Sentieri semantici della violenza: un algoritmo per l'individuazione di potenziali vittime

Fiorenza Deriu[1], Roberto Navigli[2] , Pasquale Pavone[3]

[1]University of Roma "La Sapienza" – fiorenza.deriu@uniroma1.it

[2]University of Roma "La Sapienza" – navigli@di.uniroma1.it

[3]Sant'Anna School of Advanced Studies – pasquale.pavone@santannapisa.it

Abstract

This paper concerns a preparatory study that is part of a larger research project entitled "Active Web against Violence", funded by the International House of Women in Rome. The project aims at preventing violence against women, reaching potential victims through the web, so trying to facilitate their determination to ask for help (disclosure). Evidences show that a very low percentage of women suffering from violence ask for help (Istat, 2014). The objective of this study is to produce an integrated system of linguistic resources on violence against women preliminary to building up an algorithm that, similarly to what already happens in the marketing field, allows to identify, through the analysis of the implicit and explicit language used by women on the web, the women potentially interested in the problem. The project is developed in several phases and in this paper some preliminary results are presented.

The study is based on a large corpus of 36,241 electronic printed articles, provided by the Monrif group with reference to the last 10 years, for a total of 12,932,187 occurrences and 170,552 graphic forms (hapax 39.87%), selected starting from the keyword "violence". The corpus was preliminarly analysed using automatic lexical-textual techniques (Bolasco, 2013), and then submitted to an unsupervised classification in order to identify some main word communities specifically concerning violence against women. At this aim, the corrispondence analisys and a clustering alghorithm were used.

Four classes of specific words concerning violence against women were identified. At this point, using the powerful multilingual semantic web of Babelnet, we tried to reconnect the single words, the syntagmas and the polithemes, contained in the 4 extracted vocabularies, to some concepts implicitly connected to them. This process enabled the creation of a knowledge graph representing the different aspects of violence in a multilingual perspective. Thanks to these new graphs, it will be possible to analyse new textual documents without looking for specific terms used in the corpus, because the algorithm will enable the researcher to identify anyway the semantic correlations to the different violence topics.

The next step of the project is aimed at the definition of a new algorithm of conversion of the implicit language on violence into an explicit one.

Keywords: violence against women, text mining, semantic web

Riassunto

Questo studio preparatorio si inserisce in un più ampio progetto di ricerca dal titolo "Web Attivo contro la Violenza", finanziato dalla Casa Internazionale delle Donne di Roma, con l'obiettivo di prevenire la violenza contro le donne, raggiungendo le potenziali vittime attraverso il web e cercando così di facilitarnee la determinazione alla richiesta di aiuto (*disclosure*), I dati ci mostrano infatti che solo una bassa percentuale di donne che vivono una relazione violenta chiede aiuto (Istat, 2014). Obiettivo di questo studio è costruire un sistema integrato di risorse linguistiche sulla violenca contro le donne preliminare alla costruzione di un algoritmo che, analogamente a quanto già avviene nell'ambito del marketing, consenta di individuare, attraverso

l'analisi del linguaggio implicito ed esplicito utilizzato dalle donne sul web, quelle potenzialmente interessate al/dal problema. Il progetto si sviluppa in più fasi e in questo paper si presentano alcuni risultati preliminari.
Lo studio si basa su un ampio corpus di 36.241 articoli a stampa elettronica, fornito dal gruppo Monrif con riferimento agli utimi 10 anni, per un totale di 12.932.187 occorrenze e 170.552 forme grafiche (hapax 39,87%), selezionati a partire dalla parola chiave "violenza". Il corpus è stato preliminarmente analizzato con tecniche lessico-testuali (Bolasco, 2013) e poi sottoposto a una classificazione non supervisionata per identificare alcune principali community di parole relative più specificamente alla violenza contro le donne. A tal scopo sono state utilizzate sia l'analisi delle corrispondenze semplici sia un algoritmo di cluster analysis.
Sono state identificate 4 classi di parole specifiche sul tema della violenza contro le donne. A questo punto, facendo ricorso alla potente rete semantica di BabelNet, si è cercato di ricollegare le singole parole, i sintagmi e le polirematiche, contenute nei 4 vocabolari estratti, ad alcuni concetti implicitamente ad essi collegati. Questo processo ha consentito la creazione di un knowledge graph rappresentante i differenti aspetti della violenza in un ambiente mutilingue. Grazie a questi nuovi grafi sarà possibile analizzare nuovi documenti testuali senza cercare i termini utilizzati nel corpus perché l'algoritmo consentirà al ricercatore di identificare comunque le correlazioni semantiche utili con i differenti ambiti tematici della violenza.
La prossima fase del progetto di ricerca è finalizzata alla costruzione di uno specifico algoritmo di conversione del linguaggio implicito in quello esplicito sulla violenza contro le donne.

