

Aprire la scatola nera

Corso riflessivo di sociologia computazionale

Guido Anselmi

1. Problemi epistemologici nei dati digitali

I metodi di ricerca che trattano di dati digitali arrivano in un contesto caratterizzato da una sostanziale crisi dei metodi tradizionali. È diventato particolarmente oneroso assicurare che un sondaggio selezioni un campione effettivamente casuale.

Anche in un contesto di ricerca commerciale i dati digitali rendono disponibili approcci molto meno costosi e maggiormente efficienti se comparati all'organizzazione di focus group.

Dati

I dati rappresentano la materia grezza su cui l'inchiesta scientifica formulerà poi ipotesi e testerà teorie; sono tecniche che nascono per reperire i dati.

Per Kitchin il dato rappresenta una astrazione necessaria in relazione ad una realtà empirica che resta tuttavia necessariamente fuori fuoco, il dato non rappresenta 'il reale' ma una serie di misurazioni e trascrizioni di valori che sono stati selezionati dal ricercatore, fra tutti i dati effettivamente 'possibili'.

I dati si possono definire in una griglia tassonomica:

- Per *forma* → quantitativi o qualitativi
- Per *struttura* → strutturati parzialmente, totalmente o non strutturati
- Per *fonte* → 'catturati', quindi prodotti da un sistema disegnato con lo scopo di produrli; 'exhaust', quindi generati come sottoprodotto di un'altra elaborazione; 'derivati'; 'transienti'
- Per *produzione* → primari, secondari, quindi prodotti per differenti scopi e terziari, quindi divulgati.
- Per *tipo* → indici, attributi o metadati.

Ciò che in un certo punto del processo di indagine viene ritenuto 'dato', in un punto precedente è stato prodotto a partire da scelte di metodo più o meno esplicite o più o meno riflessive.

Per analizzare in maniera critica la 'datificazione' del mondo dobbiamo comprendere come la produzione di dati debba essere analizzata tenendo contemporaneamente presente una tensione di fondo:

- l'esistenza di una realtà oggettiva che non può essere né totalmente decostruita fino ad arrivare ad un puro set di convenzioni linguistiche
- la stessa realtà empirica non può essere né totalmente compresa, né predetta.

Per Veltri esiste una distinzione fra:

- i dati provenienti da sforzi di raccolta progettati dal ricercatore: la variabile finale sarà il risultato di un processo lineare

- i dati che provengono da fonti non originariamente progettate per la ricerca: bisognerà mappare una o più variabili.

Per comprendere come i dati sono generati c'è la necessità di capire il contesto sociotecnico in cui i dati sono prodotti, come la loro produzione si inserisce nel contesto delle tecnologie disponibili, nel contesto delle relazioni economiche e nel contesto dei processi di regolamentazione predominanti.

Piattaforma

La piattaforma è un artefatto sociotecnico in cui è stabilito un set di regole che governa lo scambio di informazioni fra utenti. Questa mediazione non è neutrale perché la struttura della piattaforma permette determinate relazioni e non altre; le azioni permesse rientrano in una strategia commerciale che si sovrappone alle azioni degli utenti.

Per Ellison un social network è definito da tre specifiche caratteristiche:

1. la facoltà di crearsi un profilo personale
2. la capacità di connettersi ad altri utenti attraverso liste dedicate
3. la capacità di 'navigare' queste liste

Un ruolo non secondario lo rappresentano le metriche della reputazione e i connessi algoritmi di curatela. Ridurre i social network ad un mero artefatto tecnico sarebbe poco corretto: ciascuna piattaforma è caratterizzata da specifiche 'regole di ingaggio' e da una specifica cultura che viene prodotta dagli utenti nel corso dell'uso quotidiano.

Caliandro e Gandini definiscono anche i social network come spazi di autopresentazione, uno spazio in cui gli attori possono introdurre molteplici logiche di *self presentation*. Da questo ne derivano due considerazioni aggiuntive:

1. Diventa possibile includere le regole di presentazione del sé all'interno delle specifiche norme culturali che governano l'utilizzo delle piattaforme digitali.
2. La presentazione del *self* può essere iscritta dentro una economia della reputazione.

Le piattaforme rappresentano una fattispecie più ampia dei social, anche se ne condividono tutte o parte delle feature individuate da boyd ed Ellison. Le definizioni di cosa sia una 'piattaforma' sono molteplici.

Una piattaforma rappresenta una soluzione 'crowdsourced' al problema dei costi di ricerca nel caso dei beni differenziati. Le piattaforme operano secondo due logiche parallele:

1. Ordinando i contenuti in funzione delle 'preferenze' ricevute dagli utenti
2. Si attivano se la piattaforma offre beni molto differenziati degli algoritmi di raccomandazione in grado di consigliare il 'corretto' bene da consumare a partire da un confronto con altri beni consumati in un 'paniere' virtuale da parte di altri utenti. All'interno di questa dinamica esiste una 'corsia preferenziale' per i contenuti 'promossi'.

A fianco delle piattaforme commerciali, tuttavia, continuano ad esistere dei luoghi digitali, in cui lo scopo primario è produrre relazioni sociali fra utenti.

La sociologia definisce le 'relazioni sociali' utili ad ottenere uno scopo altrimenti non possibile come il 'capitale sociale'. È spesso palese come sulle piattaforme sociali si produca una qualche forma di capitale sociale.

Riguardo la produzione di capitale sociale a mezzo piattaforma digitale, le posizioni in letteratura si possono dividere sostanzialmente in tre macro campi:

1. Pessimisti → che rilevano come si possa registrare una diminuzione del capitale sociale disponibile in funzione della diffusione delle piattaforme,
2. Positivi → sottolineano le potenzialità delle piattaforme digitali per produrre nuove connettività
3. Critici → si riconosce la creazione di capitale sociale ma si sottolineano gli aspetti problematici di questa produzione

Tecniche di analisi

Il campo epistemologico ora contiene tutti i tool concettuali necessari a passare in rassegna le tecniche di analisi dati.

Nel corso degli ultimi decenni le scienze sociali si sono spesso divise riguardo alla famiglia di metodi adottata: le scienze economiche sono state sempre più quantizzate; l'antropologia e alcuni filoni della geografia si sono sempre più caratterizzati per una posizione 'interpretativista'; nel mezzo la sociologia, senza una collocazione collettiva su questo continuum.

Nel contesto della ricerca digitale spesso è stata proposta l'equivalenza fra l'immaginario dei 'big data' e la ricerca sociale quantitativa ma non è possibile ridurre lo studio dei dati digitali a questa semplice dimensione; è necessario comprendere sia il contesto in cui i dati 'quantitativi' sono stati prodotti, sia darne una lettura più profonda per comprenderne il senso.

Spesso si ignora come 'a valle' di tecniche di classificazione 'non supervisionate' diventi necessario effettuare un lavoro di interpretazione pena la totale impossibilità di leggere i risultati; oppure, si ignora come applicare tecniche quantitative anche su un piccolo set di dati possa dare risultati eccellenti.

Big data

L'analisi di dati mediante strumenti informatici è uno dei campi in più veloce espansione nelle scienze sociali. Nota con il nome di 'big data analysis', sarebbe da preferirgli il più corretto termine 'computational social science' o 'scienza sociale computazionale'.

Il termine 'big data' trova la sua logica di applicazione a partire da considerazioni avvenute all'interno dell'industria informatica verso gli inizi degli anni 90. Sono caratterizzati da:

- Volume
- Velocità
- Varietà

Boyd e Crawford individuano i big data come un fenomeno a cavallo fra: tecnologia, tecniche di analisi e mitologia.

Secondo Kitchin i big data sono caratterizzati, oltre che alle tre V da:

- Esaustività → l'analisi 'big data' si prefigge lo scopo di analizzare l'intero universo
- Micro-risoluzione → i dati sono raccolti alla più piccola granularità possibile
- Relazionalità → i dati presentano una variabile indice che permette l'interoperabilità con altri set di dati
- Flessibilità → i dati sono in grado di espandersi facilmente sia in volume sia in profondità.

Possiamo considerare che gli approcci principali in relazione ai dati digitali sono sostanzialmente due:

1. Un gruppo di persone che, riconoscendo l'intrinseca problematicità delle relazioni causali nelle scienze sociali, scelgono di ignorarle
2. Quanti rifiutano di rinegoziare il bilanciamento fra descrizione e studio delle relazioni causali, rivendicando un ruolo chiave dell'elaborazione teoretica, del disegno della ricerca e del processo di creazione e testing delle ipotesi.

Cambiamento e studio dei big data

Le scienze in passato si sono sempre basate sugli 'small data'. A partire da questo problema fondamentale evolve il metodo scientifico.

Secondo gli autori che considerano i big data una sostanziale discontinuità epistemologica, il miglioramento esponenziale delle nostre capacità di cattura dati rende non necessario il processo di theory testing.

