counter-speech un’indagine sui contenuti che contrastano l’estremismo online

Jamie Bartlett
Alex Krasodomski-Jones

Ottobre 2015

DEMOS

Pubblicato da Demos, Ottobre 2015.

Unit 1, Lloyds Wharf
2-3 Mill Street
London
SE1 2BD

hello@demos.co.uk
www.demos.co.uk

Questo articolo è stato supportato da Facebook. Le opinioni espresse sono quelle degli autori e non rispecchiano necessariamente quelle di Facebook. Tutti gli errori e le omissioni sono da imputare agli autori.

INDICE

INTRODUZIONE

Circa 1,5 miliardi di persone utilizzano Facebook su scala globale. Sebbene la maggioranza lo usi a scopi positivi, ce ne sono alcune che usano questa piattaforma in modo negativo. Partendo da questo presupposto, Facebook ha creato delle politiche, ossia gli Standard della Comunità, che elencano quale tipo di contenuto possa essere condiviso, e quale no. Ad esempio, Facebook proibisce e rimuove i commenti che incitano all'odio, definendoli "contenuti che attaccano direttamente una persona (o un gruppo di persone) in virtu' di: razza, etnia, nazionalità di origine, affiliazione religiosa, orientamento sessuale, sesso, genere o identità di genere, disabilità o malattia". Nonostante i contenuti che incitano all'odio non siano permessi, a volte le persone condividono contenuti sgradevoli o inquietanti che, tuttavia, non violano le politiche di Facebook.

Al fine di contrastare questo genere di contenuti estremisti e spiacevoli, Facebook ha dichiarato pubblicamente che il *counter-speech* (N.d.T.: il replicare, il controbattere), insieme ai mezzi che la piattaforma prevede per incoraggiarlo, debba rivestire un ruolo importante. Facebook ritiene non solo che, potenzialmente, tale approccio possa essere più efficace per affrontare il problema, ma che possa anche essere quello vincente nel lungo periodo.

Il *counter-speech* è una risposta frequente ai contenuti che incitano all'odio o all'estremismo, affidata agli utenti del web (N.d.T.: *crowd-sourced*). I *post* (N.d.T.: messaggi) eccessivi si scontrano spesso con il pubblico disaccordo, la derisione e campagne a loro detrimento. Questo metodo per combattere l'estremismo ha i suoi vantaggi: è più veloce, più flessibile ed efficiente, capace di affrontare l'estremismo in ogni lingua e ovunque ci si trovi, mantenendo saldo il principio dello spazio pubblico aperto e libero per il dibattito. Ciò nonostante, le forme che il *counter-speech* assume sono varie quanto l'estremismo che cerca di mettere in discussione. Inoltre, è anche probabile che non sia sempre efficace quanto ci si aspetterebbe, senza contare che certi tipi di *counter-speech* potrebbero addirittura essere potenzialmente controproducenti.

Poiché Facebook crede fortemente nel potere del *counter-speech* e, in misura sempre crescente, vuole sviluppare un approccio rigoroso, basato sull'evidenza dei fatti, per comprendere meglio il fenomeno, ha chiesto a Demos di condurre una serie di ricerche che esaminino in che misura le diverse forme di *counter-speech* sono prodotte e condivise su Facebook.

Questa breve relazione provvisoria fornisce, in sintesi, le conclusioni cui si è giunti nella "Fase I", che si è concentrata sul modo in cui sono prodotti e condivisi gli interventi che criticano le pagine della destra populista in Europa. Le relazioni successive, di questa stessa serie, esaminano gli interventi e i contenuti che mettono in discussione l'ideologia islamica nel Regno Unito, e nel resto del mondo.

E' molto difficile creare una definizione oggettiva di "pagina che incita all'odio", e riteniamo che nessuna delle pagine incluse in questo studio incitino all'odio, o che il loro contenuto sia carico di odio. Per questo motivo abbiamo incentrato questo studio sulle pagine Facebook della destra populista che, spesso, sono accusate di essere degli spazi in cui una grande quantità di contenuti carichi di odio è postata o condivisa. Sebbene ci riferiamo a queste pagine definendole sempre "della destra populista", abbiamo scoperto che al loro interno è presente una grande varietà di contenuti che sono postati e condivisi. Nel definire le pagine del *counter-speech*, abbiamo provato a identificare chiaramente le pagine che hanno come scopo principale quello di controbattere o rispondere a tutto ciò che considerano gruppi, pagine o contenuti che incitano all'odio.

