

Un approccio integrato tra Sentiment Analysis e Social Network Analysis nell'analisi della diffusione delle opinioni su Twitter

Francesco Santelli
Domenico De Stefano

ABSTRACT

In this work, we reconstruct the tweet-retweet and tweet-reply relations of opinions about a trending topic on the Twitter platform. We propose a multi-steps approach to derive a signed network expressing the spread of contents and opinions. The first step consists in reducing data dimensionality by means of a clustering procedure on tweets able to identify the concepts they convey. In the second step, focusing on message contents, we adapt different sentiment analysis algorithms in order to determine the sign of both the original tweet (with respect to the trending topic) and the sign of the edge connecting the original tweet to the replies, conditional on the replied tweet. Each tweet will spread its concepts by means of signed retweet and reply relations. The aim is to study the different structure, in terms of both network structure and sentiment, of the signed network related to each concept. A comparative analysis will be possible as well among the various identified signed networks

RIASSUNTO

In questo lavoro ricostruiamo le relazioni tweet-retweet e tweet-reply delle opinioni su uno specifico tema di riferimento sulla piattaforma social Twitter. Proponiamo un approccio costituito da più fasi per derivare una rete “segnata” che esprima il modo in cui si diffondono contenuti e opinioni su tale piattaforma. Il primo passo della procedura consiste nel ridurre la dimensionalità dei dati attraverso il clustering (suddivisione in gruppi) dei tweet originali, in grado di rappresentare ogni gruppo di tweet attraverso il “concetto” principale che esprime. Nella seconda fase, concentrandoci sui contenuti del

messaggio veicolati dal tweet, vengono utilizzati strumenti propri della sentiment analysis in modo da determinare il segno (positivo, neutro o negativo) sia del tweet originale (rispetto al trending topic, o tema di riferimento) sia il segno del legame (link) che collega il tweet originale ai replies (le risposte al tweet). Ogni tweet diffonderà i suoi concetti originali tramite re-tweet segnati e relazioni con i replies relativi. L'obiettivo è studiare la diversa struttura, in termini sia di dinamiche di rete che di sentiment, della rete segnata relativa a ciascun concetto (derivato da ogni cluster).

KEYWORDS

Twitter, Sentiment Analysis, Signed networks, Tweets Clustering, Opinion Spread

PAROLE CHIAVE

Twitter, Sentiment Analysis, Reti segnate, Clustering, Diffusione delle opinioni

PROFILO BIOGRAFICO

Francesco Santelli attualmente è assegnista di ricerca, presso la Federico II, all'interno di un progetto PRIN incentrato sulla mobilità universitaria: "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide". Statistico di formazione (laurea magistrale in Scienze Statistiche), ha svolto esperienze di visiting presso la KU Leuven (Belgio), ha conseguito il dottorato in scienze statistiche e sociali alla Federico II di Napoli e ha svolto in seguito un periodo di ricerca all'Università di Trieste sull'analisi di dati provenienti da Social Network, in collaborazione con l'azienda di analisi e ricerche di mercato SWG con un progetto co-finanziato dalla regione F.V.G. Tra i suoi interessi figurano anche la bio-statistica e le indagini di ricerca di tipo clinico, il Text Mining e analisi dai testuali provenienti da fonti Social Media, tecniche di Unsupervised Learning come l'Analisi degli Archetipi e lo studio di fenomeni sociali legati ai concetti di mobilità e migrazione. È autore di oltre 20 pubblicazioni e numerose partecipazioni a convegni nazionali ed internazionali, e ha svolto attività di docenza a contratto, anche con titolarità di cattedra, in diversi atenei italiani. È inoltre membro dell'editorial board della rivista scientifica "Fuori Luogo".

Domenico De Stefano è Professore Ordinario di Statistica Sociale presso il Dipartimento di Scienze Politiche e Sociali dell'Università di Trieste, Italia. I suoi interessi di ricerca si concentrano principalmente sull'analisi dei social network e sulla modellazione statistica, con particolare attenzione allo sviluppo di nuovi strumenti metodologici per l'analisi di dati relazionali. Tra le sue più recenti pubblicazioni: 'Density-based clustering of social networks', *Journal of Royal Statistical Society series A*, 2022 (con G. Menardi); 'Pre-electoral polls variability: a hierarchical Bayesian model to assess the role of house effects with application to Italian elections', *Annals of Applied Statistics*, 2022 (con F. Pauli, N. Torelli); 'Community structure in co-authorship networks: the case of Italian statisticians', in: Greselin F., Deldossi L., Bagnato L., Vichi M. (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer (con S. Zaccarin e M.P. Vitale).

Negli Online Social Media Data (OSMD) un ruolo chiave è svolto dalla piattaforma Twitter, soprattutto per quanto riguarda l'analisi delle opinioni (Onorati e Diaz, 2016). Risulta piuttosto interessante, quindi, studiare le dinamiche, i tempi e i modi con cui tali opinioni si diffondono nella catena degli utenti della rete Twitter.

In questo quadro di analisi, i retweet si pongono come un elemento imprescindibile per comprendere il meccanismo e la dinamica di tale flusso di opinioni (Rossi e Magnani 2012; Suh et al. 2010).