L'idea dell'approccio di Andreson è quella di abbracciare una trasformazione radicale, passando da una logica predittiva che risulta estremamente utile nel contesto del marketing ma mostra i propri limiti nel contesto della scienza.

In confronto agli approcci 'entusiasti' gli approcci critici sottolineano da un lato come sia tuttora indispensabile il disegno della ricerca tradizionale, dall'altro lato mettono in luce gli specifici limiti metodologici che si associano al paradigma dei big data. La critica principale è l'inapplicabilità di un approccio empirico.

L'attività di cattura dei dati non è mai una pura osservazione della realtà empirica ma presuppone una serie di limitazioni indotte dal contesto in cui l'osservazione si situa.

Problemi di 'cattura' dati

Fatti salvi gli sforzi di ricerca che riescono ad effettuare un'attività di cattura 'in proprio', la stragrande maggioranza dei progetti dipende da dati generati da utenti terzi per altro scopo. Bisogna capire quali variabili considerare e quali ignorare. Va compreso inoltre come i dati sono stati originariamente generati dalla piattaforma.

Nella misura in cui il campione si approssima all'universo, un numero maggiore di associazioni tra variabili diventerà statisticamente rilevante, anche se con effetti estremamente contenuti. Il problema diventa quello di selezionare le variabili appropriate. Estremizzando il ragionamento, se tutto interagisce con tutto e tutte le relazioni sono significative, torna a essere rilevante la capacità di discriminare quali relazioni sono di interesse.

Per quanto riguarda il problema dell'opacità, sappiamo come operino collezione e analisi dei dati, quando queste sono svolte da attori privati. Sono rari i contesti in cui un attore privato fornisca accesso, tramite API, alla totalità dei dati disponibili sulla piattaforma.

In mancanza di una possibilità di accesso indipendente viene meno anche la capacità di analizzare i claim portati dalle aziende nel corso della loro attività di ricerca.

Questi problemi di trasparenza, in fase di analisi, si ripercuotono anche sul ruolo delle piattaforme in un contesto pubblico: le piattaforme hanno un problema di controllo nella misura in cui contribuiscono a formare le decisioni pubbliche in un determinato settore.

Ad esempio, nel caso di Airbnb, come diventa possibile per un Comune rendersi conto dell'impatto delle piattaforme di short term rental in termini di valori immobiliari?

Qual è il costo in termini di privacy e supremazia tecnologica di affidare tutte le email di tutte le istituzioni di istruzione superiore a Google?

Esiste anche un problema di responsabilità nei confronti del singolo cittadino o della singola azienda e non del contesto sociale più ampio.

La produzione di software sempre più spesso avviene in un contesto di machine learning; questo significa che produrre del software non significa esplicitamente scrivere delle 'linee di codice', ma significa addestrare un sistema informatico ad agire in funzione di determinati input, facenti parte di un training set che è stato conferito al sistema nelle fasi della sua creazione.

2. Tecniche di raccolta dati

Il primo passo è la cattura dei dati medesimi a partire da quanto rendono disponibile le piattaforme.

Le tecniche di estrazioni dati sono ancora estremamente 'artigianali', spesso e volentieri dipendono da tool che sono stati creati da enti terzi.

In un certo senso, la condizione di 'artigianalità' è endemica dei metodi digitali per due ragioni:

1. Le piattaforme sono scarsamente trasparenti rispetto ai dati che producono
2. La giovane natura del campo fa sì che non esistano veri e propri standard tecnici di riferimento per l'acquisizione dei dati.

I tool commerciali sono stati progettati per altri scopi rispetto alla ricerca accademica, dal punto di vista del ricercatore gran parte di questi software soffre di forti opacità. Risulta impossibile, il più delle volte, comprendere quale logica stia dietro alle procedure di analisi.

L'acquisizione dei dati digitali avviene attraverso due strade principali:

1. l'accesso autorizzato tramite API → rientrano tutti quegli approcci che sfruttano una divulgazione controllata di dati fornita direttamente dalla piattaforma.
La API più utilizzata nell'ambito delle ricerche sui social media è stata, fino al 2023, quella di Twitter.
2. *screen-scraping* → è un insieme di tecniche che mira ad acquisire dati da una pagina web eseguendo il codice HTML che governa la visualizzazione.

Screen e web scraping

Le pagine web, in teoria, sono classificabili in due diversi ambiti:

1. Statiche → tutte quelle pagine il cui contenuto non varia a seconda delle circostanze
2. Dinamiche → tutte quelle pagine il cui contenuto è generato al momento della richiesta effettuata dall'utente tramite un browser.

Se gli script sono gestiti server side è possibile procedere allo scraping senza usare un emulatore di browser, in generale utilizzando un modulo di Python che gestisce le richieste HTML senza preoccuparsi dello scripting. Sarà necessario adottare uno specifico modulo per effettuare 'parsing', ossia la 'lettura' del codice HTML.

Lo screen scraping è certamente uno strumento potente e versatile poiché ogni pagina moderna deve generare del codice HTML che può essere quindi 'parsato'. Lo scraping può essere un lavoro estremamente lento ed artigianale, ogni parser deve essere configurato a mano per uno specifico sito o piattaforma.

Il webscraping da parte di enti terzi non è una pratica particolarmente gradita da parte di chi gestisce siti e piattaforme. In parte sussistono considerazioni di natura commerciale; in parte, poiché ogni richiesta di dati da un client ha un 'costo' in termini di banda.

I rischi etici sono da vagliare separatamente rispetto alla controversia legale ed è necessario ricordare che sussistono anche per gli approcci di data capture che non comportano il webscraping.

I ricercatori sono tenuti ad utilizzare i dati esclusivamente per scopi di ricerca, ad archivarli in un ambiente sicuro e a non divulgare i dati in forma che permetta l'identificazione dei soggetti studiati. Nel caso di studi a carattere quantitativo, i dati dovranno essere riportati in forma aggregata; nel caso di studi a carattere qualitativo si dovrà ricorrere a forme di pseudonimizzazione.

Codice → [webscraping di wikipedia](#)

Lo scraping di Wikipedia rappresenta un'alternativa poco efficiente per ottenere i dati contenuti negli articoli. Wikipedia dispone di un pacchetto dedicato⁴ il quale permette di acquisire i dati interfacciandosi direttamente con le API.

Il primo passaggio consiste nell'acquisire i dati tramite Requests e passarli a BeautifulSoup; sono disponibili diversi parser a seconda della complessità del testo da parsare potrebbe essere necessario cambiare parser.

stiamo chiedendo a Beautiful Soup di individuare quel tag h1 che, come attributo 'classe' ha questa specifica stringa. Per ottenere i valori dei tag, è necessario ispezionare la pagina utilizzando gli strumenti di debug e procedere alla visualizzazione del codice HTML per individuare il preciso tag che vogliamo acquisire, per poi copiarne gli attributi.

Si può usare anche Selenium, che è un emulatore di browser che permette di visualizzare pagine dinamiche.

Alcuni dataset di tweet venivano originariamente diffusi a partire dai codici ID dei singoli tweet; conseguentemente alla trasformazione in X non è possibile rintracciare i tweet partendo dall'ID in maniera gratuita. Per acquisire i dati da YouTube si è usato pyYouTube.

Non sempre le API restituiscono l'intero numero di commenti, soprattutto nel caso di numerosi commenti inseriti in una struttura molto ramificata.

3. Tecniche di analisi di base

Che si tratti di *Natural Language Processing*, sia che si tratti di *Network analysis*, sia che si tratti di classificazione supervisionata, quello che accomuna questa 'famiglia' di metodi e la loro irriducibilità a forme 'elementari' di conteggio o analisi statistica. Nel bagaglio culturale dello scienziato sociale dovrebbero essere presenti una serie di tecniche di analisi quantitativa.

Il corretto utilizzo delle tecniche computazionali più avanzate si basa su una serie di strumenti di base che dovrebbero essere padroneggiati prima di cimentarsi con questo tipo di approcci.

Codice → dati acquisiti tramite API di YouTube e Twitter

Il dataframe sarà il nostro strumento di lavoro principale.

In un contesto in cui tutto è stato eseguito alla perfezione non avremmo necessità di svolgere controlli preliminari sui dati; tuttavia, i dati possono essere stati raccolti in più 'ondate' o essere la risultante di più dataset aggregati in un secondo tempo, oppure possono essere stati catturati a partire da una serie di parole chiavi non mutualmente esclusive. A causa di tutte queste circostanze è sempre necessario ispezionare i dati prima di effettuare analisi.

Qualora ci fossero valori missing nella colonna 'id' sarebbe indice di un errore piuttosto grave in fase di raccolta. Per contro, sappiamo che è possibile, per un tweet, non avere 'entità' attribuite, per cui la presenza di missing in questo campo è da considerarsi normale.

