Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Scienze di Internet

Analisi del comportamento e dello stato d'animo di pazienti malati del Morbo di Crohn tramite social networks

Tesi di Laurea in Multimedia e Tecnologie creative

Relatore:
Chiar.mo Prof.
Marco Roccetti

Presentata da:
Alice Casari

Correlatore:
Chiar.mo Dr.
Gustavo Marfia

Sessione I
Anno Accademico 2013/2014

Indice

Capitolo 1

Introduzione

Al giorno d'oggi una pratica molto comune è quella di eseguire ricerche su Google per cercare qualsiasi tipo di informazione e molte persone, con problemi di salute, cercano su Google sintomi, consigli medici e possibili rimedi. Questo fatto vale sia per pazienti sporadici che per pazienti cronici: il primo gruppo spesso fa ricerche per rassicurarsi e per cercare informazioni riguardanti i sintomi ed i tempi di guarigione, il secondo gruppo invece cerca nuovi trattamenti e soluzioni. Anche i social networks sono diventati posti di comunicazione medica, dove i pazienti condividono le loro esperienze, ascoltano quelle di altri e si scambiano consigli.
Molte volte il fatto di condividere le proprie esperienze in forma anonima e con persone sconosciute può portare ad una maggiore sincerità e disinvoltura nel raccontare informazioni personali che risultano quindi più veritiere.
Tutte queste ricerche, questo fare domande e scrivere post o altro ha contribuito alla crescita di grandissimi database distribuiti online di informazioni, conosciuti come BigData, che sono molto utili ma anche molto complessi e che necessitano quindi di algoritmi specifici per estrarre e comprendere le variabili di interesse.
In questo lavoro di tesi il gruppo considerato più interessante e studiato è quello dei pazienti cronici: questi pazienti vivono giorno dopo giorno i differenti aspetti della loro situazione e raccontano come evolvono nel tempo questi loro problemi e come modificano il loro stile di vita per ottimizzare i trattamenti, stare meglio e massimizzare la qualità della loro vita. Questi pazienti cronici conoscono in modo approfondito ogni aspetto della malattia che li affligge ed hanno quindi un alto grado di auto-coscienza di come trattare il loro caso medico, maggiore rispetto ai pazienti occasionali.
Per analizzare questo gruppo interessante di pazienti gli sforzi sono stati concentrati in particolare sui pazienti affetti dal morbo di Crohn, che è un tipo di malattia infiammatoria intestinale (IBD) che può colpire qualsiasi parte del

tratto gastrointestinale, dalla bocca all'ano, provocando una grande varietà di sintomi. Questo morbo fa purtroppo parte di quelle malattie croniche i cui sintomi possono essere trattati ed alleviati, ma di cui una cura totale e definitiva deve ancora essere trovata. Sono possibili periodi di remissione dalla malattia, nei quali non si presenta nessun sintomo, ma questi periodi sono destinati prima o poi a terminare e a lasciar spazio di nuovo alla ricomparsa dei problemi. Le persone affette da questo morbo trascorrono quindi la loro intera vita lottando contro questa malattia che in qualche raro caso può essere anche fatale.
Lavorando con un team di medici specializzati nel morbo di Crohn, del Policlinico Sant'Orsola-Malpighi di Bologna, è stata analizzata la terminologia esatta utilizzata in questo ambito medico ed è stato osservato il comportamento dei pazienti affetti da questa malattia, dove scrivono, cosa scrivono e come scrivono. Tra le varie fonti prese in considerazione (social network, forum e blog) ed analizzate, questa tesi si concentra sullo studio avvenuto sui social network, in particolare Facebook e Twitter.
Il primo obiettivo è quindi quello di studiare il comportamento dei pazienti sui social network, quanto scrivono, in quanti scrivono, quanto sono attivi, per quanto tempo scrivono e di che argomenti parlano.
Particolare attenzione è stata data al sentimento che i pazienti dimostrano per i diversi tipi di trattamenti possibili ed un altro obiettivo di questa tesi è quello di esaminare i sentimenti espressi dai pazienti e di determinare il sentimento prevalente per ogni specifico trattamento. Questo secondo lavoro di analisi del sentimento non era mai stato effettuato prima da precendeti lavori e ricerche ed è quindi il primo lavoro di ricerca che analizza questo aspetto del morbo di Crohn.
In sintesi il lavoro svolto in questa tesi ha l'obiettivo comprendere il comportamento e il linguaggio che pazienti malati usano online, come utilizzano i social network e come caratterizzano i loro discorsi, per questo gli algoritmi utilizzati sono rilevanti sia in ambito informatico che in campo medico.
Dopo lo studio approfondito delle metodologie più idonee ad estrarre e analizzare i risultati più rilevanti è stato possibile creare un sistema software di *data mining* in grado di riprodurre tutte queste analisi in modo automatico e su qualsiasi argomento.
Lavoro di questa tesi è stato quindi anche quello di creare questo sistema software che in base all'argomento datogli in input sulla quale fare queste ricerce, provveda in modo automatico ad estrarre i dati da Facebook e Twitter e ad elaborarli, analizzarli e presentarli nel modo indicato nelle metodologie.

Partendo da un'analisi approfondita dei lavori già svolti precedentemente in questo ambito, si è proceduto evidenziando i punti innovativi di questo

lavoro e a presentarli interamente nel capitolo dedicato alle metodologie applicate.
Un capitolo è stato dedicato alla presentazione e alla discussione dei risultati estratti con le metodologie indicate sul tema centrale del morbo di Crohn.
Infine un capitolo mostra il sistema software creato, la sua architettura, il suo funzionamento e il modo corretto di utilizzarlo.

Capitolo 2

Stato dell'arte e problema

L'evoluzione di Internet e dei social networks ha aperto la strada ai ricercatori per l'applicazione di tecniche e metodologie di data mining sull'enorme quantità di dati condivisi nei vari social.
Quello che interessa è quindi estrarre ed analizzare questa grande quantità di sapere contenuta all'interno dei social networks, dove, al giorni d'oggi, moltissimi utenti condividono le proprie esperienze, i propri pensieri e opinioni ed i propri sentimenti ed umori.
Le tecniche di data mining risultano particolarmente interessanti, in quanto, grazie a specifici algoritmi, analizzano grandi quantità di dati e restituiscono informazioni e conoscenza (non dati).
La maggior parte delle persone dedica molte ore alla navigazione sui vari social networks e alla condivisione di ormai ogni aspetto della propria esistenza.
Esistono persone più riservate, che utilizzano questi strumenti a solo scopo comunicativo, ma una grande maggioranza sembra rinunciare a buona parte della propria privacy per scrivere sui social.
Dalla nascita di questi strumenti è sicuramente interesse di molti analizzare il comportamento degli utenti e l'evoluzione di questo comportamento online.
In ambito generale si ha sempre di più la necessità, la curiosità e l'utilità di monitorare gli utenti, capire di cosa parlano, cosa fanno, cosa dicono e come lo dicono e tantissimi tipi diversi di ricerche e analisi sono state fatte sui social e continueranno ad essere fatte in futuro.
Esistono appunto tantissimi ambiti di ricerca possibili per cui può essere interessante sfruttare il grande potenziale racchiuso nei social networks e tantissimi tipi diversi di utenti da analizzare. Negli ultimi anni soprattutto si è vista l'evoluzione di un nuovo tipo particolare di utenza che ha preso sempre più piede e si tratta dei "pazienti".
Per paziente intendiamo una persona con problemi di salute, che esegue ricerche, cerca consigli e rimedi, spera in un supporto morale e racconta tutto

riguardo la sua malattia.
Grazie all'evoluzione enorme ed in pochissimo tempo di questo tipo di utenza, sono nate vere e proprie comunità dedicate a certe malattie sui vari social.
I problemi che ricevono più attenzione e che favoriscono la crescita di queste comunità sono sicuramente le malattie croniche, che interessano maggiormente gli utenti/pazienti essendo problemi sgradevoli con cui devono imparare a convivere per tutta la vita e proprio questi pazienti non smetteranno mai di essere alla continua e disperata ricerca di qualsiasi modo per alleviare i sintomi o addirittura curarli.
Proprio queste comunità di pazienti cronici, costituiscono emormi fonti di informazioni e conoscenza, come anche enormi fonti di pericoli e problemi.
Se usati in modo corretto infatti i social networks possono rappresentare luoghi di conforto, di scambio di esperienze, di consigli e di stati d'animo, ma se usati in modo scorretto, possono rappresentare luoghi di comunicazione di massa di informazioni non veritiere e di bassa qualità. Possono influenzare enormemente le persone, soprattutto quelle fragili, che non hanno molto da perdere ed è per questo che nasce anche la necessità di tenere sotto controllo e di monitorare quello che succede online.
Algoritmi e metodologie di data mining, sui dati pubblicati sui social networks, sono quindi utili sia per scoprire il comportamento degli utenti, per analizzare i discorsi più interessanti e il sentimento espresso dalle persone, sia per monitorare e controllare la situazione e le informazioni scambiate, soprattutto in ambiti delicati come quello medico.
In questo specifico caso, l'ambito medico studiato in questo lavoro di tesi è quello del morbo di Crohn, che è una malattia infiammatoria intestinale (IBD) cronica, che può colpire qualsiasi parte del tratto gastrointestinale, dalla bocca all'ano, provocando una grande varietà di sintomi [21] e in alcuni rari casi anche essere fatale.
Prima di proseguire con la descrizione delle metodologie utilizzate per ottenere i risultati voluti, è interessante analizzare i precedenti lavori, che analizzano vari aspetti dei social networks e lo stato dell'arte delle pratiche fino ad ora ritenute idonee per queste analisi.

In ambito generale, diversi studi sono stati già effettuati sul potere di informazione racchiuso nei vari social.
Già nel 2003, uno studio del dipartimento di informatica dell'Università del Massachusetts [12], mostra come si affrontavano le tecniche di data mining applicate ai social.
Questo studio indica tecniche per l'apprendimento di modelli statistici, cioè l'uso di algoritmi per imparare da dati relazionali.
Le tecniche di data mining analizzate ed utilizzate per questi scopi sono:

- modelli relazionali probabilistici;
- logica Bayesiana;
- alberi di probabilità relazionali.

Queste tecniche vengono applicate ai dati reali per costruire i modelli e i suoi parametri, in modo da poter poi utilizzare questi modelli appresi per dati futuri.

Uno studio approfondito dei social più utilizzati è stato pubblicato nel libro [1]. Qui sono analizzati tre social in particolare, Twitter, Facebook e LinkedIn e sono paragonati a vecchi metodi di comunicazione online ormai scomparsi come MSN Messenger. Notevoli sono le differenze tra questi nuovi mezzi e quelli precedenti e ognuno è differente dagli altri per vari aspetti che vengono chiaramente messi in luce. Durante la caratterizzazione dettagliata di questi strumenti specifici, viene indicato per ognuno, quali sono le informazioni che è possibile estrarre e come farlo. Sono elencati i vari tipi di analisi possibili, tra cui, analisi strutturali (a rete) basate sui collegamenti e analisi dei contenuti, analisi dimaniche ed analisi statiche. In sintesi questo libro rappresenta una vera e propria guida alla scoperta dei social e delle potenzialità in essi racchiuse.

Nel 2010 uno studio di tesi dell'Università di Bologna [15] ha studiato i profili virtuali che gli utenti creano nei vari social networks ed è stato progettato e realizzato un sistema, che per un dato utente, recupera le istanze dei suoi profili virtuali sui vari social esistenti, in modo da integrarne gli attributi e ricavarne un profilo unico che sia il più possibile completo ed esaustivo. Ogni social è specializzato in qualcosa di particolare e per questo gli utenti che utilizzano molti di questi strumenti non si comporteranno nello stesso modo e non condivideranno le stesse informazioni personali su tutti. Poter risalire ad un profilo virtuale completo ed esaustivo di ogni utente analizzato, raggruppando le informazioni fornite in tutti i social possibili crea un profilo chiamato "profilo mediale" che è estremamente utile sia per indagini di tipo economico sia di tipo sociologico.

Un tema ampiamente discusso, che ha interessato parecchie ricerche, è il tema della diffusione delle informazioni sui social.
In particolare, in uno studio del 2007 [13], sono indicati approcci per lo studio

dei dati sui social, con riferimento ai modelli del processo di ricerca dei dati ed il processo appunto di diffusione. Tutti i dati vengono espressi con reti del tipo "small-word", che sono dei tipi di grafi matematici dove pochi nodi sono direttamente collegati tra di loro, ma tutti i nodi possono comunicare con gli altri nodi con pochi passi di distanza. I social networks possiedono le caratteristiche delle reti "small-word", addirittura su Facebook è stato calcolato il grado di separazione tra tutte le coppie di individui ed in media, i gradi di separazione sono 4.74.
Questo significa che il 92% delle coppie di persone è separato da non più di 4 gradi.

Continuando ad analizzare gli studi fatti in ambito di diffusione, in [7], vengono creati altri modelli di influenza, utilizzando soltanto il grafo della rete sociale e registrando le azioni degli utenti che ne fanno parte.
Vengono proposti algoritmi per apprendere i parametri del modello sociale e testare questo modello imparato per fare previsioni sui dati futuri. Questi modelli appresi dovrebbero riuscire anche a predire il momento futuro nel quale l'utente svolgerà una certa azione.
Per validare la metodologia di questo studio sono stati analizzati i dati del social Flickr, sono stati estratti 1300000 nodi del grafo con 40000000 di legami tra questi nodi e più di 300000 azioni distinte degli utenti ed è stato dimostrato che ci sono reali influenze tra le azioni sui social.