Riteniamo che sia importante mantenere il principio della libertà di internet, che dovrebbe essere uno spazio dove le persone sentono di poter esprimere la propria opinione apertamente e liberamente. Di conseguenza, crediamo che il dibattito, il disaccordo e il mettere in discussione siano di gran lunga preferibili alla censura e alla rimozione dei contenuti, anche quando questi sono eccessivi o radicali, qualunque sia la loro origine. Tuttavia, crediamo anche che questo fenomeno possa e debba essere verificato in modo empirico, per permetterci di comprenderlo meglio e per sapere come reagire. E' ciò che cercheremo di fare attraverso questa serie di ricerche.

METODO

Attraverso l'utilizzo dell'interfaccia pubblica di programmazione dell'applicazione (*Application Programming Interface – API*) abbiamo raccolto i *post* e i dati inerenti alle interazioni su 150 pagine pubbliche di Facebook, appartenenti alla destra populista e le relative pagine del *counter-speech,* localizzate in Gran Bretagna, Francia, Italia ed Ungheria. Abbiamo utilizzato "R", un *software open source* che permette ai ricercatori di accedere ai dati disponibili pubblicamente dell'interfaccia *API.*

Abbiamo evitato di raccogliere o utilizzare informazioni personali riguardanti singoli individui. Tantomeno abbiamo cercato di identificare le singole persone. Nel caso in cui un nome utente o un identificativo sia stato inavvertitamente raccolto, è stato cancellato in seguito. Non abbiamo raccolto dati riguardanti gruppi, profili personali, nascosti o segreti. Nel complesso sono stati utilizzati solo dati provenienti da pagine pubbliche e accessibili a chiunque. Al fine di proteggere ulteriormente la *privacy* dei singoli individui, non abbiamo citato o ripubblicato nessun *post* che potesse essere direttamente collegato a qualcuno in particolare.

Le pagine scelte sono state selezionate dai ricercatori di Demos, e, per questo motivo, non devono essere considerate come un campione esauriente ai fini della ricerca. Non esiste un modo ideale per identificare le pagine rilevanti, e va ribadito che questo è un progetto pilota. I successivi articoli accademici, appartenenti a questa serie, svilupperanno ulteriormente la metodologia.

Nell'arco temporale di due mesi (1 ottobre – 1 dicembre 2014) abbiamo raccolto 27.886 *post* condivisi sulle 150 pagine, localizzate in Gran Bretagna, Francia, Italia ed Ungheria (tuttavia, gran parte dell'analisi fa riferimento alla Gran Bretagna, Francia e Italia). Quando si parla di "*post*", ci si riferisce a un messaggio di forma testuale pubblicato dall'amministratore (o dagli amministratori) della pagina stessa. Oltre ai *post,* abbiamo raccolto tutte le interazioni collegate a questi ultimi. Per interazioni si intendono i "mi piace" (N.d.T.: *like*), le condivisioni ed i commenti relativi a quel post. I dati concernenti le interazioni possono essere utili per valutare il grado di diffusione del contenuto stesso, giacché ogniqualvolta un utente interagisce con un contenuto, questo appare nella cronologia (N.d.T.: *timeline*) dei propri amici (a seconda delle impostazioni sulla *privacy* che sono state applicate). In totale sono stati rilevati 8.4 milioni di interazioni.

Abbiamo sottoposto questi dati a una serie di analisi, che comprendevano: il calcolo del numero medio di interazioni, utilizzando i risultati automatici dell'*API;* la valutazione della struttura dei tipi di dati più popolari attraverso l'uso dei risultati

automatici dell'*API;* la valutazione del tipo e dello stile dei contenuti più popolari attraverso un'analisi manuale condotta da una persona; la valutazione di tipi di discorso affrontati sulle differenti pagine attraverso un'analisi manuale condotta da una persona; la valutazione del tipo di *network* (N.d.T.: rete, connessione) tra queste pagine, attraverso un'analisi automatica del *network*; la valutazione del modo in cui i diversi tipi di contenuto sono stati condivisi su differenti pagine rispetto ai *newsfeed* degli utenti, attraverso l'uso di un'analisi automatica.

E' importante ribadire che si tratta, in molti casi, di metodologie sperimentali. Non esistono *best practice* per raccogliere ed analizzare questo genere di dati, che siano state rigidamente definite. Inoltre, il presente studio è concepito come "esplorativo", ne consegue che i relativi risultati debbano essere letti con cautela.

