L'analisi qualitativa dei media nell'era dei Big Data

Andrea Rubin
Dipartimento di Sociologia - Università Cattolica di Milano
andrea.rubin@unicatt.it

I *Big Data* sono entrati da un po' di tempo nella riflessione metodologica delle scienze sociali. In questa prospettiva e avvalendosi del contributo teorico degli *Science and Technology Studies* (STS) e degli studi sulla *public communication of science* (Pcst), la mia ricerca di dottorato si è posta l'obiettivo di indagare la relazione tra opinione pubblica, mass media e scienza nel periodo 2010-2016. In particolare si vuole tentare di dare risposta alle seguenti domande di ricerca:

RQ1: *C'è una relazione tra la copertura mediatica di un tema tecnoscientifico (rilevata con il conteggio degli articoli pubblicati) e l'opinione pubblica su quel tema?*

Questa generale domanda, che riecheggia l'annosa questione propria dei *media studies* sugli "effetti dei media", ha suggerito una seconda e più dettagliata domanda di ricerca:

RQ2: Osservare il contenuto degli articoli può aiutare a comprendere meglio il rapporto media/opinione pubblica? In altre parole, è possibile individuare un percorso di co-evoluzione tra i mass media e l'opinione pubblica?

Dalle due domande precedenti ci è sembrato opportuno dedurre anche una riflessione epistemologica e metodologica

RQ3: Ha senso avvicinare l'analisi del rapporto tra copertura mediatica e opinione pubblica alla nascente prospettiva dei 'Big Data'?

In altre parole, si intende indagare il contenuto degli articoli di stampa che parlano di cibo e alimentazione, all'interno di un *frame* tecnoscientifico, individuandone i *topic* prevalenti e ipotizzare una possibile inferenza con gli atteggiamenti dell'opinione pubblica, in un periodo compreso tra il 2010-2016.

Dati gli strumenti informatici di cui il gruppo di ricerca in cui sono inserito è fornito, si è scelto di focalizzare l'attenzione sulle versioni *online* di tre quotidiani nazionali (*Repubblica*, *Corriere*, *La Stampa*) e tre riviste di divulgazione scientifica (*Focus*, *National Geographic*, *Le Scienze*). Il criterio di selezione delle testate da considerare per la nostra analisi è stato il numero di lettori. I dati Audipress attestano *Focus*, *National Geographic*, e *Airone* tra le riviste di divulgazione più lette in Italia. Invece di considerare quest'ultima rivista, la scelta è ricaduta su *Le Scienze* in quanto la rivista della *Cairo editore* non è disponibile in formato digitale.

Perché i media online? La crescente crisi dell'editoria vede opporsi un crescente numero di utenti che accedono all'informazione istituzionalizzata dalle diverse piattaforme web. Il mondo del web, inoltre, offre la possibilità di riflettere criticamente sul crescente fenomeno della *datification*. Oltre a una prima parte di analisi più strettamente quantitativa, la ricerca prevede un'ampia parte di analisi del contenuto degli articoli selezionati.

La ricerca si basa sull'assunto proprio dell'approccio STS che vede nei *media* un attore, fra i tanti, seppur privilegiato, che concorre alla costituzione di un fatto e alla formazione delle «rappresentazioni sociali» attraverso la proposizione pubblica di «scene mediali».

Attraverso la piattaforma TIPS (*Technoscientific Issues in the Public Sphere*) messa a punto dall'unità di ricerca Pa.S.T.I.S. dell'Università di Padova è possibile monitorare in tempo reale la produzione giornalistica di 8 giornali online (*la Repubblica*, *La Stampa*, *Corriere della Sera*, *Il Mattino*, *Messaggero*, *Avvenire*, *Il Sole 24 Ore, il Giornale*). La piattaforma permette di estrarre tutti gli articoli contenenti dei termini di ricerca selezionati per la *query* e rilevanti per scienza e tecnologia.

Per ottenere il corpus sul quale successivamente ho iniziato a svolgere l'analisi testuale si è dapprima avviata una fase esplorativa che ha previsto:

1) una ricerca esplorativa negli archivi *online* di quotidiani e riviste

2) raccolta di un campione *ad hoc* di articoli necessario alla selezione delle *keywords*.

Utilizzando le *keywords cib** e *aliment** si è proceduto poi a estrarre gli articoli relativi ai tre quotidiani scelti. Gli stessi termini sono stati utilizzati per estrarre un *corpus* confrontabile dai motori di ricerca interni ai siti delle tre riviste.

Dopo l'estrazione dei due *corpora*, i dati sono stati dapprima analizzati in una prospettiva quantitativa evidenziandone l'andamento storico. Successivamente si è provveduto a sottoporre i testi degli articoli all'analisi automatica attraverso il *software* Mallett al fine di estrarre per ciascun *corpus* i 10 *topic* più rilevanti. Infine, i due *corpora* di articoli sono stati trattati con il software QDA Miner per monitorare l'*indice di rischio* in essi presente.

il presente contributo intende quindi riflettere sul ruolo che gli algoritmi possono svolgere nel supportare la ricerca qualitativa e sulle più recenti e affermate tecniche di analisi. Nel mio caso, da un punto di vista metodologico, per lavorare su un *corpus* di articoli molto grande è stato necessario avvalersi di sistemi di automazione hanno permesso lo *scraiping* automatico dei contenuti prodotti quotidianamente sul web e l'uso di algoritmi (nel mio caso si tratta dell'algoritmo LDA – *Latent Dirichelet Allocation*) che permettano l'analisi della composizione del contenuto attraverso procedure di *topic detection*, consentendo di individuare automaticamente gli argomenti (*topic*) che sono contenuti all'interno di un *corpus*. Questa tecnica di analisi testuale permette una serie di analisi qualitative approfondite su una quantità di dati altrimenti impensabili. Ma quali sono i punti di maggior criticità emersi da un simile disegno di ricerca, da un punto di vista metodologico?