In [18], si studia l'infuenza che certi nodi (utenti) possono avere su altri e per quali motivi. Ad esempio si è notato che gli amici hanno una forte influenza sulla vita privata, mentre i colleghi hanno influenza sul lavoro. Vengono distinte le varie influenze e ne viene stimata la forza, per grandi reti reali. Propongono il modello TAP (Topical Affinity Propagation) per modellare l'influenza sociale a livello di topic, cioè a livello di singoli argomenti che certi gruppi influenzano. TAP utilizza algoritmi di apprendimento distribuiti e viene dimostrata la sua efficienza in questo lavoro.

Ultimo lavoro centrato sull'influenza è del 2012 [2], quantifica l'effetto di questi mezzi per la diffusione delle informazioni e per l'identificazione di chi influenza chi.
Sono stati presi in considerazione 253 milioni di soggetti divisi tra utenti che utilizzano i social e utenti che invece non ne fanno uso ed è stato scoperto che coloro che utillizzavano queste tecnologie entravano in possesso di determinate informazioni molto prima degli altri. Altra verità scaturita da questo studio, è che, sebbene i pochi legami veramente forti presenti online siano sicuramente più influenti, la velocità di propagazione delle informazioni è do-

vuta invece in gran parte alla più elevata quantità di legami deboli presenti. Questo suggerisce che i legami deboli svolgono un ruolo più dominante per la diffusione delle informazioni.

Altro aspetto che interessa notevolmente l'analisi dei social networks è il sentimento che gli utenti mostrano nei loro discorsi. Particolarmente interessante e utile, l'umore delle persone è un'informazione da sempre ritenuta preziosa.
Metodologie interessanti sono state pubblicate in [3], dove è riconosciuto che le emozioni possono influenzare profondamente i comportamenti individuali e il processo decisionale. Si è indagato se è quindi possibile che l'umore pubblico espresso online possa essere correlato o addirittura predittivo di indicatori economici.
Sono stati analizzati grandi quantità di tweet su Twitter e si è cercato di verificare se lo stato d'animo collettivo espresso fosse correlato al valore dell'indice "Dow Jones Industrial Average (DJIA)" e alla sua evoluzione nel tempo.
Per estrarre il sentimento sono stati utilizzati due diversi strumenti:

- OpinionFinder, che mappa il sentimento in tre polarità, come positivo, negativo o neutro;
- Google-Profile of Mood States (GPOMS), che misura l'umore in 6 dimensioni (calma, allerta, sicurezza, vitalità, gentilezza e felicità).

Per validare l'utilizzo di questi sentimenti, come predittori di altri eventi, è stata misurata la relazione tra le serie temporali degli umori del pubblico nel periodo prima delle elezioni presidenziali del 2008 e i risultati di queste elezioni ed è stato mostrato come le distribuzioni fossero altamente correlate al periodo ed agli eventi che stavano accadendo.
Utilizzando le analisi di causalità di Granger è stato allora utilizzato il sentimento estratto dai tweet per predire l'andamento del titolo DJIA e la precisione delle predizioni si è aggirata attorno all'87.6%.

Poco prima che uscisse lo studio precedente era stato già mostrato interesse in questa stessa direzione, e l'articolo [6] mostra come già si cercasse una relazione tra l'umore espresso online e l'andamento dei titoli finanziari.
In questo caso le emozioni stimate erano ansia, preoccupazione e paura ed oltre 20 milioni di post hanno confermato che l'umore "ansia" Granger causa

l'abbassamento dell'indice S&P 500.

Fino ad ora sono stati analizzati i lavori svolti in ambito generale sui social networks, ma è anche presente un'ampia bibliografia di studi effettuati in ambito medico, che come è stato detto in precedenza, diventa sempre più importante ed interessante, sia per medici che per informatici.
Presentiamo quindi le ricerche che hanno aperto la strada alle analisi mediche sui social ed insieme ad una breve descrizione di tutti questi lavori verranno sottolineate le differenze e gli aspetti innovativi di questo lavoro di tesi.
L'evoluzione di strumenti elettronici di sorveglianza medica (EHR) ha contribuito insieme ai social networks alla creazione di grandi quantità di dati (BigData) con alto potenziale. Nel seguente studio del 2013 [16], sono indicati gli approcci innovativi utilizzati per l'archiviazione dei dati, la loro pre-elaborazione, l'analisi e la visualizzazione dei dati e delle informazioni.

Molto interesse è stato mostrato sugli effetti collaterali dei farmaci e le interazioni tra questi, sul comportamento del paziente e sulla rilevazione ed il monitoraggio di infezioni, interesse che è mostrato sia da scienziati che dai media che hanno sfruttato per questo Internet ed altri strumenti informatici.
Per esempio, nel 2012 Google pubblicò una sua ricerca [5], dove mostrava la possibilità di monitorare la diffusione dell'influenza negli Stati Uniti senza il bisogno di eseguire check-up medici.
Google ha impiegato un solo giorno ad eseguire questo monitoraggio, mentre il "Centro di controllo e prevenzione malattie" (CDC) ci ha impiegato più di una settimana. Il metodo utilizzato da Google si basava sull'analisi di correlazione tra quello che le persone cercavano online e se digitavano sintomi influenzali.
Un commento sulla precisione, velocità ed economicità del metodo trovato da Google è espresso nel seguente articolo [11].

In [20], è stato ipotizzato che gli utenti online possono fornire i primi indizi sugli eventi avversi dei farmaci tramite le loro ricerche di informazioni. Sono stati analizzati i termini utilizzati nelle ricerche web per monitorare e rilevare gli effetti collaterali di farmaci. In pratica, quando un paziente scopre di avere un problema inzia a fare ricerche online correllate ai sintomi ed i disturbi che presenta ed ai farmaci che utilizza.
Sfruttando questo modello di comportamento è stato condotto uno studio su larga scala nel corso del 2010 ed è stato scoperto un effetto negativo, in particolare l'iperglicemia, provocato dall'interazione tra due diversi farmaci: Paroxetina e Prevastatina.
Un anno dopo la pubblicazione di questo studio, una ricerca medica approfon-

dita ha confermato questa interazione negativa, validando il modello usato. Questo fatto dimostra come le attività di ricerca online possano contribuire a sorvegliare la sicurezza dei farmaci.

Un ulteriore studio in [17], invece, ha analizzato l'uso di Twitter per monitorare le malattie nel tempo, in particolare le misure di sanità pubblica. Sono stati misurati i fattori di rischio comportamentali, localizzando le malattie per area geografica e sono stati analizzati i sintomi e l'uso di farmaci. Utilizzando il modello "Ailment Topic Aspect Model" sono stati analizzati oltre un milione e mezzo di tweet riguardanti la salute e con questo lavoro di sorveglianza è stato fatto per scoprire malattie, conoscere i sintomi e le eventuali associazioni tra sintomi e trattamenti.
È interessante notare che gli autori hanno individuato e seguito sette categorie diverse: allergie, insonnia, obesità, lesioni, problemi respiratori, problemi dentali e il dolore. Per ciascuna di queste categorie sono stati individuati i termini generici correlati (ad esempio, gli occhi, il sonno, il sangue, ecc) nonché i termini più comuni utilizzati per parlare di determinati sintomi e trattamenti.
Da questa ricerca è emerso che Twitter è un ottimo strumento per ricerche sulla salute pubblica.

Weitzman et al. in [19], hanno analizzato, monitorato e controllato le informazioni scambiate online sui social networks in luoghi focalizzati a trattare tutto quello che riguarda la malattia diabete. Sono stati estratti 28 indicatori di qualità e sicurezza dei siti rivolti a questo argomento che ne attestavano:

- l'allineamento dei contenuti con le informazioni ufficiali della scienza sul diabete e delle raccomandazioni cliniche;
- le procedure di sicurezza, come la reivsione dei contenuti pubblicati, la moderazione e la trasparenza;
- l'accessibilità alle politiche di privacy e le comunicazioni e il controllo dei rischi legati alla privacy;
- la condivisione centralizzata dei dati degli utenti e il controllo di questa.

I risultati di queste analisi hanno rivelato un 50% dei casi allineati con le raccomandazioni fornite dalla scienza sul diabete. Solo il 20% dei casi ha utilizzato sistemi di revisione esterna, mentre il 70% si basava su revisione

interne (come per esempio la moderazione). Le politiche di privacy in quasi tutti i casi hanno riscontrato scarsa leggibilità che impedisce la comunicazione del rischio. In conclusione, la qualità e la sicurezza riscontrata sui social è variabile e questo suggerisce l'addozione di meccanismi di miglioramento.

Altro studio centrato sul diabete, in [9], si concentra sulla valutazione della qualità delle comunicazioni tra i membri di comunità su Facebook dedicate al diabete. Sono stati analizzati i quindici gruppi maggiori trovati su Facebook ed i risultati mostrano dati interessanti: 480 utenti diversi che intervengono, 690 commenti analizzati, il 67% dei messaggi riguardano la condivisione non richiesta di strategie, oltre il 13% dei post rappresenta invece un feedback specifico richiesto da altri utenti, il 29% dei post rappresenta un sostegno emotivo e il 27% dei post riguarda attività promozionali come testimonial pubblicitari.
In conclusione Facebook rappresenta un luogo di segnalazione di esperienze personali, di domande e di comunicazione diretta con altre persone affette dal diabete. Tuttavia, l'attività promozionale e la raccolta dei dati personali è presente senza alcuna misura di sicurezza o di controllo di autenticità.

Tutti gli studi indicati affrontano quindi sfide generiche legate alle potenzialità di conoscenza racchiuse all'interno dei social networks e sfide specifiche legate alla discussione delle informazioni mediche sul web, ma nessuno, per una data malattia, affronta un'analisi completa che comprende diversi aspetti:

- Come differiscono le informazioni condivise su un social rispetto ad un altro?
- Quali argomenti sono più discussi?
- Quali sono le relazioni tra i diversi argomenti?
- Qual'è l'umore generale trovato su un social?

Tutti questi aspetti vengono considerati e analizzati in questo lavoro di tesi che si differenzia quindi dai precedenti lavori.

Anche questo lavoro ha già portato alla scrittura di un draft che potrà essere pubblicato a breve [4] e che mostra le metodologie e i risultati più salienti riguardo al tema medico specifico del morbo di Crohn.
I social su cui ci si è focalizzati sono Facebook e Twitter ed in questi sono stati estratti i dati riguardanti il morbo e sono stati analizzati i comportamenti dei pazienti, i discorsi fatti e quindi gli argomenti più discussi. Le informazioni estratte sono state mappate in topic particolari, come cause, sintomi, trattamenti ed effetti collaterali e particolare attenzione è stata data alle relazioni trovate tra questi topic.
Può essere molto rilevante in campo medico scoprire come le persone legano le possibili cause del morbo ai sintomi che gli si presentano, o come trattano i vari sintomi, con quali tecniche e quali medicinali ed infine come reagiscono alle varie cure.
Per terminare quest'analisi approfondita, molto lavoro è stato fatto per poter estrarre il sentimento dei pazienti nel discutere i vari aspetti della malattia. Come parlano dei vari trattamenti? Quali sono più apprezzati?
Anticipando brevemente i risultati più salienti, è stata trovata una grande attività su Facebook e Twitter riguardo al morbo di Crohn e questo indica che è un tema particolarmente interessante ed ampiamente discusso già anche sui social. Mentre su Facebook l'indagine ha portato all'analisi dei dati di tre anni, da ottobre 2011 a ottobre 2013, per Twitter le indagini si riferiscono a pochi mesi, da maggio 2013 a ottobre 2013, per via dei limiti delle API di questo strumento.
Per quanto il tema sia ampiamente discusso, purtroppo la costanza degli utenti non sembra elevata, nel periodo preso in analisi sono presenti pochissimi autori che scrivono continuatamente e tanto, mentre la maggior parte ha scritto solo qualche volta. E' stato però rilevato che quasi tutte le volte che un utente riscrive lo fa nello stesso giorno del post precedente o il giorno dopo, pochissime volte si riscrive con una distanza di più di una settimana, la costanza di chi scrive più volte è quindi elevata.
I temi discussi su Facebook e Twitter presentano alcune differenze, in particolare su Facebook gli utenti discutono enormemente di sintomi e trattamenti, con una leggera predilizione per i trattamenti, mentre in bassissima percentuale parlano di cause ed effetti collaterali. Su Twitter invece è presente un forte bias verso i sintomi che occupano quasi interamente l'interesse delle persone.
Ma gli utenti che discutono di questi temi, possiamo chiamarli "pazienti"?
Esaminando le caratteristiche dell'utenza è stato scoperto che su Facebook la quantità di pazienti è molto elevata e le persone che scrivono di più sono quasi tutte affetti dal morbo. Su Twitter invece la quantità di pazienti è molto ridotta e a scrivere sono persone che fanno pubblicità, parlano di donazioni

ed eventi legati al morbo e pubblicano riferimenti ad articoli interessanti.
La qualità dei dati trovati su Facebook è quindi nettamente superiore rispetto a quella trovata su Twitter.
Su Facebook oltretutto le distribuzioni delle relazioni tra i termini seguono un andamento "power law" che sappiamo essere tipico dei linguaggi naturali, mentre su Twitter ci sono delle interferenze dovute al meccanismo di retweet che permette di ripubblicare stessi tweet senza modifiche nel testo e viene troppo spesso usato.
Tra le relazioni fra i vari termini dei quattro topic, estratte sia con metriche di distanza sia con calcolo di probabilità condizionate, ne elenchiamo alcune particolari per Facebook essendo qualitativamente migliore:

- la malattia e i problemi vengono collegati a cause genetiche, al sistema autoimmune e a batteri;
- alcuni cibi peggiorano i sintomi relativi a problemi di digestione e di dissenteria;
- con problemi d artrite viene discusso il farmaco Methotrexate;
- nausa e febbre sono associati a Mercaptopurine;
- la presenza di sangue nelle feci necessita di eseguire una colonscopia;
- problemi al retto e fistole richiedono interventi chirurgici, tra cui la colostomia;
- alta è la discussione di trattamenti naturali con fermenti, calcio, aloe e vitamine, soprattutto in presenza di carenze dovute alla frequente dissenteria;
- del farmaco Entocort vengono discussi gli effetti collaterali.