RISULTATI DELLO STUDIO

Dati complessivi sulla dimensione e sulla portata dei contenuti che incitano all'odio e del *counter-speech*

Nel complesso, abbiamo identificato 124 pagine della destra populista nei quattro paesi, ma solo 26 pagine del *counter-speech* (per le ragioni precedentemente menzionate). Analogamente, c'erano molti più *post* sulle pagine della destra populista (25.522) che su quelle del *counter-speech* (2364). Non sorprende, dunque, che lì vi fossero anche molte più interazioni: sulle pagine della destra populista ce ne sono state in totale 7.8 milioni, mentre sulle pagine del *counter-speech* 546.000.

I dati includono molte più pagine della destra populista che del *counter-speech*, a causa del modo in cui sono stati raccolti. Questo non implica necessariamente che ce ne siano di più ma, più semplicemente, che le pagine del *counter-speech* sono più difficilmente identificabili. Di conseguenza, al fine di fornire un quadro più accurato della situazione, abbiamo calcolato la media dei *post* e delle interazioni per *post* in ogni paese e su ogni tipo di pagina. Si è costatato che, in generale, in Gran Bretagna le pagine del *counter-speech* sono meno e sono più limitate nella loro attività, tuttavia, esse raggiungono un numero maggiore di condivisioni e interazioni che le pagine della destra populista.

Tabella n.1: Numero medio d'interazioni per singolo *post* nei diversi paesi

Paese	*Counter-speech*	Destra populista
Francia	126	285
Italia	37	235
Gran Bretagna	402	325

Analisi della struttura del *network* e della *membership* e come ciò influenza la condivisione dei contenuti

Al fine di comprendere meglio le informazioni e il flusso di idee in queste pagine, abbiamo selezionato quelle della Gran Bretagna ed abbiamo analizzato fino a che punto le persone che commentavano in una delle pagine, commentassero anche in un'altra pagina. Questo dato è stato estrapolato dall'interfaccia *API*, attraverso "R", e visualizzato con Gephi, uno strumento *open source* per l'analisi dei *network*.

Per quanto riguarda le pagine della destra populista, ne abbiamo analizzate 92. In totale 54.495 singoli utenti hanno dato il loro contributo a queste pagine postando 159.437 commenti nell'intervallo temporale esaminato (si è giunti a questo risultato raccogliendo i commenti e calcolando il numero di identificativi dei singoli utenti

che hanno contribuito a questo insieme di dati). Emerge, così, che il 16,2 % degli utenti sono attivi su due o più pagine mentre l'1,3% è attivo su quattro o più pagine.

In seguito abbiamo analizzato 21 pagine del *counter-speech*. In totale 116.534 singoli utenti vi hanno contribuito attraverso le 135.842 interazioni reciproche. Nonostante abbiano più utenti, queste pagine hanno un'inferiore densità di *network:* il 12,7 % degli utenti è attivo su due o più pagine del *counter-speech*, mentre l'1,15 % è attivo su quattro o più pagine.

Questi dati suggeriscono che le pagine della destra populista con sede in Gran Bretagna hanno un *network* leggermente più concentrato: hanno meno utenti che, però, sono più attivi. Le pagine del *counter-speech* hanno degli utenti più attivi, i quali, però, singolarmente non producono altrettanti contenuti. Tuttavia, questo studio deve essere ripetuto su più ampia scala perché si possa giungere a dei risultati indicativi.

Il tipo di *post* che raggiunge più efficacemente un vasto pubblico

Al fine di determinare quale tipo di *post* abbia efficacemente raggiunto un vasto pubblico, abbiamo esaminato a) il tipo di *post* e b) il contenuto e il tono dei *post*. I dati dell'interfaccia *API* di Facebook permettono ai ricercatori di determinare il tipo di *post*: *link*, foto, status o video. La tabella 3 mostra quale tipo di *post* sia stato il più utilizzato dalle pagine in questione.

Tabella 2 Tipo di contenuto suddiviso per paese (percentuali arrotondate)

	Paese	Link	Foto	Status	Video
Counter-speech	Francia	64	15	14	6
	Ungheria	68	5	26	2
	Italia	29	45	17	9
	Gran Bretagna	40	34	18	8
Destra populista	Francia	26	43	17	14
	Ungheria	10	83	2	5
	Italy	59	27	9	5
	Gran Bretagna	57	23	10	10

Per quanto riguarda la modalità di condivisione dei contenuti, si rilevano notevoli differenze a seconda dei paesi e delle pagine. Ad esempio, in Francia, sulle pagine del *counter-speech* i più popolari sono stati i *link*, mentre in Italia sono state le foto.