Il retweet è, infatti, una delle proxy più adottate in letteratura per approfondire i meccanismi della catena che porta un messaggio veicolato da un tweet da un utente all'altro (Stieglitz e Dang-Xuan, 2012). Tale catena è fortemente condizionata dal livello di "influenza" dell'utente che ha scritto il tweet originale. È probabile che le opinioni condivise da utenti con un numero molto elevato di followers vengano diffuse in una vasta rete di retweeting. Tale concetto si lega spesso alla dimensione dell'*engagement*, per cui l'obiettivo di taluni utenti è far sì che i propri tweets siano re-twittrati il numero maggiore di volte e che le loro opinioni possano diffondersi velocemente sulla piattaforma.

Tuttavia, anche altri tipi di interazioni, come menzionare un utente o rispondere (reply) in riferimento a un tweet, possono in un certo senso innestare una trasmissione di informazioni o opinioni su un argomento specifico (Kwak et al., 2010).

La rilevanza delle risposte (replies) rispetto a ciò che accade nei retweet, e quindi la peculiarità dei replies nella viralità e nei meccanismi di diffusione delle opinioni sui social media non è stata ancora così ampiamente affrontata (Kim e Jaebong, 2012; Sousa et al., 2010). Esistono tuttavia contributi volti ad analizzare le relazioni tra strategie di comunicazione politica e uso di retweet e replies (Kim e Jaebong, 2012).

Un reply, nello specifico, si ottiene quando un utente risponde al tweet originale, lasciando la propria opinione personale come commento. Mentre per il retweet il contenuto del messaggio è solitamente fisso e non viene aggiunto altro testo ma solo ricondiviso il messaggio originale, per quanto riguarda i reply sussiste relazione semantica specifica, di tipo condizionale, tra la risposta e il concetto originario espresso nel tweet. In altre parole, non è possibile (o quantomeno non è utile) analizzare il contenuto o il "sentiment" del reply svincolandolo del tutto dal tweet originario. In tal senso, è da qui che nasce lo spirito

principale di questo contributo, dato che l'interpretazione del significato semantico condizionato è una questione, anche dal punto di vista statistico e dell'analisi delle reti, tutt'altro che banale. D'altronde, a volte gli utenti rispondono per esprimere il loro disaccordo con il significato originale del tweet. Essi possono però anche rispondere per legittimare il concetto/argomento/opinione del tweet. Le risposte che possono essere considerate completamente neutre non sono così comuni, ma talvolta è possibile che ciò accada. In mezzo a questi estremi, ci sono ovviamente una moltitudine di sfumature.

In questo contributo presentiamo e ricostruiamo le relazioni tweet-retweet e tweet-reply delle opinioni su uno specifico argomento di riferimento o trending topic, utilizzando come base dati le interazioni sulla piattaforma Twitter, in particolar modo soffermandoci sul valore aggiunto delle relazioni "social" in termini di contenuto semantico (Saif et al., 2012).

La prospettiva da noi adottata in questo contesto è basata sui messaggi (tweet originali e reply) e sul loro contenuto, e viene tralasciato invece la dimensione riguardante gli utenti e le loro caratteristiche. Focalizzandoci sui contenuti del messaggio, nell'analisi della diffusione delle opinioni, adattiamo diversi algoritmi di sentiment analysis (Agarwal et al., 2011) in modo da determinare sia il segno del tweet originale (rispetto al topic di riferimento, e cioè il livello di agreement/disagreement del tweet originario) sia il segno del legame che collega il tweet originale ai reply, condizionando tale segno al concetto di contenuto semantico del tweet a cui si fa riferimento.

Per i dati di Twitter, sono state sviluppate diverse procedure e algoritmi per analizzare dal punto di vista del sentiment (Go et al., 2009) una base di dati testuali, e in questo contributo associamo il sentiment relativo a ciascun reply in modo che esso dipenda in primo luogo dal topic di riferimento, ma anche dal concetto espresso nel tweet iniziale. In questo approccio, il livello di "influenza" dell'utente originario può essere considerato come una sorta di variabile di controllo, ma questo ulteriore sviluppo è lasciato a lavori successivi e verrà tralasciato a questo stadio della procedura.

L'obiettivo è trovare un modello semantico nel meccanismo di retweeting/risposta che sintetizzi, in termini statistici, la diffusione di alcune determinate opinioni. Infatti, verranno valutati il numero di retweet e anche il tipo di reply che il tweet originario ha prodotto nel periodo di tempo considerato.

Il lavoro è suddiviso come segue: nella successiva sezione si elencheranno le varie interazioni che gli utenti su un dato social network

possono effettuare per diffondere o rinforzare determinate opinioni; successivamente saranno illustrate gli step della procedura discutendo la fase di riduzione della dimensionalità dei dati e il ruolo della sentiment analysis; sarà mostrato poi l'utilizzo della procedura per l'analisi di un caso studio di interesse locale e nazionale e infine il contributo si chiude con delle osservazioni conclusive.

IMPORTANZA DELLE INTERAZIONI TWEETS, RETWEETS E REPLIES NELLA DIFFUSIONE DELLE OPINIONI

Nell'analizzare la diffusione delle opinioni adottiamo un'analisi "message based" volta a ricostruire le relazioni tweet-retweet e tweet-reply delle opinioni su un trending topic sulla piattaforma Twitter implementando alcuni elementi relativi all'analisi del campo semantico dei tweet (Saif et al., 2012).

