Come riportano i quotidiani le notizie di femminicidio? L'analisi di lessico e sintassi attraverso tecniche di text mining

Selene Bianco[1], Marina Penasso[2], Marta Girotto[1], Raphael Zanotti[3], Luisa Mondo[1], Marco Dalmasso[1], Carlo Mamo[1]

[1] Servizio sovrazonale di Epidemiologia ASL TO 3, Grugliasco (TO)
[2] Dors, Centro Regionale di Documentazione per la Promozione della Salute, Regione Piemonte, Grugliasco (TO)
[3] Giornalista de La Stampa

Per comunicazioni: selene.bianco@unito.it

Perché analizzare le notizie di femminicidio

La violenza sulle donne è uno dei risultati più radicali delle differenze di genere nella salute, perché deriva sia dalla propensione alla prevaricazione maschile nei rapporti di potere che della condizione femminile di vittima nella nostra società.
La crescente attenzione rivolta al problema, anche dal punto di vista legislativo (Legge 15 ottobre 2013, n. 119 [1]), si associa ad una maggiore attenzione verso le tecniche di comunicazione, che risultano espressione del contesto sociale e culturale. Nel caso degli organi di informazione a maggiore diffusione (tra cui i quotidiani a maggiore tiratura), il linguaggio utilizzato per trattare la notizia può essere considerato una valida espressione della visione e percezione del problema nella società.

Cos'è il text mining

All'interno dei differenti metodi di data mining, finalizzati all'analisi ed alla strutturazione automatica di conoscenze a partire dalle fonti informative disponibili, il Text Mining si concentra nell'estrazione di informazioni implicite a partire da testi non strutturati, trasformandoli in dati analizzabili tramite un'elaborazione del linguaggio naturale (NLP) ed altri metodi analitici. Le tecniche di text mining effettuano un'analisi sintattica e semantica del testo, calcolano la frequenza delle parole utilizzate, individuano e visualizzano pattern e associazioni tra parole [1] [3].
L'utilizzo di questa tecnica permette di analizzare non solo dati strutturati ma di ricavare conoscenza ed informazione anche dalla grande mole di dati disponibili in forma non strutturata (testi, conversazioni, social media, …), consentendo di scoprire ed interpretare conoscenze e significati altrimenti nascosti all'interno della forma testuale.
Nell'ambito delle scienze della comunicazione i metodi di text mining sono utilizzati per estrarre da un testo i lemmi e le associazioni che, ricorrendo con maggiore frequenza, ne definiscono il messaggio trasmesso [4][5][6].

Tipi di testi reperiti e criteri di selezione

Si è ricostruito l'elenco delle vittime di femminicidi avvenuti in Piemonte tra il 2005 ed il 2010. Per far questo, si è partiti dai dati riportati dalla Casa delle donne [7], collegandoli alle schede di morte di fonte Istat. Dalle schede di morte si sono recuperati i principali dati socio-demografici, quali l'età, lo stato civile, la nazionalità, la condizione professionale, nonché informazioni più dettagliate sul giorno e sulla causa di morte. Le vittime individuate nel periodo studiato sono state 42, di cui 33

italiane e 9 straniere. Più della metà delle vittime (22) erano coniugate, 15 nubili, 4 vedove e solo una divorziata.
I testi da analizzare sono stati tratti dai due quotidiani di rilevanza nazionale più letti nella regione: La Stampa e La Repubblica. Nei casi in cui nessuno dei due quotidiani avesse riportato articoli sulla specifica uccisione, la ricerca è stata estesa a testate locali (queste sono state le uniche fonti su cui abbiamo trovato degli articoli per 5 delle vittime). Dopodiché è stato scelto il primo articolo pubblicato sull'uccisione per ciascuna testata. Questa scelta è motivata dal fatto che generalmente nel primo articolo sono riportati i dati anagrafici in modo più dettagliato rispetto agli articoli successivi.
Per circa il 45% delle vittime (il 48% di quelle italiane ed il 33% di quelle straniere) sono stati trovati articoli sia su La Stampa che su La Repubblica, mentre nell'88% dei casi (85% delle italiane e 100 % delle straniere) su almeno una delle due testate.

Metodologia

Per analizzare i testi, per prima cosa bisogna riportare il testo ad una versione normalizzata, avendo cura di correggere spazi, accenti, virgolette e punteggiatura in maniera che le parole ed i periodi siano riconoscibili dal software di analisi.
Dopodiché, bisogna riportare ciascuna parola al lemma corrispondente, per esempio gli aggettivi alla forma singolare maschile ed i verbi all'infinito. Per far questo, si è utilizzato il software open source Freeling [8], versione 2.2 [9]. Questo software associa ciascuna parola o simbolo ad un lemma corrispondente, riportando la classificazione grammaticale e la probabilità della correttezza dell'associazione.
In questo modo, si sono potute raggruppare le parole per gruppi semantici ed effettuare un conteggio di frequenze per categoria. Le frequenze possono essere rappresentate mediante istogrammi o word clouds. Inoltre, può essere svolto un confronto con l'italiano standard, basandoci sul Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS) [10].
A questo punto il testo si può ricodificare con la sequenza dei lemmi. Per i nostri scopi, risultava utile una codifica che non tenesse conto di elementi sintattici frequenti, come articoli, preposizioni, verbi ausiliari, connettivi testuali e punteggiatura, che abbiamo raccolto in una lista di *stop words*. A partire da questa trasposizione del testo, abbiamo costruito dei grafi illustranti la frequenza del collegamento tra parole nella stessa frase.
Questa metodologia è stata applicata sia alla totalità degli articoli, sia ad alcuni sottogruppi, in modo da poter comparare i risultati. Gli articoli sono stati quindi suddivisi per: testata (La Stampa e La Repubblica), stato civile della vittima (coniugata o non coniugata), età (<40 o >=40 anni) e cittadinanza (italiana o straniera).