Oltre ai valori nulli è possibile che siano presenti valori duplicati; la conoscenza del dataset è necessaria per comprendere se la presenza di valori duplicati costituisce un problema o meno.

Prima di contare il valore medio nel dataframe sono stati eliminati i duplicati.

Calcolare gli autori più retwittati presuppone l'utilizzo di una funzione creata appositamente. Potremmo essere interessati a conoscere solo il 'dominio' degli url poiché questo può, ad esempio, fornire informazioni circa il tipo di sito che viene condiviso dagli utenti.

Il primo network che si tende a fare a partire da dei dati provenienti da Twitter è generalmente un network di Retweet, assumendo che, la suddivisione in claque omogenee evidenzia le linee di frattura nell'opinione pubblica e i vari discorsi che la animano. Queste reti, una volta caricate su Gephi potranno essere manipolate e soggette ad analisi.

Si potrebbe pensare di ignorare il fattore tempo; questa cosa pone dei problemi se l'operazione viene svolta per effettuare un'operazione di community mapping.

Un secondo tipo di rete che risulta essere estremamente utile per effettuare analisi esplorative è l'hashtag network. Nell'hashtag network le dimensioni degli hashtag sono proporzionali al degree. La distanza tra nodi si calcola generalmente utilizzando ForceAtlas2.

A seconda della struttura del nostro disegno di ricerca possiamo estrarre diversi tipi di sample come ad esempio randomico.

Per l'analisi etnografica esistono pratiche maggiormente efficienti per l'estrazione di dati; è possibile estrarli dalla 'testa' o dalla 'coda'.

Queste particolari configurazioni estrattive permettono all'analista di apprezzare la differenza fra i discorsi distribuiti lungo la 'legge di potenza' che governa i retweet, allo scopo di comprendere se esiste una particolare differenza fra i discorsi degli 'influencer' e quelli operati dalla massa di utenti.

Le analisi dati che il notebook di accompagnamento propone consistono nella misurazione delle varie grandezze che definiscono un video di YouTube. Possiamo misurare come sono distribuiti like o commenti fra i video, oppure possiamo aggregare queste grandezze a partire dal titolo del canale. YouTube permette di catalogare il contenuto di un video a partire dai tag.

Incrociando la distribuzione con altre variabili è possibile comprendere se esiste un'evoluzione dei temi.

4. La reputazione digitale

Le dichiarazioni pubbliche di Elon Musk su twitter hanno influenzato il prezzo di scambio delle criptovalute con la possibilità di pagare una Tesla in Bitcoin. Dopo questo annuncio il prezzo è salito e, quando Musk ha cambiato idea, è sceso di molto.

Questa serie di eventi permette di rendersi conto della misura in cui la reputazione personale diviene immediatamente monetizzabile nel momento in cui viene digitalizzata.

La reputazione online rappresenta una realtà quotidiana per la quasi totalità dei lavoratori es. impiegati valutato anche per il profilo LinkedIn.

Le riflessioni relative all'interconnessione tra reti digitali e reputazione sono ben più antiche. Negli anni '90 Sassen rilevava come la digitalizzazione del lavoro di ufficio e la centralità delle telecomunicazioni abbiano prodotto una nuova centralità delle relazioni personali. Anche Castells individua nella reputazione 'in rete' e nelle capacità di autoaffermazione di identità.

In conseguenza della proliferazione delle piattaforme, molte più professioni e figure sociali si sono trovate in qualche modo coinvolte, volenti o nolenti, in processi di curatela della propria reputazione digitale.

La possibilità di lavoratori e cittadini di connettersi in rete e di essere valutati, in quella sede, in maniera 'imparziale' ha rappresentato una delle narrazioni utopiche del XXI secolo.

La diffusione delle piattaforme digitali si è accompagnata ad un incremento della ricerca empirica sulle tematiche della reputazione, soprattutto nel contesto del *gig work* e della *sharing economy*.

Quello che risulta rilevante, in questa sede, è osservare come, lungi dall'essere 'neutra', la produzione di reputazione dipende fortemente sia dalle affordances della piattaforma, sia dalle sue dimensioni.

Esempio Airbnb

Altri autori rilevano come i servizi offerti dalla piattaforma Airbnb fungano da volano per la creazione di capitale reputazionale sulla piattaforma: il ruolo di superhost richiede di allestire le foto dell'appartamento in una certa maniera, oppure di farsi servire da un fotografo pre-approvato da Airbnb, richiede di definire le proprie disponibilità in funzione dei target promossi dalla piattaforma, nello specifico di fornire la propria disponibilità per un certo numero di giorni all'anno. Al tempo stesso non necessariamente le affordances delle piattaforme 'sovrascrivono' gli stigmi sociali.

Informazioni

Emerge una seconda linea di riflessione che tocca i problemi della reputazione online: la diffusione di informazioni affidabili.

Le opinioni di un cittadino si formerebbero grazie all'intermediazione di un 'leader d'opinione' che 'garantirebbe' la veridicità e la salienza dell'informazione. L'avvento dei social media ha sconvolto questa specifica rappresentazione.

la partecipazione in rete fornirebbe ai singoli utenti gli strumenti per acquisire in maniera autonoma le proprie fonti di informazione, rendendo, in qualche misura, obsoleta la reputazione 'tradizionale' detenuta da opinion leader e legacy media.

In una direzione simile si muove il ‘solco di letteratura’ delle *digital crowds*, in cui si sostiene come i social media abbiano abilitato il passaggio da una forma strutturata di pubblico, caratterizzata da una serie di rapporti collocati in gerarchia, a una serie di interazioni effimere tra utenti.

Esempio directioners e attori dotati di ‘voice’

Arvidsson e colleghi, analizzando la mobilitazione dei 'directioners', rivelano come non esistano particolari 'figure di riferimento' nel fandom, ma come i fan si limitino ad usare il pannello dei trending topic di Twitter per coordinare le proprie azioni.

Si nota una centralità della logica di piattaforma rispetto alla gestione della reputazione.

Controversie più recenti hanno analizzato come nelle controversie gli attori dotati di ‘voice’ corrispondano sempre più spesso ad attori che anche fuori dalla rete godono di una certa reputazione. Assistiamo a un effetto combinato in cui la logica della piattaforma nutre la visibilità garantita ad alcuni attori, all'interno del sistema dei media tradizionali.

Network analysis

La reputazione può essere un sistema di rating, o anche una metrica di citazione.

La network analysis rappresenta una tecnica versatile; è possibile utilizzarla per effettuare analisi del testo, per comprendere le reti di diffusione di una fake news o la capacità di una bot network di agire all'unisono.

La network analysis si occupa di analizzare le relazioni fra enti, o, 'nodi', connessi in vario modo fra di loro a mezzo di 'ponti'; la rappresentazione formale della connessione tra nodi e ponti è denominata 'grafo'. Nella network analysis oggetto dell'indagine sarà il posizionamento relativo dei nodi e la loro interconnessione all'interno del grafo.

es. network di amicizia → possiamo descrivere ogni individuo come un nodo e ogni rapporto fra persona A e B come un ‘ponte’ che collega A e B. nella misura in cui B ha rapporti con molte persone, la sua posizione ‘migliorerà’.

Tipi di connessioni e misure

Le connessioni stabilite dai ponti possono (o meno) avere un verso. Se una connessione fra A e B non dà necessariamente origine ad una connessione fra B ed A possiamo parlare di una 'direzione' tra le connessioni e quindi di un grafo 'direzionato'.

La misura della distanza non è generalmente utilizzata nell'ambito delle analisi di rete più comuni che possono essere effettuate su dati provenienti da social network.

In-degree e l'out-degree sono dei parametri per descrivere le connessioni ‘entranti’ in un nodo e quelle di ‘uscita’; il secondo è un attributo tipico dei nodi estremamente attivi. Altre due misure sono:

- *Betweenness centrality* → misura quanto spesso un determinato nodo si trova sul segmento più breve che connette due nodi
- *Closeness centrality* → descrive la velocità con cui da un determinato nodo è possibile raggiungerne altri.

Codice → network reputazionali in Accademia

I dati vengono presi da Scopus che permette l'estrazione di 20.000 elementi al massimo.

I network di co-authorship ci permettono di esplorare queste strutture emergenti senza fare particolari assunti circa la loro natura. Al posto di esplorare quello che gli autori di osservanza 'antropologica' o 'sociologica' hanno scritto, prendiamo un insieme più vasto e procediamo ad individuare dei sottografi 'densi' che rappresentano, idealmente, i singoli campi disciplinari; questa specifica procedura ci permetterà di individuare gli autori di 'snodo' fra un field e l'altro.

Dall'analisi dei grafici è possibile comprendere come i giornali che accolgono il maggior numero di contributi siano giornali *open publish*.

