

APPENDICE III

NORME DI QUALITA' SOCIAL MEDIA RESEARCH E OSSERVAZIONE DIGITALE (Revisione ratificata in sede di Assemblea degli Associati del 15 giugno 2022)

INDICE

- 1. INTRODUZIONE**
- 2. PRINCIPI FONDAMENTALI**
 - 2.1. Essere conformi con la legge
 - 2.2. Consenso e notifica
 - 2.3. Protezione dei dati personali
 - 2.4. Assicurarci di non arrecare danno
 - 2.5. Reputazione del Cliente e del Settore
- 3. ELEMENTI TECNICI E STRUTTURALI SPECIFICI DELLA SMR E DELL'OSSERVAZIONE DIGITALE**
 - 3.1. Definizioni specifiche
 - 3.2. Strumenti di ricerca
 - 3.3. Analisi volumetrica
 - 3.4. Tool informatici e metodologie di analisi
 - 3.5. Pulizia e modifica dei dati
- 4. PROGETTO DI RICERCA DELLA SOCIETÀ**

1. INTRODUZIONE

Il documento in essere rappresenta il codice etico e le norme Assirm per la conduzione di ricerche di mercato, sociali e di opinione che hanno come oggetto i contenuti generati dai vari stakeholder sui social media in relazione a determinati argomenti o fenomeni (da qui in poi genericamente indicate come “Social Media Research”).

La continua evoluzione tecnologica e la viralità che lo caratterizzano, hanno reso il web un canale privilegiato per la rilevazione di dati ed informazioni. La SMR (Social Media Research) coinvolge un complesso di strumenti e metodi di ricerca in costante evoluzione,

Le indicazioni contenute nel documento si ispirano agli standard di settore ed offrono per la SMR un approccio esemplificativo degli aspetti tecnici per la realizzazione di un buon lavoro di ricerca.

I **Social media** sono definiti come piattaforme e tecnologie presenti in internet che permettono l’interazione degli utenti e/o facilitano la creazione e lo scambio dei contenuti generati (messaggi, immagini, video, azioni, ecc.). Nonostante il raggio d’azione continui ad espandersi, attualmente gli ambiti più frequenti sono:

- forum/ community di discussione tematica;
- blog;
- social network (ad esempio Facebook) / piattaforme di collaborazione e/o comunicazione di gruppo o pluri-personale (ad esempio Twitter);
- social news, ovvero testate giornalistiche online che permettono l’inserimento di commenti da parte degli utenti, in calce all’articolo del giornalista;
- siti di condivisione di video/foto (ad esempio YouTube) e/o recensioni relative a prodotti, servizi, esperienze, ecc. (ad esempio Tripadvisor);
- siti di raccolta di domande e risposte tra utenti (ad esempio Yahoo Answers).

La **Social Media Research (SMR)** comprende tutte quelle tecniche d'indagine che rilevano dati sulle fonti online di natura principalmente social, ovvero piattaforme virtuali, generalmente non finalizzate alla ricerca di mercato, che gli utenti adottano per creare e condividere contenuti testuali, immagini, video e audio.

Fanno parte della SMR:

- **Social Media Monitoring o Social Media Listening** → insieme delle tecniche ed attività sistematiche e pianificate di ascolto del passaparola che ha luogo unicamente sui social media, al fine di comprendere e misurare quando, quanto e come gli utenti parlino di un'azienda, brand, prodotto, personaggio, settore o tema (Brand Awareness and Identity, Brand Reputation, etc.);
- **Web Content Analysis** → insieme delle tecniche ed attività sistematiche e pianificate di rilevazione di tutti i contenuti presenti in rete, interattivi (social) e non (fonti normative, news non commentabili, siti istituzionali, etc.) inerenti un'azienda, brand, prodotto, personaggio, settore o tema;
- **Netnography** → tecnica per l'individuazione di specifiche e circoscritte fonti social come web tribe, community, etc. e lo studio dei meccanismi di comunicazione, relazione, potere, linguaggio, pregiudizi, preferenze ed opinioni dei loro utenti.