Il motivo per cui un tipo di *post* è preferibile a un altro, presumibilmente, dipende da diversi fattori. Tuttavia, calcolando il numero medio di interazioni per ogni tipo di *post*, una delle conclusioni che si può trarre è che, in linea di massima, è più probabile che siano le foto a generare interazioni con gli utenti (vedi tabella 4, di seguito).

Tabella 3 Tipo di contenuto usato più frequentemente, suddiviso per paese (la maggioranza delle interazioni totali)

	Francia	Ungheria	Italia	Gran Bretagna
Destra populista	Foto	Video	Foto	Foto
Counter-speech	Foto	Foto	Foto	Foto

Analisi del contenuto dei *post*

Per comprendere in maniera più articolata quali siano il tipo e il tono di *post* più popolare, ne abbiamo analizzati 1000 selezionati in maniera casuale sulle pagine della destra populista (suddivisi tra Francia, Italia e Gran Bretagna). Abbiamo fatto altrettanto per quanto riguarda le pagine del *counter-speech* Abbiamo suddiviso questi *post* nelle due seguenti categorie: "contenuto" del *post* e "tono" del *post*. Queste categorie sono state selezionate dai ricercatori e sono descritte di seguito, attraverso degli esempi. Abbiamo valutato la popolarità di un *post* sulla base delle interazioni prodotte.

Ne è derivato che il contenuto più popolare dei *post* sulle pagine della destra populista è quello che noi descriviamo come commento (2524 interazioni in media).[1] Questi *post* hanno generato circa il doppio delle interazioni rispetto ad altri contenuti.

Per quanto riguarda i *post* sul *counter-speech*, il contenuto più popolare è rappresentato dalle domande (10934 interazioni in media, sebbene questo dato sia stato distorto da un ristretto numero di contenuti molto popolari). Il secondo contenuto più popolare dei *post* presenti sulle pagine del *counter-speech* sono i commenti, sebbene in Francia la categoria più popolare sia stata quella degli "attacchi".[2]

Sulle pagine della destra populista il tono più popolare dei *post* è stato quello "celebrativo", come nel caso di *post* in memoria delle vittime di guerra o dell'orgoglio patriottico (6607 interazioni in media).[3] Il contenuto rabbioso è stato il secondo più popolare (4093).[4]

Sulle pagine del *counter-speech* nei tre paesi, il tono dei *post* più frequente è stato quello spiritoso o ironico (2,717).[5]

Per analizzare al meglio il tipo di *post* più popolare, ne abbiamo esaminati 100 sulle pagine del *counter-speech* di natura satirica, poiché questo è stato il tipo di *post* a riscuotere più successo. Abbiamo fatto questo solo per quanto riguarda l'inglese.

Per quanto riguarda le interazioni, i temi dell'immigrazione, della razza e della religione ne hanno registrate, in media, più di 4000 e hanno rappresentato i tipi di contenuto più popolare. Il loro fine era, solitamente, quello di prendere in giro il linguaggio eccessivo utilizzato in riferimento ai suddetti temi sulle pagine incitanti all'odio. La richiesta di un reggimento dell'esercito di rimuovere da "*Britain First*", la maggiore pagina della destra populista della Gran Bretagna, un'immagine che lo celebrava, ha rappresentato il contenuto più popolare. Lo *screenshot* della richiesta ha ricevuto più di 9000 "mi piace".

Un'analisi del tipo e della natura dei commenti

Diversamente dai *post*, pubblicati dagli amministratori della pagina, i commenti ai *post* non sono soggetti a restrizioni e chiunque può pubblicarli. Abbiamo analizzato 1000 commenti riguardanti *post* sulle pagine della destra populista, scegliendo i commenti e le pagine casualmente (suddivisi tra Francia, Italia e Gran Bretagna). Abbiamo fatto altrettanto sulle pagine del *counter-speech*. Queste categorie sono state scelte dai ricercatori. Abbiamo determinato la popolarità di un commento sulla base del numero di "mi piace" ricevuti (questo è l'unico tipo di interazione disponibile in relazioni ai commenti).

Sulle pagine della destra populista, il 9% del totale dei commenti è stato classificato come *counter-speech*, ossia si trattava di commenti che dissentivano dal *post* o che proponevano un messaggio alternativo e positivo (il 17% in Francia, il 7% in Italia e il 5% in Gran Bretagna). L'estrapolazione dei dati suggerirebbe che, ogni mese, ci sono circa 25.000 commenti di *counter-speech* postati sulle pagine della destra populista nella sola Gran Bretagna.