Consideriamo, come caso di studio, il dibattito online sull'istituzione della flat-tax in Italia nel avvenuto nel corso del 2019. In questo contesto, le interazioni sia nel caso di retweet che di reply sono cruciali nella comprensione odierna dei fenomeni virali, a causa dell'influenza di Twitter sulle opinioni comuni. In sostanza, lo studio di ciò che avviene all'interno della piattaforma non rimane necessariamente all'interno di essa, ma ha ricadute verificabili anche nel dibattito al di fuori della comunità social.

Come sottolineato in alcuni lavori (Murthy, 2018), la comunicazione moderna via social è allo stesso tempo individualistica e comunitaria. Essa è in grado a dare agli utenti una possibilità di esprimere il proprio pensiero individuale attraverso un post ad esempio, impegnandosi però allo stesso tempo in un'attività comunitaria e in dinamiche di gruppo. In questa prospettiva, lo sviluppo di un approccio integrato, semantico per il livello "individuale" e analisi delle reti sociali per la dimensione "comunitaria", è la sfida della procedura proposta, e cioè tenere contemporaneamente in considerazione i due aspetti delle interazioni proprie di Twitter.

In particolare, proponiamo di utilizzare strumenti di analisi sia delle reti sociali che della sentiment analysis, organizzandoli in una procedura multi-step per derivare e analizzare le reti segnate (*signed networks*), ovvero strutture relazionali legate alla diffusione di contenuti e opinioni. Tali reti derivano da due azioni principali: retwittare e rispondere (effettuare reply) rispetto ad un tweet originario che veicola un certo "sentiment" rispetto ad un dato tema.

L'analisi del sentiment determina il segno di ogni collegamento tra il tweet originario e altri contenuti scaturiti da interazioni social.

Il primo passo consiste nel ridurre la dimensionalità dei dati attraverso una procedura di clustering sui tweet in grado di identificare i concetti sottostanti che essi esprimono, in modo da poterli raggruppare in sotto-insieme ragionevolmente coerenti e coesi. In particolare, a tal fine, utilizziamo un algoritmo di *Community Detection* (Fortunato, 2010) denominato *fastgreedy* (Clauset et al., 2004)

Nella seconda fase, concentrandoci sui contenuti semantici del messaggio, si propone una analisi del sentiment nel senso di individuare il segno del concetto espresso del tweet originale (rispetto al trending topic), ma soprattutto il segno del legame che collega il tweet originale ai reply, subordinando tale analisi semantica a ciò che era espresso nel tweet replicato.

Ciascun tweet diffonderà i concetti da esso espressi attraverso le interazioni citate in precedenza (retweet e reply) in un modo descritto dalla configurazione della rete segnata derivata dalle relazioni tweet-retweet e tweet-reply.

Queste reti segnate possono assumere diverse configurazioni strutturali legate alle modalità di diffusione di ciascun concetto. Ciò consente di effettuare un'analisi comparativa tra le varie reti di quelle che definiamo "concept-signed network" riportate nel caso studio specifico relativo al dibattito sull'istituzione della flat-tax.

RIDUZIONE DELLA DIMENSIONALITÀ DEI TWEET

L'approccio per ridurre la dimensionalità e la complessità del corpus dei tweet originali è composto da diverse fasi. Il primo passo è stato quello di selezionare alcuni hashtag relativi a un argomento specifico e poi in seguito, utilizzando questi hashtag come queries, raccogliere tweet e i retweet, ottenendo un corpus omogeneo in quanto gli hashtag identificano in maniera piuttosto soddisfacente i topic trattati nei tweets.

A partire dal corpus di tweets, viene estratta una rete di hashtag (*hashtags network*), organizzandoli in una matrice di adiacenza simmetrica "hashtag X hashtags". Tale rete è pesata (numero di volte in cui gli hashtag si verificano nello stesso tweet o retweet, quindi le co-occorrenze degli hashtags fungono da peso), non orientata in quanto non esiste una direzione da un hashtag all'altro ma essi sono in relazione bi-univoca, ed essa è simmetrica in quanto le righe e le colonne si equivalgono. In questa rete così definita, viene eseguito un algoritmo di community detection

detto fastgreedy e vengono rilevati così dei cluster di hashtag altamente connessi al loro interno, mentre sono meno legati (co-occorrenze più basse) con hashtag al di fuori della loro community di provenienza.

Ogni gruppo di hashtag definisce quello che noi definiamo essere un “concetto”, o almeno un argomento che emerge con chiarezza rispetto agli hashtag originari che definivano il campo di analisi. Tweet e retweet sono quindi contrassegnati dai concetti che contengono al loro interno. Alcuni tweet includeranno un solo concetto, ma alcuni ne includeranno diversi. In letteratura è stata proposta una procedura di clustering gerarchico applicata ai tweet una volta opportunamente descritti dai concetti (Balbi et al., 2018). In questo lavoro, dato che la procedura automatica è abbinata ad alcune speculazioni qualitative, abbiamo deciso di assegnare ad ogni tweet un solo concetto, così come definito da un gruppo di hashtag, se nel tweet sono inclusi un numero elevato di hashtag appartenenti ad una data community. In questa fase si passa alla riduzione della dimensionalità del corpus originario: da un numero molto e elevato di N di tweet a un numero di k di cluster di tweet descritti da concetti, con $k < N$.

In seguito, nella procedura viene desunta la rete di tweet-retweet e tweet-reply. I tweet originali apparterranno a cluster diversi e ogni cluster sarà denotato, come detto, dal proprio concetto di riferimento. Verrà mostrato che diversi gruppi di tweet presenteranno strutture dissimili in termini di retweet e risposte, e che il concetto espresso inizialmente sarà cruciale nel prevedere le interazioni social scaturenti.