Pur rimanendo di tipo qualitativo e non inferenziale, questi strumenti possono potenzialmente raccogliere grande quantità di dati e tracciare così l'andamento quantitativo dei fenomeni sui social media. I flussi di discussione rilevati nella SMR vengono di fatto approcciati sia attraverso tecniche di analisi quantitative, utili per misurare i volumi e gli andamenti, sia attraverso tecniche qualitative, impiegate al fine di comprendere a fondo il contenuto dei messaggi raccolti, coglierne i driver sottostanti e trarne insight utili per il cliente.

Per estensione, e data l'omogeneità delle metodologie e degli strumenti tecnologici impiegati, questo documento riguarda tutte le altre forme di Osservazione Digitale, anche non direttamente legate ai Social Media.

Tali fonti possono comprendere, ad esempio, corpus di testi non disponibili sul web o non pubblici, materiale originato da basi utenti o basi clienti, comunicazioni email, dati provenienti da registrazioni e sensori, crowdsourcing digitale.

Nell'intero documento si utilizza il verbo "devono" per descrivere un principio che i ricercatori sono obbligati a seguire se vogliono attenersi alle Norme di Qualità Assirm. Il verbo "dovrebbero" è invece utilizzato per descrivere l'adempimento ad un principio etico. Questo utilizzo distinto dei due verbi riconosce ai ricercatori la possibilità di scegliere di adempiere a dei principi etici in modi diversi in base al progetto di ricerca.

Queste norme devono essere lette insieme alle Norme di Qualità Assirm sulla Ricerca di Mercato.

2. PRINCIPI FONDAMENTALI

I requisiti generali contenuti nelle Norme di Qualità Assirm si applicano anche alle diverse tecniche di Social Media Research.

Questo paragrafo descrive la maniera in cui tali requisiti dovrebbero essere resi operativi nella raccolta di dati provenienti dai social media per scopi esclusivi di ricerca di mercato, sociale e di opinione, riconoscendo l'esistenza e distinguendosi dalle diverse e molteplici attività rese possibili dai social media (ad esempio l'assistenza ai clienti, la gestione delle pubbliche relazioni, ecc).

2.1. Essere conformi con la legge

Accedendo virtualmente ai vari siti/piattaforme online, i ricercatori sono soggetti ai Termini di Utilizzo (TdU) e alle Condizioni dei proprietari di tali siti/piattaforme, come le clausole sui diritti di proprietà intellettuale che proibiscono esplicitamente la copia non autorizzata di materiale, o i divieti di qualunque tipo di raccolta ed analisi dei dati senza averne il permesso.

I TdU di solito trattano argomenti come le questioni di copyright e l'uso di materiale reperibile dal sito. Spesso, il proprietario del sito web è anche il proprietario del

copyright. Ad ogni modo le leggi sul copyright sono complesse e variano a seconda della giurisdizione. I ricercatori devono esserne a conoscenza e devono rispettare le leggi vigenti in materia.

Nella maggior parte dei casi, tuttavia, i dati per la Social Media Analysis vengono reperiti tramite cosiddetti software di crawling, ossia programmi volti a intercettare e scaricare i flussi di discussione sui social media corrispondenti a determinati parametri di ricerca. Oggi la rete rende disponibili molteplici piattaforme di crawling, anche gratuite, che si prestano a questo compito di raccolta. Tutti i contenuti ai quali i software di crawling accedono online e rendono fruibili per i ricercatori sono rigorosamente contenuti pubblici, ossia pubblicati dagli utenti in ambienti di pubblica visibilità. Non saranno recuperati in automatico dati provenienti da profili privati e chiusi dei singoli utenti oppure da siti per i quali è previsto un login per l'accesso. Vi possono essere tuttavia casi in cui è necessario accedere a contenuti provenienti da una Community per la quale è prevista la registrazione. Tale procedura è possibile solo previo accordo con l'amministratore della Community stessa, che procederà a regolarizzare e tutelare la privacy dei propri iscritti.

È responsabilità dell'amministratore del crawling inoltre controllare le singole condizioni che si applicano al contenuto di un determinato social media che viene utilizzato per la raccolta di dati e rispettare le richieste di privacy (incluse le richieste di file robot.txt, pagine sicure, ecc.). Essi devono chiedere il permesso di acquisire contenuti da qualunque fonte e nei casi in cui ciò possa costituire una violazione dei TdU devono sempre e comunque rispettare le regole di quel servizio.