L'ultima analisi, innovativa in ambito medico, riguarda l'umore degli utenti nella discussione dei vari trattamenti.
Per fare questo, facendo riferimento alla letteratura in materia di analisi semantiche e di causalità tra distribuzioni, si è scelto di utilizzare OpinionFinder per estrarre il sentimento dei post e la Granger causality per confrontare il sentimento con i farmaci e trovare possibili causalità tra questi.
I cluster di farmaci analizzati sono: Metilprednisolone, Prednisone, Beclometasone dipropionato, Budesonide, Mesalazina, Azatioprina, Metotrexate,

Mercaptopurine, Infliximab, Adalimumab, Certulizumab e Natalizumab. In ogni cluster sono presenti tutti i nomi possibili utilizzati per questi farmaci con questi principi attivi.
In estrema sintesi, sono stati notati i seguenti sentimenti per i trattamenti:

- Infliximab, Mathotrexate, Azatioprina e Mesalazina presentano sentimenti positivi, ma Infliximab è legato anche a quelli negativi;
- Certulizumab e Prednisone sono legati a sentimenti negativi;

Dopo questa prima panoramica, nei capitoli successivi verranno mostrati approfonditamente tutte le metodologie utilizzate, i relativi risultati ottenuti sul tema specifico del morbo di Crohn e la progettazione e realizzazione di un sistema software in grado di eseguire le tecniche, le metodologie e gli algoritmi in modo automatico e generale su qualsiasi tema voluto.

Capitolo 3

Metodologia

La metodologia utilizzata in questo lavoro di tesi è partita dal tentativo di rispondere a tre domande fondamentali:

- Qual'è la popolarità della malattia di Crohn all'interno di determinati social network? In altre parole quanto è attivo il dibattito su Crohn online?
- Quali sono gli argomenti più popolari online? Di cosa si parla di più?
- Qual'è lo stato d'animo di chi scrive online? Cosa pensa dei vari trattamenti provati?

La prima decisione fondamentale presa è stata quella della rete sociale dentro la quale fare questo studio e grazie ad un processo iniziale di analisi si è scelto di concentrarsi su due social network principali, Facebook e Twitter, e di analizzare le persone che parlano in inglese avendo molto più materiale a disposizione.
La scelta dei social è giustificata dal fatto che Facebook e Twitter sono i due social networks più popolari ed utilizzati al momento ed anche se questi non sono forum specializzati, dedicati esclusivamente alla discussione di tale disturbo, queste sono le reti sociali che oggi ricevono la maggior quantità di attenzione da parte degli utenti online.
La scelta della lingua inglese è giustificata dal fatto che solamente nel nord America le persone colpite dal morbo di Crohn sono in più di 600000 [14].
Tutti i dati presentati saranno quindi mostrati in lingua inglese.

Dopo aver stabilito la rete sociale di riferimento vediamo in dettaglio la metodologia utilizzata per rispondere alle tre domande poste.

3.1 Analisi dell'attività su Facebook e Twitter

Al fine di individuare i post che trattano la malattia di Crohn sono stati adottati due approcci diversi per i due social.
Per Twitter, abbiamo cercato tutti i tweet che includevano nel testo la parola crohn o l'hash tag #crohn ed abbiamo ripetuto questa ricerca ogni settimana per cinque mesi. Il motivo di queste continue estrazioni di tweet è che le API di Twitter hanno un limite temporale di soli 9 giorni per estrarre i tweet scritti in questo arco temporale.
Per Facebook, invece, sono state cercate tutte le pagine pubbliche che trattano il morbo di Crohn (contengono il termine crohn nel nome della pagina) e sono stati estratti ed analizzati tutti i post trovati all'interno di queste pagine il cui limite è di al massimo 5000 post estratti per pagina che risulta molto meno vincolante.
Durante questo lavoro di estrazione, i dati che sono stati salvati sono:

- Il testo del post o tweet;
- L'identificativo del post o tweet;
- L'identificativo dell'autore del post o tweet;
- La data in cui è stato scritto il post o tweet;
- Il numero totale di parole utilizzate in quel post o tweet;
- L'identificativo della pagina Facebook dove si trova il post;
- L'identificativo del primo post scritto su Facebook di cui questo post è una risposta.

Grazie a tutti questi dati estratti e salvati è possibile effettuare molte statistiche su quanto sia attivo il dibattito su Facebook e Twitter, per esempio:

- il numero totale di post, tweet e relativi commenti scritti in un determinato arco temporale;
- il numero di autori diversi che scrivono;
- il numero di post o tweet scritti da uno stesso autore;

- la distribuzione temporale con cui ogni autore riscrive;
- il numero di post pubblicati (non commenti);
- il numero di commenti ad altri post;
- la caratterizzazione degli autori più prolifici;

Tutte queste prime statistiche sono fondamentali per inquadrare la popolarità della malattia di Crohn su Facebook e Twitter ed i risultati relativi verranno illustrati nel capitolo successivo dedicato a tutti i risultati ottenuti.

3.2 Analisi dei topic

Per comprendere la terminologia e gli argomenti legati alla malattia di Crohn è stata effettuata una collaborazione con un team di medici del Sant'Orsola e sono stati individuati quattro gruppi di argomenti interessanti che sono diventati i topic fondamentali al centro delle analisi:

- **Cause** - tutto quello che potrebbe influire sulla comparsa e sul peggioramento del morbo
- **Sintomi** - tutto quello che si può manifestare quando si ha il morbo
- **Trattamenti** - qualsiasi farmaco, intervento o tecnica volta a diminuire i sintomi del morbo
- **Effetti collaterali** - qualsiasi sintomo negativo portato non dal morbo ma dal trattamento effettuato

Questi gruppi di argomenti hanno delle relazioni implicite tra di loro, rappresentate nel grafo orientato in figura 3.1. La logica espressa è che alcuni tipi di cause possono influenzare i sintomi del morbo, con il manifestarsi di alcuni sintomi si ricorre a trattamenti specifici e la pratica di qualche trattamento può portare al manifestarsi di qualche effetto collaterale.

Figura 3.1: Relazioni tra i topic legati al morbo di Crohn

Come misuriamo la presenza o meno di queste relazioni? La misura delle relazioni è basata sulla frequenza con cui due argomenti si trovano insieme negli stessi post e quindi sulla loro co-locazione come espresso in figura 3.2.

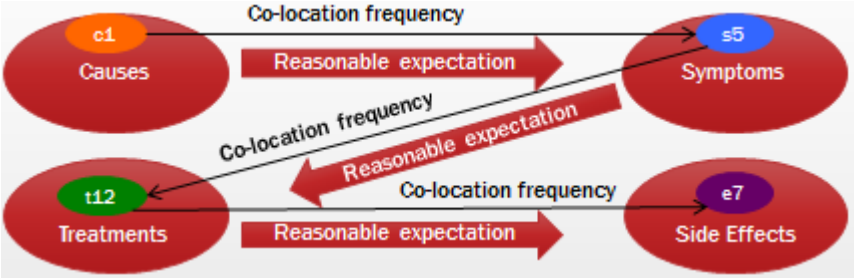


Figura 3.2: Frequenza di co-locazione come metodologia

Risulta quindi necessario stabilire per primo come affermare se un determinato post parla o meno di uno o più dei quattro argomenti. Per fare questo sono stati costruiti quattro dizionari distinti (uno per ogni topic di interesse) ed è stata modellata la presenza/assenza dei termini nel post per dedurre se tratta o meno i vari argomenti.
In figura 3.3, sono elencati tutti i termini trovati nei due social e considerati membri dei quattro argomenti fondamentali.
Come previsto, sono stati trovati più termini per Facebook rispetto a quelli di Twitter, ma questo è giustificato dal fatto che i tweet sono limitati in lunghezza e potendo essere lunghi solo 140 caratteri si riassumono tutti i concetti utilizzando meno parole.
Per quanto riguarda il dizionario delle "Cause" (Causes) del Morbo si nota come spesso vengano fatti riferimenti nutrizionali (per esempio latte, uova, cioccolata, etc.) e come sappiamo ci sono appunto alimenti che possono influenzare negativamente i sintomi della malattia.
Nel dizionario dei "Sintomi" (Symptoms) sono presenti i vari effetti che si possono manifestare con la malattia, troviamo infatti termini che riguardano specifici sintomi (per esempio febbre, diarrea, etc.), termini che riguardano parti del corpo (per esempio retto, intestino, etc.) e termini che riguardano stati di dolore e di angoscia (per esempio soffrire, depressione, etc.).
Per quanto riguarda la categoria "Trattamenti" (Treatments) vengono utilizzati termini specifici di nomi di farmaci (per esempio Azatioprina, Infliximab, Humira, Pentasa, etc.), termini che indicano interventi chirurgici (per esempio ileostomia, resezione, etc.) e vengono anche utilizzati termini più generici relativi alle cure mediche (per esempio diagnosi, visita, specialista, dose, intervento, etc.).

In fine, il dizionario degli "Effetti collaterali" (Side effects) presenta meno termini rispetto agli altri topic e per lo più generali (per esempio effetti, allergia, complicazioni, etc.) e solo per Facebook viene utilizzato un termine specifico di patologia (lupus).

	Facebook	Twitter
Causes	Food, cause, sugar, milk, bacteria, autoimmune, virus, smoke, chocolate, meat, coffee, wine, drink, tnf, alcohol, cell, syndrome, honey, gene, butter, cake, map, pasta, eggs, lactose.	Smoking, cause, virus, food, eat, cows, bovine, celiac, parkinson, bacterium, tourette, lupus, meat, milk, epstein, gluten, stress.
Symptoms	Disease, colitis, bowel, mouth, flare, symptom, pain, arthritis, severe, fistula, diarrhea, intestine, gastro, liver, ulcer, bleed, belly, cramp, illness, physical, damage, fever, rectum, suffer, bone, muscle, abdomen, constipation, exhausted, vomit, hurt, grow, blood, agony, leg, abscess, toilet, psoriasis, nausea, bladder, inflamed, anal, attack, stress, depression, anxiety, flu, infection, scar, sleep, fever, tired, digestion, appetite, disorder, deficiency, diabete.	Disease, colitis, ulcerative, suffer, symptom, bowel, inflammation, pain, cancer, diagnosed, stomach, flare, cobblestone, chronic, intestine, purple, blood, gi, disorder, leukemia, issues, problem, arthritis, fibromyalgia, severe, bipolar, colon, mouth, syndrome, autoimmune, fibrosis, esophagus, epilepsy, dysfunction, cystic, ms, colorectal, disability, illness, irritable, digestive, bathroom, depression, appetite, weight, flu.
Treatments	Diet, surgery, hospital, cure, humira, doctor, vitamin, infliximab, medication, ileostomy, test, scd, gp, treat, natural, colonoscopy, infusion, resection, azathioprine, probiotic, nutrition, surgeon, antibiotic, injection, imuran, pentasa, prescription, aloe, operation, colostomy, therapy, asacol, kefir, paleo, cannabis, ginger, methotrexate, healthcare, prescribed, steroid, entocort, oral, solution, dose, mercaptopurine, marijuana, prednisolone, enzyme, morphine, powder, ferment, remedy, capsule, diagnosis, specialist, visit, calcium, drain, mri, transplant, fda, organic, reversal, colectomy, medicine, cimzia, pill.	Marijuana, remission, cannabis, vitamin, cure, diet, treat, inhaled, fight, check, surgery, drug, supplement, healthy, hemp, benefit, therapy, soligenix, fioricet, care, endoscopy, weed, med, sgx203, mmj, infliximab, placebo, biologic, vaccine, qu, healing, glpg, humira, adalimumab, pill, clinical, ileostomy.
Side effects	Skin, effect, reaction, allergy, complications, head, lupus.	Effect, allergy.

