JCOM

BIG DATA AND DIGITAL METHODS IN SCIENCE COMMUNICATION RESEARCH: OPPORTUNITIES, CHALLENGES AND LIMITS

Old media and new opportunities for a computational social science on PCST

Federico Neresini

Abstract

Anche se con una certa ritrosia, le scienze sociali sembrano ormai aver raccolto la sfida lanciata dalla crescente digitalizzazione della comunicazione e dal conseguente flusso di dati che scorre sul web. Sono infatti ormai numerose le ricerche empiriche che utilizzano le tracce digitali lasciate dalla miriade di interazioni che avvengono attraverso i social media, e questa tendenza riguarda anche la ricerca nell'ambito della PCST. Risulta invece meno sfruttata l'opportunità offerta dalla digitalizzazione dei testi continuamente prodotti dai mass media tradizionali, in particolare dai quotidiani. Partendo dall'esperienza maturata nell'ambito del progetto TIPS, questo contributo discute i vantaggi e i limiti della computational social science sulla PCST che utilizza i quotidiani come principale fonte di dati. Vengono inoltre affrontati una serie di problemi di carattere metodologico, nell'intento di suggerire un uso più consapevole di tali dati e dei numerosi strumenti disponibili per analizzarli.

Keywords

Public perception of science and technology; Representations of science and technology; Science and media

Sono passati giusto dieci anni da quando Savage e Burrows pubblicarono un articolo che ha avuto un grande successo ed ha avviato nell'ambito delle scienze sociali un dibattito non ancora concluso. "The Coming Crisis of Empirical Sociology" — questo il suo titolo — proponeva una tesi che possiamo riassumere nel modo seguente: i metodi di ricerca che finora avevano assicurato alla sociologia una posizione di primo piano nell'ambito della ricerca e della riflessione teorica sui fenomeni sociali corrono il rischio di essere messi fuori gioco dalla crescente disponibilità di dati digitali prodotti come effetto secondario dalle innumerevoli interazioni che avvengono sul web (i cosiddetti "transactional data") e dalla diffusione di strumenti per la raccolta, l'elaborazione e l'analisi di tali dati allo scopo di sfruttarne le enormi potenzialità commerciali. La strategia suggerita da Savage e Burrows per far fronte a questa situazione consiste nell'invito a usare i *transactional data*, in primis quelli prodotti dai social media, anche per la ricerca sociale, adeguando metodi e tecniche d'analisi alla nuova realtà digitale [Savage e Burrows, 2007]. Qualche anno dopo, Rogers lancia un'analoga proposta, suggerendo di sviluppare i *digital methods* per arricchire e rinnovare l'apparato

strumentale della ricerca sociale [Rogers, 2013]; nel frattempo, l'invito a "sporcarsi le mani" con questi dati non esplicitamente prodotti per la ricerca sociale è stato ripetuto anche da numerosi altri ricercatori.

Senza dover entrare nel dibattito suscitato dall'articolo di Savage e Burrows possiamo utilmente accogliere la loro argomentazione come un buon punto di partenza per riflettere sul rapporto fra Public Communication of Science and Technology (PCST) e ricerca sociale, con particolare riferimento ai recenti sviluppi di quest'ultima proprio nella direzione auspicata dai due sociologi inglesi. Infatti, se il ragionamento di Savage e Burrows vale per la sociologia in generale rispetto al gigantesco flusso di dati generato dai social media, questo vale ancora di più per la ricerca sulla PCST, ovvero un complesso di attività che sempre più spesso utilizzano i social media e che riguardano un oggetto — la scienza e la tecnologia — molto presente nel flusso della comunicazione digitale.

Un primo aspetto su cui vale la pena di soffermare l'attenzione riguarda la preminenza accordata ai "transactional data" derivanti dalle interazioni che avvengono quotidianamente sulle piattaforme di e-commerce e sui social media. Si tratta, questo è fuori discussione, di un patrimonio di dati di grande interesse per la PCST, per esempio per lo studio delle controversie pubbliche che riguardano la scienza e la tecnologia (dai vaccini al cambiamento climatico, dalla cosiddetta "web-democracy" agli effetti più o meno desiderabili dei motori di ricerca), senza poi dimenticare che oggi è praticamente impossibile realizzare qualsiasi attività di PCST senza fare ricorso anche ai social media. Tuttavia non andrebbe trascurato il fatto che i *transactional data* presentano due importanti limitazioni. Da un lato, infatti, non sono sempre di facile accesso per la ricerca sociale, dall'altro hanno, per così dire, il fiato corto. Per quanto concerne il primo aspetto, sarà sufficiente ricordare che questi dati sono soggetti a politiche di forte privatizzazione che ne riducono moltissimo la disponibilità a scopo di ricerca, spesso con la scusa di tutelare la privacy dei loro produttori/consumatori (i cosiddetti *prosumer*, nella fortunata definizione di Ritzer e Jurgenson [2010]). A proposito del secondo aspetto, questi dati tendono poi a rimanere troppo schiacciati sul presente, sia perché i social media e le piattaforme di e-commerce sono creazioni piuttosto recenti, sia perché sono esposti ai cicli simmeliani della moda (What'sApp ha in parte sostituito FaceBook, analogamente sta accadendo con SnapChat; e poi, chi si ricorda di SecondLife?), sia perché tendono a selezionare diversi tipi di utenze e a ridurne di conseguenza la rappresentatività, e ancora perché le abitudini dei loro utilizzatori cambiano nel tempo grazie a processi di naturalizzazione e riflessione collettiva che le modificano (oggi per esempio si sta diffondendo una maggiore consapevolezza dei problemi relativi al trattamento dei dati che sta ridefinendo cosa passa attraverso il web, come testimonia la nascita e l'espansione del cosiddetto "deep-web"). A tutto ciò si dovrebbe inoltre aggiungere che l'eccessiva aderenza al presente può essere un problema per la ricerca sociale, la quale deve mantenere una necessaria distanza dai fenomeni che intende analizzare, quel "distacco" che deve essere sapientemente dosato insieme all'altrettanto necessario "coinvolgimento", come suggeriva Elias [2007], e su cui ha recentemente richiamato l'attenzione anche Frade [2016]. Non si dovrebbe infine cadere nella trappola della neutralità tecnologica: così come il web non è un semplice sostrato su cui viaggiano contenuti e interazioni, ma interviene a modellare tanto gli uni quanto le altre, allo stesso modo social media e piattaforme di e-commerce sono attori dei processi che