Parole chiave: violenza contro le donne, text mining, semantic web

1. Introduzione

La violenza contro le donne è un fenomeno ancora oggi in grande misura sommerso. Uno dei maggiori ostacoli al contrasto della violenza, soprattutto quando si realizza entro le mura domestiche, consiste nella vergogna e l'ingiustificato senso di colpa che le vittime avvertono nel rivelare tale esperienza a chi è loro vicino. A ciò si aggiunga la scarsa conoscenza dei, o mancanza di fiducia nei confronti dei, soggetti istituzionali e non, che dovrebbero occuparsi di proteggerle e tutelarle.

In base ai più recenti dati Istat, il 28% delle donne che subiscono violenza da partner e il 25,5% di quelle che la subiscono da non partner non parla con nessuno della violenza subita (2014). Il tasso di denuncia è ancora più basso, se si considera che si assesta al 12% per le vittime di violenza da partner e al 6% per quella da non partner. Ancor meno sono le donne che, in caso di violenza, si rivolgono a un Centro Antiviolenza o a un centro specializzato per chiedere aiuto: non raggiungono il 4% (Istat, 2014). Un dato ancor più allarmante consiste nella disinformazione di molte donne che solo nel 12,8% dei casi sanno dell'esistenza dei Centri Antiviolenza, di sportelli o servizi dedicati a questo tipo di supporto e assistenza. Ciò avviene nonostante a partire dagli anni Novanta l'attenzione dei media su questo fenomeno sia andata crescendo e la sensibilità culturale verso queste manifestazioni dell'asimmetria di potere tra uomini e donne sia aumentata.

Le donne non chiedono aiuto per vari motivi. Temono di essere messe sotto accusa proprio da chi dovrebbe ascoltare e credere nei fatti descritti, a causa di una cultura che non è riuscita ancora a prendere le distanze da una serie di pregiudizi e stereotipi che vedono nella donna la causa stessa della violenza (doppia vittimizzazione - *secondary victimization*) (Baldry, 1996); hanno vergogna di mostrare il fallimento di una relazione in cui hanno creduto e in cui ancora credono, nonostante tutto; temono di perdere i propri figli, nel momento in cui i servizi di assistenza sociale dovessero prendere atto del contesto famigliare violento, perché pensano di non essere delle madri adeguate; temono di non avere i mezzi per poter andare avanti, perché spesso sono dipendenti economicamente dal partner violento.

Se, dunque, le donne vittime di violenza non chiedono aiuto, è necessario cercare di raggiungerle per dare loro il coraggio necessario a uscire dal proprio isolamento e da quella "spirale della violenza" da cui è difficile, ma possibile, liberarsi (Walker, 1979), se accompagnate da operatrici appositamente formate a tale scopo.

Questo studio si inserisce in un più ampio progetto interdisciplinare di ricerca[1], dal titolo "Web Attivo contro la Violenza"[2], finanziato dalla Casa Internazionale delle Donne di Roma, finalizzato in ultima analisi alla costruzione di un algoritmo di conversione del linguaggio implicito in esplicito (DePaNa[3] algorithm), attraverso il quale individuare e raggiungere le potenziali vittime di violenza attraverso il web, per stabilire un primo contatto e facilitarne la determinazione alla richiesta di aiuto (*disclosure*), propedeutica al successivo accompagnamento nel processo di uscita dalla relazione violenta. L'algoritmo costruito a fini sociali, opera analogamente a quelli utilizzati a fini commerciali e di marketing, consentendo di individuare, attraverso l'analisi del linguaggio implicito ed esplicito utilizzato dalle donne sul web, quelle potenzialmente interessate al/dal problema. La decodifica dei contenuti testuali riconducibili a fenomeni di violenza, attiva l'invio alle donne che li hanno pubblicati di un link a una piattaforma informatica di informazione e orientamento nella ricerca di aiuto in rete.