Risultati

Un primo risultato che si ottiene è quello relativo alla frequenza dei termini usati. La parola "anno", essendo prevalentemente legata all'età di vittima ed aggressore, compare come la più frequente. Inoltre, così come accade per gli ausiliari "essere" ed "avere", che avevamo incluso tra le *stop words*, anche il verbo "fare" è molto comune e quindi è normale che sia diffuso nei testi.
Tra i lemmi più frequentemente utilizzati negli articoli dei femminicidi, troviamo parole relative alla morte ed al ritrovamento del cadavere (casa, carabiniere, uccidere,…) ed agli attori dell'omicidio (donna, uomo, madre, moglie,…), come possiamo osservare nell'istogramma in Figura 1 ed nel *word cloud* in Figura 2.

Figura 1. Lemmi più frequenti tra gli articoli analizzati: istogramma

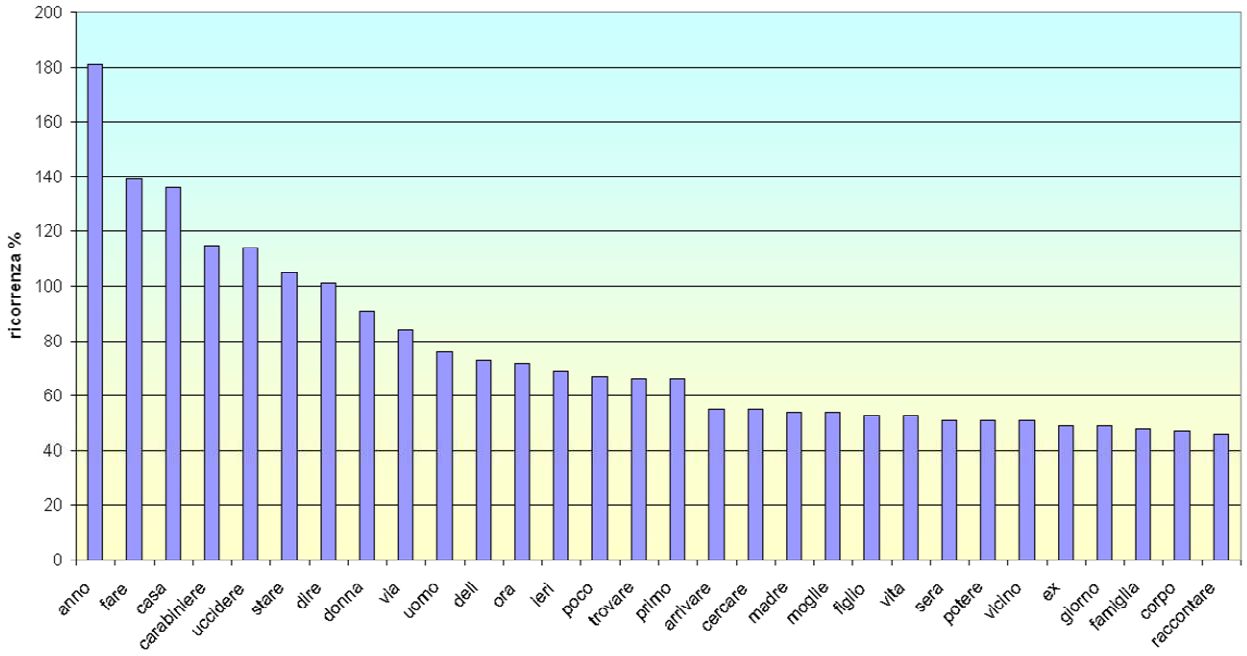


Figura 2. Lemmi più frequenti tra gli articoli analizzati: word cloud

Se invece riportiamo le sequenze più comuni, notiamo una notevole frequenza dell'associazione tra verbi legati all'uccisione e le parole "ex" e "fidanzata", come si può vedere in maniera numerica nella tabella 1 ed in maniera grafica nelle Figure 3 (tutti gli articoli), 4 (La Repubblica) e 5 (La Stampa).