Figura 3.3: Dizionario per Crohn su Facebook e Twitter

La presenza/assenza di tutti questi termini dei vari topic, è stata modellata attraverso la costruzione di un vettore binario per ogni termine. Ogni vettore binario è lungo quanto il numero di post analizzati, ogni riga del vettore rappresenta quindi un unico post e viene mappato 1 se in quel post il

termine considerato è presente, 0 altrimenti. E' possibile osservare un esempio concreto nella figura 3.4.

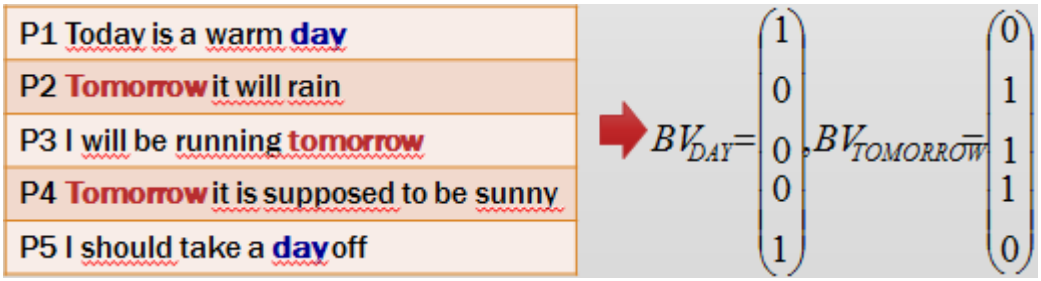


Figura 3.4: Esempio del modello a vettori binari

Grazie a questa modellazione dei dati è possibile effettuare molte statistiche sulla popolarità dei topic e sui post e tweet, per esempio:

- La distribuzione mensile della media giornialiera di post scritti con indicazione del rapporto tra gli argomenti trattati;
- La distribuzione delle discussioni classificate nei topic indicati, cioè qual'è il topic più discusso;
- Come varia la distribuzione dei topic se consideriamo i discorsi dei soli utenti più attivi sui social;
- Come varia la distribuzione dei topic se consideriamo i discorsi dei soli utenti mediamente attivi sui social;
- Come varia la distribuzione dei topic se consideriamo i discorsi dei soli utenti pochissimo attivi sui social;
- Come varia la distribuzione dei topic se consideriamo i soli post scritti per primi (cioè non i commenti);
- Come varia la distribuzione dei topic se consideriamo i soli post che sono risposte ad altri post;
- I termini del dizionario quante parole coprono nei discorsi degli utenti.

Tutte queste statistiche sono fondamentali per inguadrare la popolarità dei topic su Facebook e Twitter ed i risultati relativi verranno illustrati nel

capitolo successivo dedicato a tutti i risultati ottenuti.
Oltre alla popolarità dei topic, come già detto, l'obiettivo di questa tesi è anche quello di analizzare le relazioni che ci sono tra le varie categorie e per fare questo sono state utilizzate due analisi fondamentali: la distanza entropica e le reti bayesiane.
Infine per dimostrare che l'estrazione dei termini del dizionario mantiene invariate le proprietà del linguaggio naturale viene analizzata la distribuzione delle relazioni tra i termini e viene confrontata con la legge di potenza (power law).

3.2.1 Distanza entropica

Facendo ricorso alla grande quantità di letteratura esistente sulla teoria dell'informazione, la metrica accettata per misurare la relazione tra distribuzioni è quella di distanza.
La definizione classica di distanza si basa sulla definizione di entropia, ma questo non sembra soddisfare le nostre necessità, è stata quindi modificata questa definizione di distanza in modo da adattarsi meglio al nostro caso.

La distanza classica si comporta in modo che due vettori binari complementari presentino distanza pari a zero (per esempio, D1([0,1],[1,0]) = 0) e questo per noi non è corretto.
Nella formula della distanza classica 3.1, T1 e T2 sono i due termini di cui vogliamo misurare la relazione (sono quindi due vettori binari), H(x) è l'entropia della distribuzione x e I(x,y) è la mutua informazione tra le due distribuzioni x e y.

$$D1(T1, T2) = H(T1) + H(T2) - 2I(T1, T2) \tag{3.1}$$

La figura 3.5 mostra il comportamento della distanza classica per distribuzioni dove N è il numero di post diversi, si vede appunto che vettori binari complementari hanno distanza zero e che questa cresce all'aumentare del numero di termini che si sovrappongono.

La nostra nuova formula di distanza soddisfa la disuguaglianza triangolare, perchè $D2(x, y) \geq 0$ per ogni x,y e $D2(x, y) = D2(y, x)$, $D2(x, y) = 0$ se e solo se x=y.

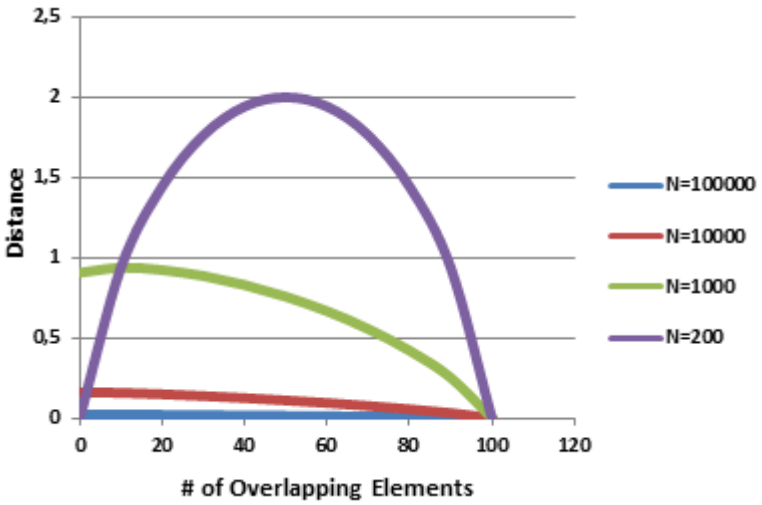


Figura 3.5: Distanza classica

Nella formula della nuova distanza 3.2, S(x) è la sparsità della distribuzione x e C(x,y) è il coseno di similitudine (o cosine similarity) tra le due distribuzioni di cui si sta studiando la relazione.

$$D2(T1, T2) = (1 + S(T1) + S(T2)) * C(T1, T2) \tag{3.2}$$

La figura 3.6 mostra il comportamento della nostra nuova distanza e si vede come la distanza diminuisca monotonicamente all'aumentare della compresenza dei termini nei post. Fissando il numero di sovrapposizione dei post, la distanza diminuisce all'aumentare del numero totale di post.
Grazie a questa nuova formula è possibile calcolare la distanza tra tutti i termini di un topic e tutti i termini di un altro topic e trovare le relazioni più forti tra diverse categorie.
Seguendo la logica di relazioni già introdotte ed indicate in figura 3.1 sono state calcolate le distanze tra i termini della categoria cause ed i termini della categoria sintomi, tra i termini della categoria sintomi ed i termini della categoria trattamenti, tra i termini della categoria trattamenti ed i termini della categoria effetti collaterali.
Se la distanza che lega due termini è minore di una certa soglia, allora si può affermare una relazione diretta tra i due termini dei due topic differenti. I termini legati da una distanza bassa saranno mostrati nel capito dedicato ai risultati.

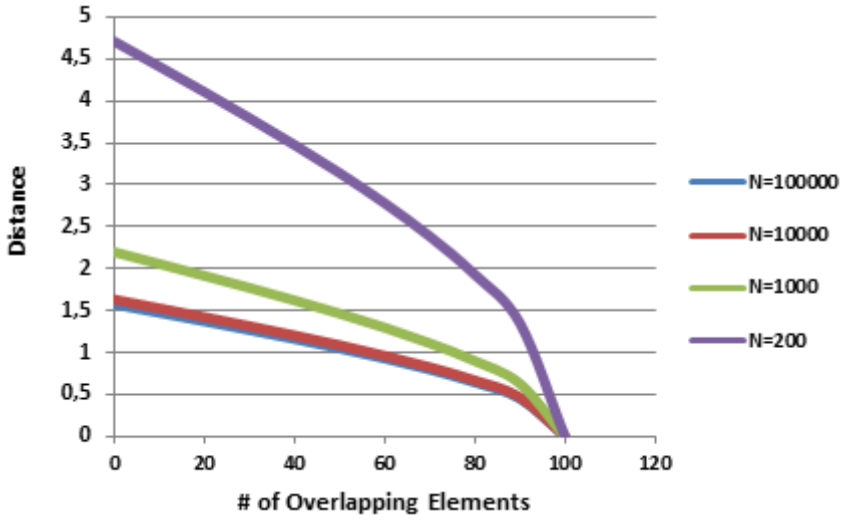


Figura 3.6: Distanza nuova

3.2.2 Reti Bayesiane

Seguendo sempre il pattern di relazioni espresso in figura 3.1, si è scelto di ricorrere ad un approccio Bayesiano, che consiste in un modello probabilistico che rappresenta un insieme di variabili stocastiche con le loro dipendenze condizionali.
In pratica l'approccio scelto è fondato sul calcolo di quanto è frequente la comparsa di un termine, sapendo che è presente un termine di un'altra categoria. Questo calcolo è più comunemente conosciuto come probabilità condizionata.
Si è quindi proceduto con il calcolo della probabilità dei termini contenuti nel topic "Syntoms", condizionati dalla presenza dei termini contenuti nei topic "Causes". In ugual modo si è verificata la probabilità dei termini contenuti nel topic "Treatments", condizionati dalla presenza dei termini del topic "Symptoms", ed infine, si è calcolata la probabilità dei termini del topic "Side effects", condizionati dalla probabilità dei termini del topic "Treatments".
I risultati di probabilità condizionata ottenuti sono considerati rilevanti per stabilire una relazione tra due topic, solamente se superano un valore di soglia di 0,25.
Questi risultati sono molto utili per comprendere quali tipi di informazioni vengono ricercati dai pazienti.
I risultati di questa analisi saranno sempre mostrati nel capitolo dedicato ai risultati.

3.2.3 Power Law

Le distribuzioni delle scienze sociali o degli eventi correlati all'attività umana sono note per seguire un tipo di andamento iperbolico specifico che è la Legge di Zipf.
Il linguaggio naturale dovrebbe quindi seguire questa legge, che è un tipo particolare di legge di potenza o power law.
Dal momento che in questo lavoro sono stati estratti termini ritenuti fondamentali per comprendere i discorsi degli utenti online, è possibile verificare se le relazioni tra i topic qui studiati seguono la power law e mantengono tutte le caratteristiche tipiche del linguaggio naturale.
Una distribuzione power law è del tipo $P(k) \sim k^{-y}$, è asimmetrica, con una lunga e pesante coda destra.

Per fare questo è stata creata una rete "termini-termini", che è una matrice quadrata simmetrica che ha per righe e per colonne i termini più significativi dei topic che si sono analizzati ed i valori della matrice stanno ad indicare quante volte il termine rappresentato dalla riga viene utilizzato insieme ad il termine indicato dalla colonna. Si tratta quindi di una matrice di co-occorrenza.
Osservando per ogni coppia di termini il loro valore di co-occorrenza, dovrebbe verificarsi che pochissime coppie presentano un valore elevato e tantissime coppie presentano un valore basso.
Vengono ordinate le coppie secondo il loro valore e viene stimata la frequenza con la quale si ripetono i diversi valori di co-occorrenza.
Per dimostrare questo andamento, oltre a graficare la distribuzione di frequenza, viene graficato anche un istogramma, su scala logaritmica log-log in modo da ottenere una linea retta e se i dati combaciano con questa allora si ha una power law.
Ultima conferma si ha utilizzando la funzione di stima di curve del software di statistica SPSS che stima l'andamento di una curva e lo paragona ad i dati osservati e volendo ovviamente controllare che la nostra distribuzione sia power law la curva scelta per stimare i dati raccolti è quella di potenza.
Tutti i risultati dell'analisi di distribuzione power law sono riportati nella sezione dedicata ai risultati e verranno mostrati sia per Facebook che per Twitter, cercando di capire se tra i due social network si presentino delle differenze.