li coinvolgono, se non altro in ragione degli algoritmi che li fanno funzionare e che partecipano attivamente alle interazioni a cui danno luogo [Gillespie, 2014].

Insomma, le caratteristiche che rendono attraenti i *transactional data* — soprattutto la loro incessante produzione e la loro genuinità, cioè il fatto di essere generati nel corso di normali interazioni sociali e non per rispondere a stimoli appositamente pensati per rispondere a problemi di ricerca, come nel caso delle domande di un questionario — sono almeno in parte depotenziate da una serie di aspetti che ne riducono la rilevanza per la ricerca sociale.

Fermo restando l'indubbio interesse rappresentato dai *transactional data* per la ricerca sociale, in generale, e per quella relativa alla PCST, in particolare, esiste un'altra opportunità recentemente messa a disposizione e ancora poco esplorata per raccogliere l'invito al rinnovamento della sociologia proveniente da Savage e Burrows. La digitalizzazione dei media tradizionali, come per esempio i quotidiani, apre infatti nuovi scenari di grande rilevanza non solo per lo studio della copertura mediale della scienza e della tecnologia, ma anche per affrontare problemi come quelli delle rappresentazioni sociali della tecnoscienza e del suo rapporto con l'opinione pubblica.

Come nel caso dei contenuti dei social media, anche quelli dei quotidiani sono testi prodotti "naturalmente", nel senso che non sono generati appositamente per rispondere alle domande di un questionario o di una intervista, né vengono costruiti all'interno di contesti artificiali come possono essere quelli di un focus group o di un'osservazione etnografica, nel corso della quale, come è noto, i soggetti sono ben consapevoli di essere sotto osservazione. L'utilizzo dei quotidiani in versione digitale come base di dati per la ricerca sociale sconta per di più minori barriere di accesso e consente una prospettiva analitica di medio lungo periodo, dal momento che gli archivi delle principali testate, almeno per quelle italiane, permettono di andare all'indietro fino ai primi anni '90 del secolo scorso.[1]

Anche nel caso dei quotidiani, restano, ovviamente, una serie di problemi tutt'altro che secondari, e che qui vale la pena di richiamare, seppur brevemente.

In primo luogo, resta aperta la questione relativa alla corrispondenza fra ciò che si può osservare analizzando i quotidiani - e in generale i mass-media — e il resto della realtà sociale, tenendo presente, fra l'altro, che i primi non si possono certo considerare come meri canali di trasmissione o semplici strumenti di rappresentazione dei fenomeni sociali dal momento che essi stessi contribuiscono attivamente alla loro costruzione. Vale tuttavia la pena di sottolineare che lo stesso problema riguarda in modo analogo la relazione fra on-line e off-line, fra quel che avviene e che si può osservare sul web e quanto avviene fuori dal web. Anche in questo caso abbiamo a che fare con un dibattito di grande intensità e di lunga durata, sia quando il termine di riferimento sono i media cosiddetti tradizionali — si pensi per esempio al problema tutt'altro che risolto degli effetti dei media,

[1] In alcuni casi, per esempio l'archivio de La Stampa e quello del Corriere — due fra i principali quotidiani italiani per longevità e per diffusione — è disponibile l'intera collezione dei numeri pubblicati, ma si tratta di archivi costituiti da documenti in un formato non adatto all'analisi automatica, se non a costo di faticose conversioni OCR (optical character recognition) che spesso danno risultati di qualità troppo bassa.

nonostante la ricca messe di riflessioni teoriche e di ricerche empiriche prodotte a tal proposito — sia quando lo sono i nuovi media, social compresi.[2]

In secondo luogo, i contenuti dei quotidiani non sono prodotti da cittadini qualunque, come accade nel caso dei social media, ma dai giornalisti e questo potrebbe ridurre notevolmente la possibilità di utilizzare i media tradizionali per la comprensione dei fenomeni sociali. Tuttavia, mentre il carattere selettivo e parziale della narrazione mediale della realtà rimane fuori discussione, non dovremmo però cadere nella trappola, oggi quanto mai insidiosa, di credere che quanto avviene sui social media sia una rappresentazione più diretta e dunque più fedele della realtà in cui viviamo: anche quel che scrivono e fanno i *prosumer* quando usano i social media è infatti frutto di processi di interpretazione che partono da un particolare punto di vista.