Tabella 1. Ricorrenza delle sequenze di tre parole più frequenti in tutti gli articoli

Frequenza	*Serie di tre parole*		
8	ex	uccidere	anno
8	legale	medico	corpo
8	fidanzata	ex	uccidere
8	senza	corpo	vita
7	via	casa	anno
7	fidanzata	ex	anno
7	colpo	sparare	pistola
6	sera	ieri	anno
6	fidanzata	sparare	ex
6	togliere	vita	uccidere
5	marito	ex	anno
5	morire	poco	anno
5	uomo	uccidere	anno
5	ieri	uccidere	anno
5	fidanzata	uccidere	anno
5	figlia	vita	anno
5	letto	madre	carabiniere
5	passo	poco	casa
5	poco	via	casa
5	medico	corpo	primo
5	legale	corpo	primo
5	legale	medico	primo

Figura 3: Grafo delle parole più frequenti. Tutti gli articoli

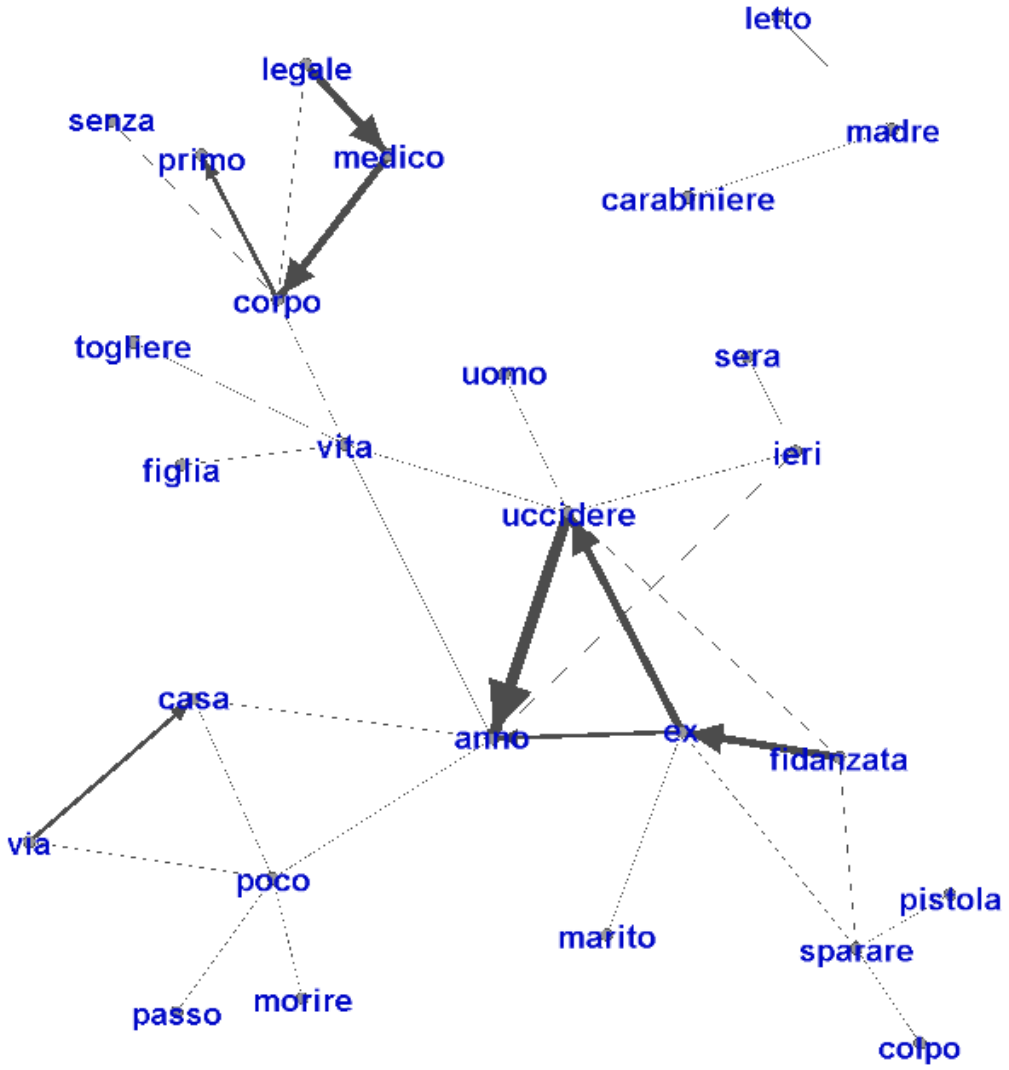


Figura 4: Grafo delle parole più frequenti. La Repubblica

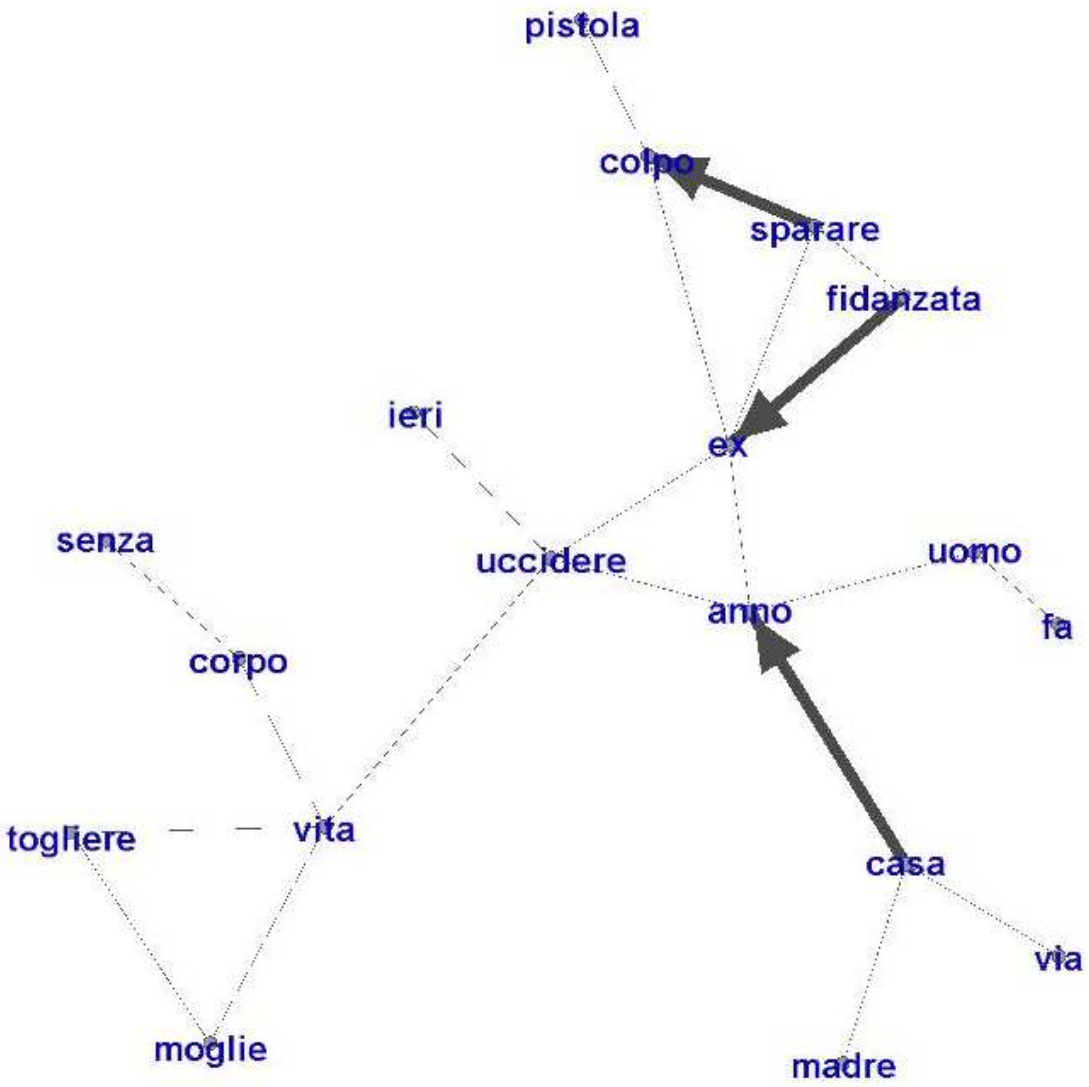
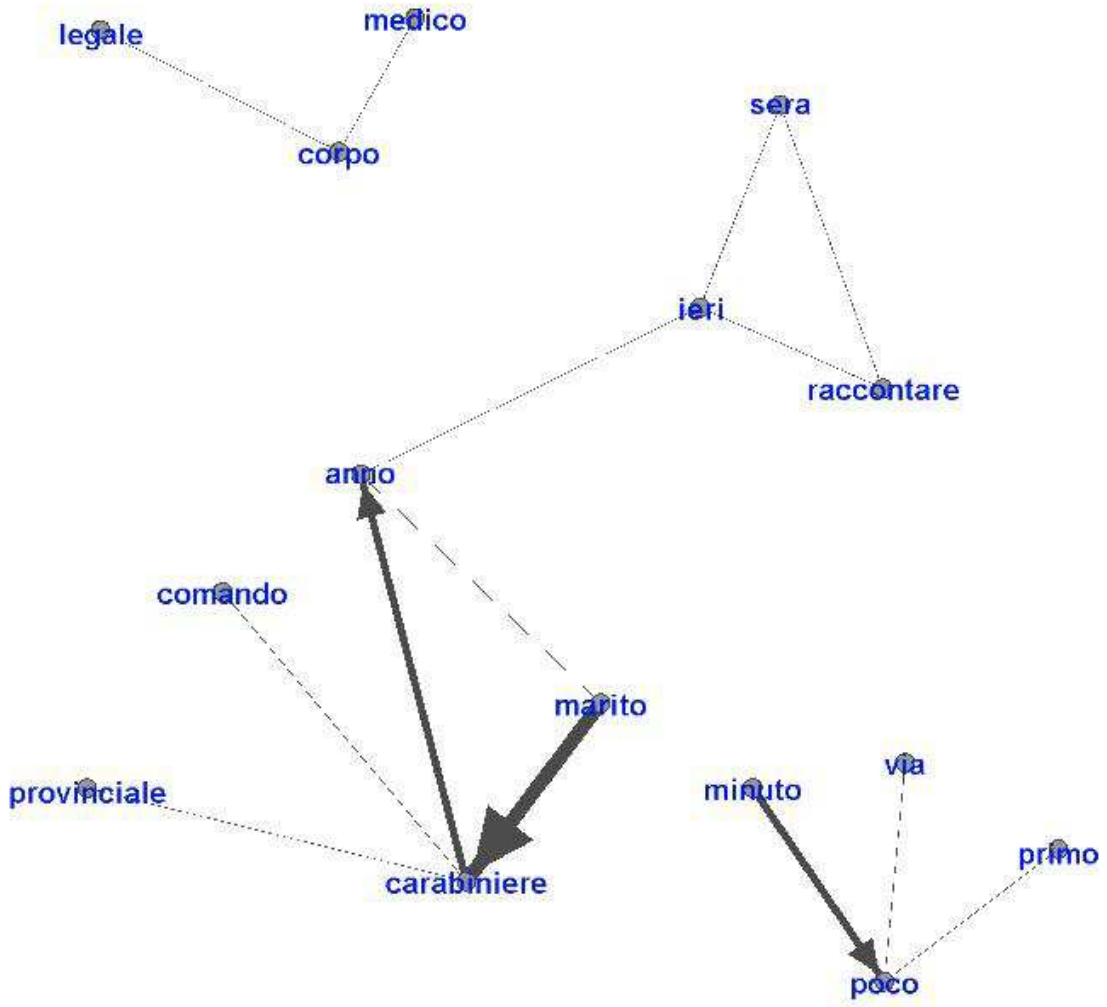