2.2. Consenso e notifica

Le Norme di Qualità Assirm affermano che la cooperazione dei partecipanti alla ricerca deve essere basata su informazioni adeguate circa lo scopo e la natura della ricerca (4.3.2. Informativa al rispondente) e che bisogna ottenere il consenso di questi a partecipare.

Nonostante sia potenzialmente semplice ottenere il consenso dei membri delle community a scopo di ricerca di mercato, sociale e di opinione, il problema si pone per altri social media, i cui utenti generalmente non vengono informati in anticipo,

o non hanno acconsentito all'uso dei dati per delle ricerche, a meno che ciò non sia esplicitamente menzionato nei TdU.

Data la non consapevolezza degli utenti web ad essere “ascoltati” da parte di società di ricerca, l'accorgimento che i Social Media Analyst devono sempre seguire è trattare tutti i dati raccolti in modo anonimo. Ogni riferimento a contenuti prodotti dagli user all'interno di report di ricerca per i clienti devono nascondere i dati sensibili e anagrafici degli utenti (vedi oltre, tecniche di “masking”). Gli screen-shot tratti dal web riprodotti nei report per il cliente sono dunque sempre anonimizzati, eliminando nome e cognome dell'interessato.

Si evidenzia che i sistemi di crawling in commercio e in uso non consentono di impostare filtri e, pertanto non consentono in automatico di “rimuovere i dati identificativi quali nome, foto, collegamenti al profilo dell'utente, ecc.”.

Per quanto riguarda il rispetto della normativa sul copyright (diritti d'autore) e dei termini di utilizzo (TdU), si rinvia al caso specifico del servizio utilizzato. In ogni caso, la documentazione relativa all'utilizzo di tali servizi deve essere disponibile nella cartellina di ricerca o in un'area comune, con le relative analisi legali e organizzative.

Per quanto riguarda la riservatezza della ricerca, la protezione dei dati personali e i cookie, si veda la normativa vigente, come richiamata anche al punto 2.2 delle Norme di Qualità Assirm.

2.3. Protezione dei dati personali

La legislazione sulla privacy si applica solo ai dati personali identificativi e non ai dati tramite i quali è impossibile identificare un soggetto.

Le piattaforme di social media offrono molte opportunità di visualizzazione di dati personali identificativi. Alcuni soggetti postano informazioni che svelano apertamente la propria identità; essi ne sono consapevoli e hanno delle aspettative limitate in materia di privacy. Altri, invece, non sono consapevoli del fatto che i servizi che stanno usando siano aperti ad altri, anche per la raccolta di dati, oppure pensano di non essere riconoscibili, mascherando la propria identità con un nickname o uno pseudonimo.

Ad ogni modo, oggi esistono servizi online che rendono possibile in molti casi l'identificazione di chi posta un determinato contenuto, ad esempio grazie allo user name o ai commenti pubblicati in seguito al post, e possono collegare l'utente a molti altri dati identificativi, inclusi l'indirizzo, il numero di telefono, il reddito approssimativo e i dati socio-demografici.

Detto ciò, appare chiaro che i dati non possano essere resi anonimi al 100% su internet, semplicemente rimuovendo lo user name e l'URL collegato al commento. Quindi, se i ricercatori intendono citare all'interno di un report dei commenti pubblici, essi devono prima controllare che l'identità dell'utente non sia facilmente rintracciabile usando i suddetti servizi disponibili in rete. Se questa ipotesi è verosimile, i ricercatori devono sforzarsi di ottenere il permesso dall'utente, al fine di poterlo citare, o, in caso contrario, di mascherare il commento, in modo che l'identità di chi lo ha postato non possa essere svelata.

Il "masking" è una tecnica tramite la quale i dati originali vengono talmente modificati da rendere difficilissima la ricerca degli stessi in rete, tramite un motore di ricerca. In tal modo, anche l'identità dell'utente che li ha originati è difficile da scoprire. Questa è una tecnica molto utile per assicurarsi che l'anonimato delle persone che commentano sia rispettato, nei casi in cui:

1. il ricercatore non ha cercato di ottenere il permesso;
2. il commento sarebbe facilmente rintracciabile tramite un servizio di ricerca.