Ci sono inoltre buone ragioni per considerare quel che si trova nei quotidiani come un buon indicatore di quanto avviene nel più ampio contesto sociale, nel quale i mezzi d'informazione si trovano immersi, respirando il clima culturale che, del resto, essi stessi contribuiscono a costruire, trasmettere, riprodurre e trasformare. Come ha bene argomentato Scheufele, per esempio, i frames comunicativi con cui i media, tradizionali e nuovi, presentano i loro contenuti derivano dal contesto all'interno del quale si trovano a operare e, nello stesso tempo, lo condizionano [Scheufele, 1999]. Pur con tutte le cautele del caso, ha senso dunque utilizzare i quotidiani come *proxy* dell'opinione pubblica, e nel caso di questioni tecnoscientifiche controverse si può verificare una corrispondenza per certi versi sorprendente [Neresini e Lorenzet, 2016].

Il progetto TIPS come esempio di *computational social science* appicato alla PCST

All'interno del quadro appena tratteggiato si colloca il progetto TIPS (Technoscientific Issues in the Public Sphere), nato proprio con l'obiettivo di analizzare il discorso mediale a proposito della scienza e della tecnologia, sia per monitorarne l'evoluzione, sia per sfruttare "the social life of data" prodotti dai quotidiani nella loro versione on-line — dunque i testi degli articoli — come proxy dell'opinione pubblica.[3] Il suo carattere intrinsecamente multidisciplinare fa sì che alla sua realizzazione contribuiscano ricercatori provenienti da diversi ambiti: sociologia, ICT, statistica, psicologia sociale e linguistica, cercando di integrare conoscenze, prospettive teoriche e metodologie di ricerca che fanno capo a diversi ambiti delle scienze sociali, con particolare riferimento agli science and technology studies (STS), alla content analysis, alla social representations theory e alla computer science.

La strategia interdisciplinare perseguita dal gruppo di ricerca che ha avviato e continua a sviluppare il progetto TIPS si può dunque riassumere nel modo seguente: utilizzare le potenzialità derivanti dalla digitalizzazione dei quotidiani per studiare il rapporto fra tecnoscienza e società, secondo una logica molto vicina

[2]Sulla possibilità di utilizzare l'on-line come base di dati per studiare l'off-line si vedano, fra gli altri, Wouters et al. [2013], Murthy [2008] e Rogers [2013].

[3]L'idea di monitorare la copertura della scienza da parte dei quotidiani ha preso spunto originariamente dal progetto SAPO [Vogt et al., 2012]. Il progetto TIPS è stato avviato come un'evoluzione del progetto SMM (Science in the Media Monitor), realizzato da Observa sotto il coordinamento scientifico di Federico Neresini.

a quella suggerita da Savage e Burrows, ma cercando di aggirare alcuni dei limiti derivanti dall'uso dei transactional data.

Nell'ambito del progetto TIPS è stata messa a punto una piattaforma che risponde a tali esigenze, sviluppando, sperimentando e implementando procedure automatiche per l'acquisizione, la classificazione e l'analisi di contenuti digitali disponibili sul web — principalmente da news, ma non escludendo quelli dei social media — allo scopo di monitorare la presenza e l'evoluzione di tematiche legate alla scienza e alla tecnologia. Tale piattaforma raccoglie e organizza in un data-base accessibile via web i documenti prodotti da una serie di fonti che attualmente sono le 8 testate giornalistiche italiane più significative su scala nazionale (il Corriere della Sera, La Repubblica, la Stampa, il Sole24Ore, Avvenire, il Giornale, il Messaggero, il Mattino), 5 testate in lingua inglese (NYTimes, Guardian, Mirror, Telegraph, Times of India), 6 testate in lingua francese (Figaro, Lacroix, Le Monde, Les Echos, Liberation, Parisien), oltre ai 100 blog italiani di maggiore impatto e circa 100 accounts di Twitter, sempre in italiano.[4]

All'inizio del 2017 il data-base della piattaforma TIPS conteneva oltre 1.300.000 articoli pubblicati dai quotidiani italiani (la collezione completa dal gennaio 2010 per il Corriere della Sera, La Repubblica, la Stampa, il Sole24Ore; per le altre 4 testate la collezione completa parte dal 2013), oltre 750.000 articoli in inglese (collezione completa per le 5 testate da gennaio 2014); oltre 700.000 articoli in francese (collezione completa per le 6 testate da gennaio 2014); un campione di 162.000 articoli pubblicati dal Corriere della Sera e dalla Repubblica nel periodo 1992–2012; oltre 500.000 post pubblicati sui blog e alcune migliaia di twitts.

La consistente quantità di dati già disponibile, e comunque destinata a crescere, rende ovviamente impossibile analizzarli manualmente e di conseguenza il progetto TIPS privilegia il ricorso a metodologie di elaborazione e analisi automatiche, utilizzando quelle disponibili, magari adattandole, oppure cercando di svilupparne di nuove.