IL RUOLO DELLA SENTIMENT ANALYSIS

Il passo successivo consiste nel determinare, a partire dal campo semantico rilevato dei concetti allegati ai tweet originali, la rete segnata derivata dal sentiment (condizionale) incorporato sia nei retweet che nei reply. In sostanza, bisogna assegnare una valutazione sul “segno” delle interazioni social. Dalla fase di riduzione della dimensionalità abbiamo ottenuto comunità (concetti) di hashtag incorporati nei tweet originali. Gli hashtag, uguali ai tweet originali o anche diversi, possono apparire anche nelle risposte. Gli hashtag relativi ai tweet originali possono avere, infatti, un significato diverso rispetto a quanto accade nel contesto di eventuali reply. Tuttavia, per la valutazione del sentiment, utilizzeremo tutti i singoli hashtag (indipendentemente da dove appaiono) a cui assegneremo un segno positivo (+), neutro (\emptyset) o negativo (-). Pertanto ogni hashtag avrà un segno.

Ogni concetto, essendo composto da più hashtag, avrà un punteggio complessivo dato dalla somma algebrica di tutti i propri hashtag; quindi ogni tweet appartiene a un concetto (o a un gruppo di concetti) in un'ottica simile a quella delle procedure di clustering. Analogamente, una simile procedura per identificare gruppi di risposte sarà effettuata considerando gli hashtag.

Quindi, coerentemente con la logica dell'obiettivo di riduzione della dimensionalità, l'analisi del sentiment condizionale che porta alla definizione di una rete segnata non verrà eseguita utilizzando ciascun tweet come elemento singolo, ma verrà effettuata utilizzando invece la relazione concetto originale espresso nel tweet-concetto espresso nel reply. Dunque, dopo aver classificato i tweet originali, grazie ai gruppi di hashtag è possibile identificare concetti anche nei reply del tweet originale. Essi potranno avere una attitudine positiva, negativa o neutra nei confronti del concetto veicolato nel tweet originale.

Come è di tutta evidenza, le relazioni tweet-retweet diffonderanno invece gli stessi concetti incorporati nel tweet originale (che, come detto, può essere una accezione positiva, neutra o negativa rispetto all'argomento considerato). Mentre, nel caso dei reply, il segno (sentiment) non sarà necessariamente lo stesso del concetto originale del tweet replicato. È infatti abbastanza comune esprimere un forte disaccordo all'interno di una risposta; anzi, tendenzialmente è più comune esprimere disaccordo, dato che è possibile ricondividere un concetto tramite un retweet piuttosto che rafforzarne il sentiment con un reply.

Pertanto, la rete segnata risultante includerà sia i retweet che le risposte: ogni retweet manterrà il segno originale del tweet, mentre il sentimento condizionale determinerà il segno della risposta, discretizzato nelle possibili modalità di accordo (+), disaccordo (-) o atteggiamento neutrale.

Otterremo, alla fine della procedura, un numero di reti segnate per ogni tweet che trasmette un dato concetto o gruppo di concetti; ogni rete è una sorta di catena di tweet-retweet-risposta, un albero in cui il concetto espresso dal tweet originario viene trasmesso di ramo in ramo, dove ogni ramo è l'interazione social di riferimento. Un ulteriore passo consisterà nella descrizione e confronto delle diverse strutture di rete segnate ottenute in questa fase. Ogni struttura di rete incorpora in sé le informazioni e la struttura di come si diffondono concetti, siano essi positivi, negativi o neutri.

Per dare un'idea di come funzioni nella sua interezza la procedura proposta, verrà presentata un'applicazione utilizzando dati reali. Verrà approfondito un tema ampiamente dibattuto in Italia tra i mesi di aprile e maggio 2019, ovvero il tema della possibile istituzione di un sistema di tassazione denominato flat-tax.

Un sistema di questo tipo, in sintesi, applica la stessa aliquota fiscale a tutti i cittadini indipendentemente dalla fascia di reddito. Il governo italiano si è posto come obiettivo politico quello di riuscire a stabilire tale sistema fiscale, principalmente sotto la pressione della Lega Nord e del suo leader politico Matteo Salvini, che hanno da tempo espresso di sostenere con forza tale proposta, al punto da farne un cavallo di battaglia in campagna elettorale. Il sistema ovviamente non è facilmente applicabile a causa del suo costo economico per la spesa pubblica, e questo è uno dei motivi principali per cui c'è un forte dibattito pubblico intorno ad esso. In teoria, poter abbassare le tasse trova l'opinione favorevole di gran parte dei cittadini (utenti): tuttavia, il costo in termini di spesa pubblica resta un tema dibattuto e divisivo. Esso è, probabilmente insieme al Reddito di Cittadinanza proposto principalmente dal Movimento 5 Stelle, il tema politico-economico più divisivo in Italia negli ultimi mesi. Per poter collocare correttamente nel tempo tale dibattito, sia nell'opinione pubblica che nei social media online, si tenga presente che nel Documento di Economia e Finanza del 9 aprile 2019, il sistema di tipo flat-tax stato in qualche modo introdotto ufficialmente nel sistema italiano, anche se non nella misura del tutto prevista da Matteo Salvini. Per questo, la maggior parte dei giornali ha scritto che questa flat-tax è ora nella sua "fase embrionale" e che probabilmente, se le condizioni politiche ed economiche lo avessero consentito, sarebbe stata introdotta gradualmente e lentamente nel sistema in futuro in modalità più estese.