3.3 Analisi del sentimento

I social networks sono luoghi dove non solo è possibile valutare quali temi sono più interessanti per le persone, ma anche come (e in che termini) le persone parlano di questi argomenti.
Quello che in questo lavoro di tesi si è voluto analizzare è quindi anche lo stato d'animo dei pazienti affetti dal Morbo di Crohn ed in particolare le considerazioni positive e negative che vengono fatte sui vari trattamenti possibili.
Grazie alla collaborazione con i medici sono stati individuati i trattamenti farmacologici possibili per il morbo in questione e sono stati raggruppati in categorie diverse in base ai principi attivi, in un unica categoria sono quindi presenti tutti i farmaci che utilizzano lo stesso principio attivo o che vengono commercializzati con nomi diversi.
I gruppi di farmaci analizzati sono:

- **Metilprednisolone**: methylprednisolone, a-methapred, artisone, besonia, depo-medrol, dopomedrol, esametone, firmacort, lemod, medesone, medixon, medlone, medrate, medrol, medrone, mesopren, metastab, methyleneprednisolone, methylprednisolone, methylprednisolonum, metilbetasone, metilprednisolona, metilprednisolone, metrisone, metrocort, metysolon, moderin, nirypan, noretona, promacortine, reactenol, sieropresol, solomet, solu-medrol, summicort, suprametil, urbason, urbasone, wyacort;

- **Prednisone**: adasone, ancortone, apo-prednisone, bicortone, cartancyl, colisone, cortan, cortidelt, cotone, dacorten, dacortin, decortancyl, decortin, decortisyl, dekortin, dellacort, delta cortelan, delta-cortisone, delta-dome, deltacortene, deltacortisone, deltacortone, deltasone, deltison, deltisona, deltra, diadreson, econosone, encorton, encortone, enkorton, fernisone, fiasone, hostacortin, in-sone, incocortyl, juvason, liquid pred, lisacort, lodotra, lodtra, me-korti, metacortandracin, meticorten, nisona, nizon, novoprednisone, nurison, orasone, origen, panafcort, panasol, paracort, parmenison, pehacort, predeltin, prednicen-m, prednicorm, prednicort, prednicot, prednidib, prednilonga, prednison, prednisona, prednisone, prednisonum, prednitone, prednizon, prednovister, presone, pronison, rayos, rectodelt, retrocortine, servisone, skprednisone, sone, sterapred, supercortil, ultracorten, ultracortene, winpred, wojtab, zenadrid;

- **Beclometasone dipropionato**: aerobec, alanase, aldecin, aldecina, aldecine, atomase, beclacin, beclate, beclazone, beclazone 250, beclo-

forte, beclomet, beclometasone, beclometasone dipropionate, beclometasone dipropionato, beclomethasone dipropionate, beclorhinol, becloturmant, becotide, korbutone, vanceril, viarox, clipper, topster;

- **Budesonide**: bidien, budeson, budesonide, budesonido, budesonidum, entocort, micronyl, preferid, pulmicort, pulmicort flexhaler, pulmicort respules, respules, uceris, entocir, rafton;

- **Mesalazina**: mesalazina, azulfidine, lialda, apriso, delzicol, mesavancol, pentacol, pentasa, enterasin, enteraproct, claversal, asavixin, asamax, asalex, asacol, asacol hd, dipentum, colazal, sulfazalazine, rowasa, canasa;

- **Azatioprina**: azamun, azanin, azasan, azathioprin, azathioprine, azathioprinum, azatioprin, azatioprina, azothioprine, immunoprin, imuran, imurek, imurel, muran, purine, rorasul;

- **Metotrexate**: a-methopterin, a-methpterin, abitrexate, alpha-methopterin, amethopterin, amethopterin l-, antifolan, brimexate, emtexate, emthexat, emthexate, farmitrexat, fauldexato, folex, lantarel, ledertrexate, lumexon, maxtrex, medsatrexate, metatrexan, metex, methoblastin, methotrexat-ebewe, methotrexate, methotrexate lpf, methotrexate l-, methotrexatum, methylaminopterin, methylaminopterinum, metotressato, metotrexato, metrotex, mexate, mexate-aq, novatrex, otrexup, rheumatrex, texate, tremetex, trexeron, trixilem, metotrexate;

- **Mercaptopurine**: 6mp, 6 mp, 6-mercaptopurin, 6-mercaptopurine, 6-merkaptopurin, 6-mp, 6-purinethiol, 6-thiohypoxanthine, 6-thiopurine, 6-thioxopurine, hypoxanthine thio, ismipur, leukerin, leupurin, mercaleukin, mercaptopurin, mercaptopurina, mercaptopurine, mercaptopurinum, mercapurin, merkaptopuryna, mern;

- **Infliximab**: infliximab, avakine, remicade, remsima, inflectra;

- **Adalimumab**: adalimumab, humira, trudexa;

- **Certulizumab**: certulizumab, cimzia;

- **Natalizumab**: natalizumab, tysabri

Per ogni gruppo elencato è stata creata una serie temporale che indica quante volte vengono menzionati i membri dei vari gruppi nei post. Viene

quindi utilizzata la stessa metodologia precedentemente indicata per i quattro topic fondamentali, che prevede la costruzione di vettori, questa volta non binari ma valorizzati, essendo possibile trovare nominati più farmaci di una stessa categoria in un unico post.
Ora, mentre è chiaro come si è proceduto per misurare la frequenza con cui gli utenti citano i vari farmaci, il problema di valutare l'umore degli utenti non è così banale. Al fine di effettuare questo tipo di analisi si è quindi deciso di fare ricorso a tecniche di analisi del sentimento già note ed ampiamente utilizzate.
Lo strumento utilizzato è OpinionFinder, un sistema in grado di elaborare un corpus di testo ed identificare la soggettività all'interno delle frasi, comprese le opinioni, le espressioni soggettive dirette, gli eventi linguistici e le espressioni del sentimento [22].
In breve, durante l'elaborazione di un post, OpinionFinder restituisce il sentimento di ogni frase che lo compone, classificandolo come neutro, positivo o negativo. Utilizzando OpinionFinder è quindi possibile creare la distribuzione del sentimento dei vari post, creando un vettore per il sentimento negativo ed uno per quello positivo.
Avendo la distribuzione del sentimento per tutti i post, è possibile misurare l'evoluzione dello stato d'animo generale degli utenti che affrontano questo tema del morbo di Crohn, e l'evoluzione del sentimento specifico espresso nei post dove si parla di un certo gruppo di terapie farmacologiche. Per visualizzare l'evoluzione del sentimento per ogni categoria in un unico grafico, si è scelto di calcolare il valore del sentimento come la differenza tra i sentimenti positivi e quelli negativi espressi in un arco di tempo (per esempio, giornalmente o settimanalmente). Questo vuole dire che se in un determinato arco di tempo OpinionFinder trova 10 sentimenti positivi e 4 sentimenti negativi, il valore del sentimento per quel periodo sarà 6.
Analizzando i grafici generati, quindi, un valore positivo corrisponde ad un sentimento prevalentemente positivo durante quell'arco di tempo, mentre un valore negativo corrisponde a sentimenti prevalentemente negativi.
Grazie ad un vocabolario di termini positivi e negativi è stato possibile anche analizzare quali sono i termini di entrambe le polarità più utilizzati in correlazione con i vari gruppi di farmaci.

Anche tra i diversi trattamenti e le possibili polarità di sentimento può essere interessante stimare una "distanza" per catturare delle possibili relazioni tra le due distribuzioni. La "distanza entropica" usata come metrica di distanza tra i termini delle categorie (cause, sintomi, trattamenti ed effetti collaterali) ed indicata precedentemente, non è idonea, così com'è, ad essere applicata anche in questo caso, ma sono necessarie piccole modifiche che

vengono precisate nella sezione apposita di "distanza entropica" applicata a questo nuovo scopo.

Un ulteriore passo avanti, dopo aver analizzato le opinioni delle persone sui vari trattamenti ed aver stimato le distanze è quello di verificare se la discussione di determinati trattamenti influenza la comparsa di certi sentimenti. Quello che in questo lavoro si è voluto quindi studiare è se è possibile affermare che certi trattamenti "causano" sentimenti positivi o negativi, cosa che non è possibile con il solo calcolo delle distanze. Le distanze trovano relazioni generali, ma non relazioni di causalità.
Questa analisi può essere molto importante in campo medico, perchè può rispondere ad una domanda che i medici si pongono quando suggeriscono un nuovo farmaco ad un paziente, ed è: con questo trattamento, quale sarà la condizione del paziente in futuro?
Per poter fare questo studio è necessario correlare due grandezze: la distribuzione del sentimento (positivo e negativo) e la distribuzione di ogni singolo trattamento considerato.
La correlazione però non prova la causalità, per questo si è mostrata la necessità di utilizzare uno strumento matematico che permettesse di stabilire l'esistenza di un rapporto di causalità tra due vettori, quello dei sentimenti e quello dei trattamenti.
E' stato deciso di utilizzare un metodo di analisi già noto, l'analisi di causalità di Granger, per valutare se le informazioni fornite da una distribuzione possono predire le informazioni su un'altra distribuzione.

3.3.1 Distanza entropica

La nuova formula di "distanza entropica" precedentemente mostrata, è ottimizzata per distribuzioni binarie, dove i valori sono solo 0 o 1, in quanto il termine compariva o non compariva in ogni post o tweet.
In questo caso, trattandosi di cluster di farmaci, più farmaci, appartenenti ad uno stesso cluster, possono essere menzionati nello stesso post, come a sua volta più sentimenti positivi o negativi possono essere trovati. Questo modifica le nostre distribuzioni da binarie a valorizzate ed obbliga ad una revisione della formula per meglio adattarla a questo caso.
La soluzione trovata è una via di mezzo, tra la classica formula di distanza espressa nell'equazione 3.1 e la nostra precedente nuova formula di distanza estressa nell'equazione 3.2.
La nuova "distanza entropica" applicata quindi a questo caso è mostrata nel-

la seguente equazione 3.3.

$$D2(T1, T2) = (1 + H(T1) + H(T2)) * C(T1, T2) \tag{3.3}$$

Come si può notare, la differenza sta nel sostituire il calcolo della sparsità $S(T)$ delle singole distribuzioni, con l'entropia classica $H(T)$ e di mantenere invariato invece il calcolo della cosine similarity.
Il motivo di questo scambio è che in questo caso è più utile misurare la quantità di incertezza o informazione presente in un segnale aleatorio più che la sparsità di una distribuzione che sappiamo già essere molto "sparsa".
Dal momento che anche l'entropia sappiamo funzionare meglio con le distribuzioni binarie, è stato deciso di considerare la nostra distribuzione valorizzata, come binaria per il calcolo dell'entropia (0 se non compare nessun termine di quel cluster, 1 se ne compare almeno 1), mentre è stata mantenuta la distribuzione valorizzata per il calcolo della cosine similarity.
In più la distribuzione usata per il calcolo della cosine similarity è stata normalizzata prima di eseguire il calcolo.
In questo modo, con questa nuova formula di distanza ibrida, i risultati, che saranno mostrati successivamente, risultano molto più accurati e precisi.

3.3.2 Causalità di Granger

La causalità di Granger, determina in maniera statistica una causalità tra variabili, e fornisce un'indicazione di quanto una serie temporale possa predire un'altra serie [8], [10].
La logica di questa analisi temporale si fonda sul fatto che se un evento Y accade prima di un evento X, allora è possibile che Y causi X, ma non è possibile che X causi Y. La causalità di Granger utilizza infatti modelli di regressione e grazie a questi afferma che X "Granger causa" Y se i valori passati di X possono spiegare quelli di Y.
Quando si esegue un'analisi di causalità di Granger bivariata, la dipendenza di una variabile su un'altra è stabilita prima di tutto grazie alla costruzione di due diversi modelli di regressione lineare.
Il primo modello di regressione lineare calcola il valore di X al tempo t (cioè $X(t)$) e questo valore dipende solamente dalla storia di X e da un rumore η.
Questo primo modello è espresso nell'equazione 3.4.

$$X(t) = \sum_{m=1}^{M} a_m X(t - m\Delta t) + \eta(t) \tag{3.4}$$

Il secondo modello di regressione lineare calcola il valore di X al tempo t dipendentemente da Y e da un rumore ν. Questo secondo modello è espresso nell'equazione 3.5.

$$X(t) = \sum_{m=1}^{M} a_m X(t - m\Delta t) + \sum_{l=1}^{L} b_l Y(t - l\Delta t) + \nu(t) \tag{3.5}$$

Dopo aver costruito i due modelli, è possibile affermare che Y causa secondo Granger X, se $Var(\nu) \ll Var(\eta)$, questo significa che quando i valori passati di Y sono considerati come predittori dei valori correnti di X, la varianza del rumore si riduce enormemente e quindi aumenta l'accuratezza con cui è stimata X.
I risultati ottenuti con un'analisi di causalità di Granger, comprendono analisi fatte con il metodo dei minimi quadrati e con il test-F. Questi metodi vengono utilizzati per testare la significatività statistica dei risultati generati e quindi servono per scartare l'ipotesi nulla. In particolare, l'indice analizzato è il p-value, che indica la probabilità di ottenere un risultato pari o più estremo di quello osservato, anche chiamato livello di significatività. Per poter respingere l'ipotesi nulla, questo livello di significatività deve essere almeno inferiore al 5% per poter affermare che la causalità secondo Granger è presente.
Chiaramente questo valore di affidabilità dipende dall'unità temporale scelta per la distribuzione Δt e per coerenza con le analisi precedentemente svolte tutto il lavoro di analisi della causalità di Granger è stato effettuato sia su distribuzioni settimanali che su distribuzioni giornaliere.
Per eseguire il test di causalità di Granger sono state costruite le distribuzioni necessarie in vari modi differenti.
Per quanto riguarda i trattamenti, per ogni singolo tipo di trattamento è stata costruita la distribuzione giornaliera e settimanale di quante volte, in questo arco di tempo, il trattamento veniva nominato nei vari post.
Per quanto riguarda il sentimento è stata costruita la distribuzione giornaliera e settimanale dell'umore trovato da OpinionFinder su due diversi set di post:

- tutti i post di Facebook estratti;
- solo i post di Facebook dove compariva almeno un trattamento farmacologico nominato.