Innanzitutto occorre precisare che, a parer nostro, l'analisi seppur automatizzata dei testi rientra a pieno titolo nella ricerca qualitativa. Uno strumento informatico, seppur

complesso, quale valore aggiunge alla ricerca sociale? È chiaro che anche i software più aggiornati presentano dei limiti evidenti per questo campo d'indagine: essi non sono programmati per rispondere a tutte le domande che possono essere formulate dai ricercatori. Nella fase *ante* e *post* elaborazione rimane dunque essenziale il ruolo del ricercatore. Il confronto con realtà di dati di grande dimensione pone ai ricercatori nuove sfide e richiedono nuove competenze. Gli strumenti informatici presuppongono nuove capacità riflessive degli scienziati sociali: *in primis*, la conoscenza (seppur a livelli superficiali) degli strumenti utilizzati; e, ovviamente, di una consapevolezza delle scelte (sempre arbitrarie) che si svolgono. Il pericolo principale è l'accettazione degli *output* come oracoli. L'interpretazione dei risultati ottenuti con i più disparati algoritmi oggi disponibili rischia di creare una ricerca sociale permeata sui dati (*data-driven research*), sancendo al tempo stesso la fine della teoria e della ricerca qualitativa. Potremmo anche spingerci a chiederci a questo punto, provocatoriamente, se la ricerca qualitativa è ancora necessaria? Per quanto riguarda il mio lavoro di ricerca la risposta è: certamente sì. Dovremmo invece chiederci se siamo in grado, creando team multidisciplinari, di costruire piattaforme con capacità di elaborazione necessaria alle nostre esigenze di studiosi della realtà sociale.

Sono questioni non solo di carattere eminentemente pratico ma anche lasciano aperte strade di riflessione epistemologica (com'è possibile conoscere) e ontologica (cos'è in realtà). Il tema della riproducibilità dei risultati – necessari per un lavoro scientifico – rimane ancora una volta, quanto mai attuale.

Concludendo, potremmo dire che gli aspetti critici promossi da ricerche in simili ambiti disciplinari e con questi strumenti possono nascondere alcune insidie. Possiamo elencare:

1. Arbitrarietà. All'inizio e alla fine del processo è indispensabile il lavoro "umano" (in fase di *input* del campione necessario a istruire il *software* e in fase di *output* per l'interpretazione). Tale "libertà" non deve però essere vissuta dal ricercatore in modo acritico.

2. La non esaustività di tutte le domande di ricerca possibili: si tratta di analisi che presentano, inevitabilmente, delle lacune. I software, ad esempio, non sono in grado di interpretare il significato del complesso e articolato uso del linguaggio naturale.

3. Tra le nuove sfide poste alle scienze sociali vi è certamente l'uso consapevole degli strumenti informatici da parte dei ricercatori: cioè l'uso competente, critico e guidato dall'immancabile guida fornita dalla domanda di ricerca.

Una conclusione ci pare però giungere dall'analisi testuale automatizzata dei testi: essa infatti fornisce la base empirica per il superamento di limiti disciplinari che non solo non hanno a nostro avviso più senso di esistere ma rischiano di precludere importanti strade alla ricerca sociologica e non solo. Il superamento di un dualismo metodologico tra ricerca quantitativa e ricerca qualitativa che pare trovare fondamenta nella proposta dell'approccio *mixed methods*.

Bibliografia

Cardano, M. (2011) *La ricerca qualitativa*, Bologna, Il Mulino.

Corbetta, P. (2003) *La ricerca sociale: metodologie e tecniche*, Bologna, Il Mulino

Giardullo, P. (2015) *Does 'bigger' mean 'better'? Pitfalls and shortcuts associated with big data for social research*, in «Quality and Quantity», DOI 10.1007/s11135-015-0162-8.

Graham, S.; Weingart, S.; Milligan, I. (2012) *Getting Started with Topic Modeling and MALLET*, in «Programming Historian», reperibile online: http://programminghistorian.org/lessons/topic-modeling-and-mallet.html.

Krippendorf, K. (2013) *Content Analysis* (third edition), London: SAGE.

Moscovici, S. (1961) *La psicanalisi, la sua immagine e il suo pubblico*, Milano, Unicopli.

Neresini, F. e de Leonardis, O. (a cura di), (2015) Rassegna Italiana di Sociologia, 3-4, luglio-settembre, Bologna, Il Mulino.

Tipaldo, G. (2014) *L'analisi del contenuto e i mass media*, Bologna, Il Mulino.

Tuzzi, A. (2003) *L'analisi del contenuto*, Roma, Carocci.