Un ristretto numero di articoli è depositario della stragrande maggioranza delle citazioni, mentre una enorme 'coda lunga' di articoli può contare su pochissime citazioni ciascuno. Qualora fossimo interessati a quali sono i giornali più prestigiosi, una metrica più interessante della semplice conta del numero di articoli presenti per ciascun journal sarebbe un raffronto tra somma e media delle citazioni, per articolo, per ciascun journal.

Se consideriamo la media di citazioni per articolo, le riviste che assicurano la migliore performance sono generalmente delle raccolte di atti da conferenze o di area 'scienze della formazione' o di area informatica. Se analizziamo l'andamento del numero degli articoli nel tempo, possiamo osservare come il field delimitato dalla nostra query sia in fase sostanzialmente calante dal 2020.

Probabilmente, questo fenomeno è da riscontrarsi nell'alterazione delle keyword con cui i ricercatori definiscono un approccio basato sui dati digitali nel loro specifico field. Nella misura in cui ciascun subfield evolve una propria terminologia decrescerà l'uso di termini generici.

Il clustering medio può essere espresso come un valore compreso fra 0 e 1: più i valori si approssimano allo zero più ciascun nodo sarà reciprocamente disconnesso dai nodi a lui vicini, più si avvicinano a 1 più il grafo sarà 'completo', ossia più esisterà un legame per ciascuna coppia di nodi.

L'assortatività misura la proporzione di nodi che si associano a nodi con identiche misure di centralità. Per 'componente' si intende una sotto porzione 'disconnessa' di un determinato grafo in cui tutti i nodi risultano raggiungibili da tutti quanti gli altri.

Quello che descrivono queste misure è un grafo estremamente disconnesso popolato da un largo numero di microcomponenti che rappresentano le claques di autori più avvezzi alla collaborazione reciproca. All'interno di queste claques si svolge un rapporto denso e reciproco fra gli autori.

Sostanzialmente quasi ogni componente del grafo rappresenta una comunità a sé stante. Una volta individuate le appartenenze diventa possibile isolare le top 5 comunità ed effettuare un'analisi del contenuto.

Vengono riportate 5 wordcloud. Analizzando i termini associati a ciascuna delle comunità è possibile farsi un'idea del loro contenuto.

Queste tecniche potrebbero essere utilizzate per analizzare dei network di reputazione in contesti radicalmente differenti.

5. Le controversie digitali

Uno dei fenomeni più indagati nelle scienze sociali che si occupano di 'cose digitali' è sicuramente quello delle controversie online. Per controversie si intende una disputa pubblica relativa all'interpretazione di alcuni fenomeni. es. la pandemia da COVID-19 ha fornito un discreto numero di esempi di controversie.

Gli addentellati epistemologici e metodologici fra il controversy mapping, la sociologia computazionale ed i metodi digitali sono potenzialmente molti e molto fruttuosi.

Actor Network Theory (ANT)

Il punto fondamentale dell'*Actor Network Theory* (ANT) è la caratterizzazione dei fenomeni sociali. Se la sociologia classica adotta un approccio che fa discendere i fenomeni sociali a dei costrutti astratti, l'ANT adotta un percorso inverso. Le 'forze sociali' vengono ricostruite come la risultante di una serie di relazioni tra attori umani e non umani.

Con 'attori non umani', generalmente, ci si riferisce a delle pratiche professionali, ma molto facilmente si potrebbe classificare sotto questo taxa anche delle idee, delle 'strutture psichiche' o dei preconcetti sufficientemente diffusi.

Precisamente in quell'area in cui gli artefatti tecnici escono dal 'laboratorio' ed incontrano l'uso quotidiano risiede la sovrapposizione tra ANT e metodi digitali.

Le affordances rappresentano una manifestazione, estremamente palese, di quello che nell'ANT sarebbe definito come 'attore non umano'. Sotto l'etichetta delle affordances troviamo:

- La struttura 'tecnica' della piattaforma
- Le rispettive culture d'uso delle piattaforme si sono strutturate in relazione sia alla struttura di interfaccia sia in relazione a sé stesse e alle passate esperienze d'uso

Controversie in ambiente digitale

Si tratta di tematiche intrinsecamente conflittuali. Per quanto la comprensione di una controversia digitale richieda una profonda conoscenza del contesto socio/tecnico/politico in cui è immersa. È possibile usare almeno due 'famiglie di metodi':

1. Topologico → tutti quei metodi che si basano sulla modellizzazione delle relazioni fra attori a partire da una struttura a grafo

Esempio lavoro topologico

es. lavoro di Yasseri relativi a Wikipedia sulle 'edit war', ossia una serie di edit successivi di una specifica voce in cui un determinato gruppo di utenti 'cancella' o 'emenda' i lavori di un altro gruppo.

Mappare le controversie a partire da un approccio topologico rappresenta un metodo relativamente semplice e indipendente dal linguaggio; tuttavia, questo tipo di approccio è soggetto a potenziali errori che però è relativamente facile controllare. Andrebbe controllato il 'fattore tempo': una controversia è tale se si situa in un periodo preciso. Accrescere il periodo di osservazione potrebbe portare, a seconda delle circostanze, a 'falsi positivi'.

Nonostante la natura formalmente 'language agnostic' dell'approccio topologico, qualora un determinato tema sia stato sviluppato su più linguaggi, le relazioni di mutua citazione tratteranno i confini di comunità linguistiche e non di due lati della stessa controversia.

2. Semantico → si basa su tecniche di Natural Language Processing. Una prima tecnica che si può adottare è il topic modeling.

Tutti gli algoritmi rappresentano altre tecniche per associare tra di loro le parole che co-occorrono più spesso e dissociare quelle che co-occorrono meno spesso.

Le tecniche di topic modeling produrranno dei 'cluster' di parole che catturano le dimensioni semantiche latenti dell'insieme di documenti, varianti di questa specifica tecnica sono la *word network analysis* e lo *structural topic modeling*.

Le tecniche di topic modeling rappresentano un ottimo strumento per 'leggere' un testo qualora sia impossibile farlo in maniera tradizionale. È necessario ricordare che i risultati del topic modeling necessitano di interpretazione per essere significative.

Un ulteriore metodo per assicurarsi la presenza di una controversia online passa dall'individuazione di metriche di permeabilità fra 'cluster'. Avendo a disposizione un criterio di clustering come ad esempio una *community detection*, diventa possibile calcolare il numero di citazioni incrociate tra cluster.

Un ulteriore tipo di tecnica non supervisionata, basata sul natural language processing che può essere utile applicare all'analisi delle controversie, può essere la funzione TF-IDF → lo score viene calcolato a partire dalla frequenza del dato termine in un documento e dalla frequenza inversa di quel termine fra tutti i documenti. La TF-IDF si utilizza quando c'è chiarezza relativamente alla suddivisione dei cluster che compongono la controversia.

Analisi delle controversie

Nel contesto dell'analisi delle controversie può essere utile conoscere quali sono i nuclei di senso che vengono portati da quali attori e in quali configurazioni. Realisticamente le tecniche vengono utilizzate in combinazione:

Il primo passo è acquisire la storia completa degli edit di una pagina Wikipedia, dati messi a disposizione da uno specifico set di API. L'output prodotto dalle API è organizzato in un formato JSON in cui a ciascun utente corrisponde il testo effettivamente editato.

wiki event → elabora un file XML acquisito da una pagina per trarne un network di eventi che collegano tutti gli utenti che hanno scritto modifiche al testo di quella pagina. Gli eventi possono essere di tre tipi:

1. aggiunta di testo
2. cancellazione di testo aggiunto da un determinato utente
3. ripristino del testo aggiunto da un utente A e cancellato da un utente B.

Codice → esempio controversia sulla pagina Wikipedia di George W. Bush

Poiché abbiamo necessità di una voce controversa, procediamo a scaricare il file xml relativo alla pagina di George W. Bush.

Per scaricare il dump xml ci sono due modi: software che possono scaricare il file per noi o acquisire i file manualmente. Sarà necessario importare il file xml in Wikievent chiedendo al software di estrarre gli eventi.

La creazione di una variabile datetime per descrivere la successione delle recensioni e una di duplicazione degli eventi.

Il numero di singoli eventi può iniziare a darci un'idea della conflittualità presente in una determinata voce, mentre invece il grafico della distribuzione temporale ci permette di comprendere quando effettivamente si sono svolti gli eventi di editing. In condizioni 'normali' questo grafico dovrebbe registrare molta attività nel periodo di creazione della voce e poi una stabilizzazione degli interventi verso numeri più bassi: la presenza di eventuali picchi di attività in periodi successivi può indicare o revisioni della voce.