Il masking può essere applicato in vari gradi, che vanno dal semplice cambiamento di qualche parola tramite l'alterazione di alcuni elementi chiave di un commento, fino all'abbreviazione. È responsabilità del ricercatore decidere quale grado sia il più appropriato. Alcuni fattori da tenere in considerazione al riguardo sono:

- se l'argomento di discussione è delicato o personale;
- se vi è uso di linguaggio offensivo;
- se ci sono riferimenti contro la legge;
- se ci sono elementi che potrebbero creare imbarazzo o avere un impatto negativo sulle opportunità di carriera;
- se ci sono informazioni identificative;

- se ci sono dati a proposito di altre persone che non siano già stati resi pubblici.

Nel caso di foto o video pubblici bisognerebbe tenere in considerazione alcune tecniche come ad esempio la dissimulazione dei volti tramite l'uso di pixel, nei casi in cui il masking è necessario.

È importante considerare che il masking probabilmente non sarà sufficiente in molti contesti business-to-business oppure quando la ricerca viene condotta all'interno di un gruppo molto ristretto, visto che l'identificazione è molto semplice, nonostante si cerchi di mascherare i dati identificativi.

Quando i ricercatori inseriscono un commento in un report per un Cliente è necessario indicare chiaramente se sono state utilizzate delle tecniche di masking. Se il ricercatore decide di chiedere il permesso dell'utente per citarlo, egli deve rispettare le normative vigenti nel paese dove è residente l'utente; deve inoltre assicurare l'utente sul fatto che queste vengano rispettate e deve spiegare chiaramente e onestamente lo scopo del suo lavoro. Si dovrebbe dare l'opportunità all'utente di verificare la buona fede del ricercatore, se lo desidera, prima di autorizzarlo a procedere.

Nei casi in cui i ricercatori utilizzano dei servizi per arricchire i commenti dell'utente con dati demografici ad esso associati (ad esempio informazioni tratte dai profili personali), essi dovrebbero usare queste informazioni solo per scopi di classificazione della ricerca. Visto che questi servizi spesso forniscono informazioni identificative personali (come i numeri di telefono e gli indirizzi e-mail), questi potrebbero permettere un collegamento non intenzionale dei dati di ricerca ai dati personali e ciò non deve assolutamente essere usato nell'analisi o trasmesso a terzi.

2.4. Assicurarsi di non arrecare danno

Un principio chiave stabilito e condiviso dalle Norme di qualità Assirm è che i rispondenti non debbono mai ed in alcun modo subire conseguenze negative, come risultato diretto della loro partecipazione ad una ricerca.

Il rischio maggiore nella ricerca tramite i social media è collegare alla rivelazione accidentale l'identità degli utenti, che non avevano compreso di partecipare ad una ricerca e che quindi non si aspettavano di poter essere identificati.

Ancora una volta, per assicurarsi che gli utenti non subiscano danni a causa delle attività di ricerca, bisogna rispettare un principio di cautela, rimuovendo ogni dato personale identificativo al più presto possibile e prendendo in considerazione l'attuazione dei controlli di qualità necessari. In tal senso i ricercatori devono essere particolarmente cauti quando conducono ricerche in cui sono coinvolti bambini o minori (4.3.3. Raccolta dati da minori o da rispondenti vulnerabili).

Questa è una questione particolare nel caso di piattaforme social, alle quali anche i bambini possono prendere parte. Quando dei dati provengono probabilmente da un bambino, i ricercatori devono prestare particolare attenzione alla dissimulazione delle risposte, al fine di assicurare che l'utente non possa essere identificato, oppure ottenere il permesso da un genitore o dal tutore legale per raccogliere ed usare i dati identificativi.