Per poter costruire una base dati relativa a questo fenomeno di carattere sociale, i tweet vengono raccolti utilizzando come chiave (query) l'hashtag #flattax: nell'analisi vengono inclusi solo i tweet contenenti questo hashtag. La finestra temporale del caso studio qui illustrati è stata circoscritta alle prime ore del mattino del 1 maggio 2019 fino al pomeriggio del 13 maggio 2019. I tweet che includono l'hashtag #flattax in questo intervallo vengono raccolti tramite sviluppi (insieme di funzioni A.P.I.) in ambiente R versione 3.6., in grado di utilizzare le

potenzialità di Twitter dal lato sviluppatore (developer) e di poter quindi, attraverso il pacchetto “rtweet”, ottenere una matrice dati coerente con gli obiettivi della ricerca (Kearney, 2018). Tali *app*, gratuite, sfruttano le potenzialità dell'utilizzo delle API ma al tempo stesso le relative limitazioni: in particolar modo, non è possibile recuperare tweet risalenti a una settimana precedente l'interrogazione, e i dati raccolti non sono un archivio completo ma solo un campione casuale. Di tale campione non sono note tutte le informazioni riguardanti la selezione. Inoltre, solo una piccola percentuale dei risultati sono tweet e/o reply, mentre la stragrande maggioranza sono retweet.

Visto lo scopo di questo lavoro, che è anche incentrato sulla relazione tweet-reply, abbiamo deciso di ‘forzare’ l'API gratuita a restituire più risposte utilizzando la seguente procedura: dati i tweet originali, abbiamo estratto l'id utente associato a ogni risposta, e nelle query successive abbiamo filtrato i nuovi reply associandoli all'ID del tweet originale, ottenendo così risposte solo a quei tweet già nei dati ottenuti come risultato della prima query.

Il numero totale di tweet nel *corpus* è 6006, inclusi retweet e reply. I reply sono 360 e i tweet originali sono 438. Come accennato, le API gratuite collegate a software open source come R Project di solito consentono di ottenere un numero di retweet molto alto rispetto alla mole complessiva dei dati. La struttura dei dati è evidenziata nella Figura 1, dove si può notare che alcuni punti (nodi) che occupano una posizione centrale all'interno della rete sono circondati da diversi (anche migliaia) di punti verdi, che è il colore usato per identificare i retweet. Significa che questa è l'interazione più comune, portando i retweet a svolgere un ruolo chiave e diretto nella diffusione delle informazioni (così come evidenziato da altri autori, vedi Stieglitz e Dang-Xuan, 2012). Quando gli utenti retwittano un post su twitter, essi esprimono in modo diretto e semplice il loro accordo con il contenuto originale; quindi, una volta analizzato il concetto espresso nel tweet di partenza, non sono necessarie ulteriori analisi per ottenere una “stima” del sentiment espresso nei relativi retweet.

Da questo momento in poi, l'hashtag #flattax verrà eliminato da ogni analisi. È la chiave (query) che abbiamo utilizzato per interrogare il software ed ottenere un corpus, quindi essendo presente per definizione in tutti i tweet abbiamo deciso di rimuovere tale hashtag.

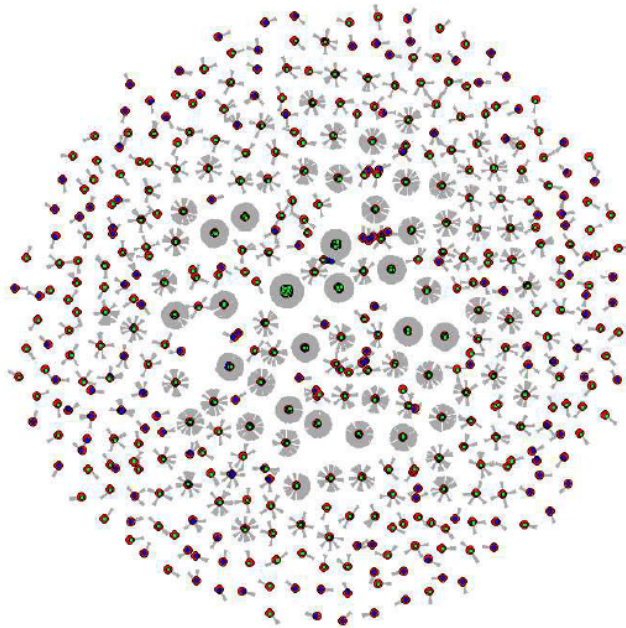


Figura 1: struttura di rete complessiva nei dati. Punti rossi sono tweet originali, punti blu sono reply e i punti più piccoli verdi sono i retweet. Maggiore è l'alone grigio intorno ad ogni tweet, maggiore è il numero di interazioni ottenute.

Per quanto riguarda gli hashtag, escludendoli, le prime 26 occorrenze sono evidenziate nella Tabella. I primi cinque sono legati, sostanzialmente, al mondo della Lega Nord. Il leader politico è Salvini, Siri fa riferimento ad Armando Siri, sottosegretario al ministero dei Trasporti.