Questa distinzione è stata fatta pensando che limitare l'analisi ai soli post che discutono dei trattamenti, avrebbe portato ad un più forte rapporto predittivo.
Il calcolo della causalità di Granger ha un'ulteriore parametro necessario che corrisponde ai passi massimi eseguiti dal calcolo, cioè il valore L dell'equazione 3.5. I valori $b1 = ... = b_L = 0$ sono i valori controllati per dichiarare l'ipotesi nulla e questi valori si ottengono con probabilità p-value.
Per questa analisi si è deciso di provare ad eseguire l'analisi di Granger con valori di passi da 1 a 5.
I risultati, che come per gli altri si trovano nel capitolo dedicato, mostrano, per ogni esecuzione, i passi scelti, il p-value, il valore dell' F-test, il valore R^2 della regressione e le due distribuzioni interessate in questo calcolo di causalità.

Capitolo 4

Risultati

Nel capitolo precedente sono stati indicati i metodi utilizzati per rispondere a tre domande fondamentali, in che modo si sono affrontati i vari obiettivi e quali tipi di risultati si sarebbero mostrati.
In questo capitolo verranno presentati e analizzati i risultati ottenuti e le risposte alle tre domande iniziali.

4.1 Quanto è attivo il dibattito su Facebook e Twitter?

Dalle pagine pubbliche di Facebook che trattano il morbo di Crohn è stato possibile estrarre 31163 post, scritti tra il 27/10/2011 ed il 26/10/2013. In tabella 4.1 e 4.2 sono elencate le pagine pubbliche analizzate su Facebook ed il link per visualizzarle.
Da Twitter sono stati estratti 26737 tweet contenenti il termine "Crohn", scritti tra il 30/04/2013 ed il 26/10/2013. Il numero inferiore di tweet e l'arco temporale inferiore è dovuto al fatto che, come già anticipato, le API di Twitter non permettono di estrarre dati che risalgono a più di 9 giorni di distanza dalla data di ricerca.
Questi post sono stati scritti da 6815 diversi autori, in media si parla quindi di 4.57 post per autore, mentre per i tweet hanno partecipato 12213 diversi autori, 2.19 tweet per autore. Notiamo quindi già che su Twitter gli autori scrivono meno frequentemente rispetto a Facebook.
Il primo dato di interesse abbiamo detto essere il numero totale di post e tweet scritti in un determinato arco temporale, il risultato di questa analisi per Facebook è mostrato in figura 4.1, mentre per Twitter è mostrato in figura 4.2.

Facebook page
Crohns Disease
Crohns Disease Awareness
The Great Bowel Movement - Awareness for Crohn's Colitis
Crohns Disease with Ladyzeebs.
Wanted: Crohn's End
Crohn's Growth Foundation
Crohns Disease
CrohnsNet
Crohn's and Colitis Foundation of America
CROHNS DISEASE
Crohn's Awareness Project
Life After Crohn's
Crohn's And Me - Make The Connection
Crohn's and Colitis Foundation of Canada
The Crohn's Journey Foundation
Crohns Moms Humor
Crohn's and IBD Support Group
National Crohn Colitis Day
Crohn's Breakthrough Blog
CCNZ (Crohn's Colitis New Zealand)
Strong People Fighting Crohn's Disease and Colitis
Crohn's and Colitis UK
Team Challenge for Crohn's Colitis (CCFA)
Never Leave Home Without It Crohns and Colitis
Crohns Disease Support
Take Steps Be Heard for Crohn's Colitis (CCFA)
The Crohn's and Ulcerative Colitis Diaries: Living with IBD
Crohns Disease Support
Crohnsforum.com
Crohn's Colitis Foundation of America - New England Chapter
Crohn's Colitis Australia
Crohn's Disease Sucks
Worldwide Crohn's and Colitis Community
My Crohns Doctor
My Stomach Hurts - Life with IBS, Crohn's Disease Ulcerative Colitis

Tabella 4.1: Pagine su Facebook inerenti al Morbo di Crohn

Facebook page
Crohn's and Colitis Foundation of America - Michigan Chapter
Crohn's Zone
Crohn's Colitis UK - Edinburgh Group
The Community Crohn's Foundation
Crohn's Colitis Foundation
Cure Crohn's and Colitis
Crohn's Colitis Foundation of America - Illinois Chapter
Crohnology
Crohn's and Colitis Research UK Desire for Life
NoMoreCrohns.com
Crohn's Cookbook
Crohn's and Colitis Research UK Desire for Life
Crohn's and Colitis Foundation of America - North Florida Chapter
The Crohn's Awareness Global Engine
Crohn's Help
Crohn's and Colitis Foundation of America - Greater New York Chapter
Living with Crohn's Disease - Diary of a Crohnie
World Crohn's and Colitis Day

Tabella 4.2: Altre pagine su Facebook inerenti al Morbo di Crohn

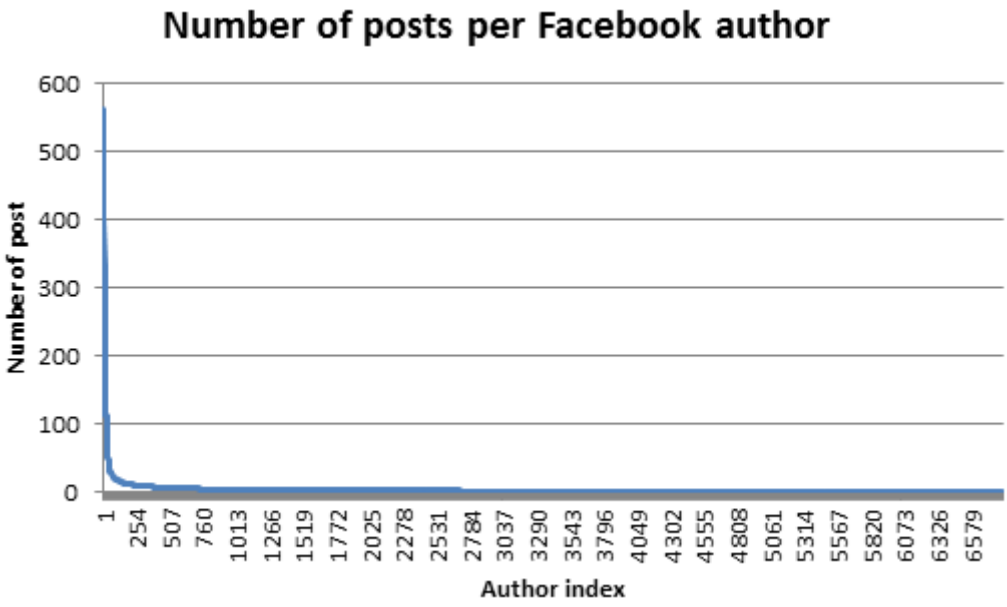


Figura 4.1: Numero di post per autore su Facebook

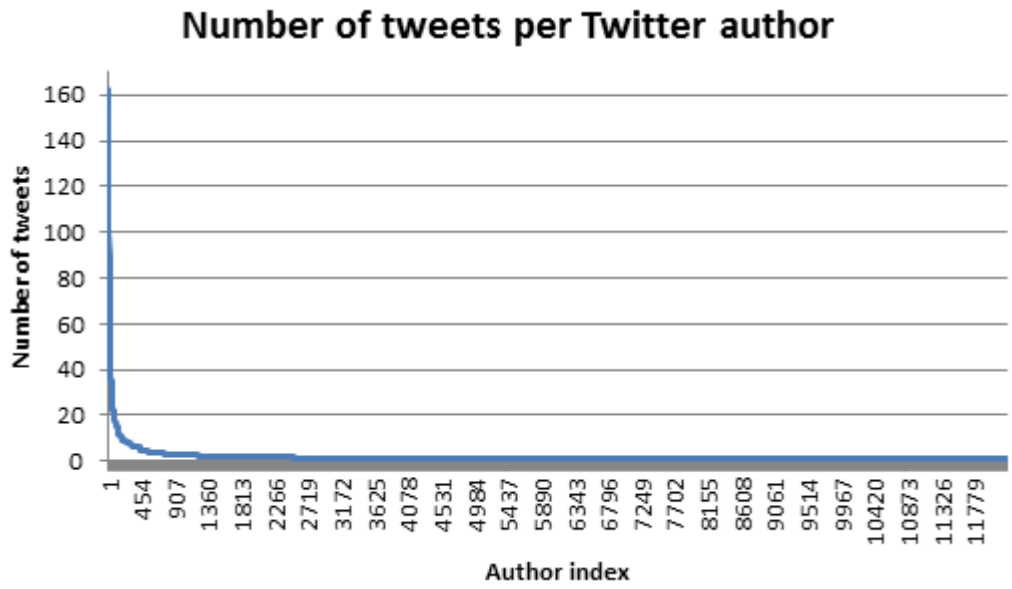


Figura 4.2: Numero di tweet per autore su Twitter

Si vede chiaramente il fatto che una piccolissima percentuale di autori scrive ripetutamente tantissimi post e tweet, mentre la maggior parte degli autori scrive pochissimo. In particolare i soli primi 10 utenti più attivi su Facebook hanno scritto 4050 post cioè circa il 13%, ed i primi 10 utenti più attivi su Twitter hanno scritto 1366 tweet quindi circa il 5%.

Una seconda metrica di interesse che mostra come gli utenti sono attivi è il tempo che passa tra i post di uno stesso autore, cioè quanto tempo un utente attende per scrivere un nuovo post dopo averne scritto uno precedente. In figura 4.3 sono mostrati i risultati per Facebook e come si vede la successione di messaggi è concentrata nella stessa giornata. Questo dato può voler dire che la discussione su Facebook è molto accesa sull'argomento e che questo porta gli utenti a riscrivere nella stessa giornata e solo in minor parte con molti giorni di distanza.
Per Twitter in figura 4.4 vediamo lo stesso fenomeno di distribuzione concentrata nello stesso giorno, che mano a mano cala con l'aumentare dei giorni.

Durante l'estrazione dei post di Facebook è stato possibile ricavare se questi post fossero una risposta ad un post precedente (con informazione dell'identificativo del post a cui si risponde) oppure un primo post pubblicato. E' interessante analizzare per ogni utente quanta percentuale di post scritti è un commento ad un post precedente.

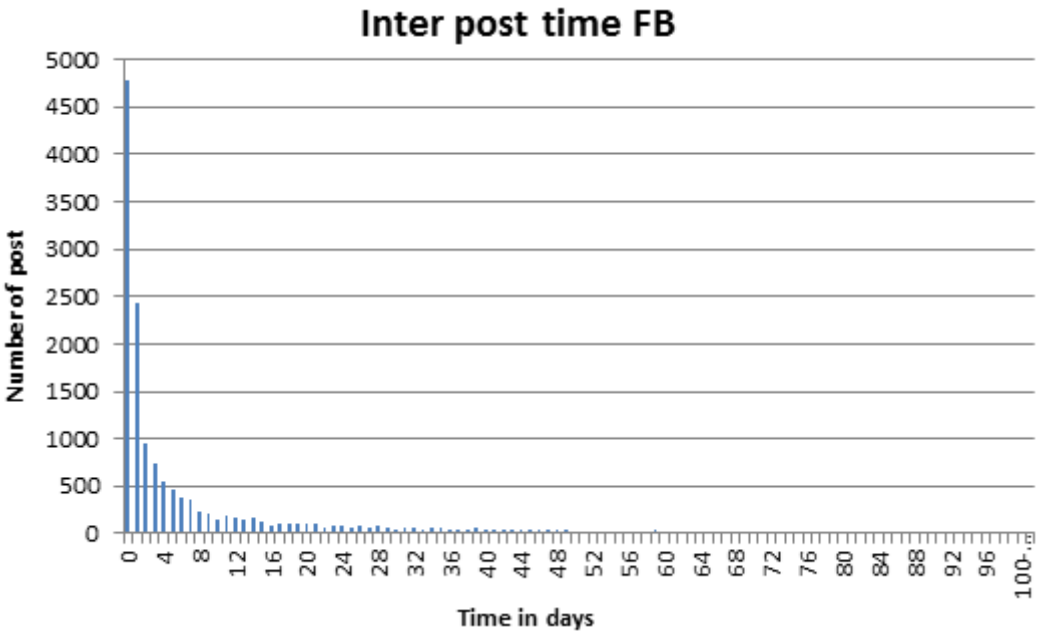


Figura 4.3: Tempo trascorso tra post di uno stesso autore

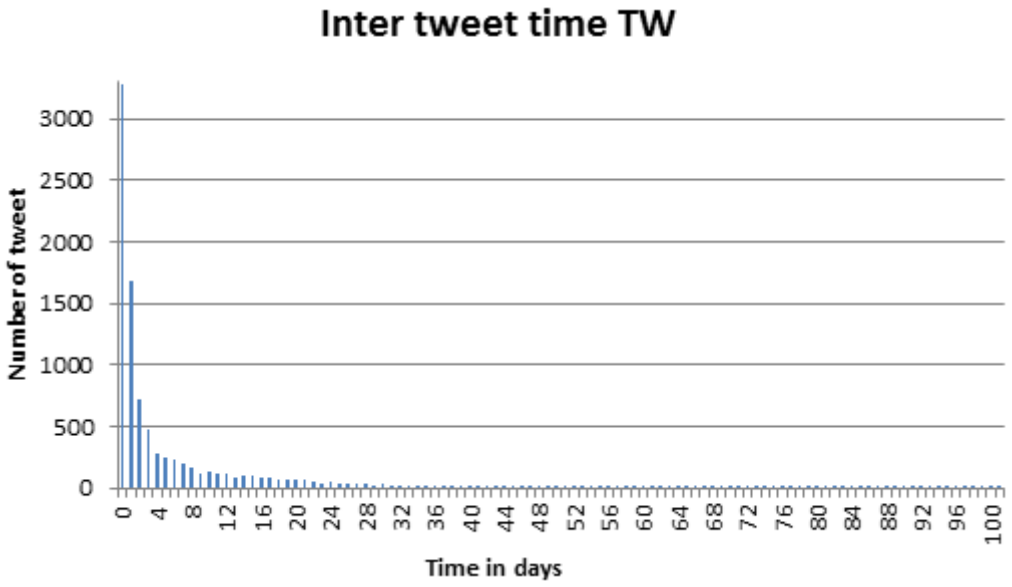


Figura 4.4: Tempo trascorso tra tweet di uno stesso autore

In figura 4.5 sono riportati i risultati per tutti gli autori ordinati per numero di post scritti di quanti post abbiano scritto per primi e quanti siano invece commenti ad altri post. Si vede chiaramente come gli utenti che scrivono tanto scrivono soprattutto primi post, mentre chi scrive poco lo fa soprattutto come risposta ad altri post.