Per questa ragione, oltre a raccogliere in modo sistematico i documenti, la piattaforma TIPS permette di analizzarne automaticamente il contenuto mediante l'applicazione di classificatori e indici appositamente sviluppati; gli indici già disponibili sono i seguenti:

- *salience* = rapporto fra il numero di articoli classificati come pertinenti rispetto a una data tematica e il totale degli articoli pubblicati nello stesso arco di tempo dalla medesima fonte;
- *prominence* = rapporto fra il numero di articoli classificati come pertinenti rispetto a una data tematica pubblicati in home-page e il totale degli articoli pubblicati in home-page nello stesso arco di tempo dalla medesima fonte;

[4]La struttura modulare della piattaforma consente di aggiungere nuove fonti qualora risultasse necessario. Recentemente, per esempio, sono state aggiunte anche El Pais (Spagna) e Jornal de Noticias (Portogallo), mentre è allo studio la possibilità di arricchire ulteriormente il quadro con 6 testate sudamericane. La selezione delle varie testate avviene sulla base di due criteri: diffusione e rappresentatività delle diverse linee editoriali presenti nel panorama della stampa di ciascun paese.

– *general framing* = distribuzione degli articoli pertinenti nelle varie sezioni dei quotidiani (per esempio, oltre all'home-page, politica, cronaca, economia, sport, cultura, . . .);

– *risk* = presenza nel testo di un articolo di un insieme di parole associate alla dimensione del rischio.

Seguire la tecnoscienza sui quotidiani

Nell'intento di mostrare qualche risultato ricavabile dalla piattaforma TIPS possiamo iniziare con l'andamento della salienza degli articoli italiani caratterizzati per un significativo contenuto tecnoscientifico nel periodo 2010–2016.[5]

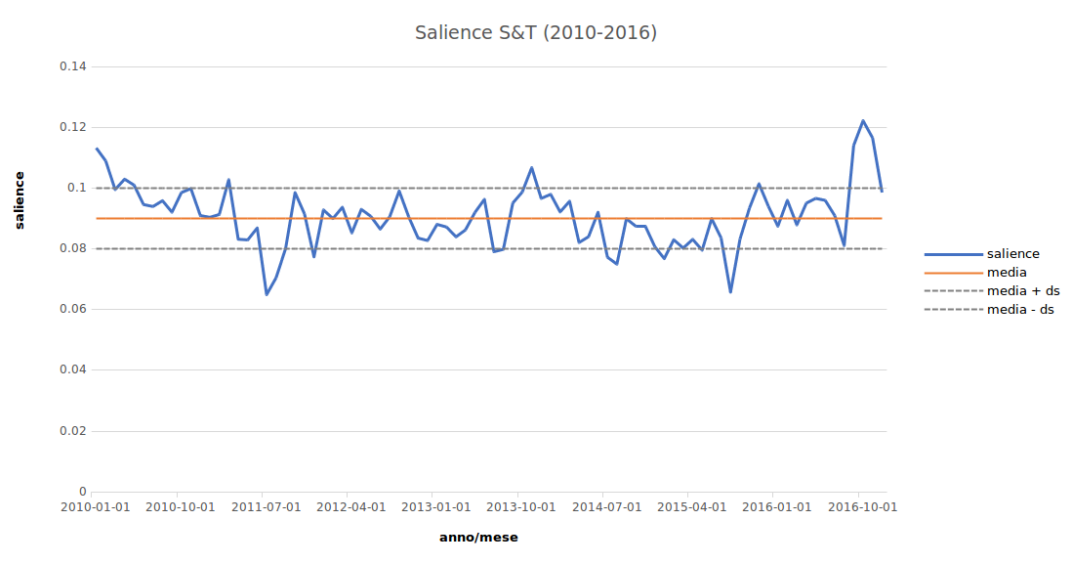

Figura 1. Andamento della salienza degli articoli pubblicati da Corriere della Sera, Repubblica, Stampa e Sole24Ore con un significativo contenuto di tecnoscienza (2010–2016).

Come si può vedere, la salienza della tecnoscienza nel discorso mediale si caratterizza per una certa stabilità, nonostante il numero complessivo degli articoli pubblicati dai quattro quotidiani presi in esame sia significativamente aumentato con l'andare del tempo. La tecnoscienza ricopre quindi un ruolo consolidato nel nostro contesto sociale e, di conseguenza, non ha alcun senso pensarla e trattarla come un mondo a sé stante; è, al contrario, parte integrante della nostra vita quotidiana, mentre la ricerca scientifica e l'innovazione tecnologica si qualificano come attività sociali di grande rilevanza.

Osservando il grafico dell'andamento della salienza (Figura 1), possiamo notare un'area compresa fra le due linee tratteggiate, che corrisponde al range di

[5]La piattaforma TIPS utilizza una serie di classificatori automatici in grado di stabilire se un articolo si caratterizza o meno per contenuti rilevanti rispetto a un determinato ambito tematico. Quello utilizzato per selezionare gli articoli con un *significativo contenuto tecnoscientifico* si basa, per esempio, su un algoritmo che attribuisce a ciascun articolo un punteggio in funzione della presenza nel testo di una serie di keyword opportunamente pesate e combinate; se il punteggio risulta superiore a un valore soglia appositamente determinato, allora l'articolo viene classificato come "rilevante" per la tematica tecnoscienza. La piattaforma permette di avere a disposizione vari classificatori da utilizzare in modo intercambiabile così che, sempre a titolo esemplificativo, si possono selezionare gli articoli pertinenti anche per altri ambiti tematici come sicurezza alimentare, scienza, nanotecnologie, biologia sintetica. Altri classificatori si possono aggiungere a seconda delle necessità di ricerca.

oscillazione della salienza stimabile come "fisiologico".[6] La definizione di una simile zona di normale variabilità permette di identificare alcuni picchi, vale a dire periodi all'interno dei quali la tecnoscienza ha inciso maggiormente sul fluire del discorso mediale.