Il progetto si sviluppa in più fasi e in questo paper si presentano alcuni risultati preliminari, relativi alla costruzione di alcune risorse linguistiche caratterizzanti quattro specifici domini della violenza sulle donne e la loro integrazione nelle tassonomie del web semantico di BabelNet (Navigli & Ponzetto, 2012), una rete semantica multilingue e una ontologia lessicalizzata. Tuttvaia, sebbene si tratti di lavoro preliminare alla costruzione dell'algoritmo di traduzione del linguaggio implicito sulla violenza in quello esplicito, in esso sono già contenute alcune importanti potenzialità di disambiguazione dei testi oggetto di studio. Infatti, già in questa fase, grazie alla integrazione delle risorse linguistiche prodotte attraverso l'analisi lessico-testuale-prima e di classificazione poi, con le risorse multilingua del web semantico di BabelNet, è stato possibile sfruttare un algoritmo di disambiguazione dei significati delle parole (Word Sense Disambiguation) che collega i termini ai concetti presenti nella rete semantica di BabelNet. Per ciascun documento, l'algoritmo utilizza i termini estratti e crea un grafo dei potenziali significati e dei collegamenti semantici tra essi. Pertanto, già allo stato attuale, qualora si volessero studiare nuovi documenti, i concetti di interesse non dovrebbero necessariamente apparire nei testi degli articoli, perché sarà il grafo a rivelarli, indipendentemente dalla lingua in cui siano espressi. Infatti, i grafi prodotti da Babelnet possono essere già validamente utilizzati per determinare l'attinenza di futuri testi (ricerche online, documenti esaminati, ecc.) scritti da donne potenzialmente o effettivamente vittima di violenza ai diversi aspetti di violenza identificati in questo studio preliminare.

In questo articolo, dopo la descrizione del corpus dati utilizzato e dei metodi di analisi prescelti, si procede nel paragrafo 3 alla descrizione dei processi che hanno condotto alla identificazione delle diverse classi tematiche relative al concetto di violenza in generale e alle

[1] Il gruppo di ricerca è costituito da una sociologa, uno statistico sociale e uno spin-off Sapienza, Babelscape, diretto da un informatico, creatore di BabelNet.

[2] Il progetto nasce da un'idea della web-designer Lorella Muzi, brevettata nella sua formulazione embrionale.

[3] L'acronimo dell'algoritmo indica le iniziali dei cognomi dei tre studiosi che lo hanno ideato e costruito (Deriu, Pavone e Navigli).

4 classi specifiche relative alla violenza contro le donne; nel successivo paragrafo 4 è descritta la fase di integrazione delle risorse linguistiche acquisite con la cluster analysis nel web semantico multilingue di Babelnet e la successiva creazione di knowledge graphs capaci di individuare una serie di connessioni concettuali nascoste, latenti, sottostanti le parole delle risorse linguistiche considerate. Nelle conclusioni i principali risultati già conseguiti in questa fase preliminare della ricerca.

2. Dati e metodi

Il corpus in analisi, fornito dal gruppo Monrif[4], è composto da 36.241 articoli a stampa elettronica, riferiti agli ultimi dieci anni, selezionati in base alla presenza nel testo del termine «Violenza» e relativo a tre testate giornalistiche: Il Giorno, il Resto del Carlino e La Nazione. La dimensione totale del corpus è di 12.932.187 occorrenze, relativo a 170.552 forme grafiche diverse di cui 39,87% hapax.

Nel presente lavoro viene applicata un'analisi automatica lessico-testuale (Bolasco, 2013) e una classificazione non supervisionata per individuare le principali tematiche che caratterizzano la narrazione della "Violenza" presente all'interno del corpus in analisi. Con il fine di lasciare il più possibile aperta la definizione degli argomenti da selezionare, vengono adottate le tecniche esplorative multidimensionali in modo da poter far emergere gli argomenti sulla base delle caratteristiche dei testi in analisi (Bolasco, 1999).

L'individuazione delle principali tematiche contenute nel corpus è effettuata attraverso l'analisi delle corrispondenze e l'analisi dei cluster (Benzécri, 1992; Greenacre, 2017). L'analisi multidimensionale permette di osservare la somiglianza degli articoli in base al loro contenuto lessicale. La fase di clustering basata sulle parole è una classificazione non supervisionata che riflette la somiglianza semantica tra i record. Questa omogeneità concettuale rende esplicito il tema o il tratto semantico prevalente in un gruppo di articoli; tale tratto semantico può essere riassunto in una categoria non definita a priori ma ottenuta attraverso l'analisi. Ciascun cluster viene quindi etichettato attraverso la lettura esperta dei dizionari associati a ciascun gruppo di articoli.

L'analisi del corpus è stata effettuata utilizzando quattro software. TaLTaC2 (Bolasco, 2010) è stato utilizzato per: l'analisi lessico-testuale; l'identificazione delle unità di analisi lessicale (intesa sia come forme semplici che come espressioni multiword presenti nel corpus); e la definizione delle matrici lessicali e testuali (da esplorare attraverso l'analisi multidimensionale). Spad® è stato utilizzato per l'analisi multidimensionale e l'analisi dei cluster. Attraverso il software IRaMuTeQ (Ratinaud, 2009) è stato possibile definire un'analisi di rete delle co-occorrenze interne a ciascun gruppo tematico individuato, mentre con il software Gephi (Bastian et al., 2009) sono stati elaborati i Grafi prodotti e sono state individuate le diverse "communities" di parole.