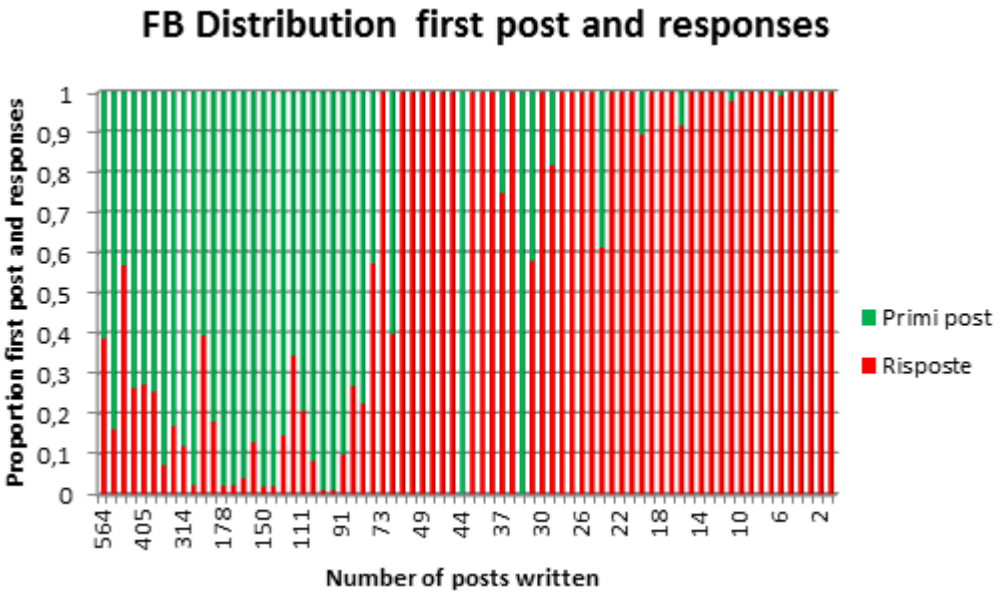


Figura 4.5: Distribuzione tra primi post e risposte ad altri post su Facebook

Questa prima analisi che si basa esclusivamente su semplici indicatori, quali il numero di autori, il numero di messaggi e le frequenze, mostra dei risultati alquanto sbilanciati sia per Facebook che per Twitter, risulta infatti che solo 261 utenti Facebook e solo 201 utenti Twitter ha pubblicato almeno 10 messaggi.

Indagando maggiormente sui post e tweet scritti è stato possibile caratterizzare maggiormente gli autori più prolifici, in particolare i primi 20 utenti che hanno scritto di più, cercando di determinare chi sono queste persone e in che modo sono legate alla malattia.
In tabella 4.3 sono mostrati i risultati per Facebook, mentre il tabella 4.4 sono mostrati quelli per Twitter.
Un risultato molto importante è scaturito da questa analisi, che mostra chiaramente la differenza di utenza che scrive sui due social. Su Facebook sono quasi tutti pazienti che raccontano quindi la loro esperienza diretta cercando

di autarsi a vicenda, ovviamente a scrivere di più sono gli amministratori delle pagine pubbliche analizzate, mentre su Twitter, sono pochissimi i pazienti e si tratta soprattutto di pubblicità, condivisione di articoli, donazioni ed eventi.
Sembrano essere quindi molto più affidabili i dati su Facebook, rispetto a quelli estratti da Twitter.

Type of Author	**Period of Activity**
Admin, Patient	18/11/2012 - 25/10/2013
Admin, Patient	04/05/2012 - 08/10/2013
Admin, Patient	27/10/2011 - 25/10/2013
Admin, Patient	16/05/2012 - 25/10/2013
Admin, Patient	29/06/2012 - 18/10/2013
Admin, Patient	09/11/2012 - 25/10/2013
Admin, Scientific Articles, Advertisement	26/03/2012 - 23/10/2013
Admin, Patient	24/09/2012 - 25/10/2013
Admin, Scientific Articles, Advertisement	31/10/2011 - 25/10/2013
Admin, Patient	17/07/2012 - 25/10/2013
Admin, Event	15/03/2012 - 13/10/2013
Admin, Patient	02/11/2011 - 01/10/2013
Admin, Patient	14/02/2012 - 16/10/2013
Admin, Event	27/10/2011 - 25/10/2013
Admin, Donation	19/05/2011 - 09/10/2013
Admin	23/11/2011 - 21/10/2013
Admin	03/08/2012 - 19/10/2013
Admin	18/09/2012 - 24/10/2013
Admin	02/05/2012 - 15/10/2013
Admin, Patient	27/10/2011 - 23/10/2013

Tabella 4.3: I 20 autori più prolifici su Facebook

Type of Author	**Period of Activity**
Awareness, Advertisement	30/04/2013 - 25/10/2013
Donations, Advertisement	04/06/2013 - 05/08/2013
Scientific Literature, Advertisement	30/04/2013 - 26/10/2013
Scientific Literature, Advertisement	30/04/2013 - 06/08/2013
Scientific Literature, Advertisement	30/04/2013 - 06/07/2013
Donations, Advertisements	03/05/2013 - 06/08/2013
Blogger, Advertisements	11/05/2013 - 14/07/2013
Donations, Advertisements	30/04/2013 - 05/08/2013
Patient	12/05/2013 - 24/10/2013
Celebrities suffering Crohn's disease	01/05/2013 - 26/10/2013
Donations, Event	30/04/2013 - 25/10/2013
Scientific Literature	08/05/2013 - 23/10/2013
Patient, Scientific Literature	30/04/2013 - 19/10/2013
Literature, Advertisement	10/05/2013 - 23/10/2013
Scientific Literature	01/05/2013 - 26/10/2013
Donations, Event	19/05/2013 - 02/10/2013
Scientific Literature	01/05/2013 - 18/07/2013
Scientific Literature	18/05/2013 - 24/10/2013
Event	30/04/2013 - 14/06/2013
Scientific Literature	30/04/2013 - 09/06/2013

Tabella 4.4: I 20 autori più prolifici su Twitter

Si prosegue l'analisi entrando ancora di più nel dettaglio ed esplorando quali argomenti interessano di più gli autori.

4.2 Quali sono gli argomenti più popolari?

Con le metodologie precedentemente indicate è stato possibile ricavare diversi risultati che mostrano il rapporto tra gli argomenti trattati sui social riguardo il morbo di Crohn e come questi evolvono nel tempo.
Un riassunto delle analisi che mostra già chiaramente il quadro della situazione è riportato in figura 4.6 e in figura 4.7, dove viene mostrata per i due social la distribuzione mensile della media giornaliera di post o tweet scritti con indicazione del rapporto tra i diversi argomenti trattati.
Da queste immagini si notano due informazioni importanti, la prima rivela come su Facebook stia aumentando sempre di più la discussione del morbo di Crohn mentre su Twitter stia calando.

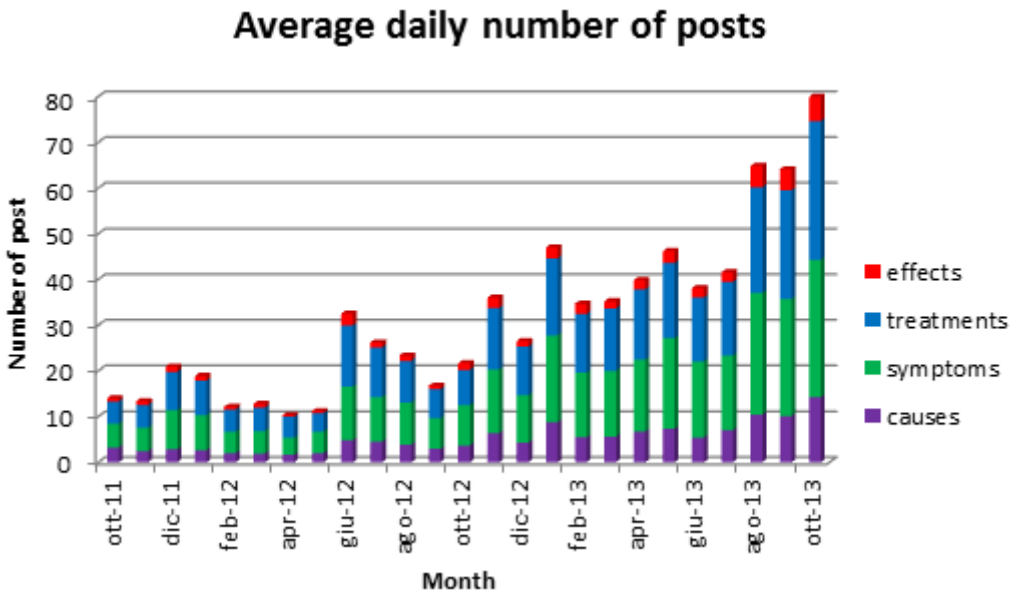


Figura 4.6: Media giornaliera dei post e argomenti su Facebook

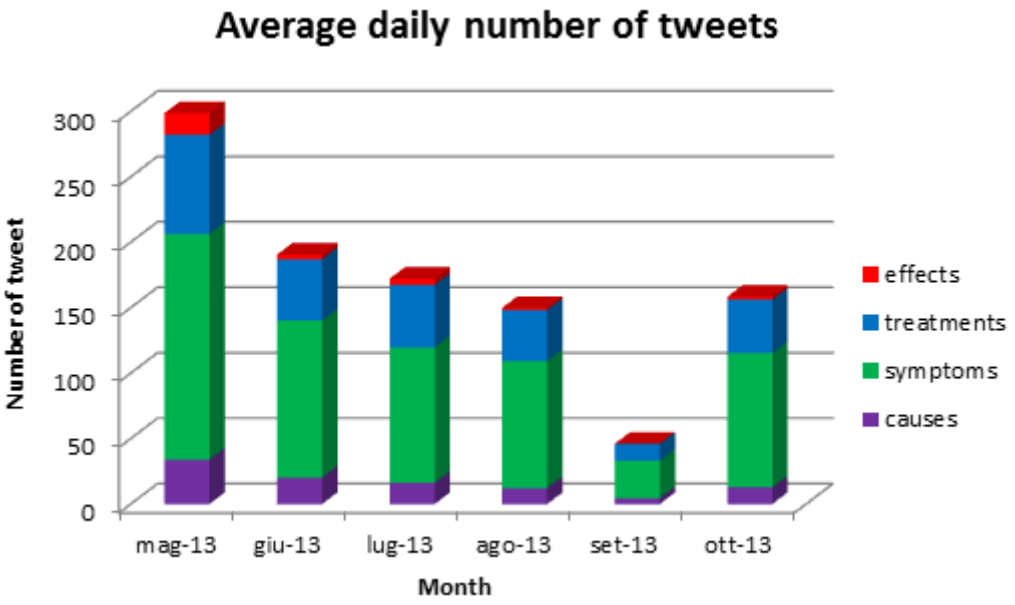


Figura 4.7: Media giornaliera dei post e argomenti su Twitter

La seconda informazione che si può immediatamente notare è che su Twitter c'è una chiara preferenza per il topic "symptoms", che in percentuale è nettamente superiore rispetto agli altri, mentre su Facebook entrambi i topic "symptoms" e "treatments" vengono trattati ampiamente e prevalgono sugli altri due.
Il rapporto tra gli argomenti trattati è approfondito meglio nei prossimi grafici, che mostrano proprio come la discussione sui social è distribuita tra le quattro categorie.
La figura 4.8 rivela che su Facebook circa il 60% dei post contiene termini relativi ai trattamenti e sintomi, mentre solo il 25% circa parla di cause e il 10% circa parla di effetti collaterali.