Sulle pagine del *counter-speech*, il "*counter-speech* costruttivo" è stato il tipo di commento più efficace e popolare (una media di 5,9 "mi piace" per commento).[6] Immediatamente dopo, nella classifica, si trova la "discussione costruttiva" (una media di 5,3 "mi piace" per commento).[7] Al contrario, il "*counter-speech* non costruttivo" ha ricevuto mediamente 3,3 "mi piace" per commento, e i "*fact check*" 3,8 "mi piace" per comment.[8] Tuttavia, nonostante la popolarità del *counter-speech* costruttivo, esso ha rappresentato solo il 6% del totale dei contenuti, rispetto al 20% del *counter-speech* non costruttivo.

Abbiamo analizzato 100 commenti riguardanti "discussioni costruttive" fatti su pagine del *counter-speech* in inglese (anche in questo caso, le categorie sono state

selezionate dai ricercatori) per comprendere meglio quale sia il tipo specifico di commento più popolare. Ne è derivato che le "discussioni su argomenti di politica generale" hanno la maggior parte dei "mi piace" (13,2), seguite dalle "domande su dettagli riguardanti le politiche dei partiti/movimenti" (8,7). Ciò suggerirebbe che i commenti su temi politici specifici, su dettagli sulle politiche o quelli volti a chiedere informazioni, raggiungono efficacemente un vasto pubblico. Ciò nonostante, essi costituiscono una parte davvero piccola del numero totale dei commenti (rispettivamente il 6% e il 3%).

Analisi dei contenuti: dove e da chi sono condivisi e discussi

Abbiamo esaminato il modo in cui il contenuto si è diffuso dalle pagine prese in considerazione, ai *newsfeed* degli utenti. Abbiamo, inoltre, cercato di stabilire se esista una diversa modalità di interazione con tale contenuto, nel caso in cui esso sia visto da altri. Abbiamo analizzato le interazioni concernenti 4388 *post* di *counter-speech* e 75132 *post* della destra populista, calcolando la proporzione di *post* che è stata condivisa a partire dalla pagina originale o da un altro utente. Ciò si basa su un insieme di dati più grande rispetto a quello usato per le ricerche menzionate in precedenza: sebbene le pagine siano le stesse, il periodo di raccolta dei dati è stato più esteso. Questo insieme di dati è stato raccolto solo con l'ausilio del *Data Science Team* di Facebook, ed è costituito da dati storici pubblici aggregati non sperimentali.

Tabella 4 Gli spazi dove hanno luogo le interazioni relative ai *post*

	Post della destra populista			*Post* di *counter-speech*		
	% dalla pagina originaria	*% da parte di un altro utente*	*N=*	*% dalla pagina originaria*	*% da parte di un altro utente*	*N=*
Like	94	6	4,138,425	82	19	56,884
Condivisioni	97	3	474,558	99	1	13,430
Commenti	73	27	2,173,678	60	41	16,232

Questo suggerisce che, sul totale dei "mi piace", in percentuale, le pagine del *counter-speech* ricevono più "mi piace" e commenti per i *post* condivisi (N.d.T.: *reshares*), li ricevono, cioè, da persone che hanno interagito col contenuto sul *newsfeed* di un altro utente, anziché sulla pagina originaria (benché possano mettere "mi piace" anche alla pagina sulla quale il *post* è stato originariamente pubblicato). Ciò significa che il contenuto, potenzialmente, potrebbe diffondersi ulteriormente, se le persone lo condividessero di più con i propri amici. Inoltre, siamo fermamente convinti che possano esserci molti commenti / discussioni riguardanti *post* condivisi, ad esempio sui *newsfeed* delle altre persone, tra i propri amici. Crediamo che, probabilmente, questo sia il luogo dove avviene molto del *counter-speech*, anche se, per ragioni di riservatezza, non abbiamo raccolto questi dati.

Impiegando lo stesso metodo, abbiamo anche esaminato il modo in cui i diversi tipi di contenuto raggiungono gli utenti cui non piace la pagina dove questo era stato originariamente postato (in altre parole, questo significa che il contenuto si diffonde al di là degli utenti che hanno messo "mi piace" alla pagina). Abbiamo ottenuto questo risultato calcolando la percentuale di persone che hanno messo "mi piace" o hanno lasciato un commento ai *post* che esprimevano il mancato gradimento della pagina su cui il *post* era stato originariamente pubblicato.

Tabella 5 Grado di diffusione dei contenuti, al di là delle persone che mettono "mi piace" alle pagine

	% di *like* di persone cui non piace la pagina		% di commenti di persone cui non piace la pagina	
	Destra populista	*Counter-speech*	*Destra populista*	*Counter-speech*
Link	50	18	66	32
Foto	52	5	65	25
Video	68	26	75	48
Status	21	7	41	23

Ciò dimostra che le pagine della destra populista sono molto più efficaci nello scrivere contenuti che si diffondono al di fuori del proprio *network* di fan. Nel caso delle pagine del *counter-speech* (e di quelle della destra populista) i video sono il tipo di contenuto che raggiunge più efficacemente un ampio pubblico.