Rank	Hashtag	N	Rank	Hashtag	N
1	salvini	54	14	salariminimo	11
2	siri	51	15	isf	10
3	donato	46	16	dimarted	9
4	albania	43	17	europree2019	9
5	lega	41	18	festadellavoratori	9
6	dimaio	22	19	governodelfallimento	9
7	facciamorete	22	20	macron	9
8	1maggio	21	21	primomaggio	9
9	m5s	19	22	apl	8
10	dimartedi	15	23	csg	8
11	redditicittadinanza	15	24	festadellavoro	8
12	calenda	14	25	governo	8
13	quota100	11	26	retraite	8

Tabella 1: top 26 hashtags nei tweets originali e loro frequenza

Armando Siri è stato posto sotto inchiesta in un'indagine per corruzione. Il dibattito pubblico sulla sua indagine e sul suo ruolo politico è stato intenso nei primi giorni di maggio; in particolare c'è stata una forte e aspra dialettica politica tra Movimento 5 Stelle e Lega Nord sul suo ruolo politico; in particolar modo è stato Luigi Di Maio, Leader del Movimento 5 Stelle, a chiedere a Siri di dimettersi "per motivi etici e politici". Donato fa riferimento a Francesca Donato, leghista e fondatrice di Eurexit Project, movimento culturale e politico che mira a "ritirarsi dall'euro e ripristinare la democrazia in Italia". Il 1 di maggio, durante un dibattito pubblico televisivo svoltosi sul canale LA7, "Di Martedì", ha apertamente affrontato con toni forti il tema della flat tax e della sua istituzione in Albania con Carlo Calenda, rappresentante all'epoca del Partito Democratico. Carlo Calenda risulta infatti nella lista degli hashtag più popolari, così come quello della trasmissione Di Martedì. Ciò significa che i dibattiti sui social media come Twitter sono strettamente legati alle notizie degli ultimi giorni e a ciò che accade sulle tv nazionali. Se c'è un dibattito pubblico che si trasforma in una discussione aperta e franca nei salotti televisivi, è probabile che le persone siano più disposte a postare, ritwittare ed effettuare reply su ciò che hanno appena visto, piuttosto che soffermarsi sulle teorie sottostanti e sui documenti ufficiali.

Vale la pena notare che c'è un impatto molto diverso tra i due partner della coalizione di governo dell'epoca, Lega e Movimento 5 Stelle. Mentre il primo, Matteo Salvini e la Lega, erano dati in costante aumento nei sondaggi sulle intenzioni di voto e i loro membri-supporters hanno sempre mostrato la loro approvazione all'istituzione della flat-tax, il secondo (Luigi Di Maio), nonostante avesse quasi il doppio dei parlamentari della Lega, ha sembrato svolgere un ruolo secondario nel dibattito sui social media sulla flat tax, restando quasi in sordina.

Luigi Di Maio è stato nominato con il proprio hashtag 22 volte, meno della metà di Matteo Salvini, a conferma di questa forbice. Altri hashtag sono legati alla Festa del Lavoro che si svolge il 1 maggio: primomaggio o 1maggio, salariominimo, festadailavoratori. Altri hashtag hanno una chiara connotazione politica avversa al governo dell'epoca: ad esempio, il governodelfallimento.

In seguito, è stata analizzata la struttura semantica dei tweet attraverso l'analisi delle relazioni tra gli hashtag in un'ottica di rete (mediante gli strumenti della Social Network Analysis) a partire dalla matrice di co-occorrenze, e dunque capire le dinamiche ricorrenti in hashtag che frequentemente appaiono congiuntamente.

Salvini. Orban, primo ministro ungherese, è usato in questo contesto come paragone negativo. C'è poi un riferimento a #vincisalvini: è una sorta di gioco presentato e diffuso da Matteo Salvini sui suoi account social, utilizzato come spunto per dare visibilità ai contenuti pubblicati sulle sue pagine. In quei giorni è stato ampiamente accusato dai suoi rivali di utilizzare qualunque mezzo per scopi di propaganda, e i tweet che utilizzano questi hashtag hanno tono sarcastici e irriverenti nei confronti del Leader della Lega.

Al centro della rete in Figura 2 si identificano due comunità principali: quella gialla e quella verde. La prima contiene diversi hashtag e, anche se non tutti con un chiaro atteggiamento prettamente positivo nei confronti del governo e della flat tax, il concetto principale espresso non è sicuramente negativo considerato nel suo complesso. Questa è l'unica comunità identificata, anche se grande, ad utilizzare sia hashtag propriamente relativi all'argomento e allo stesso tempo non esprimendo un forte disaccordo in merito. La comunità verde presenta un gran numero di riferimenti al concerto del 1 maggio e al mondo dei lavoratori in generale. Da questo punto di vista, risulta essere un gruppo di hashtag legati al mondo politico schierato a sinistra. Inoltre, nella parte alta gli hashtag utilizzati diventano via via più rudi, addirittura offensivi in alcuni casi. Ad esempio, *governodelfallimento* e *trota*. Quest'ultimo hashtag si riferisce al figlio di Umberto Bossi, fondatore della Lega Nord negli anni '90: un soprannome del genere è usato dai suoi avversari per evidenziare un'ipotetica mancanza di intelligenza e di vivacità intellettuale da parte del figlio del noto politico. Inoltre, è stato anche accusato di aver "comprato" la sua laurea in Albania, ed è per questo che è nella rete occupa un posto molto vicino a Calenda, DiMartedì, Donato, per la citata polemica televisiva sulla flat-tax in Albania. Potrebbe essere un indicatore del fatto che gli oppositori della Lega abbiano trovato un filo conduttore tra le due questioni.