Anche la SMR condotta nel campo farmaceutico necessita di particolari procedure. La normativa sulla farmacovigilanza prevede infatti che ogni cosiddetto "evento avverso" su un farmaco rilevato all'interno delle conversazioni web debba essere segnalato agli appositi uffici della casa farmaceutica che ha commissionato l'indagine e proprietaria del farmaco in questione. Per evento avverso si intende qui un qualsiasi fenomeno clinico spiacevole che si presenta durante un trattamento con un farmaco, ma che non abbia necessariamente un rapporto di causalità (o di relazione) con il trattamento stesso. Le aziende hanno diretta responsabilità nel monitorare ciò che avviene sulle proprie properties digitali. Nel momento in cui invece si affidano a società di ricerca per il monitoraggio del conversato sui farmaci su forum e social network, diviene responsabilità della Società stesso attenersi alle procedure della farmacovigilanza, così come comunicate e inoltrate dall'azienda. Le conversazioni online sono infatti spesso

ricche di dettagli sul contesto della reazione avversa, sulle comorbilità, sugli usi off label, sulle interazioni tra farmaci. Eventuali post o commenti che riportino perciò un evento avverso, un reclamo su prodotto o altro che l'azienda ritiene rilevante, devono essere inoltrati da parte del ricercatore all'azienda farmaceutica che provvederà ad attivare la segnalazione al reparto di farmacovigilanza, seguendo le tempistiche e le modalità previste dalla legge. Nel caso di rintracciamento di un evento avverso sarà peraltro richiesto e concesso al ricercatore di raccogliere e inoltrare all'azienda farmaceutica le informazioni personali (se disponibili) dell'autore del post o del commento essenziali per poter procedere con la segnalazione.

2.5. Reputazione del Cliente e del settore

La Società risponde interamente di tutti i servizi svolti in relazione alla ricerca, incluse le parti del lavoro o servizio affidate a fornitori, singoli professionisti o società (Outsourcing) per le seguenti categorie di servizi, con impatto determinante sulla qualità della ricerca, tranne i casi in cui la scelta del fornitore sia del Cliente:

- Traduttori
- Centri di trattamento dati (Data Entry, etc.)
- Tool e software per il crawling, la normalizzazione e l'elaborazione dei dati
- Altre Società di ricerca
- Consulenti

La Società deve documentare per le categorie di servizi sopra indicate le modalità per:

- selezionare e qualificare i fornitori;
- stabilire i rapporti contrattuali con essi;
- fornire agli stessi le necessarie istruzioni per l'esecuzione del servizio loro affidato;
- controllare la qualità del servizio fornito;
- valutare le prestazioni dei fornitori con cadenza almeno annuale.

3. ELEMENTI TECNICI E STRUTTURALI SPECIFICI DELLA SMR E DELL'OSSERVAZIONE DIGITALE

Come precedentemente indicato la SMR e l'Osservazione Digitale comprendono tutte quelle tecniche d'indagine che rilevano dati sul web, condivisi dalle persone, e dati provenienti da altre fonti digitali testuali o convertibili in formato testo (voice recognition).

3.1. Definizioni specifiche

- **Crawling** o Scraping è il processo di estrazione dei dati dalle fonti originarie a scopo di analisi. I dati raccolti dai social media, in particolare, si riferiscono ai contenuti (post, commenti, foto, ecc.) che gli utenti generano, arricchiscono o condividono mentre interagiscono sui social media. La raccolta di dati include spesso dati identificativi degli utenti.
- **Post** è ogni messaggio originario pubblicato su un social media (può essere un post su Facebook, un post di un blog, un articolo sul sito interattivo di un quotidiano online, un tweet, un post su Instagram, etc.), e per analogia, ogni messaggio unitario trasmesso su altri mezzi (email, richiesta di assistenza, commento, contributo).
- **Interazioni** sono le attività che gli utenti possono effettuare su un post originario, quali like, commenti, risposte o condivisioni.
- **Engagement** è una misura di quanto un post originario è stato in grado di coinvolgere altri utenti attraverso le loro interazioni.
- **Social Properties** sono i canali ufficiali delle aziende o brand o personaggi sui social network.
- **Masking** è una tecnica con la quale i dati originari come i post, i commenti, le foto o i video, sono alterati e anonimizzati fino al punto in cui non possono più essere tracciati o attribuiti all'utente che li ha pubblicati in origine.
- **Community di ricerca** è una community online creata con lo scopo specifico di ospitare ricerche di mercato, sociali e di opinione.
- **Utente** è un soggetto o un'organizzazione da cui si raccolgono informazioni per scopi di ricerca di mercato, indipendentemente dal fatto che esso ne sia

consapevole o inconsapevole. Nel contesto della SMR, l'utente può anche essere menzionato come Autore, Membro, Fan, Follower o Poster (colui che posta un contenuto su un social media), termini equivalenti a quello di Rispondente o Partecipante per altre modalità di ricerca di mercato, sociale e di opinione di tipo tradizionale. Nel contesto dell'Osservazione Digitale, può essere un Dipendente o Collaboratore, un Cliente, un Ascoltatore, un Passante eccetera.