Figura 5: Grafo delle parole più frequenti. La Stampa



Come si può notare dai grafi, a seconda della testata possono cambiare le associazioni di parole più frequenti. Se si riportano le 20 parole più frequenti nei testi de La Repubblica ed in quelli de La Stampa, si può osservare come alcune parole siano molto più usate da una testata rispetto all'altra, sia in riferimento al totale dei lemmi (Figura 6), sia come numero medio per articolo (Figura 7). Per esempio, parole come "uccidere", "uomo", "ex" e "trovare" sono più frequenti in La Repubblica, mentre "raccontare e "carabiniere" in La Stampa.
A livello dei grafi, le differenze che appaiono evidenti tra le due testate sono la centralità del lemma "carabiniere" in La Stampa, mentre gli articoli de La Repubblica sembrano avere il fuoco nella parola "uccidere".

Figura 6: Le 20 parole più frequenti negli articoli de La Repubblica e La Stampa: frequenza lemmi

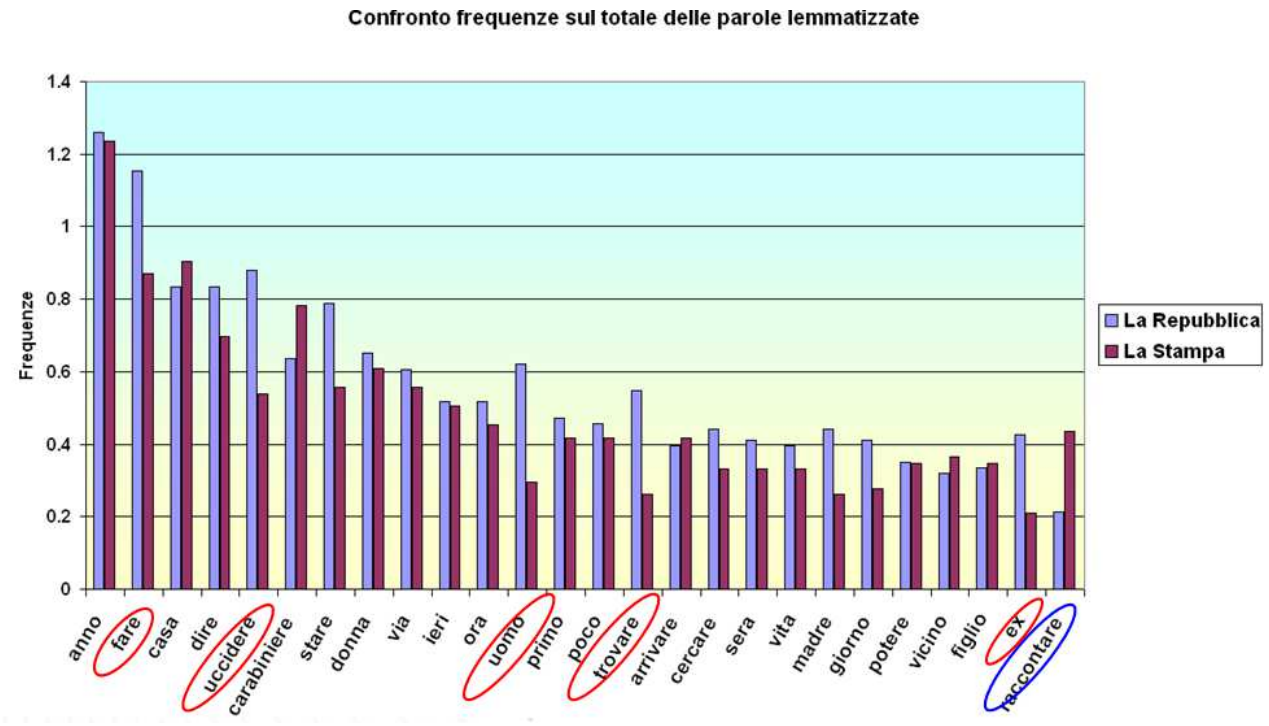
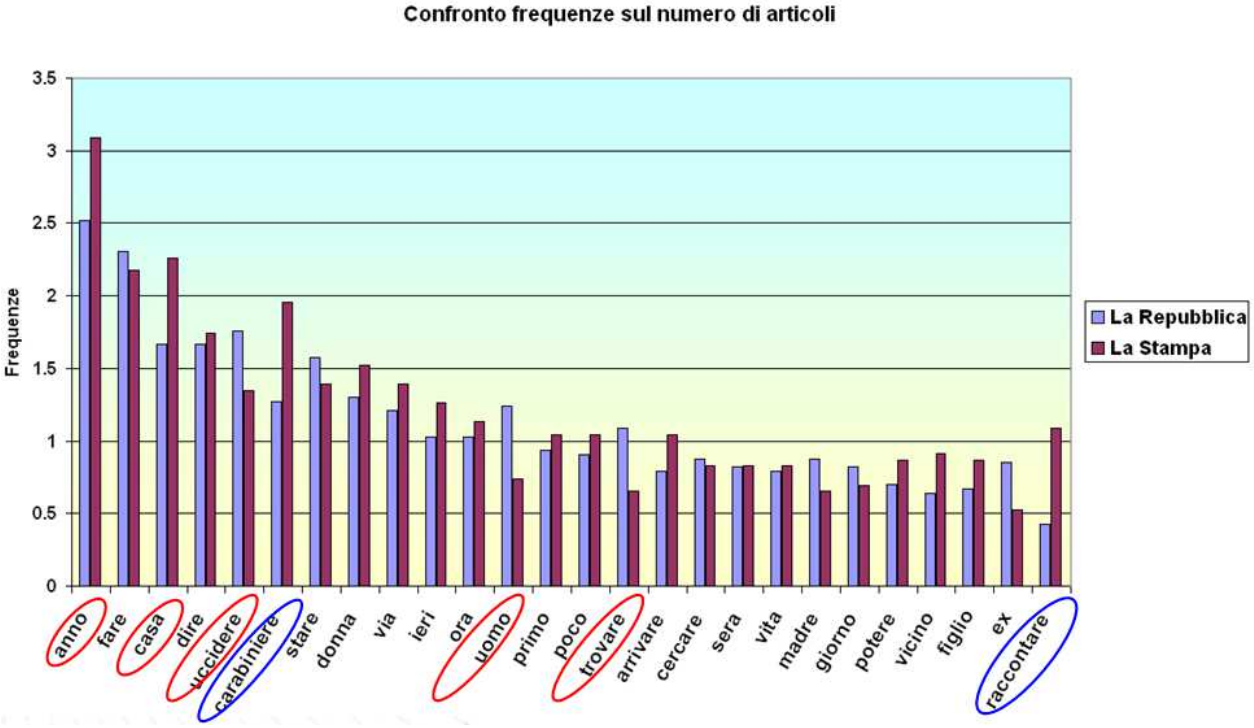


Figura 7: Le 20 parole più frequenti negli articoli de La Repubblica e La Stampa: media frequenza per articolo



Se si dividono gli articoli per categorie di vittima, possiamo confrontare le frequenze dei termini più ricorrenti. Alcune differenze sono naturalmente intrinseche alla categoria (per esempio tra le coniugate sono più ricorrenti termini quali “moglie” e “marito”, mentre tra