Come previsto il picco delle review si situa a ridosso del 2003, in coincidenza con l'invasione dell'Iraq. Se consideriamo il numero assoluto di reviews, il punto di massima della linechart si situa negli ultimi mesi del 2002; se consideriamo l'estensività degli interventi, il picco coincide con i mesi immediatamente successivi all'invasione.

A partire dal network di conflitto poi è possibile calcolare un network di 'mutuo conflitto', ovvero un grafo che connetta utenti che hanno effettuato modifiche negative verso lo stesso nodo: in pratica se A e B hanno cancellato del testo di C, il codice successivo crea una relazione fra A e B.

Dopo averne calcolato l'estensione, sarà necessario visualizzarlo tramite Gephi per poter utilizzare ForceAtlas2 per poter comprendere i confini delle comunità di affinità. Pur senza conoscere gli effettivi contenuti delle review, si può già comprendere come il grafo assuma una forma, almeno parzialmente, polarizzata.

A partire dalla suddivisione in comunità diventa poi possibile calcolare gli score di TF-IDF relativi alle singole comunità.

6. Il problema della concentrazione

Il web di cui abbiamo esperienza nel contesto contemporaneo è tutt'altro che decentrato.

Quello che si viene a creare è una concentrazione di risorse connaturata con l'attività delle piattaforme che ospitano i diversi contenuti. Quello che a noi interessa è sia comprendere dove si situi lo scarto fra la predizione e la realtà contemporanea, sia acquisire gli strumenti per misurare la concentrazione nei contesti in cui si manifesta.

Quando parliamo della relazione tra piattaforme e concentrazione di risorse possiamo riferirci a due differenti scenari:

1. La piattaforma favorisce la concentrazione di risorse che preesistono alla piattaforma. Qui si collocano le piattaforme che intermediano relazioni economiche di vendita o affitto di beni.
2. Sia la risorsa che si concentra che il sistema adottato per distribuirla, esistono esclusivamente all'interno della piattaforma digitale, la piattaforma controlla totalmente sia la risorsa, che il sistema per distribuirla. Questo scenario è tipico del social network mainstream.

Web 'libero' e nodi

Nel caso di piattaforme digitali o del web 'libero', la metafora della rete descrive, con il concetto di 'nodi', una serie di documenti contenenti informazioni; con il concetto di 'link' un collegamento fra documenti che indicizza informazioni presenti nel documento A ma necessarie nel documento B.

Nella misura in cui il documento A è molto più visibile di altri godrà di una rendita di posizione: in altre parole i futuri creatori di documenti, avendo necessità di reperire informazioni circa argomenti coperti da A, avranno molte più probabilità di collegarsi ad A piuttosto che ad un qualsiasi altro documento che tratti la stessa materia. Nel web contemporaneo la visibilità di una pagina è determinata dal suo posizionamento nei motori di ricerca. Uno dei fattori che favorisce un buon piazzamento è il numero di connessioni verso la pagina da indicizzare.

Modalità di visualizzazione, ranking e vendita

Lo score attribuito dagli utenti costituisce la base per poter decidere una gerarchia di visualizzazione dei contenuti. A sua volta, la gerarchia privilegerà i contenuti con un ranking migliore. Gli utenti con un ranking migliore. Bisogna ora considerare che gli utenti che decidono di acquistare un bene o un servizio sostengono un costo di ricerca informazioni circa lo stato o le qualità dell'oggetto, questo costo sale nella misura in cui è difficile stimare di prima mano le qualità del bene.

Risulta difficile stimare le qualità di un bene quando esiste un'asimmetria informativa tra acquirente e offerente; le piattaforme digitali rappresentano un tentativo di colmare questa asimmetria informativa utilizzando informazioni disponibili.

Il costo è la possibilità che il bene sia non conforme alle aspettative di chi lo richiede; diventa comprensibile dal punto di vista dell'utente orientarsi verso beni che abbiano già un pool di recensioni all'attivo.

Se prendiamo una relazione di 'follow' esiste, anche in questo caso, un costo sostanziale che è quello dell'acquisire informazione di cattiva qualità, o comunque non conforme alle proprie aspettative. Il giudizio degli altri utenti viene funzionalizzato per risolvere questo problema. Gli account che possono

contare su un pool di condivisioni preesistenti saranno realisticamente gli account che accumulano più connessioni.

Il marketplace digitale

Uno dei primi ambiti ad essere oggetto di analisi è stato quello dei P2P marketplaces. Le prime descrizioni presenti in letteratura si focalizzano su come le piattaforme riescano a diminuire radicalmente i costi di partecipazione al mercato.

Chris Anderson rileva come i marketplace digitali distribuiscono le preferenze degli utenti su una coda lunga. Questo ha un effetto di de-concentrazione molto forte.

Esempio Airbnb

Esito simile stanno avendo alcune stime circa la distribuzione delle review su Airbnb. Picascia e colleghi misurano la concentrazione del fatturato in 10 città italiane stimando score ben superiori a .5 sull'indice di Gini per tutte le città coinvolte. Il livello di concentrazione aumenta al crescere degli utenti che utilizzano la piattaforma, almeno fino al 2019. L'offerta di alloggi su Airbnb è estremamente concentrata nello spazio.

La concentrazione indotta delle piattaforme digitali potrebbe essere analizzata anche in funzione della dimensione degli attori che utilizzano la piattaforma. Cocola Gant e Gago, utilizzando un mix di tecniche qualitative e quantitative, dimostrano come a Lisbona Airbnb abbia favorito la penetrazione da parte di attori transnazionali nel locale settore del real estate.

Non solo le piattaforme producono concentrazione tramite le metriche di reputazione, ma magnificano fenomeni di concentrazione preesistenti nel capitalismo contemporaneo. La logica è da ricercarsi nella natura dei mercanti immobiliari.

Airbnb, e in generale le altre piattaforme di short term rental, fungono da layer di compatibilità fra un contesto locale opaco e le necessità di standardizzazione proprie dei capitali transnazionali. Possiamo dividere il problema della concentrazione in due assi:

1. Stabilendo se esiste concentrazione o meno
2. Stabilendo chi viene avvantaggiato da un'eventuale concentrazione delle metriche social.

È necessaria un'ulteriore premessa: la concentrazione riguarda risorse scarse, che nel primo scenario sono più facili da visualizzare. In questo caso, la risorsa da concentrarsi è rappresentata dall'attenzione degli utenti.

Esempio Twitter

Nella letteratura empirica la maggior parte dei lavori cita Twitter per ovvie ragioni di disponibilità dei dati. Le riflessioni sulla concentrazione sono coeve alla nascita del network.

In condizioni di 'baseline' la distribuzione delle misure di centralità è sostanzialmente più 'equa' rispetto a ciò che avviene durante i media events, in cui un ristretto numero di élite digitali ottiene un gran numero di menzioni da parte dell'utenza generale.

Nel caso di facebook abbiamo indicazioni circa la concentrazione delle economie della reputazione, anche se in questo caso l'impossibilità di ottenere facilmente una rete di relazioni rende oggettivamente più complicato misurare la concentrazione di metriche di network.

Concentrazione nei social network

Un secondo aspetto connesso al problema della centralizzazione relativo alle economie dell'attenzione è legato alla classe di persone che ne beneficia.

La 'rappresentazione pop' dei social network e molta della retorica dell'ideologia californiana vuole che essi siano un luogo in cui il consenso si forma tramite un variegato 'mercato delle idee' in cui un buon dilettante può competere ad armi pari con i professionisti della produzione di contenuti e dove il retroterra di credibilità e capitale sociale abbia un'importanza più contenuta. I primi lavori che hanno investigato la comunicazione politica su Twitter sembrano fornire una conferma di questo stato di cose.

es. le conversazioni collegate ad #egypt rilevano come gli attori più influenti di una RT network siano cittadini comuni.

Esiste un sostanziale numero di contributi recenti che ribaltano questa prospettiva: es. rifugiati in Europa, rilevano come la distribuzione a legge di potenza di Twitter favorisca account collegati al sistema dei legacy media.

La discrepanza fra queste due prospettive non implica necessariamente che una di esse sia errata; è necessario considerare come gli studi che osservano una predominanza degli attori comuni siano tutti dislocati nell'infanzia di Twitter.

Legacy media e politici hanno imparato ad usare Twitter facendone un cardine della propria strategia editoriale.

Codice → calcolare score di concentrazione economica su Airbnb

L'indice di Gini è calcolato per numero di listing e introiti. Inizialmente si scaricano i dati relativi al numero di prenotazioni e al numero di recensioni da Inside Airbnb; i dati utilizzati sono provenienti da Milano.

Si caricano i due file contenenti i listing e le reviews supponendo che a ogni review corrisponda una effettiva permanenza.