Riprenderemo più avanti la riflessione su alcuni aspetti metodologici riguardanti l'analisi longitudinale della copertura assicurata dai quotidiani alla tecnoscienza, mentre per ora vale la pena di dedicare una breve analisi ai picchi appena individuati.[7] Così, per esempio, l'aumento della salienza registrato nel marzo 2011 è dovuto in buona parte all'incidente della centrale nucleare di Fukushima e alla presentazione del nuovo iPad. Nel Novembre del 2013, invece, il picco risulta più difficilmente riconducibile a eventi precisi, rendendo in questo modo evidente che la presenza della tecnoscienza sui quotidiani dipende soprattutto dal contributo di numerose notizie singole, a cui si uniscono poi eventi di portata significativa ma non di grande rilevanza. In questo caso, sempre a titolo esemplificativo, possiamo contare, fra gli altri, il rientro dalla stazione spaziale dell'astronauta Parmitano, la controversia relativa al caso Stamina, il conflitto sui brevetti fra Samsung e Apple. Due anni dopo, sempre a Novembre, troviamo di nuovo due eventi che hanno catalizzato l'attenzione dei quotidiani: la conferenza sul clima di Parigi e il dibattito sulla destinazione dell'area Expo a Milano come cittadella della scienza.

Ma come viene percepita la tecnoscienza nella sfera pubblica? Mettendo a confronto il peso e l'andamento dello risk indicator negli articoli riguardanti la tecnoscienza e negli altri articoli possiamo ricavare qualche interessante indicazione a questo proposito. Per prima cosa possiamo notare che nel discorso mediale relativo alla tecnoscienza il risk indicator tende a diminuire nel tempo. Tuttavia, potendo comparare questo dato con quello calcolabile analogamente su tutti gli articoli pubblicati nello stesso periodo, diventa immediatamente evidente che si tratta di un trend generale e che quindi non sarebbe corretto interpretare il dato relativo alla tecnoscienza come il segnale di una percezione pubblica meno incline a valutarla in termini di rischio.

Se poi compariamo l'andamento del risk indicator distinguendo fra articoli che riguardano la tecnoscienza e articoli che sono invece più strettamente riconducibili alla sola scienza — nella Figura 2 rispettivamente S&T e S — notiamo che le due curve sono praticamente sovrapposte. Di conseguenza, che si parli di scienza o che si parli di tecnologia il riferimento alla dimensione del rischio sostanzialmente non cambia. Ciò pone seri dubbi sull'idea — invero piuttosto consolidata - che scienza e tecnologia si debbano invece trattare separatamente, quanto meno nella sfera pubblica.

Alcune questioni epistemologiche e metodologiche

Ma, al di là dei risultati che si possono ottenere, l'esperienza accumulata nel corso degli anni lavorando secondo la logica suggerita dal progetto TIPS permette di fare una serie di considerazioni di carattere metodologico che possono rivestire una certa importanza per poter valutare appieno le potenzialità, ma anche i limiti a cui si va incontro nell'affrontare la ricerca sulla PCST volendo sfruttare nella

[6] Dal punto di vista operativo, TIPS delimita la zona di normale oscillazione come quella compresa fra i valori pari a media + standard deviation (sd) e media — sd.

[7] Vedi la successiva sezione 3.2.

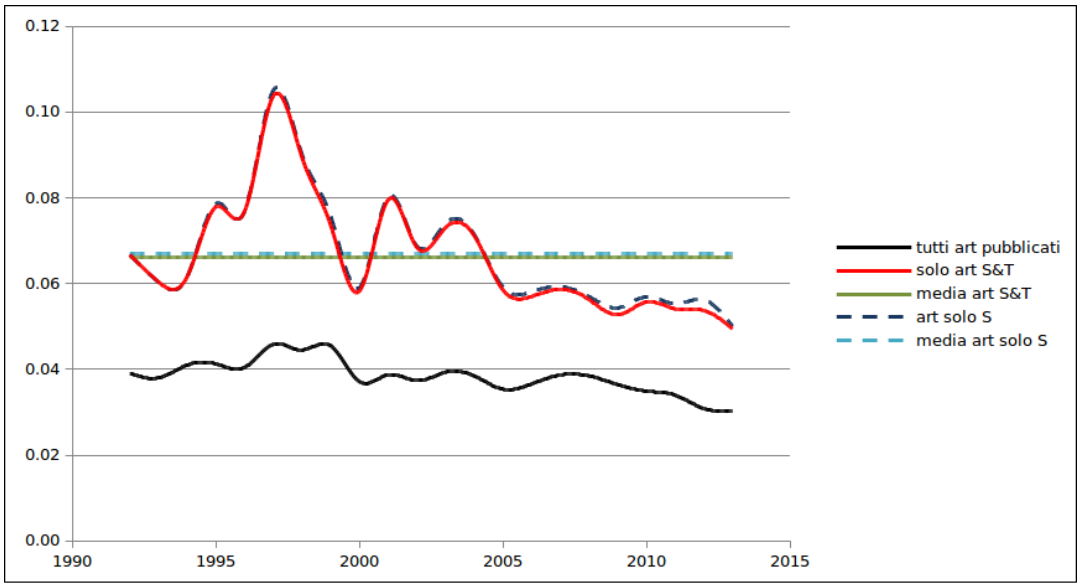

Figura 2. Trend del risk indicator applicato a corpora selezionati sulla base di differenti criteri (Il Corriere della Sera and La Repubblica — sample *artificial week*, 1992–2013; n=160.451).

prospettiva di una computational social science l'enorme quantità di dati testuali oggi disponibili grazie alla digitalizzazione dei media tradizionali.