RACCOMANDAZIONI

Alla luce dei risultati sopramenzionati, si può affermare che le pagine del *counter-speech* non sono attive quanto quelle della destra populista. In Gran Bretagna hanno successo in termini di interazioni medie, mentre ne hanno meno in Francia ed in Italia. Per raggiungere un numero più ampio di persone, le pagine del *counter-speech* italiane e francesi dovrebbero produrre più contenuti. In Gran Bretagna, le pagine del *counter-speech* hanno più collaboratori che però interagiscono meno frequentemente dei collaboratori delle pagine della destra populista.

A quanti desiderino incoraggiare la diffusione del *counter-speech*, questo studio suggerisce che potrebbe essere d'aiuto un cambio di approccio. Nello specifico:

- Gli amministratori dovrebbero postare più foto e video in proporzione al totale dei contenuti prodotti. Dovrebbero essere contenuti capaci di diffondersi al di fuori del *network* di persone cui piace la pagina.
- Coloro che commentano dovrebbero preferire il "*counter-speech* costruttivo" al "*counter-speech* non costruttivo", e fare più commenti su temi politici specifici.
- I collaboratori delle pagine del *counter-speech* dovrebbero incoraggiare i propri amici a condividere più contenuti con i rispettivi amici.
- Nel complesso: se gli amministratori e gli utenti delle pagine del *counter-speech* fossero più attivi, e modificassero leggermente i propri contenuti, il grado di diffusione dei loro *post* aumenterebbe notevolmente.

POSSIBILI SOLUZIONI

Il *counter-speech* è un fenomeno più complesso di quanto possa sembrare. E' importante che il *counter-speech* sia considerato come molto di più del semplice dissentire o confrontarsi con un contenuto su una pagina pubblica. A volte esso è esplicito, come quando controbatte le opinioni nel momento in cui queste appaiono su un *newsfeed*, o nel caso in cui cerca attivamente il contenuto incitante all'odio criticandolo in maniera diretta. Altre risposte sono meno esplicite: alcune possono denunciare i discorsi incitanti all'odio, bloccare o disattivare l'utente, o manifestare il disaccordo in un messaggio privato. In altri casi, vengono creati gruppi spiritosi o gruppi seri che si contrappongono ad una pagina o ad un individuo. Su queste pagine, abbiamo trovato discussioni costruttive e dibattito. In altre occasioni, abbiamo riscontrato un *counter-speech* meno costruttivo, come ad esempio insulti e minacce aggressive.

Abbiamo chiaramente identificato: *counter-speech* costruttivo; *counter-speech* non construttivo, *fact checking*; discussion costruttive, pagine satiriche d'opposizione, pagine d'opposizione dal tono più serio. Alcuni di questi tipi di *counter-speech* andrebbero maggiormente incoraggiati rispetto agli altri tipi.

E' difficile determinare ciò che costituisce un risultato positivo, poiché ci sono tanti tipi differenti di *counter-speech*. In questo momento esistono opinioni divergenti sulle metriche essenziali, ma potrebbe essere utile individuarne tre tipi (riteniamo che possano tutti essere misurati, con diversi gradi di difficoltà).

Metriche quantitative

Di recente, l'analisi quantitativa dei contenuti dei *social media* è divenuta un *business* notevole. Centinaia di società hanno elaborato migliaia di strumenti che cercano di quantificare l'influenza dei *social media*. L'analisi accademica non è applicabile alla maggior parte di questo studio, ma molti dei suoi concetti centrali sono validi in questo caso:

- **La partecipazione è un'unità di misura dell'interazione dell'utente.** Ad esempio, potrebbe essere il rapporto tra gli utenti che hanno visitato la pagina e quelli che si sono registrati, o il numero degli utenti che hanno visualizzato un contenuto per poi condividerlo o mettere un "mi piace". Se da un lato si tratta di uno strumento inadeguato, dall'altro dà un valore indicativo del grado di diffusione potenziale. Siamo stati capaci di misurare questo aspetto in modo efficace.
- **Volume ed esposizione mediatica.** Quanti *post* sono stati prodotti e quanti singoli utenti stanno discutendo su uno specifico argomento? Ci sono picchi

quando a un dibattito partecipano più utenti del normale (N.d.T.: *discourse density*)? Ciò può essere valutato ma è necessario accedere a una quantità maggiore di traffico di dati, piuttosto che ai soli dati specifici riguardanti la pagina.