- **Influencer** è un utente specifico che ha una capacità di impatto sul conversato oggetto di studio in termini di attività svolta (numero di post scritti sull'argomento) e di livello di engagement prodotto (interazioni ricevute).
- **Sentiment** è la valenza positiva/negativa o neutra che il messaggio rilevato contiene in termini di opinione espressa dall'autore sull'oggetto di studio.

3.2. Strumenti di ricerca

Al pari di questionari, piani di campionamento, tracce dei focus group nella ricerca tradizionale tali elementi della SMR e dell'Osservazione Digitale, devono essere dettagliati, verificati e validati, internamente alla Società e, dove richiesto, dal Committente mediante approvazione formale.

Come detto, nella maggior parte dei casi i ricercatori raccolgono i dati rilevanti per il monitoraggio attraverso il ricorso a specifici software di crawling o scraping. La maggior parte dei tool gratuiti sono dedicati ad analizzare i flussi e le metriche di un particolare tipo o sottoinsieme di sorgente. Le versioni più evolute sono a pagamento e consentono di andare oltre il dato grezzo e fornire intelligence alle aziende, leggendo dentro e oltre una conversazione, un commento, un like o un retweet.

Ovviamente anche le piattaforme di crawling a pagamento differiscono sulla base di una serie di parametri, quali ad esempio la tecnica di data capture. Ci sono due approcci al data collection, ognuno dei quali genera impatto sulla qualità e quantità dei dati restituita. Vi sono piattaforme che lavorano su White List, ovvero su un perimetro prestabilito di siti legittimati sui quali vengono effettuate le ricerche. Tale approccio garantisce da un lato dati mirati e puliti, ma limita i risultati in

termini di quantità, creando in altre parole il rischio di “perdersi qualcosa”. Le piattaforme cosiddette di crawling spediscono i crawlers su un database di dati registrati attraverso lo spazio dei pubblici domini online. Tale approccio consente di costruire un record ampio di discussioni online, seppur aumenti il rischio di catturare anche contenuti spam o di bassa qualità. Vi è anche un tipo di raccolta di dati dai social media, che si aggancia ai risultati ottenuti dai motori di ricerca (Google, Yahoo, ...). Gli spider lavorano utilizzando i primi risultati ottenuti dai motori di ricerca web (as Google). Restituisce gli hot topic legati ai ranking dei motori di ricerca in un preciso momento temporale, legandosi dunque in maniera parziale ai criteri di ranking dei motori, nonché al limite impostato nello scaricare i risultati (solitamente le prime decine di pagine di risultati).

I software di crawling vanno in ogni caso impostati dai ricercatori attraverso la definizione di una serie di parametri, quali:

- **IL PERIMETRO DELLE FONTI:**

la scelta, la numerosità e le peculiarità delle fonti monitorate, in particolare quando si vuole focalizzare l’attenzione su alcune fonti specifiche della rete in cui rilevare i comportamenti e/o le reazioni.

- **GLI STIMOLI:**

la scelta, la sequenza e gli scopi degli stimoli somministrati, in particolare quando si vogliono analizzare le differenze fra i contenuti spontanei e sollecitati degli utenti, per rilevarne insight specifici.

- **LE QUERY:**

l’interrogazione da parte del ricercatore di uno o più database o software di crawling attraverso la combinazione dei seguenti strumenti:

- **KEYWORD:**

identificazione delle parole chiave più appropriate e pertinenti l’oggetto d’indagine, compresa la gestione di omonimi, sinonimi, misspelling, neologismi e ambiguità linguistiche

- **SINTASSI**

Spesso le keyword vengono combinate tra di loro attraverso l'uso di una sintassi specifica che ricorre ad esempio all'uso di operatori booleani (AND, OR, NEAR, AND NOT).