Procederemo in modo schematico, nell'intento di fornire una panoramica delle possibili questioni chiamate in causa da una prospettiva di ricerca di questo genere piuttosto che discutere in modo approfondito ciascuna di esse.

3.1 La costruzione dell'oggetto d'analisi, ovvero la selezione del corpus

Potrebbe sembrare ovvio, oppure banale, ma un problema decisivo per una corretta ed efficiente ricerca sulla PCST utilizzando i quotidiani riguarda proprio la costruzione del *corpus*. Gli aspetti da prendere in considerazione a tale proposito sono almeno due: il reperimento degli articoli e la loro selezione, così da ottenere un corpus pertinente rispetto alla domanda di ricerca per la quale si intende utilizzarlo.

Per quanto concerne il reperimento, dobbiamo tenere presente che, nonostante la digitalizzazione dei contenuti testuali veicolati dai media abbia consentito il superamento di numerosi vincoli che ne rendevano complicato l'utilizzo nella loro versione analogica, l'acquisizione degli articoli pubblicati dai quotidiani richiede ancora un certo impegno. La barriera più grande rimane ovviamente l'accesso agli archivi esistenti o, in alternativa, la costruzione di appositi archivi, ma non va trascurato il problema dei formati, per esempio nel caso di testi trasmessi o conservati come immagini, ovvero in un formato che rende molto complicato il loro trattamento con strumenti di analisi automatica del contenuto. Il problema del reperimento si collega poi a quello della selezione. L'esperienza di TIPS dimostra che per avere una corretta visione del discorso mediale sulla tecnoscienza, tanto considerandola in termini generali quanto riferendola a tematiche specifiche, è

necessario costruire sempre un'analisi di tipo comparativo: articoli che trattano contenuti tecnoscientifici *versus* articoli che trattano altri argomenti, articoli che affrontano una questione rilevante per la tecnoscienza (biotecnologie, politiche della ricerca, bioetica, ...) *versus* l'insieme degli articoli che nello stesso periodo presentano contenuti rilevanti per la tecnoscienza generalmente intesa. Solo una visione comparativa permette infatti di cogliere l'andamento della copertura riservata alla tecnoscienza dai quotidiani, per esempio di non scambiare l'evidente crescita nel tempo del numero di articoli che hanno a che vedere con la tecnoscienza con l'aumento della loro rilevanza sulla scena mediale: l'indicatore della salienza mostra infatti che questa è rimasta tutto sommato costante nel tempo.[8] Allo stesso modo, solo in chiave comparativa possiamo capire se le caratteristiche osservate a proposito di una certa questione la distinguono dalle altre oppure no.

Ma il problema della selezione tocca anche quello, ancora più complicato, della definizione dell'oggetto che si intende studiare. La faccenda sembra facilmente risolvibile nel caso di tematiche relativamente specifiche: se voglio studiare il discorso mediale sulle nanotecnologie, sarà sufficiente selezionare tutti gli articoli che contengono il termine "nanotecnolog*". In realtà non è proprio così, sia perché spesso si usano come sinonimi l'abbreviazione "nano" oppure altri termini come "nanoscienza", sia perché si può parlare di nanotecnologie anche senza nominarle direttamente grazie a perifrasi come "il mondo dell'infinitamente piccolo" o "la ricerca a livello atomico". Se la faccenda è già meno semplice di quel che potrebbe sembrare con oggetti relativamente specifici, è facile immaginare che diventi estremamente complessa quando l'obiettivo dell'analisi ha contorni più sfumati, come nel caso della tecnoscienza. Cosa intendiamo infatti per "tecnoscienza"? Scienza e tecnologia sono da trattare separatamente? E in caso affermativo, dove passa la linea di distinzione? Non sono questioni solamente filosofiche o sociologiche, poiché dalla risposta a tali domande dipende la procedura operativa che poi si segue per selezionare gli articoli e dunque per costruire la base dei dati (corpus) su cui poi verranno sviluppate le nostre analisi. La soluzione adottata da TIPS, per esempio, si potrebbe definire un approccio pragmatico fondato su presupposti derivanti dagli STS: la scienza è un'attività sociale, dunque fatta da qualcuno, all'interno di un'organizzazione, mediante l'utilizzo di alcuni strumenti, strutturata in campi disciplinari con una specifica nomenclatura; è inoltre un'attività che produce contenuti trasmessi mediante riviste o in particolari occasioni (convegni, conferenze, ...). Se in un articolo sono presenti almeno un paio di riferimenti a tali risvolti pragmatici — per altro facilmente identificabili - della ricerca scientifica, allora possiamo dire che abbiamo a che fare con un contenuto rilevante per la nostra analisi della tecnoscienza nella sfera pubblica.

3.2 Granularità, periodizzazione e distorsioni campionarie

Il secondo aspetto di carattere metodologico che possiamo brevemente discutere riguarda l'organizzazione della dimensione temporale dell'analisi. Come abbiamo sostenuto all'inizio, la possibilità di ricostruire l'evoluzione del discorso mediale sulla tecnoscienza lungo archi temporali di medio/lungo periodo costituisce uno degli elementi che qualificano in senso positivo la scelta di utilizzare i quotidiani anziché i transactional data.