le non coniugate sono più frequenti “ex”, “amico”,…). Le uniche categorie in cui si presentano differenze più marcate sono quella delle vittime italiane versus quelle straniere. Tra gli articoli legati alle vittime italiane, la parola “violenza” è usata molto di rado (occorre nel 0,05% delle parole lemmatizzate, contro lo 0,33% di quelle relative alle straniere). Inoltre, aggettivi come “sessuale” e “passionale”, abbastanza frequenti quando si parla di vittime straniere, non occorrono mai negli articoli relativi alle vittime italiane. Possiamo notare questa differenza dal *word cloud* in Figura 7.

Sfruttando inoltre l’informazione sulla classificazione grammaticale dei lemmi, possiamo confrontare le forme verbali in cui è coniugato il verbo uccidere. In media, la forma passiva, in cui la vittima diventa soggetto (es. “Maria è stata uccisa”), ricorre nel 26% delle occorrenze. Questa proporzione resta fondamentalmente la stessa al variare della testata giornalistica, dell’età e dello stato civile della vittima.
Se si considerano però le classi “italiane” e “straniere”, vi è una notevole disparità. Negli articoli riguardanti le vittime italiane la forma passiva passa al 30%, mentre per le vittime straniere arriva appena al 9%, facendo così apparire l’uomo il soggetto dell’uccisione nel 91% delle occorrenze.

Figura 7: Aggettivi più frequenti per le vittime italiane e straniere.