3. Risultati dell'analisi lessico-testuale

Nel presente paragrafo sono esposte le principali fasi dell'esplorazione del corpus avvenute attraverso l'utilizzo di alcuni strumenti dell'analisi automatica dei testi. Partendo dall'individuazione delle unità di analisi lessicale fino alla definizione dei percorsi semantici interni a gruppi omogenei in termini di tematica generale.

[4] http://www.monrifgroup.net/

3.1. Identificazione delle unità di analisi lessicale

Attraverso il processo di annotazione grammaticale delle forme grafiche del Vocabolario e l'applicazione di un modello lessico-testuale (Pavone, 2010, 2018) per la ricerca delle strutture sintattiche è stato possibile identificare le *Multiword Expressions* (MWEs) di tipo nominale presenti nel corpus. Viste le grandi dimensioni del corpus, si è proceduto a riconoscere come unità di analisi solamente le MWEs con almeno 50 occorrenze, sono state quindi lessicalizzate e riconosciute 1.304 MWEs. Di seguito vengono riportate le venti MWEs più occorrenti: *violenza sessuale* (2649 occ.), *sostituto procuratore* (1862 occ.), *anni di reclusione* (1474 occ.), *colpi di pistola* (1455 occ.), *arresti domiciliari* (1442 occ.), *rito abbreviato* (1379 occ.), *parte civile* (1255 occ.), *omicidio volontario* (1229 occ.), *squadra mobile* (1030 occ.), *Corte d' Assise* (985 occ.), *mesi di reclusione* (873 occ.), *sequestro di persona* (858 occ.), *colpo di pistola* (812 occ.), *primo grado* (811 occ.), *anni di carcere* (809 occ.), *omicidio colposo* (806 occ.), *luogo del delitto* (706 occ.), *abusi sessuali* (695 occ.), *pubblico ufficiale* (681 occ.), *piede libero* (679 occ.). Come si può notare il riconoscimento delle MWEs permette di disambiguare semanticamente le forme semplici[5], consentendo di identificare la terminologia giornalistica della narrazione degli eventi violenti.

Grazie alla classificazione grammaticale delle forme grafiche è stato possibile anche distinguere le forme attive (intese come parole di contenuto: sostantivi, verbi, aggettivi e avverbi) e le forme supplementari (in letteratura definite anche come *stop words* o parole di struttura del discorso: congiunzioni, preposizioni, articoli ecc.). anche in questo caso sono state considerate come unità di analisi esclusivamente le forme attive con almeno 50 occorrenze nel corpus. Sono state pertanto identificate 9.569 forme grafiche attive[6] (semplici e MWEs) e si è prodotta una matrice *Testi* × *Forme Selezionate* (36.241 × 9.569)

3.2 Identificazione delle tematiche principali

Attraverso l'analisi delle corrispondenze effettuata sulla matrice testuale e la cluster analysis applicata ai risultati dell'analisi fattoriale, è stato possibile classificare in modo univoco i diversi articoli in analisi. Attraverso lo studio del dendrogramma risultante dall'analisi, vengono riconosciuti dieci gruppi principali. Per ognuno di essi viene definita la tematica prevalente della loro similarità, sulla base della lettura esperta dei loro dizionari di termini caratteristici. Nella figura 1 è riportata la distribuzione del lessico sul piano fattoriale *f1f2.*

[5] A titolo di esempio riguardo il processo di disambiguazione semantica, si riportano di seguito tutte le 74 diverse MWEs identificate relative alla parola « violenza »: *violenza sessuale*, (2649), *episodi di violenza*, (513), *violenza carnale*, (485), *violenza privata*, (484), *episodio di violenza*, (394), *accusa di violenza sessuale*, (332), *atti di violenza*, (273), *accusato di violenza sessuale*, (194), *casi di violenza*, (142), *violenza sessuale di gruppo*, (132), *violenza subita*, (124), *segni di violenza*, (121), *violenza inaudita*, (120), *forma di violenza*, (118), *atto di violenza*, (116), *tentativo di violenza*, (110), *violenza sessuale aggravata*, (105), *accusa di violenza*, (103), *violenza fisica*, (99), vittima della violenza, (96), escalation di violenza, (89), accuse di violenza sessuale, (83), esplosione di violenza, (81), reato di violenza sessuale, (78), caso di violenza, (74), storia di violenza, (67), tentativo di violenza sessuale, (65), violenza gratuita, (61), ondata di violenza, (58), violenza politica, (58), vittime di violenza, (57), violenza domestica, (54), reati di violenza sessuale, (52), violenza psicologica, (51), spirale di violenza, (51), vittime della violenza, (50).

[6] Volendo procedere al riconoscimento della similarità dei testi in base ai contenuti, in questa fase sono state considerate come forme attive unicamente i sostanti e gli aggettivi.