[8] È per questa ragione che la piattaforma TIPS raccoglie e conserva tutti gli articoli pubblicati dai quotidiani che vengono monitorati.

Per sfruttare al meglio questa possibilità vanno definiti due aspetti che risultano, anche in questo caso, solo apparentemente di secondaria importanza. Il primo riguarda la cosiddetta granularità, ovvero l'unità temporale con cui i dati vengono strutturati, analizzati e visualizzati. È chiaro che se l'arco temporale considerato è relativamente breve — poniamo un anno — la distribuzione dei contenuti presentati dagli articoli si può suddividere per mese, oppure per settimana, mentre quando si prende in considerazione un periodo più lungo — per esempio una decina d'anni — le opzioni disponibili contemplano anche unità temporali più lunghe (anni, semestri, quadrimestri, . . .), fermo restando che, comunque, l'unità di analisi più piccola nel caso dei quotidiani rimane il giorno. A questo riguardo, non sembra esistere un criterio di scelta ottimale, se non quello della praticità; tuttavia non si dovrebbe dimenticare che la granularità prescelta incide non marginalmente su quello che si può osservare e dunque ricavare dall'analisi. Diverso, invece, il problema della periodizzazione. Per consuetudine culturalmente consolidata si ricorre normalmente a periodizzazioni che assumono come unità di riferimento archi temporali analoghi a quelli utilizzati per la granularità. Tale soluzione, pur rimanendo la più semplice e diretta, nasconde però un'insidia, dal momento che non tiene conto di una caratteristica cruciale della comunicazione mediale. Nei media, tradizionali o nuovi che siano, la presenza di determinati contenuti si presenta infatti sia in termini "puntuali" — un articolo sull'energia nucleare oggi, uno a distanza di un mese, e così via — sia come "storie" agganciate a eventi più o meno rilevanti — l'annuncio di una nuova strategia energetica da parte del governo, un incidente in una centrale nucleare, . . . - e dunque come narrazioni con cicli di vita più o meno lunghi e più o meno ricorrenti.

Per questa ragione è evidente che una periodizzazione suddivisa per unità di tempo fisse — di nuovo un anno, un mese, una settimana, . . . — finisce per interrompere artificiosamente tali cicli, restituendo così un'immagine distorta del discorso mediale. Sulla base di simili considerazioni sarebbe quindi più opportuno pensare a forme di periodizzazione organizzate in modo più aderente alla dinamica della comunicazione mediale. Nel caso di TIPS la soluzione adottata (vedi Figura 1) è quella di considerare l'oscillazione attorno al valore medio della salienza come la zona di normale oscillazione e, di conseguenza, i valori della salienza al di fuori di tale zona come "picchi" del discorso mediale.

Vale la pena di notare inoltre che la dinamica che caratterizza la comunicazione mediale e che rende quanto meno discutibile una scansione temporale a periodi fissi costituisce anche il motivo per cui non può essere soddisfacente la costruzione di corpora mediante procedure di campionamento, soprattutto tenendo presente che quasi sempre l'unità di campionamento corrisponde al giorno. Se infatti vengono selezionati casualmente giorni che cadono all'interno di un picco di copertura, anziché durante un periodo di normale oscillazione attorno alla media, la situazione cambia drasticamente, ottenendo una sovra- o una sotto-rappresentazione del discorso mediale sulla tecnoscienza.

3.3 Validità

Il terzo aspetto metodologico riguarda infine un problema di portata generale per la ricerca, ovvero quello della validità. Come possiamo infatti essere ragionevolmente sicuri che stiamo osservando proprio il fenomeno che ci interessa?

Nel caso della discorso pubblico sulla tecnoscienza analizzato mediante i quotidiani, come possiamo sapere che gli articoli selezionati per il nostro corpus sono effettivamente pertinenti rispetto al nostro oggetto di studio? In parte questo problema ci riporta a quello dell'appropriata costruzione del corpus che abbiamo già affrontato in precedenza, in parte mostra un'insidia che si nasconde pericolosamente nelle strategie di ricerca che usano grandi quantità di dati, come accade nel progetto TIPS. È chiaro, infatti, che quando si lavora con centinaia di migliaia di articoli la verifica della validità non può avvenire manualmente e che di conseguenza dobbiamo affidarci a test compiuti su quantità molto ridotte. A questo proposito esiste fortunatamente una lunga esperienza maturata nell'ambito della ricerca informatica che si occupa di data retrieval, e che qui ovviamente non iniziamo nemmeno a discutere. Ma il punto da sottolineare è piuttosto un altro: nessuna tecnica di trattamento automatico è in grado di risolvere il problema della validità se non passando attraverso il giudizio dei ricercatori, il quale, inevitabilmente, introduce una serie di scelte, assunti e opacità di carattere epistemologico. Non è possibile fare diversamente, e la consapevolezza di tale ineludibilità non ci dovrebbe mai abbandonare, nonostante l'apparente oggettività che viene facile attribuire agli strumenti di analisi automatica.

Un'ultima considerazione: le domande di ricerca prima di tutto!