Nel corso degli anni l'indice di Gini è cresciuto molto velocemente e poi è rimasto sostanzialmente stabile. Qualora volessimo calcolare invece l'indice di Gini per i guadagni sarebbe necessario implementare una piccola modifica al codice. Sommare il ricavato ottenuto da ogni host disponibile nella colonna 'price' del nostro dataset. È necessario pulire il dataset prima di manipolarlo.

Per calcolare i valori dell'indice di Gini su base annuale servirà una tabella pivot. Dopo aver pulito i dati si creerà, per ciascun anno, una tabella pivot in cui l'indice sarà costituito dagli id degli host. Verranno sommati tutti gli introiti ottenuti da un determinato host.

Airbnb risulta essere particolarmente adatto per esplorare il problema della concentrazione, essendo una piattaforma che media servizi di domanda e offerta alloggi; cognitivamente è molto semplice prefigurare il ruolo della concentrazione. Il problema dell'iperconcentrazione delle risorse tocca tutte le piattaforme digitali ed è quindi possibile calcolare l'indice di Gini anche per i retweet.

7. Gli attori non umani

Uno dei campi di battaglia del conflitto Russo-Ucraino del 2022 è sicuramente quello dei social media. Tutti i partecipanti al conflitto hanno utilizzato i social per disseminare le proprie posizioni in termini di propaganda; i social sono diventati una vera e propria arma.

Il punto più rilevante è l'utilizzo massiccio di account automatizzati per diffondere le notizie 'di parte'. Una analisi della rete di retweet caratterizzati dalla presenza degli hashtag #istandwithrussia e #istandwithputin ha individuato alcuni cluster linguistici che potevano essere definiti 'sospetti', in generale composti da account con, storicamente, poche interazioni ed impegnati a retwittare massicciamente alcuni nodi centrali che fungevano da produttori di contenuto virale come ad esempio meme pro-invasione → un attore con un'agenda politica utilizza un certo numero di account 'fittizi' per spingere alla creazione di un consenso a lui favorevole.

Account automatizzati

L'individuazione degli account 'automatizzati' rappresenta un problema metodologico di sicuro interesse e di una discreta complessità. L'uso di bot è ben diffuso anche al di fuori delle operazioni di propaganda bellica.

Tecniche anti-bot

Nel corso dell'ultimo decennio le tecniche di detezione dei bot hanno subito un'enorme diffusione.

La stragrande maggioranza degli approcci riguarda Twitter. Tuttavia, soprattutto dal rilascio di CrowdTangle, è stato possibile sviluppare approcci dedicati anche a Facebook o Instagram. Tre diversi approcci alla Bot Detection:

1. Utilizzo di tecniche di machine learning per stimare una serie di feature in grado di designare lo status di 'bot' o meno di un determinato account.
2. Utilizza la coordinazione fra account per stabilire lo status di account 'sospetto': nella misura in cui due o più account condividono lo stesso contenuto in un determinato 'buffer' temporale; questi account saranno 'sospetti' di essere operati 'in serie' da un operatore umano o da un vero e proprio automatismo.
3. Utilizza le metriche di rete: un gruppo di account che si condivide il contenuto a vicenda tende ad attivarsi esclusivamente in precisi momenti temporali e che presenta anche un basso numero di interazioni con altri account che non ne condividono le caratteristiche specifiche; sarà con ogni probabilità una 'bot network'.

Botometer (approccio 1)

Botometer si usa per individuare gli account automatizzati su Twitter. Il cuore del sistema è composto da un modello previsionale che utilizza una serie di stimatori per comprendere se un determinato account è un bot o meno.

Fra le feature ritroviamo: feature linguistiche, caratteristiche specifiche del profilo quali la bio, la foto profilo, la data di apertura dell'account, il numero di post effettuati in un dato arco temporale, caratteristiche connesse alla rete di contatti. Il programma restituisce uno score da interpretarsi come una percentuale di possibilità che l'account esaminato sia o meno automatizzato.

L'approccio di Knauth parte da presupposti simili rispetto a Botometer ma è 'language agnostic', ossia fra le feature necessarie per predire la presenza o meno di un bot non si considerano caratteristiche riconducibili al natural language processing.

Comparare delle stime derivate a partire da tecniche di NLP fra diversi gruppi linguistici vuol dire, infatti, comparare degli approcci tecnici caratterizzati da gradi di precisione molto dissimili.

CooRnet (approccio 2)

CooRnet è un pacchetto per R, il cui funzionamento si basa sulla verifica delle occorrenze temporali nella condivisione di un determinato set di link. Si interfaccia con le API di CrowdTangle e richiede quindi un accesso 'approvato' da parte del team di gestione della piattaforma.

A partire da un gruppo di link il pacchetto computa il 'coordinated sharing' per ciascun link. A partire da questi risultati diventa possibile dedurre 'a ritroso' le potenziali bot network.

Un ulteriore approccio rileva lo svolgimento di azioni simili in una finestra temporale estremamente ristretta.

Python ddna

Python ddna mira a individuare l'azione coordinata di account su Twitter ma si basa sulla somiglianza rispetto a tutte le altre. È il più indicato a scoprire gli account 'ibridi'.

(approccio 3)

L'utilizzo delle metriche di rete per individuare le caratteristiche specifiche sia degli account sospetti, sia delle reti che questi generano.

Esistono anche alcuni approcci che mirano a individuare bot o 'fake' accounts su Instagram.

Ciascuna di queste tecniche incontra dei limiti piuttosto onerosi nella capacità di accedere ai dati tramite API: diventa impossibile classificare dei profili che vengono rimossi dalla piattaforma poiché sospettati di essere account automatizzati.

Codice → individuare bot su Twitter

Alcune delle tecniche più diffuse per individuare i bot sono, recentemente, rese quasi inservibili dalla chiusura delle API di Twitter.

Partendo da due insiemi di tweet preesistenti si istruirà un classificatore in grado di individuare se un tweet proviene da un bot o se proviene da un utente regolare. I due dataset che utilizzeremo sono:

1. una raccolta di tweet effettuati da account riconducibili al Internet Research Agency
2. un training set utilizzato per la sentiment analysis che contiene tweet estratti randomicamente nel 2008.

Selezionando un numero di tweet molto piccolo rispetto ai dataset. Rimuoviamo url e mentions da entrambi gli insiemi. Vorremmo evitare di istruire il modulo nell'individuare gli account dei personaggi politici e degli influencer 'intercettati' dai bot, vorremmo invece basarci sul puro testo condiviso nei tweet.

Successivamente al primo round di pulizia i due insiemi di tweet saranno conferiti ad un singol dataframe su cui poi effettueremo il target. Importiamo l'utility test train split che ci permetterà di prendere un insieme di dati e dividerlo in una componente di train ed una di test.

La fase di testing ritornerà poi un accuracy score che varia da 1 a 0 che rappresenta la proporzione fra il totale degli elementi di test e le predizioni corrette. In pratica, nel nostro caso, la fase di test 'predice' il valore della colonna 'flag' basandosi sul testo presente nella colonna text e il valore di accuracy è semplicemente il rapporto fra le predizioni corrette e il totale dei tweet.

La funzione produce quattro variabili e richiede quattro parametri:

1. i testi dei tweet
2. sequenza delle label che designano se il contenuto è stato scritto da un bot o da un utente umano e costituisce ciò che l'algoritmo di classificazione dovrà predire.
3. Porzione di dati che utilizzeremo per comporre la componente di test nel machine learning
4. Il seed della relazione randomica dei singoli tweet

Il train prende i due insiemi di variabili su cui operare la predizione e li divide in quattro variabili differenti. Una proporzione di X e Y verrà usata come input per addestrare il classificatore, successivamente si procederà a predire in serie il valore di Y_test utilizzando le feature contenute in X_test ed il modello calcolato a partire dai valori di train.

Nel nostro caso specifico vogliamo verificare se un tweet appartenga a un bot dell'IRA, o meno; si tenga presente però che queste sono feature 'minimali' su cui trainare un modello. I modelli effettivamente utilizzati per fare bot detection utilizzano un vettore di feature decisamente più nutrito.

La variabile binaria sarà la classificazione in bot e 'normali' dei tweet e la distribuzione delle parole definita dal vettorizzatore funge da parametro.

Ciascuna porzione inizializza il classificatore secondo l'algoritmo prescelto, poi istruisce il classificatore utilizzando le variabili di train e calcola l'accuracy dello specifico classificatore.

Scelto il modello con la migliore performance si procede alla classificazione di nuovi tweet e lo si farà utilizzando la variabile che contiene il classificatore.

8. *L'analisi multiplatforma*

C'è da considerare, in prima battuta, un problema di accesso ai dati, in seconda un problema di interoperabilità dei dati.

Lo stato dell'arte per quanto riguarda il data capture su Facebook è stato caratterizzato dall'utilizzo di un mix di API e scraping. L'unico modo di accedere ad Instagram è stato l'utilizzo di screen scraper. Per quanto riguarda piattaforme di natura più commerciale e meno social l'unico modo di acquisire dati è lo screen-scraping.