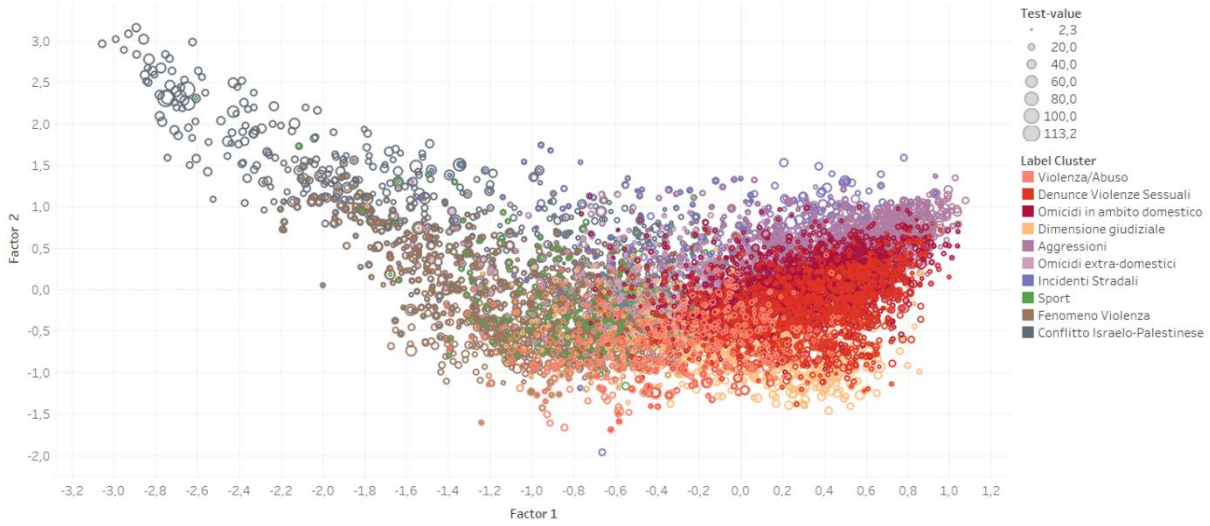

Figura 1 Distribuzione sul Piano fattoriale f1f2 delle forme caratteristiche dei dieci gruppi di articoli individuati. Le dimensioni dei punti sono proporzionali al test-value, il colore indica l'appartenenza al gruppo individuato.

L'analisi delle corrispondenze ci consente di evidenziare le relazioni esistenti nell'intero corpus, producendo la migliore rappresentazione simultanea dei profili riga e colonna della matrice e rappresenta l'informazione in termini di similarità fra gli elementi della matrice. La principale differenza degli argomenti trattati all'interno del corpus della violenza, osservabile sul primo fattore, consiste nella polarizzazione (da destra verso sinistra) fra la narrazione della violenza come fenomeno collettivo che interessa grandi aggregati di individui (guerra, concetto generale di violenza, violenza nello sport) rispetto agli eventi di violenza individuale, che interessano singole persone. L'analisi dei dieci gruppi individuati attraverso la cluster analysis ha permesso di distinguere fra gli articoli direttamente collegati all'ambito della violenza sulle donne, rispetto alle altre narrazioni di violenza collegate ad altri ambiti. Nella Figura 2 sono riportate le principali forme grafiche caratteristiche dei dizionari dei quattro gruppi collegati alla violenza sulle donne.

Il primo gruppo concerne la *violenza domestica*, che si realizza in ambito famigliare e che mette a tema i casi di maltrattamento e abuso di donne e minori. Questo dominio concettuale include quindi anche la *violenza assistita* riferita minori che vivono in nuclei violenti, spettatori e testimoni, più o meno consapevoli dei maltrattamenti del proprio padre nei confronti della loro madre. Ricorrono tra le forme grafiche semplici le parole *donne*, *minori*, *bambini*, *abusi* e *maltrattamenti*. Tra le MWEs si evidenziano *abusi sessuali*, *donne maltrattate*, ma anche *violenza domestica*, *mura domestiche*, *ambito familiare*, *interno della famiglia*.

Il secondo gruppo concerne le denunce di *violenze sessuali e stupri*. Si tratta di forme di violenza che interessano soprattutto le giovanissime, talvolta minori, che però riescono a superare la paura e la vergogna dell'accaduto, e con l'aiuto e l'assistenza di genitori o assistenti sociali attivano il processo di denuncia della violenza. Tra le parole semplici ricorrono *ragazzina*, *bambina*, *ragazza*; così come *genitori*, *denuncia* e *accuse*. Tra le MWEs si segnalano *violenza sessuale*, *accusa di violenza sessuale* e *assistenti sociali*.