La questione della validità non è affatto banale, anche perché si pone in stretto collegamento con il rischio di utilizzare come vere e proprie black-box i software oggi disponibili per il trattamento di grandi quantità di dati testuali, sovente applicando algoritmi di complessità spesso piuttosto elevata. Certo non è pensabile che gli scienziati sociali diventino anche statistici competenti e bravi informatici, non si tratta di questo; si tratta piuttosto di perseguire un utilizzo consapevole e possibilmente critico di tali strumenti, conoscendone, almeno in linea di massima, gli assunti su cui si reggono e le procedure che li fanno funzionare, così da poter scegliere quelli più appropriati rispetto alla domanda di ricerca su cui si sta lavorando, avendo presenti i loro limiti principali e sapendone interpretare correttamente gli output.[9]

In questo modo si dovrebbe anche acquisire qualche salutare anticorpo nei confronti di quella che potremmo chiamare una specie di "bulimia del dato", dovuta alla smisurata disponibilità di dati non solo facilmente accessibili, ma anche trattabili con una relativa facilità grazie ai software che permettono di ridurne la complessità e dunque di ottenere risultati significativi al di là dell'apparente caoticità dei dati di partenza. Rimanere impigliati nel desiderio di poterne elaborare quantità sempre maggiori, e magari a ritmi sempre più veloci, nasconde infatti il pericolo di trascurare ciò che dovrebbe rimanere sempre in primo piano: la centralità della domanda da cui muove la nostra ricerca, la necessità di non fermarci sul piano meramente descrittivo, ma di affrontare sfide conoscitive che, mentre ci espongono al rischio del fallimento, lasciano aperta la possibilità di comprendere qualcosa di più del mondo in cui viviamo.

[9] Un'interessante discussione a questo proposito è stata fatta da Hoffmann [2013], con riferimento alla computational history.

Riferimenti bibliografici

Elias, N. (2007). Involvement and Detachment. Dublin, Ireland: University College Dublin Press.

Frade, C. (2016). 'Social Theory and the Politics of Big Data and Method'. *Sociology* 50 (5), pp. 863–877. DOI: 10.1177/0038038515614186.

Gillespie, T. (2014). 'The Relevance of Algorithms'. In: *Media Technologies*. A cura di T. Gillespie, P. J. Boczkowski e K. A. Foot. Cambridge, MA, U.S.A.: The MIT Press, pp. 167–194. DOI: 10.7551/mitpress/9780262525374.003.0009.

Hoffmann, L. (2013). 'Looking back at big data'. *Communications of the ACM* 56 (4), pp. 21–23. DOI: 10.1145/2436256.2436263.

Murthy, D. (2008). 'Digital Ethnography: An Examination of the Use of New Technologies for Social Research'. *Sociology* 42 (5), pp. 837–855. DOI: 10.1177/0038038508094565.

Neresini, F. e Lorenzet, A. (2016). 'Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power'. *Public Understanding of Science* 25 (2), pp. 171–185. DOI: 10.1177/0963662514551506.

Ritzer, G. e Jurgenson, N. (2010). 'Production, Consumption, Prosumption: The nature of capitalism in the age of the digital 'prosumer''. *Journal of Consumer Culture* 10 (1), pp. 13–36. DOI: 10.1177/1469540509354673.

Rogers, R. (2013). Digital methods. Cambridge, MA, U.S.A.: MIT Press.

Savage, M. e Burrows, R. (2007). 'The Coming Crisis of Empirical Sociology'. *Sociology* 41 (5), pp. 885–899. DOI: 10.1177/0038038507080443.

Scheufele, D. A. (1999). 'Framing as a theory of media effects'. *Journal of Communication* 49 (1), pp. 103–122. DOI: 10.1111/j.1460-2466.1999.tb02784.x.

Vogt, C., Castelfranchi, Y., Righetti, S., Evangelista, R., Morales, A. P. e Gouveia, F. (2012). 'Building a science news media barometer SAPO'. In: *The culture of science. How the public relates to science across the globe*. New York/London: Routledge, pp. 400–417.

Wouters, P., Beaulieu, A., Scharnhorst, A. e Wyatt, S., cur. (2013). Virtual Knowledge. Cambridge, MA, U.S.A.: The MIT Press. URL: https://mitpress.mit.edu/books/virtual-knowledge.

Autore

Federico Neresini (Ph.D. in Sociologia e Ricerca Sociale) insegna "Scienza, Tecnologia e Società" e "Sociologia dell'Innovazione" presso l'Università di Padova, dove coordina l'Unità di Ricerca PaSTIS (Padova Science, Technology and Innovation Studies — www.pastis-research.eu).

Il suo principale ambito di ricerca riguarda la sociologia della scienza e della tecnologia, in particolare la comunicazione pubblica della scienza, le rappresentazioni sociali della scienza e l'analisi delle trasformazioni del lavoro di ricerca nei laboratori. Si è occupato soprattutto di biotecnologie, con particolare attenzione alla fecondazione artificiale e alla clonazione, e di nanotecnologie. Negli ultimi anni ha iniziato a lavorare anche sul rapporto fra big-data e attività di ricerca scientifica, oltre che sulle implicazioni per le scienze sociali derivanti dalla disponibilità di grandi quantità di dati grazie al web.

Ha pubblicato numerosi saggi e articoli anche su riviste internazionali (fra cui Nature, Science, Public Understanding of Science, New Genetics Society, Science Communication), oltre ad alcuni volumi fra i quali, più recentemente, "Il nano-mondo che verrà. Verso la società nanotecnologica" (il Mulino, 2011) e "Il valore dell'incertezza" (Mimesis, 2015, insieme a Paolo Vidali).
E-mail: federico.neresini@unipd.it.

How to cite

Neresini, F. (2017). 'Old media and new opportunities for a computational social science on PCST'. *JCOM* 16 (02), C03_it.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it