Ben più spinoso è il problema dell'interoperabilità dei dati e della loro operativizzazione. Misure apparentemente simili possono essere lievemente diverse: un hashtag su Twitter ha un ruolo radicalmente diverso da un hashtag su Instagram.

Comparazione

Entro quest'ambito rientra l'avvertenza formulata da Rogers riguardo alle 'device cultures': nell'interpretare i dati provenienti dalle diverse piattaforme è importante considerare come ciascuna imponga differenti regole di visibilità.

È necessario ottenere una padronanza delle 'affordances' di ciascuna piattaforma prima di effettuare paragoni e confronti.

Un ragionamento molto simile si dovrebbe fare nel comparare le varie forme di 'reaction' o 'share' disponibili sulle piattaforme.

Un ulteriore punto critico nel comparare piattaforme differenti è rappresentato dall'associazione delle identità digitali, dal tentativo di comprendere se l'account A su Instagram corrisponde all'account B su Facebook.

Raffinando ulteriormente queste posizioni possiamo appoggiarci a quanto riscontra Noortje Marres analizzando i bias impliciti nei dati digitali. Individua tre ambiti problematici nella ricerca digitale.

Sia che si parli di tool precompilati, sia che si parli di utilizzo delle API, ciascun tool possiede delle proprie idiosincrasie che possono significativamente alterare la rilevazione e l'analisi dei dati. Se queste circostanze sono relativamente gestibili in una ricerca su una singola piattaforma, a patto che si riconosca e problematizzi il bias, rischiano di diventare estremamente più problematiche per un disegno di ricerca comparativo tanto da poter inficiare la comparazione stessa.

es. un disegno cross-platform per confrontare Twitter con Mastodon, poiché le API del primo prevedono la possibilità di effettuare una ricerca per parole chiave, e poiché le API del secondo non prevedono quest'ultima possibilità le chances di effettuare un confronto metodologicamente solido fra piattaforme relativamente simili decrescono sostanzialmente,

Cross-platform analysis

La Cross-platform analysis non si basa su una specifica tecnica ma su una particolare forma di disegno della ricerca.

La comparazione fra piattaforme si può articolare su tre assi principali, in funzione dell'unità di analisi prescelta. Una comparazione tra differenti piattaforme si può articolare:

1. sui soggetti/utenti → il più facile da concettualizzare ma il più ostico da seguire; in assenza di qualsiasi forma di interoperabilità standardizzata diventa difficoltoso effettuare un confronto in cui il comportamento del singolo su piattaforme differenti costituisca l'unità di analisi fondamentale.

Nel caso in cui fosse possibile tracciare la vita di determinati utenti su più piattaforme sarebbe necessaria una speciale cautela etica.

2. per giustapposizione tra piattaforme differenti → metodologicamente e praticamente il più semplice. L'unità di analisi è generalmente un fenomeno dai confini non eccessivamente definiti.

La strategia comparativa prescinde dall'analizzare i singoli utenti nelle piattaforme che compongono la comparazione e si limita a comparare dei valori aggregati. È possibile superare questo specifico stile di disegno di ricerca in due alternative:

- a. simmetrica → disegni di ricerca che separano il focus analitico più o meno in parti uguali fra le piattaforme;
 - b. asimmetrica → pur considerando differenti piattaforme, si focalizzeranno in maniera esplicita su una di queste lasciando sullo sfondo le altre. È rappresentata dalla natura specifica del caso a cui si dedica maggiore attenzione. Una buona pratica di disegno della ricerca richiederebbe di essere particolarmente espliciti da questo punto di vista
3. per comparazione 'innestata' → più ostico da inquadrare da un punto di vista metodologico in quanto utilizza l'interazione fra diverse piattaforme come sito di osservazione privilegiata: come unità di analisi è possibile adottare:
 - a. individui → rientrano tutti gli studi empirici sulla 'dieta mediale' digitale degli individui o alcuni studi sugli effetti della diffusione delle 'fake news'.
 - b. un determinato fenomeno digitale → rientrano tutti quegli studi che, vogliono provare come l'utilizzo di un determinato social influenzi le dinamiche d'uso di un secondo social 'generalista'. Le piattaforme non sono né 'comparate' né 'giustapposte', ma sono 'innestate'. Riconoscere gli effetti di interazione tra differenti contesti digitali.

Esperimento Caliendo

Il lavoro di Caliendo segue le pratiche digitali di 30 soggetti anziani a partire dal loro utilizzo dello smartphone. Lo smartphone rappresenta per sua stessa costituzione un mezzo dedicato alla fruizione di numerose piattaforme. A seconda del contesto utilizzando differenti app in maniera da definire differenti 'sfere di privacy' e modalità di interazione.

Comparazione innestata

Considerando la comparazione innestata, il lavoro di Burgess e Fernandez rappresenta un eccellente esempio di taxa: gli autori indagano in dipanarsi delle narrazioni sul #gamergate fra tre distinte piattaforme.

È interessante la strategia di cattura dati e analisi: a partire dai dati acquisiti da Twitter, riusciranno ad individuare una forte frequenza delle condivisioni su YouTube nel dataset, a partire da questa scoperta

andranno 'nel dettaglio' ad elaborare un video network a partire dai video originariamente condivisi su Twitter.

La visibilità dei differenti artefatti mediatici non corrisponde tra differenti piattaforme; infatti, video prominenti non necessariamente hanno la stessa importanza nel video-networks su YouTube.

Altre strategie possibili di 'comparazione innestata' prevedono studiare l'effetto dei media tradizionali sulla condivisione di una data notizia, oppure diventa possibile investigare come la produzione di contenuti da blog impatta il discorso pubblico su un determinato tema. Questo tipo di strategia di ricerca presuppone riconoscere l'ibridità del sistema dei media.

Codice → analisi sul vaccino COVID su Twitter e Reddit

Adotteremo la strategia di comparazione più semplice, la giustapposizione: considereremo infatti quelle che sono le menzioni dei vaccini per il COVID su Twitter e su Reddit.

Il dataset di Reddit, in particolare, è molto pesante, contiene circa 17 milioni di commenti ma i criteri di sampling non sono completamente trasparenti; il dataset di Twitter contiene 228207 tweet in lingua inglese raccolti a partire da un gruppo di parole chiave connesse ai vaccini per il COVID; si tratta di un sotto campione.

Per quanto riguarda Reddit si è deciso di utilizzare l'indice di Gini sullo score attribuito a ciascun commento, per quanto riguarda Twitter si è deciso di misurare la concentrazione dei retweet fra i differenti post.

Il processo di comparazione necessita costantemente di tenere sotto controllo le specifiche caratteristiche delle fonti dati.

Si è scelto di paragonare 'score' e 'retweets', assumendo che il secondo fosse una misura 'più forte' della reputazione. I valori dell'indice di Gini per le due grandezze nelle due rispettive piattaforme sono molto simili. Non sembra quindi esistere un 'effetto piattaforma'.

TextBlob incorpora anche un approccio 'rules based', il che implica che il modulo è in grado di riconoscere delle regole grammaticali di base ed utilizzarle per produrre un calcolo più accurato del tono emotivo di una frase. La capacità di TextBlob di riconoscere il sentiment di una frase è disponibile solo in inglese.

Oltre al sentiment TextBlob è in grado di stimare anche la 'soggettività' di un testo. Come il sentiment anche la soggettività è misurata su una scala da 0 a 1.

Tuttavia, il sentiment aggregato ha scarsa utilità, risulta più interessante declinare una sentiment analysis in corrispettiva di una serie di covariate.

Dal trend temporale a livello di piattaforma è possibile notare come i tweet siano lievemente più positivi dei messaggi su Reddit, esiste quindi un effetto piattaforma ma è di senso inverso rispetto a quello che ci saremmo attesi, poiché il consenso generale è che Twitter sia una piattaforma che facilita il conflitto e la creazione di 'flame'.

Su Twitter i vaccini con maggiore sentiment positivo sono Sinofarm e Covaxin ed Astrazeneca; Pfizer e Moderna hanno un sentiment inferiore.

In generale, i tweet negativi sembrano avere a che fare con le circostanze burocratico-amministrative che sostengono il processo di vaccinazione e i tweet positivi in generale con il processo di vaccinazione astratto da queste circostanze.

Le word cloud per Reddit sono meno definite, in generale sembra emergere tuttavia un focus sulle conseguenze immediate del vaccino. Le word cloud per piattaforma non producono nessun risultato apprezzabile.

Non tutte le piattaforme hanno delle API 'generose' nel restituire risultati di ricerca a partire da una query. Fuori dai temi più frequentati, anche trovare un dataset pre-catturato è un'eventualità altamente improbabile. Una volta acquisiti i dati è necessario poi operativizzare correttamente e assicurare che ci siano effettivamente delle variabili comparabili.