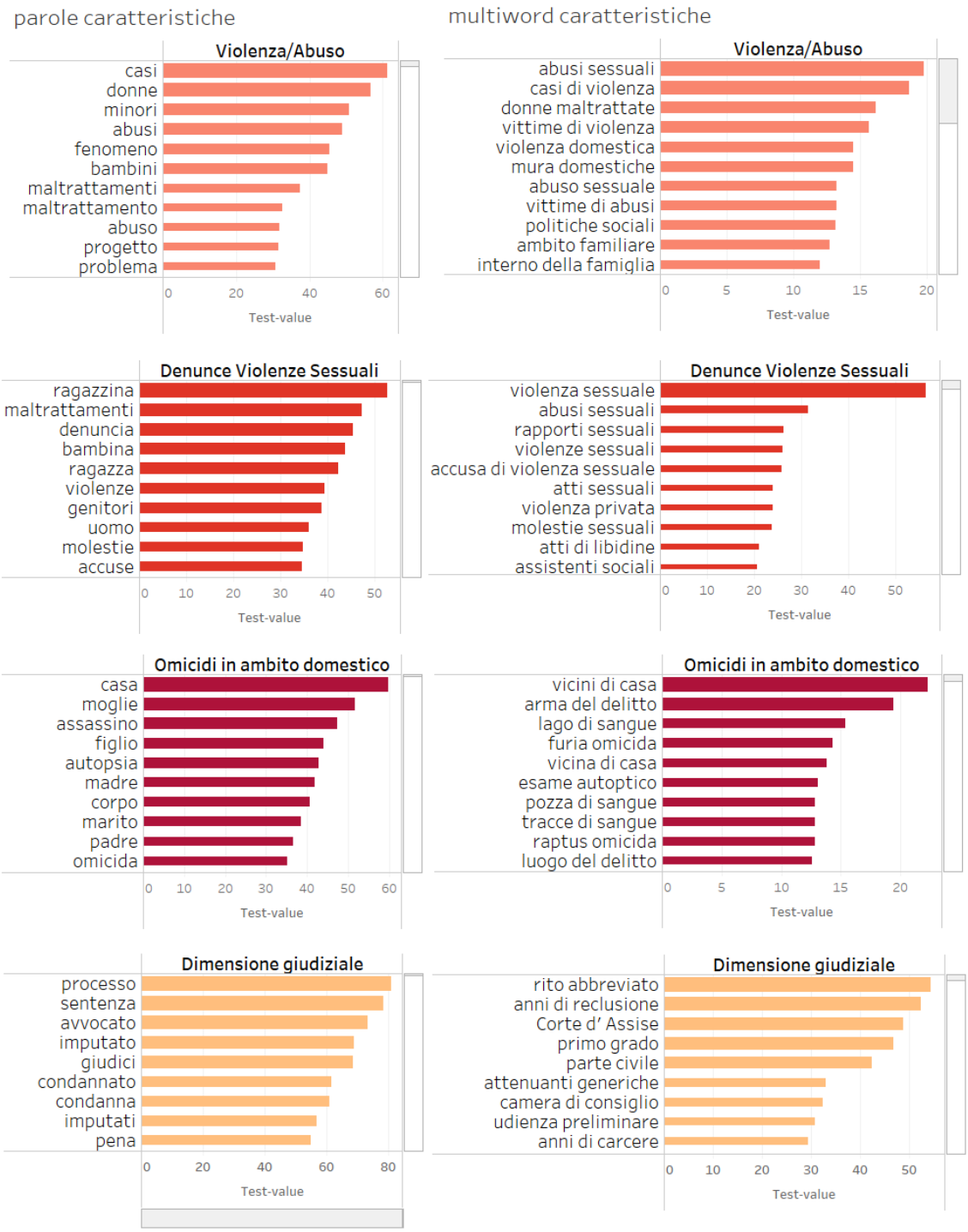

Figura 2 Dizionari Caratteristici dei gruppi di articoli collegati alla narrazione della violenza sulle donne

Il terzo gruppo connota chiaramente il *femminicidio*, l'omicidio di donne in quanto donne. Le forme grafiche semplici evocano un registro narrativo, tipico della cronaca giornalistica, in cui domina la descrizione dettagliata degli attori (*moglie*, *marito*, *padre*, *madre*, *figlio*, *assassino*, *omicida*), dei luoghi (*casa*, *cucina*, *camera*), degli oggetti (*coltello*, *pistola*, *arma*) e dei tempi della violenza. A queste forme grafiche si aggiungano MWEs come *furia omicida*,

raptus omicida che evidenziano la tendenza da parte della stampa a ricondurre la violenza sulle donne a forme estemporanee di perdita del controllo, a forme di disagio e stress psicologico e mentale, tralasciando che la violenza sulle donne si realizza invece senza soluzione di continuità nel quotidiano. E ancora espressioni come *lago di sangue*, *pozza di sangue*, *tracce di sangue*, *esame autoptico* riconducibili alla narrazione, talora *voyeuristica*, della scena del delitto e dell'azione investigativa.