- **Il grado di diffusione misura la propagazione di una conversazione nata sui *social media*.** Qual è l'ampiezza del pubblico? I contenuti incitanti all'odio sono limitati a delle comunità isolate (dalle comunità stesse o dagli algoritmi delle personalizzazioni dei contenuti)? Quanto spesso raggiungono i *feed* degli utenti esterni a queste comunità? Il grado di diffusione può essere una metrica potente, ad esempio, se utilizzata in combinazione con la partecipazione. In questo studio abbiamo trovato alcuni criteri utili a tal riguardo.

Metriche del contenuto

Le metriche del contenuto tentano di valutare il tipo e la natura del contenuto piuttosto che il volume. Ciò, spesso, richiede un'analisi manuale o un'analisi sofisticata basata sull'esame del linguaggio naturale (N.d.T.: *Natural Language Processing - NLP*; sistemi capaci di elaborare il tipo di linguaggio che due interlocutori "umani" usano normalmente per comunicare), che sono entrambe possibili, ma con diversi gradi di accuratezza. Uno dei punti di forza di questi approcci è che possono essere estesi e applicati velocemente a un enorme volume di contenuti.

- La *sentiment analysis* (N.d.T.: si tratta di un metodo di analisi che raccoglie in tempo reale le reazioni degli utenti o trend davanti a un qualsiasi evento) è una tecnica che, una volta fissato il quadro analitico, ci permetterà di determinare se un contenuto può essere classificato come incitante all'odio, *counter-speech* o nessuno dei due. Potrebbe anche essere possibile valutare quanto un contenuto sia eccessivo analizzando il linguaggio utilizzato. Riteniamo che si tratti di un risultato difficilmente raggiungibile ma plausibile.
- Nello specifico, i contenuti incitanti all'odio e il *counter-speech* potrebbero essere classificati attraverso l'utilizzo del *NLP* in varie categorie minori: istigazione all'azione *off-line* (N.d.T.: nella realtà), discussioni genuine, insulti, etc… Riteniamo che anche questo sia un risultato difficilmente raggiungibile ma plausibile.
- Una domanda in sospeso è la misura in cui una discussione su Facebook si evolve o si risolve: il fatto di scrivere un contenuto di *counter-speech* ha un impatto sul resto della discussione? Riteniamo che sia difficile farlo utilizzando tecniche automatiche, ma che sia plausibile se fatto manualmente, da analisti in carne e ossa.

Metriche del mondo reale

Si cerca di valutare se il *counter-speech online*, a lungo termine, abbia un impatto sugli atteggiamenti e i comportamenti *off-line*. Questo è molto difficile da studiare con una certa precisione, così com'è difficile determinare se i contenuti incitanti all'odio *online* abbiano un effetto sugli atteggiamenti e i comportamenti degli utenti. Per determinare delle metriche ponderate e giustificabili in tal senso, sarebbero necessarie ulteriori ricerche

Questo studio esplorativo ha anche individuato alcune aree riguardo alle quali si potrebbe migliorare la nostra comprensione, grazie ad una piccola quantità di ricerche supplementari.

In primo luogo, è importante ideare un metodo più efficace per misurare il volume totale dei contenuti della destra populista e del *counter-speech*. E' molto difficile ottenere delle cifre globali affidabili relative ai diversi tipi di *counter-speech* senza avere un metodo rigoroso e scientifico per valutare oggettivamente le pagine individuate. Con un metodo di individuazione *ad hoc*, c'è il rischio di ottenere dati fuorvianti. Un approccio automatizzato basato sulle pagine più attive e più amate nei diversi paesi e tra le varie categorie, fornirebbe una misura molto più solida e affidabile.

In secondo luogo, ci sono nuovi tipi di contenuti che possono essere studiati nel dettaglio. Nella fattispecie si tratterebbe: del grado di diffusione dei diversi tipi di contenuto (piuttosto che concentrarsi solo sulle interazioni); dell'esame di interi *thread* di conversazioni relativi ad un *post* per valutare come esso cambia (esistono specifici gruppi o utenti che sono più capaci di altri di venire a capo di una discussione o di controbattere i contenuti incitanti all'odio?); dell'esame dei contenuti dei *post* condivisi sul *newsfeed* individuale (rispetto a quelli presenti sulla pagina stessa).