- **I FILTRI:**

identificazione degli ambiti di ricerca mediante filtri di natura geografica (es. mercato: Italia), linguistica (es. lingua: italiano), e di periodo (indagine retroattiva, su dati riferiti al passato, oppure ongoing, su dati aggiornati periodicamente, oppure real time, su dati aggiornati in tempo reale). È possibile anche filtrare per tipologia di fonte (es. solo social network o solo forum). In alcune piattaforme più evolute è possibile anche filtrare la ricerca solamente sulle social properties oppure sul resto del web.

3.3. Analisi volumetrica

Ciascuna ricerca produce (reperisce) una certa mole di contenuti che richiedono la presentazione di un resoconto volumetrico che ne identifichi a livello quantitativo gli output principali e la bontà della ricerca:

- quanto se ne parla (es. totale numero di messaggi rilevati);
- dove se ne parla (es. totale fonti, distribuzione per tipo di fonte e per singolo dominio);
- chi ne parla (es. numero di autori unici e identificazione autori più attivi);
- di che cosa si parla (es. identificazione di termini e concetti più ricorrenti);
- come se ne parla (es. qual è la distribuzione delle opinioni positive, negative, neutre e interrogative).

Tale analisi volumetrica dell'indagine è ulteriore elemento chiarificatore della buona impostazione e della qualità della ricerca stessa e pertanto va verificata e validata, internamente alla Società e dove richiesto dal Committente.

All'analisi volumetrica o quantitativa può essere affiancata, in alcuni progetti di ricerca, anche l'analisi qualitativa delle discussioni, tramite lettura da parte dei ricercatori di un sotto-insieme o della totalità dei messaggi raccolti. I ricercatori procedono così, attraverso una griglia di analisi tipica delle ricerche qualitative, alla interpretazione del materiale raccolto e alla stesura di un report volto alla generazione di insight per il committente.

La costruzione dell'eventuale sotto-insieme da analizzare in profondità può essere fatta attraverso tecniche di campionamento dei messaggi. Una tecnica diffusa è quella del campionamento semplice, che riproduca dunque l'universo di riferimento del totale messaggi raccolti. Ma il ricercatore, in accordo con il committente, può procedere anche nell'analisi qualitativa attraverso carotaggi specifici su alcune fonti o piattaforme.

L'estrazione del campione di messaggi ha anche la funzione di agevolare l'analisi del sentiment, qualora non si ricorra a motori automatici. Non tutte le piattaforme di crawling infatti possiedono un motore di sentiment automatico o talvolta il ricercatore ritiene più efficace procedere con il "tagging" manuale dei messaggi.

3.4. Tool informatici e Metodologie di analisi

Per definire e rendere efficace l'output della ricerca nel progetto di ricerca vanno descritti i software utilizzati e le metodologie di analisi applicate.

Si tratta di sintetiche descrizioni dei tool e delle metodologie per garantire la comprensione dell'approccio e la fruibilità dei risultati. Tali descrizioni descrivono e spiegano, ove presenti, i seguenti elementi:

- le metriche per l'analisi dei contenuti, la tipologia di motore di sentiment, le modalità di misurazione della polarità dei contenuti;
- le tecniche di analisi testuale, linguistica e semantica impiegate;
- i modelli di selezione dei contenuti e di elaborazione/interpretazione dei dati;
- gli algoritmi applicati;

- i tipi di dati che sono analizzati (per esempio frasi brevi, frasi lunghe, non frasi, emoticon, gergo, parole irriverenti, sarcasmo, acronimi, errori ortografici eccetera);
- il numero di siti, autori, commenti, pubblicazioni inclusi nell'analisi;
- se la raccolta avviene sull'universo disponibile o su un suo campione;
- il processo di valutazione dell'affidabilità e dell'accuratezza dell'analisi (per esempio la sua frequenza, il numero di risultati su cui agisce, le prove effettuate);
- le lingue escluse e incluse, le regole applicate per i messaggi multilingue;
- la possibilità di intercettare e rimediare agli errori di analisi;
- qualsiasi altra metrica che possa esprimere la qualità dell'analisi del sentimento o del testo (ad esempio, punteggi di qualità dei codificatori);
- i metadati raccolti (esempio indirizzo IP, sistema operativo, agente utente del browser, cookie, geolocalizzazione, data, ora);
- il tipo di "unicità" identificata dai metadati (cookie, browser, dispositivi, visitatori);
- come sono gestiti i rifiuti (ad esempio, rifiuto del cookie) e gli abbandoni;
- eventuali limitazioni che sono state applicate ai dati raccolti, per esempio sul periodo di conservazione o sulla capacità di validazione.