Infine, il gruppo centrato sulle diverse fasi dell'*iter giudiziale* che si avvia in seguito alla denuncia e che coinvolge *giudici*, *imputati*, *avvocati*; che si traduce in *sentenze*, *condanne*, e nella comminazione di *pene*, al termine del *processo*. Le MWEs mettono ancora meglio in luce il carattere tecnico del linguaggio espresso da questo dominio concettuale: *rito abbreviato*, *primo grado*, *udienza preliminare*, *camera di consiglio*; e ancora *anni di reclusione*, *anni di carcere*, *attenuanti generiche*.

3.3 Identificazione dei percorsi semantici

Dopo aver definito quattro diversi sub-corpora di articoli, ognuno dei quali si presenta omogeneo in termini di contesto semantico, si è proceduto a elaborare, per ciascun sub-corpus, un'analisi di rete delle co-occorrenze e una identificazione delle communities di parole in base alla modularity (Blondel et al. 2008).

Nella figura 3 sono rappresentati i grafi ottenuti dalla Network analysis applicata sui sub-corpora Omicidi Domestici (sinistra) e Violenza/Abuso (destra).

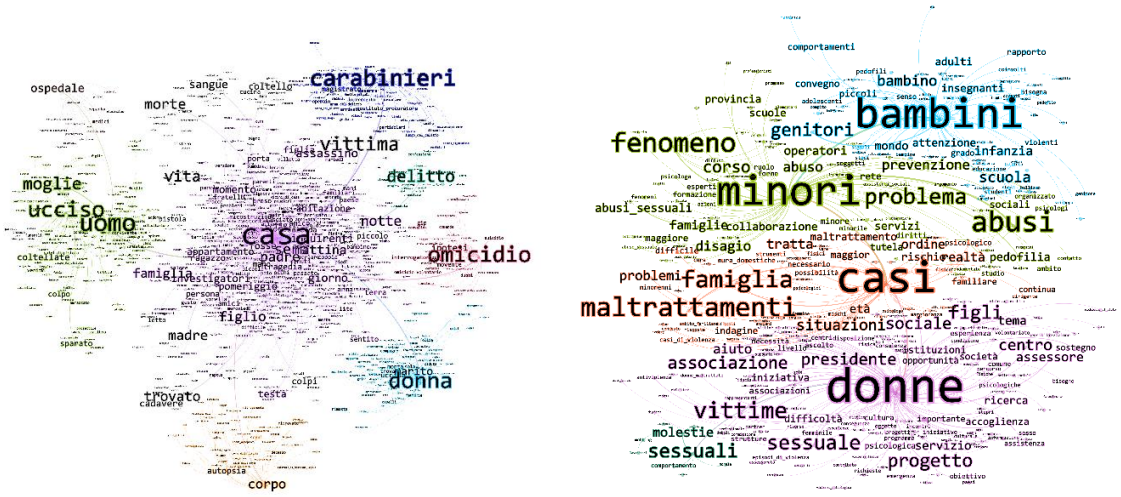

Figura 3 Analisi di rete effettuata sul sub-corpus Omicidi Domestici (sinistra) e sul corpus Abuso/Violenza (destra)

L'analisi delle co-occorrenze ha contribuito in questa fase alla ulteriore conferma e precisazione dei confini semantici dei domini concettuali individuati nella precedente analisi, nonché le parole pivot di alcuni sotto-domini semantici.

4. Disambiguazione esplicita e collegamento a BabelNet

Sebbene l'analisi delle co-occorrenze abbia fornito numerose indicazioni sulla semantica della violenza nei suoi diversi aspetti, abbiamo ulteriormente analizzato e collegato il contenuto degli articoli del quotidiano utilizzando un algoritmo di disambiguazione dei significati delle

parole (*Word Sense Disambiguation*) il cui scopo è collegare i termini ai concetti presenti nella rete semantica di BabelNet. Per ciascun documento, l'algoritmo utilizza i termini estratti e crea un grafo dei potenziali significati e dei collegamenti semantici tra essi. Tali collegamenti sono ottenuti dal grafo di BabelNet e provengono dalle diverse risorse integrate, quali WordNet, Wikipedia, WikiData e così via. Un algoritmo basato sull'identificazione dell'interpretazione più coesa (ovvero del sottografo più connesso) fornisce la scelta dei significati più probabili per i termini estratti da un singolo articolo.