Nel complesso, le comunità di hashtag sono in grado di evidenziare diversi aspetti del dibattito su Twitter sull'istituzione della di flat tax, ed esse appaiono piuttosto distinte. A partire da questa analisi descrittiva, si ottengono sufficienti elementi per assegnare a ciascuna comunità uno specifico concetto latente. Ogni community descrive un aspetto peculiare della rete, quindi ogni tweet verrà assegnato a ciascun concetto per completare le fasi di riduzione della dimensione statistica della base dati.

Nel definire ciascuna comunità, abbiamo utilizzato un modo molto restrittivo per considerare la polarità degli hashtag, semplificandolo

in sole 3 possibili modalità: in termini di valore positivo (+), neutro (Ø) o negativo (-) rispetto sempre all'argomento della flat-tax. La maggior parte dei concetti è stata assegnata ad una attitudine neutrale (Ø) soprattutto per quanto riguarda hashtag legati a cognomi o nomi propri (con alcune eccezioni notevoli), nomi di città o regioni, parole tendenzialmente neutre per loro stessa natura come hashtag legati a concetti come telegiornali, TV, informazioni e così via. Tenendo conto della somma algebrica delle attitudini degli hashtag contenuti in ogni cluster, risulta che le due grandi community nel centro della rete sono quella con l'attitudine positiva più elevato, e cioè la comunità gialla, e il più alto livello di atteggiamento negativo, e cioè la comunità verde. La comunità gialla, in particolare, ha molti hashtag che non possono essere classificati come (+) o (-): ad esempio, le persone che usano hashtag su Salvini o Di Maio non possono essere indubbiamente identificate come in accordo o in disaccordo sulle due figure politiche, se altri hashtag non sono presenti.

Dopo queste procedure di identificazione delle communities, i passaggi di riduzione della dimensionalità totale del corpus passano dalla dimensione degli hashtag a quella dei tweet. Ogni tweet è identificato, in media, da pochi hashtag, anche se in teoria ne sono ipotizzabili più di 10.

Di solito, dati i vincoli sul numero di caratteri, i tweet sono contrassegnati da 1, 2 o 3 hashtag, oltre alla chiave hashtag flattax che, come abbiamo detto, rimuoviamo dall'analisi. In questa prospettiva, assegniamo ogni tweet a una specifica community, quando possibile, utilizzando gli hashtag contenuti all'interno. Questa procedura di classificazione diventa sempre più importante con l'aumento della quantità di tweet e hashtag. Per questo specifico ambito applicativo, invece, una procedura di classificazione completamente automatica, vista la ridotta portata di hashtag e tweet, è stata affiancata da alcune speculazioni qualitative. Tali speculazioni sono state effettuate nella fase di assegnazione dei tweet ai concetti.

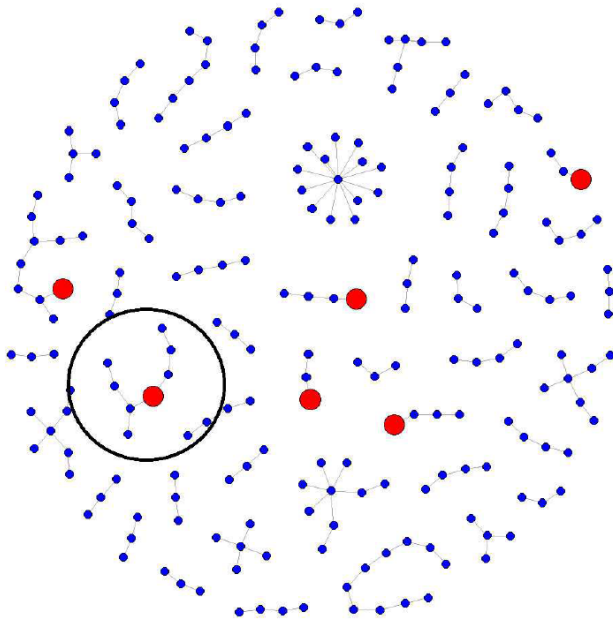


Figura 3: tweet iniziali, in rosso, generano catene di reply (blu). Nel cerchio è presente l'esempio che approfondiamo in seguito.

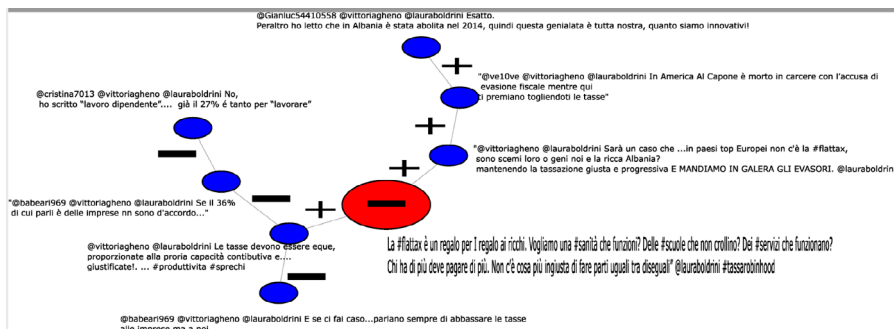


Figura 4: tweet iniziale veicolante attitudine contraria alla flat-tax. I segni + e - relativi ai legami da tweet a reply date da sentiment analysis condizionata.

