimate partner violence. Aggression and Violent Behavior, 10(3), 333-361.

220. Wulczyn, E., Thain, N., Dixon, L. (2017, April). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th international conference on world wide web (pp. 1391-1399).

221. Y. Liu, X. Huang, A. An, X. Yu, ARSA: a sentiment-aware model for predicting sales performance using blogs, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2007, pp. 607–614.

222. Y. Mehdad and J. Tetreault. 2016. Do characters abuse more than words? In Proc. of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). pages 299–303.

223. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456).

224. Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai. 2005. US Domestic Extremist Groups on the Web: Link and Content Analysis. IEEE intelligent systems, 20(5):44–51.

225. Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE

226. Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80, Amsterdam, Netherlands, September. IEEE.

227. Yun, J., Jing, L., Yu, J., Huang, H. (2010). A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications, 39(2), 2035-2046.

228. Zhang, Z., Robinson, D., Tepper, J. (2018, June). Detecting hate speech on twitter using a convolution-gru based deep neural network. In European semantic web conference (pp. 745-760). Springer, Cham.

229. Ziccardi G. (2016). L'odio online. Violenza verbale e ossessioni in rete. Milano: Raffaello Cortina.

230. Zuo Y., Zhao J. and Xu K. (2015). Word network topic model: a simple but general solution for short and imbalanced texts, London: Knowledge and Information Systems.