Tali informazioni semantiche sono state quindi aggregate in modo da ottenere un grafo semantico della violenza. Il vantaggio di tale grafo rispetto a cluster o nuvole di parole è che i concetti identificati sono per loro natura multilingui e legati a risorse simboliche computazionali create a mano, al contrario dei cluster ottenuti automaticamente. Ciascun concetto corrisponde infatti a un cosiddetto *synset* multilingue, ovvero a un insieme di sinonimi che lessicalizzano il concetto in ciascuna lingua coperta dallo stesso. Ad esempio, il concetto di casa può essere espresso anche mediante la parola abitazione. Inoltre, essendo il *synset* multilingue, esso può contenere termini anche in altre lingue, quali house, dwelling, maison, ecc. Inoltre, il concetto è dato di definizioni testuali esplicative ed è legato in BabelNet ad altri concetti mediante relazioni di iperonimia (la violenza è un atto), iponimia (un tipo di violenza è la violenza sessuale), meronimia (la sentenza fa parte del processo), olonimia (il processo ha come parte il dibattimento) e altre relazioni di correlazione semantica (ad esempio, la violenza fisica è correlata al sangue). È importante notare come tali concetti non debbano necessariamente apparire nei testi degli articoli per essere utilizzati in seguito ai fini di analisi di nuovi articoli o documenti, indipendentemente dalla lingua degli stessi. Di conseguenza, il grafo risultante costituisce una generalizzazione semantica delle rappresentazioni di superficie estratte dagli articoli sotto forma di parole e polirematiche. Inoltre essendo la semantica multilingue è possibile utilizzare il grafo semantico per indicizzare contenuti in qualsiasi lingua così come arricchirlo utilizzando informazioni provenienti da altre lingue.

Un primo grafo semantico ottenuto dalla disambiguazione dei termini più rilevanti delle quattro community identificate precedentemente ha portato alla creazione di altrettanti grafi semantici (con un'interessante sovrapposizione concettuale data dai concetti "cardine" della violenza). Notiamo che tali grafi possono essere utilizzati per determinare l'attinenza ai diversi aspetti di violenza identificati di futuri testi (ricerche online, documenti esaminati, ecc.) scritti da donne potenzialmente o effettivamente vittima di violenza.

5. Conclusioni

In questo studio sono stati presentati i primi risultati di un'analisi preliminare, propedeutica alla costruzione di un algoritmo innovativo (DePaNa), diretto alla traduzione del linguaggio implicito della violenza in linguaggio esplicito. In questa prima fase è stato necessario costruire delle risorse linguistiche specifiche di alcuni domini concettuali della violenza. L'analisi ha portato alla enucleazione da un vasto corpus testuale di quattro domini relativi alla violenza domestica e assistita; alla violenza sessuale e lo stupro; al femminicidio; infine all'iter giudiziale. Risorse linguistiche che non si limitano al vocabolario specifico di tali domini ma anche alle MWEs e ai sintagmi caratterizzanti.

Le risorse linguistiche così individuate sono state utilizzate per integrare il web semantico di BabelNet, portando alla creazione di *knowledge graph* dei diversi aspetti della violenza. Grazie a tali grafi, sarà possibile analizzare nuovi testi senza ricercare necessariamente i

termini trovati nel corpus esaminato, ma identificando comunque le correlazioni semantiche con le tematiche della violenza.

Riferimenti bibliografici

Baldry A. C. (1996). Rape victims' risk of secondary victimization by police officers. *Criminological & Legal Psychology*, vol. 25: 65–68.

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Benzécri, J. P. (1992). *Correspondence Analysis Handbook*. Dekker, New York.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment. 2008 (10): doi:10.1088/1742-5468/2008/10/P10008.

Bolasco, S. (1999). Analisi multidimensionale dei dati [multidimensional analysis of data]. Roma: Carocci.

Bolasco, S. (2010). *Taltac2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. Milano: LED.

Bolasco, S. (2013). L'analisi automatica dei testi [Automatic text analysis]. Roma: Ca-rocci.

Greenacre, M. J. (2017). *Correspondence analysis in practice*. London: Chapman & Hall.

Istat (2014). Indagine sulla sicurezza delle donne. https://www.istat.it/it/violenza-sulle-donne/il-fenomeno/violenza-dentro-e-fuori-la-famiglia/consapevolezza-e-uscita-dalla-violenza

Navigli R., Ponzetto S. P. (2012). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.

Pavone, P. (2010). Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus. In Bolasco S., Chiari I. and Giuliano L. (Eds.). Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles, Sapienza University of Rome, (pp. 131-140). Roma: LED.

Pavone, P. (2018) Automatic Multiword Identification in a Specialist Corpus. In: Tuzzi A. (ed) *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer.

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. http://www.iramuteq.org.

Walker L. (1979). *The Battered Woman*. New York: Harper and Row.