Osservando la diversa struttura in termini di diffusione delle informazioni, i tweet relativi a concetti diversi agiscono in modo molto diverso. I concetti relativi a un atteggiamento positivo (o propositivo) nei confronti della flat-tax di solito hanno più risposte che trasmettono invece disaccordo; mentre i tweet relativi a concetti fortemente contrari alla flattax hanno i retweet come principale strumento di diffusione. La

Figura 4 mostra una rappresentazione della catena di risposta (*reply chain*) associata a un tweet che trasmette un concetto negativo verso la flat-tax. In questo caso le risposte sono poche e tendono generalmente a concordare con i concetti negativi.

CONCLUSIONI

Il presente lavoro mostra come la diffusione dell'opinione sui social media, in particolare su Twitter, possa portare a reti segnate, una struttura di network piuttosto peculiare e informativa, che descrive le dinamiche delle interazioni di retweet e di reply di concetti polarizzati relativi a un argomento di tendenza.

Per costruire tale rete composta da queste catene legate ai tweet abbiamo adottato una procedura multi-step. La prima fase ha riguardato una riduzione della dimensionalità dei tweet mediante una procedura di clustering in grado di identificare i concetti sottostanti veicolati dal gruppo di tweet. Quindi, utilizzando algoritmi di sentiment analysis per determinare il segno sia del tweet originale (rispetto al trending topic analizzato) sia il segno delle connessioni tra il tweet originale, le risposte e i retweet, si ottengono le vere e proprie reti segnate.

Il risultato finale di questa procedura è la descrizione di tali catene di retweet/reply associate ad ogni concetto. Dopo la ricostruzione di tali reti, sarà possibile un'analisi anche comparativa tra le diverse reti di concetti. Come mostrato nel caso di studio presentato, ad esempio, è possibile analizzare come si diffondono i concetti polarizzati da un segno positivo o negativo, e quali sono le caratteristiche (ad esempio, il segno) delle interazioni osservate.

Una limitazione della presente procedura è che, all'interno di questo approccio, non viene considerato il livello di "influenza" dell'utente originale che produce il tweet. In futuro, si propone di includere le caratteristiche degli utenti come variabili di controllo.

Inoltre, si propone di considerare l'adozione di un approccio modellistico-statistico, come il modello "Exponential Random Graph" per le reti segnate, al fine di stimare il "processo generativo" alla base delle diverse reti segnate rilevate per ciascun concetto identificato.

Ulteriori lavori, con un corpus composto da un maggior numero di tweet, retweet e risposte, sfrutteranno sicuramente in maniera più approfondita l'esito della sentiment analysis, anche con la possibile di scegliere algoritmi più raffinati e sofisticati, ma validi in caso di numerosità più elevate.

BIBLIOGRAFIA

- A., Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau
2011, "Sentiment analysis of twitter data," In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38.
- S. Balbi, M. Misuraca, M. Spano
2018, "A two-step strategy for improving categorisation of short texts," In: Proceedings of JADT conference (JADT 2018), pp. 60–67.
- A. Clauset, M. E. J. Newman, C. Moore
2004, "Finding community structure in very large networks", *Phys. Rev. E*, 70(6), p. 066111.
- S. Fortunato
2010, "Community detection in graphs," *Physics reports* 486, pp. 75–174.
- A., Go, R. Bhayani, L. Huang
2009, "Twitter sentiment classification using distant supervision," In: CS224N Project Report, Stanford, 1(12).
- M.W. Kearney
2018, "rtweet: Collecting Twitter Data," R package version 0.6.7, <https://cran.r-project.org/package=rtweet>.
- J. Kim, Y. Jaebong
2012, "Role of sentiment in message propagation: Reply vs. retweet behavior in political communication," 2012 International Conference on Social Informatics (IEEE), pp. 131–136.
- H. Kwak, C. Lee, H. Park, S. Moon
2010, "What is Twitter, a social network or a news media?" in: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600.
- D. Murthy
2013, "Twitter: Social Communication in the Twitter Age," Polity Press Cambridge, UK.
- T. Onorati, P. Diaz
2016, "Giving meaning to tweets in emergency situations: a semantic approach for filtering and visualizing social data," *SpringerPlus*, 5(1), p. 1782.
- L. Rossi, M. Magnani
2012, "Conversation practices and network structure in Twitter," In: Sixth International AAAI Conference on Weblogs and Social Media.
- H. Saif, Y. He, H. Alani
2012, "Semantic sentiment analysis of twitter. In: International semantic web conference," Springer, Berlin, Heidelberg, pp. 508–524.
- D. Sousa, L. Sarmento, E.M. Rodrigues
2010, "Characterization of the Twitter @replies network: Are user ties social or topical?," in: SMUC'10 Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, pp. 63–70, Toronto: CIKM.
- S. Stieglitz, L. Dang-Xuan
2012, "Political communication and influence through microblogging: An empirical analysis of sentiment in Twitter messages and retweet behavior," in: R. H. Sprague Jr. (Ed.), HICSS 2012: Proceedings of the 45th Hawaii International Conference on System Science, pp. 35003509, Washington, DC: IEEE Computer Society.

B. Suh, L. Hong, P. Pirolli, E.H. Chi
2010, “Want to be retweeted? large scale
analytics on factors impacting retweet in
twitter network,” In: 2010 IEEE Second
International Conference on Social
Computing, pp. 177–184.

J. Tang, Y. Chang, C. Aggarwal, H. Liu
2016, “A survey of signed network
mining in social media,” ACM Computing
Surveys (CSUR) 49, p. 42.