L’ultima analisi che abbiamo effettuato consiste nel confronto tra il linguaggio dei femminicidi sui quotidiani e l’italiano standard scritto (banca dati CoLFIS) [10]. Se confrontiamo i verbi, aggettivi e nomi comuni con maggiore frequenza tra i lemmi degli articoli reperiti, le differenze di frequenze più significative dall’italiano standard sono prevalentemente relative all’uccisione (“morta”, ”accoltellata”, ”omicida”, ”strangolare”,…), ma è notevole anche quella relativa alla parola “romeno”, presente con una frequenza superiore di circa 1500 volte rispetto a quella standard. Le differenze di linguaggio dipendono in parte dalla testata che riporta l’articolo, ma possono essere sensibilmente influenzate dalla categorizzazione della vittima (italiana/straniera).

Commenti

Le tecniche di text mining sembrano quindi essere un utile strumento per analizzare il modo con cui la stampa quotidiana narra le notizie di cronaca. L'analisi testuale dei femminicidi appare fornire diversi spunti utili di riflessione sulle tecniche di comunicazione.
Da un punto di vista lessicale (Figure 1-2), è interessante notare come la parola più frequente (dopo "anno", "fare" e le *stop words*) sia "casa". Questo dimostra come il femminicidio, nella maggior parte dei casi, avviene proprio tra le mura domestiche.
Il termine "carabiniere", presente con una frequenza circa trenta volte superiore al suo corrispettivo "poliziotto", indica che la distribuzione dei femminicidi è omogenea sul territorio. I delitti avvengono, infatti, nei grandi centri abitati come nei piccoli. La distribuzione più capillare dei carabinieri sul territorio fa sì che la maggior parte dei femminicidi venga seguita dall'Arma.
Il termine "uccidere" si trova in misura nettamente maggiore rispetto ai suoi sinonimi (per esempio 9 volte più frequentemente di "ammazzare" e 72 volte più frequentemente di "assassinare" o "trucidare"). Questo è verosimilmente da ascrivere a due motivi. Da un lato, alcuni sinonimi vengono usati più in altri contesti tipici della criminalità comune oppure sono usciti dal linguaggio giornalistico perché considerati troppo crudi; dall'altro la parola "uccidere" ha meno caratteri rispetto, per esempio, ad "ammazzare" o "assassinare".
L'uso di un termine piuttosto che un altro è spesso dovuto a questioni di carattere pratico: con spazi sempre più ristretti, il linguaggio giornalistico tende a uniformarsi utilizzando parole più corte.

Se vogliamo concentrarci sul legame tra parole, dobbiamo concentrarci sulla ricorrenza delle loro sequenze (vedi Tabella 1) o sulle loro disposizioni all'interno dei grafi (Figure 3-5). Per esempio, possiamo notare come la parola "uccidere" si ritrovi in terzine in cui compaiono altre parole come "ex" o "fidanzata", ma non, per esempio "marito" oppure "corpo".
E' interessante notare come spesso, per indicare l'assassino e la sua vittima, non si usino le parole "uomo" o "donna", ma il rapporto sentimentale tra essi, vivo o interrotto. L'accento non è dunque posto sulla differenza di genere cosa che invece, considerando la definizione di femminicidio, dovrebbe essere fondamentale. A "uccidere" è l' "ex", molto spesso a essere uccisa è la "fidanzata". Il rapporto che viene descritto tra vittima e carnefice non è un rapporto di dominanza, ma un rapporto familiare in cui in qualche modo s'insinua, già nelle parole usate, una certa "colpa" da parte della vittima nell'aver scelto di accompagnarsi con chi, poi, l'ha uccisa.
Anche il "tipo di funzione" stabilisce l'uso delle parole da accostare. La "fidanzata" viene "uccisa" sparandole, con un "colpo" di "pistola". Alla "moglie", invece, tendenzialmente e più pietosamente si "toglie" la "vita". Questa differenza è molto evidente, per esempio, nel grafo sulle parole usate da Repubblica (Figura 4).
Se analizziamo le terzine ma anche i cluster che emergono dai grafi, scopriamo che la vittima, una volta conclusa l'azione, non ha più alcun legame con l'assassino. E' tale solo in una parte del testo, giusto il tempo di fornire una sua breve descrizione: età, professione, qualche vago accenno alla vita. Ben presto, però, lascia il campo al suo "corpo" che, in quanto tale, è ormai inanimato, senza voce. E' un oggetto da accostare asetticamente solo al "medico" "legale" per le indagini di rito. Non racconta di sé nulla, se non attraverso le ferite subite. Questo è un problema serio nel linguaggio giornalistico derivante in parte da una cultura poco attenta alla vittima, in parte da condizioni oggettive. Una vittima ha smesso di parlare, di raccontare, non offre spunti. Al contrario un assassino ha ancora una "vita giornalistica" piena: deve spiegare il suo gesto, viene interrogato, subisce delle azioni come l'arresto. Anche i racconti sulla vittima da parte di parenti, vicini di casa, colleghi spesso appaiono piatti

Se vogliamo poi confrontare la differenza del linguaggio usato dalle due testate principali, La Stampa e La Repubblica, ci possiamo affidare sia ad un confronto tra grafi che ad uno sulla ricorrenza delle parole più usate.