9. Il capitalismo della sorveglianza

La centralità delle reti di telecomunicazioni, rispetto al funzionamento del capitalismo contemporaneo, è chiara fin dai primordi della transizione digitale: secondo Sassen e Castells la capacità di mantenere reti di informazione in 'real time' costituisce una delle caratteristiche necessarie per assicurare la prosperità di un'economia basata sul settore terziario avanzato. Finanza, Immobiliare e Media necessitano sia di una costante capacità di confronto fra diversi attori.

A partire dai pattern di consumo, es. di serie televisive, diventa possibile progettare nuovi prodotti culturali che incontrino i gusti del pubblico.

Dopo la problematizzazione metodo-epistemologica seguirà un interesse specifico alle trasformazioni portate dalla 'data revolution' nel capitalismo contemporaneo.

Connettività globale

Nell'immaginario dei tecno-utopisti assistiamo ad un re-embedding dello scambio economico nelle relazioni sociali. Kostakis e Bauwens colgono una tensione problematica abilitata dalla connettività globale:

- le infrastrutture digitali promettono di disgregare il capitalismo contemporaneo in una struttura reticolare che connette micro-produttori, nessuno dei quali in grado di ottenere un sostanziale controllo sul mercato;
- la rete rischia di centralizzare da parte di alcune entità chiave favorendo l'evoluzione in quello che Bauwens e Kostakis chiamano 'neterchical capitalism'. È questa prospettiva a conoscere maggiore diffusione; mettono al centro l'enorme potere esercitato da un numero estremamente ridotto di attori dotati di potere di monopolio.

Piattaforme

Sappiamo che una piattaforma è un ente che svolge un lavoro di intermediazione in mercati caratterizzati da asimmetria informativa, fornendo informazioni su una transazione commerciale.

Questa specifica definizione poco ci dice sia di come la piattaforma possa garantirsi una posizione monopolistica sia di come si integri con il capitalismo finanziarizzato contemporaneo.

Per Langley e Leyshon la piattaforma:

- Offre un lavoro di curatela delle connessioni
- Si occupa di proporre la connessione 'adatta' ad un determinato utente
- Opera reperendo capitale finanziario a debito e mettendo 'a valore' le utilità di rete generate nel futuro

Produzione di valore nell'epoca delle piattaforme

In questo contesto si inserisce il lavoro di Shoshana Zuboff con lo scopo di chiarire come avviene la produzione di valore nell'epoca delle piattaforme. La traccia di dati prodotta sia 'volontariamente', come conseguenza dell'attività online, sia 'involontariamente' a causa dell'interazione con dispositivi digitali, produce una 'traccia' di dati che può essere estratta e valorizzata da alcuni attori.

Il dato acquisisce valore nel momento in cui è possibile metterlo in correlazione con altri dati grezzi.

Zuboff, analizzando la struttura e il funzionamento di Google, definisce la materia prima implicata in questo ciclo di valorizzazione come 'behavioral surplus': a partire dalle informazioni raccolte diventa così possibile prevedere il comportamento degli individui con un discreto grado di precisione. Il sistema economico che Zuboff analizza è 'capitalismo della sorveglianza' per la capacità di ricavare valore dall'osservazione di un determinato insieme di dati fra loro connessi. Le critiche che possono essere mosse a questo modello sono tre:

1. Teorico → Lo scopo non è quello di descrivere il funzionamento di una piattaforma digitale ma quello di registrare i mutamenti del capitalismo contemporaneo. La descrizione di Zuboff è eccessivamente 'entusiastica'.

Esiste un gap forte tra quello che le piattaforme permettono di fare e l'effettiva accuratezza delle predizioni.

Per quanto Google e affini possano avere a disposizione algoritmi che rappresentano lo stato dell'arte, difficilmente questo risolverà il problema sul breve termine.

La descrizione di Zuboff non considera in maniera approfondita la maniera in cui il capitalismo della sorveglianza è inserito in un determinato ordine geopolitico e dalla centralità del settore finanziario.

La propensione di un gruppo di utenti al consumare un prodotto stabilisce la base del calcolo di valore effettuato da Facebook; in funzione di questa abilità diventa possibile misurare i comportamenti umani.

2. Metodo → la rete degli individui connessa ad una determinata piattaforma cresce; cresce anche l'utilità dei futuri utenti e quindi la loro propensione ad unirsi alla rete.

Nella letteratura su questa specifica issue ci sono due posizioni:

- a. Teoria del valore lavoro → il surplus economico delle piattaforme sarebbe garantito dalla compressione del tempo libero a vantaggio del tempo di lavoro, il quale sarebbe garantito dalla 'compressione temporale' del capitalismo postindustriale. Questa prospettiva non riesce a fornire una via di uscita dal problema originale.

Proposta di Arvidsson: così come il valore di un titolo finanziario è definito dall'interesse che un gruppo di investitori potenziali conserva per quel dato prodotto, così il valore di qualcosa, in un ambiente digitale, deriva dalla proposizione di attenzione dedicata dagli utenti a quello specifico oggetto.

- b. Teoria del valore affettivo → Le affordances digitali forniscono un'occasione per calcolare l'interesse in un determinato ente o il coinvolgimento degli utenti in una determinata piattaforma. Fenomeno evidente sia con la presenza online dei brand commerciali sia con le criptovalute.

3. Empirico → scarsità di materiale empirico prodotto nella sua investigazione. I casi di ricerca empirica che si presentano sono spesso casi si studio svolti con metodi etnografici, analisi dei discorsi o survey data.

Resta un gap sia per quanto riguarda l'effettiva produzione di contributi empirici che analizzano il capitalismo della sorveglianza, sia, soprattutto, per quanto riguarda un'analisi che proceda utilizzando metodi 'nativi digitali'.

Codice → Analisi dei discorsi online

L'analisi dei discorsi online rappresenta un eccellente punto di partenza per cimentarci con una implementazione empirica del 'capitalismo della sorveglianza'.

Il primo passo è la cattura dati. A causa della chiusura delle API c'è stato un appoggio a Mediacloud. Oggi è relativamente più semplice operare su interi universi dati. Nonostante questo, la stragrande maggioranza dei progetti sarà portata avanti da piccoli team di ricerca con risorse minimali e budget estremamente ridotti. Si è deciso di operare secondo una logica intermedia: verranno acquisiti tutti i titoli di giornale in lingua italiana in un dato periodo.

L'acquisizione di dati dalle API social avviene per buona parte attraverso 'query'. È necessario selezionare delle parole chiave neutre che non sovra-rappresentino nessuna posizione ideologica, pena la perdita di una componente significativa dello spettro politico ed una descrizione mutilata del fenomeno.

È necessario pulire i dati da potenziali 'inquinanti'. Un successivo passaggio di pulizia può essere necessario per assicurarsi di avere articoli molto focalizzati sull'argomento.

Mediacloud seleziona contenuti a partire da archive.org, in questo modo può avere accesso anche a testo e metadata archiviati sul portale; questi contenuti non sono poi restituiti all'utente che interroga Mediacloud.

Il codice permette di misurare anche la distribuzione degli articoli sulle varie testate 'catturate'. Le successioni porzioni di codice si occuperanno di effettuare una sentiment analysis dei testi acquisiti tramite mediacloud e di declinare i risultati in funzione delle differenti categorie in cui possiamo suddividere il testo.

Il sentiment del testo è spesso negativo. Abbiamo scelto di estremizzare il valore della sentiment analysis, ignorando risultati potenzialmente 'neutri'. I termini positivi siano legati, principalmente, ad articoli promozionali che riportano occasioni, mentre invece i contenuti di natura negativa hanno più a che fare con la regolazione politico/economica del settore tech.

La word cloud per Amazon, tuttavia, riporta anche un discreto numero di termini collegabili alla dimensione della regolazione. Nel caso di Google i termini 'caratteristici' includono menzioni dei vari progetti dell'azienda o relativi alla govenance dell'azienda; per quanto riguarda Facebook e Microsoft prevalgono termini collegati alla govenance e regolazione.

All'interno di Gephi calcoliamo il grado di ciascun termine ed effettuiamo una community detection. Da Gephi sarà poi particolarmente semplice isolare le varie comunità per interpretare la coerenza interna. Possiamo notare come la comunità più grande in assoluto abbia principalmente a che fare con vendite ed occasioni commerciali, la seconda invece risulta essere connessa con il rilascio di contenuti digitali; la comunità 3 ha a che fare con i diritti dei lavoratori nel settore tech; la comunità 4 e 5 hanno a che fare con aspetti più specifici nella regolazione del settore.

Le piattaforme digitali costituiscono non tanto un 'mondo a parte' rispetto alla società 'offline' ma ne costituiscono la vera e propria spina dorsale tecno-economica.