I software e gli algoritmi che possono essere utilizzati nella SMR sono i seguenti:

- sistemi di crawler o scraping;
- sistemi di analisi del sentiment;
- sistemi di language processing e language understanding;
- sistemi di analisi statistica.

3.5 Pulizia e modifica dei dati

È necessario documentare i processi di pulizia e modifica dei dati e i controlli ad essi applicati. Nello specifico della SMR, tali attività possono riguardare, a titolo di esempio, il trattamento di:

- omonimie, che vanno ricomposte ove possibile;

- dati provenienti da clienti e concorrenti, che vanno separati da quelli dei consumatori;
- pubblicità, buoni e altro materiale promozionale;
- spam, account fake e altri contenuti di disturbo o malevoli;
- canali di notizie, link e riferimenti esterni;
- traffico non valido o non umano (bot, traffico interno, traffico simulato)

La documentazione deve descrivere e spiegare, dove presenti, i seguenti elementi:

- tipo di pulizia effettuata (manuale, automatica supervisionata, automatica non supervisionata), e, se effettuata in modalità mista, con quali proporzioni;
- effetti attesi della pulizia prevista;
- criteri di pulizia e cancellazione dei contenuti inappropriati;
- criteri di etichettatura dei contenuti come “mancanti” o “non validi”;
- eventuali metriche sulla validità dei risultati e su falsi positivi e falsi negativi riscontrati.

4. PROGETTO DI RICERCA DELLA SOCIETÀ

La Proposta della Società, documentata in forma scritta, descrive perimetro d'indagine, tecniche di rilevazione, metodologie di elaborazione ed analisi dei dati. Nel caso della SMR e dell'Osservazione Digitale, dovrebbero essere specificati i seguenti elementi:

- l'argomento trattato;
- requisiti relativi alle informazioni del cliente;
- l'entità dell'impegno richiesto (dimensioni e durata della ricerca);
- le fonti digitali da analizzare, e i criteri di selezione delle stesse;
- le dimensioni tecniche e di campionamento delle fonti da analizzare;
- il periodo di tempo dell'osservazione;
- le tecniche, le metodologie e le modalità di rilevazione (es. software di crawling);

- gli strumenti di ricerca (fonti, stimoli, query, ecc.);
- tool informatici e metodologie applicate per il trattamento dei dati;
- lo scopo, il tipo e la frequenza della fornitura dei risultati;
- modalità e durata della memorizzazione dei dati;
- la descrizione di ciò che verrà consegnato in esecuzione dell'incarico;
- il prezzo;
- la legislazione e i codici di condotta professionale applicabili.

In particolare le descrizioni relative ai tool informatici e software di ricezione, trattamento e/o analisi dei dati in forma automatica, rappresentano un manuale d'uso (funzionalità e modalità di fruizione) a disposizione dei Committenti, nel quale andrebbero descritti, ove presenti, i seguenti elementi:

- le metriche di rilevanza ed influenza adottate per l'analisi dei contenuti, la tipologia di motore di sentiment eventualmente utilizzato, l'indicatore del sentiment e le modalità di misurazione della polarità dei contenuti;
- eventuali tecniche di analisi testuale, linguistica e semantica (associazione in fattori, clusterizzazione, etc.) impiegate;
- i criteri di scelta e selezione del campione di messaggi analizzati "manualmente";
- eventuali modelli di elaborazione/interpretazione dei dati.

Infine, già dal Progetto dovrà essere indicato come il Cliente potrà alla fine disporre delle informazioni di ricerca: la base dati, i rapporti scritti, le sintesi, le presentazioni, la partecipazione a workshop interni, la consultazione dei risultati online, etc.