Partendo dai grafi, per quanto riguarda La Stampa, è evidente la centralità della parola “carabiniere”, che infatti si ritrova più frequentemente in questa testata piuttosto che in quella de La Repubblica. Questo potrebbe essere dovuto al fatto che, tendenzialmente, La Stampa utilizza giornalisti locali per seguire i femminicidi ed è facile che abbiamo rapporti maggiori con fonti all’interno dei carabinieri. Anche la parola “raccontare” risulta molto più usata da La Stampa, sia rispetto al totale delle parole lemmatizzate sia rispetto al numero di articoli.

Le differenze più marcate, però, si trovano se si confrontano gli articoli relativi a vittime italiane o straniere. E' indicativo che la parola “violenza” compaia sei volte più frequentemente nei femminicidi con vittime straniere rispetto che nei femminicidi con vittime italiane. Questo è, purtroppo, l’effetto di una distorta cultura del nostro Paese che tende a colorare di sfumature violente i rapporti tra stranieri (la totale assenza di “passionale” e “sessuale” negli articoli sui femminicidi con vittime italiane ne è la controprova) e a minimizzare la violenza nei rapporti con vittime italiane. E questo a fronte di un fatto incontrovertibile: i femminicidi con vittime italiane nel periodo preso in esame sono quasi quattro volte quelli con vittime straniere.
Interessante anche notare, dall’analisi dei due diversi *word cloud* come nel primo (vittime italiane) compaia spesso la parola “vicino”, assente nel secondo (vittime straniere) che invece riporta la parola “lontano”. E’ probabile che la parola “vicino” per le vittime italiane sia usata frequentemente perché non solo aggettivo, ma anche sostantivo. Le vittime italiane sono persone integrate nella comunità, con vicini che possono essere intervistati e possono raccontare. Cosa che accade meno con le vittime straniere.
Il fatto che il confronto tra il linguaggio negli articoli sui femminicidi e quello standard scritto faccia emergere il lemma “romeno” (una frequenza di circa 1500 volte superiore rispetto allo standard) è un altro importante indicatore. La nazionalità romena è una delle più numerose in Italia e 4 delle 42 vittime hanno in effetti questa nazionalità.
Particolarmente significativo il dato sull’utilizzo della forma passiva del verbo “uccidere”. Il fatto che solo in un caso su quattro venga utilizzata tale formula (“Maria è stata uccisa”) è un altro campanello d’allarme del fenomeno culturale che soggiace alla logica del femminicidio. Il rapporto di dominanza e di potere, vero motore del femminicidio, purtroppo si ritrova anche nel linguaggio giornalistico dove non è mai la donna uccisa la vera protagonista, ma sempre l’uomo che uccide. Inoltre, si osserva come la percentuale delle volte in cui la donna è il soggetto del verbo uccidere passa dal 30% per le vittime italiane al 9% per quelle straniere.
Il dar centralità alla vittima piuttosto che all’omicida, dovrebbe, in realtà, essere il punto di partenza per avviare un serio cambio di paradigma che riesca a modificare quei meccanismi che fanno del femminicidio la conseguenza di un radicato sentimento distorto dei rapporti di genere.

Riferimenti bibliografici

[1]. http://www.gazzettaufficiale.it/eli/id/2013/10/15/13G00163/sg
[2]. Bolasco S., *Statistica testuale e text mining: alcuni paradigmi applicativi*, Quaderni di Statistica Vol. 7, 2005, 17-53.
[3]. Gupta V. e Lehal G.S., *A Survey of Text Mining Techniques and Applications*, Journal of Emerging Technologies in Web Intelligence 2009; Vol 1, No 1, pp. 60-76.
[4]. Feldman, R., and Sanger, J. *The Text Mining Handbook*. New York: Cambridge University Press. 2006
[5]. Mehl MR. *Quantitative Text Analysis*. in: *Handbook of multimethod measurement in psychology*. pp. 141-156. Washington, DC, US: American Psychological Association. 2006
[6]. Indurkhya, N., and Damerau, F. *Handbook Of Natural Language Processing*, 2nd Edition. Boca Raton, FL: CRC Press. 2010.

[7]. http://femicidiocasadonne.wordpress.com/ricerche-pubblicazioni/
[8]. http://nlp.lsi.upc.edu/freeling/
[9]. Padrò L. e Stanilovsky E., *FreeLing 3.0: Towards Wider Multilinguality*, Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.
[10]. Laudanna A. et al. *Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente*. In S. Bolasco, L. Lebart e A. Salem (a cura di), III Giornate internazionali di Analisi Statistica dei Dati Testuali. 1995; Volume I, pp. 103-109.
[11]. http://linguistica.sns.it/CoLFIS/Home.htm

Articolo pubblicato su www.dors.it - 18 giugno 2014
Licensed under a Creative Commons Attribuzione - Non commerciale - Condividi allo stesso modo 3.0 Italia License.
Based on a work at www.dors.it.