In terzo luogo, sarebbe utile per capire in che modo Facebook, in qualità di *network* di utenti, i quali condividono contenuti della destra populista e del *counter-speech*, può e dovrebbe ispirare il nostro metodo. Ciò comporterebbe l'esame della modalità di diffusione dei contenuti. Ad esempio, tracciando un singolo contenuto possiamo utilizzare l'analisi di *network* per controllare chi lo condivide, dove è stato trovato, e dove è andato a finire in seguito. Questo potrebbe aiutarci a rispondere ad alcune domande fondamentali su come si diffondono questi tipi di contenuti: quali percorsi fanno? Contenuti diversi viaggiano nella rete in maniera differente? Quale differenza esiste tra i diversi paesi?

In quarto luogo, ulteriori ricerche potrebbero contribuire a determinare una misura ponderata e solida del successo di questi contenuti nella vita reale. Per fornire una visione più pratica e particolareggiata, sarebbe necessaria una serie di casi studio dettagliati di campagne *online* che hanno avuto un forte impatto sulla vita reale (N.d.T.: *off-line impact*).

NOTES

[1] Commento" è un'etichetta omnicomprensiva applicata al contenuto che è creato o condiviso dalla pagina che commenta una situazione senza, però, citare necessariamente fonti esterne.

[2] Un attacco è un contenuto particolarmente aggressivo, solitamente indirizzato a uno specifico gruppo o a una singola persona. Un attacco può essere indirizzato a un gruppo etnico, un personaggio politico o un'organizzazione.

[3] Il contenuto celebrativo è quello che celebra la pagina o i suoi valori.

[4] Il contenuto rabbioso può impiegare linguaggio volgare e richiedere delle misure drastiche in relazione al contenuto.

[5] Ciò fa riferimento a tutte le battute e ai contenuti ironici.

[6] Qualunque tentativo fatto per avviare una discussione seria su temi specifici, relativi a contenuti incitanti all'odio (xenophobia, immigrazione).

[7] Ci si riferisce a qualunque tentativo fatto per avviare una discussione seria su temi inerenti alla politica, all'attualità e all'interesse generale.

[8] "*Counter-speech* non costruttivo" è tutto ciò che mette in discussione i contenuti incitanti all'odio, ma in un modo non costruttivo (ad esempio: attacchi personali, offese…). '*Fact check'* è il porre domande o fare verifiche su un avvenimento o su un'affermazione fatta.

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work,or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive,perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phonorecords of, display publicly,perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised.The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute,publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phonorecord of the Work You distribute, publicly display,publicly perform, or publicly digitally perform.You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder.You may not sublicence the Work.You must keep intact all notices that refer to this Licence and to the disclaimer of warranties.You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement.The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licencor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation.The exchange of the Work for other copyrighted works by means of digital filesharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

C If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works,You must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

A By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

B except as expressly stated in this licence or otherwise agreed in writing or required by applicable law,the work is licenced on an 'as is'basis,without warranties of any kind, either express or implied including,without limitation,any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental,consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

A This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence,however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

B Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

A Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

B If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

C No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

D This Licence constitutes the entire agreement between the parties with respect to the Work licensed here.There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You.This Licence may not be modified without the mutual written agreement of Demos and You.

Il counter-speech che mette in discussione, dissente, offre un punto di vista contrastante, è, potenzialmente, una maniera importante per affrontare i contenuti offensivi o eccessivi on line. E' veloce, flessibile ed efficiente, capace di affrontare l'estremismo in ogni lingua e ovunque ci si trovi, mantenendo saldo il principio di uno spazio pubblico aperto e libero per il dibattito.

Ciò nonostante, è anche probabile che non sia sempre efficace quanto ci si aspetterebbe, senza contare che certi tipi di counter-speech potrebbero addirittura essere virtualmente controproducenti.

Questa breve relazione provvisoria fornisce, in sintesi, le conclusioni di un nuovo studio che si è concentrato sul modo in cui gli interventi che criticano le pagine della destra populista in Europa sono prodotti e condivisi su Facebook. Essa esamina come gli utenti hanno interagito con 150 pagine Facebook e come abbiano condiviso il loro contenuto nell'arco temporale di due mesi.

Questa è la prima relazione di una serie di studi sul *counter-speech,* il cui scopo è quello di contrastare il contenuto eccessivo *online*. Le relazioni successive esamineranno l'ideologia islamica, nel Regno Unito e nel resto del mondo.

Jamie Bartlett è il Direttore del Centro per l'analisi dei Social Media presso Demos. Alex Krasodomski-Jones è un ricercatore del Centro per l'analisi dei Social Media presso Demos.

ISBN 978-1-909037-93-9