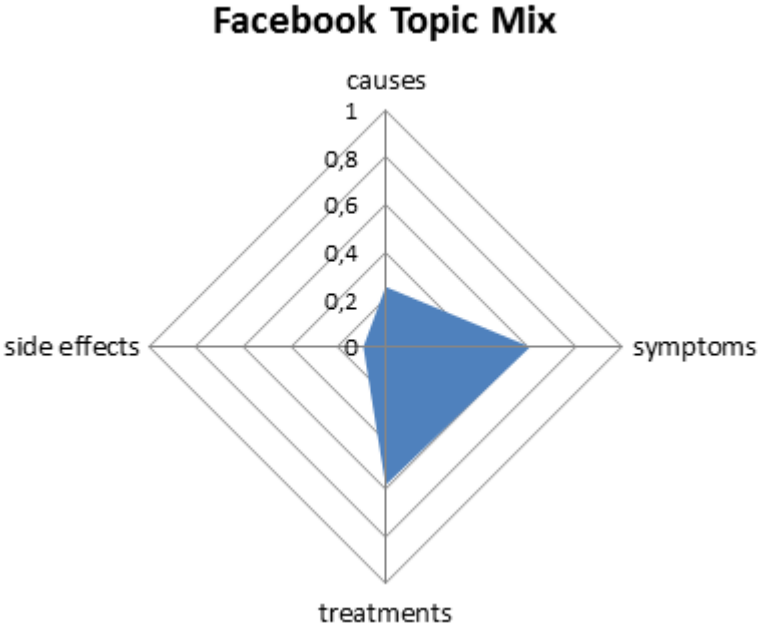


Figura 4.8: Distribuzione dei topic su Facebook

Questo risultato riguarda tutto l'insieme di post scritti da tutti gli autori, ma come cambia questo rapporto se consideriamo un insieme ristretto di autori e di post?
E' interessante analizzare questa distribuzione tra i topic per i soli autori più attivi, che hanno quindi scritto almeno 100 post, in figura 4.9 viene illustrato come questi autori maggiormente attivi si concentrino leggermente di più sul tema dei sintomi, senza però tralasciare quello dei trattamenti.
Se invece le analisi si specializzano sui soli utenti che hanno scritto pochissimo, cioè al massimo 5 post, si vede in figura 4.10 che questi parlano sia di trattamenti che di sintomi con una leggerissima preferenza per i trattamenti.

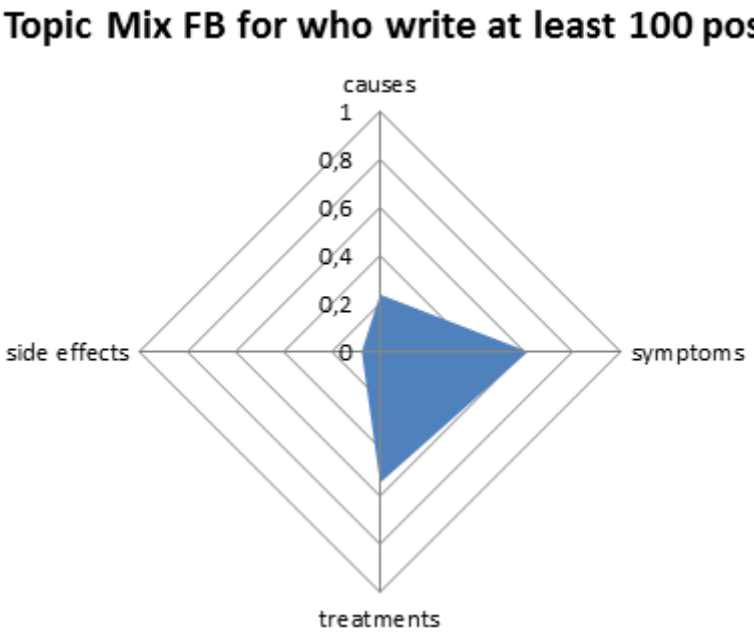


Figura 4.9: Distribuzione dei topic per gli autori che hanno scritto almeno 100 post su Facebook

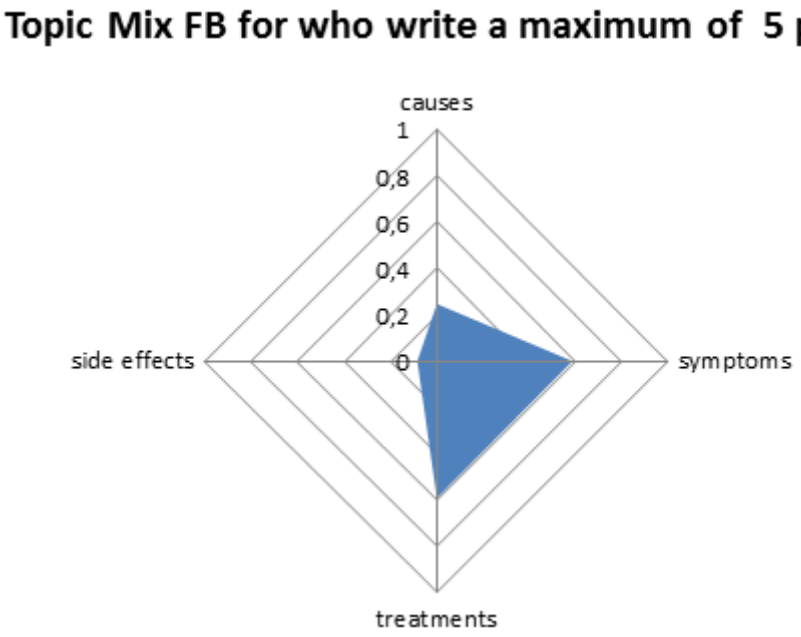


Figura 4.10: Distribuzione dei topic per gli autori che hanno scritto al massimo 5 post su Facebook

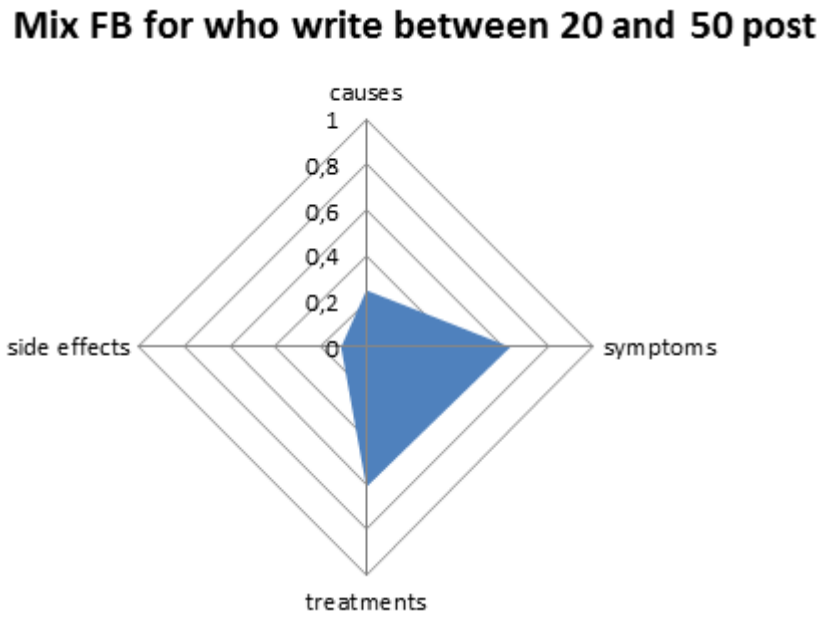


Figura 4.11: Distribuzione dei topic per gli autori che hanno scritto tra i 20 e i 50 post su Facebook

L'ultima analisi sul "topic mix" di Facebook ha studiato gli utenti che hanno scritto tra i 20 e i 50 post, cioè utenti non del tutto occasionali ma che non scrivono in modo eccessivo, in figura 4.11 viene indicato che questi utenti parlano quasi in equal misura di trattamenti e di sintomi e sembrano parlare leggermente di più anche degli effetti collaterali.
Queste diverse analisi che considerano separatamente gli utenti in base al loro grado di attività sul social, non hanno portato a risultati particolari. Sembra quindi che la proporzione tra gli argomenti trattati sia costante per tutti gli utenti, indipendentemente da quanto sono attivi e per tutti risulta una grossa predilizione per i sintomi e i trattamenti del morbo.

I risultati per Twitter, invece, mostrano un chiaro bias verso i sintomi portati dal morbo, sembra quindi che gli utenti su questo social tendano a sfogarsi maggiormente, lamentando i problemi sorti con questa malattia.
Considerando l'intera comunità di utenti che ha scritto su Twitter riguardo Crohn, in figura 4.12, risulta che circa il 90% dei tweets contiene almeno un termine presente nel dizionario dei sintomi, si scende fino al 40% per quanto riguarda i trattamenti, il 20% prende in considerazione le cause e soltanto il 10% gli effetti collaterali.
La discussione di cause ed effetti nei due social ha la stessa importanza e mo-

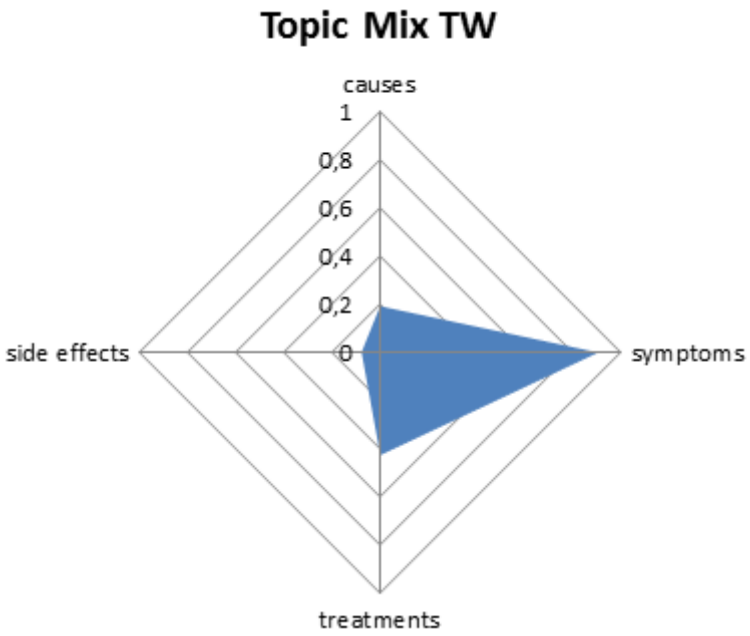


Figura 4.12: Distribuzione dei topic su Twitter

stra quindi un comportamento in comune che hanno gli utenti che scrivono, quello che risulta invece altamente differente è la proporzione tra la discussione dei sintomi e dei trattamenti e questo dimostra che il comportamento degli utenti varia al variare del social sul quale interagiscono.
Approfondendo lo studio per Twitter, in figura 4.13 sono riportati i risultati che considerano soltando gli utenti più attivi, che hanno scritto almeno 100 tweets. Per questi utenti vediamo che addirittura il grafico si "schiaccia" ancora di più verso il bias dei sintomi, presentando un valore prossimo al 100% e abbattendo ancora di più i valori restanti. La discussione dei trattamenti schende infatti drasticamente ad un 25% circa, le cause si riducono al 10% e gli effetti collaterali si annullano del tutto.
C'è quindi un alta specializzazione di argomento tra gli utenti più attivi.
Analizzando invece gli utenti meno attivi, che scrivono al massimo 5 tweets, in figura 4.14, ricompare la situazione già visualizzata per la globalità di utenti. La quantità nettamente maggiore di utenti che scrivono poco, rispetto agli utenti che scrivono molto, porta a far prevalere la loro distribuzione di argomenti, che rispecchia le percentuali già indicate precedentemente.
Considerando in fine gli utenti mediamente attivi, che scrivono tra 20 e 50 tweets, viene rilevato, in figura 4.15, un comportamento simile agli utenti poco attivi.

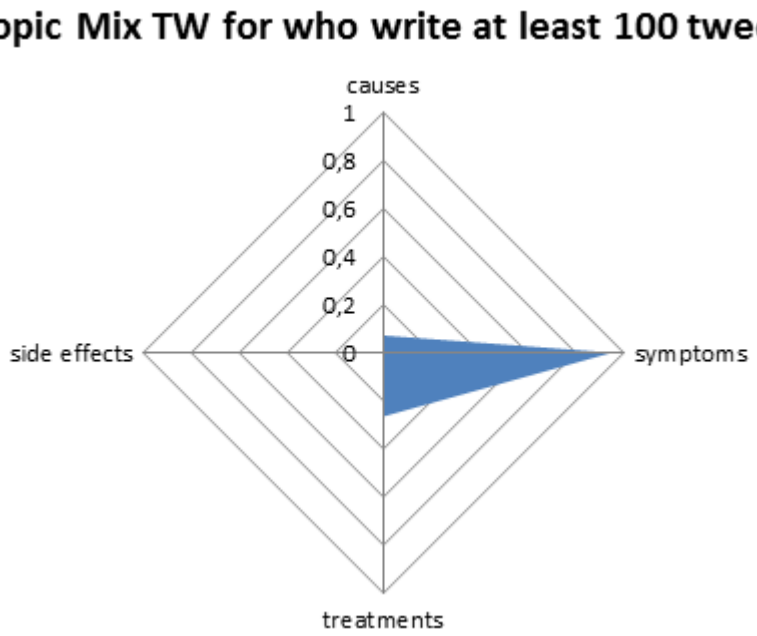


Figura 4.13: Distribuzione dei topic per gli autori che hanno scritto almeno 100 tweet su Twitter

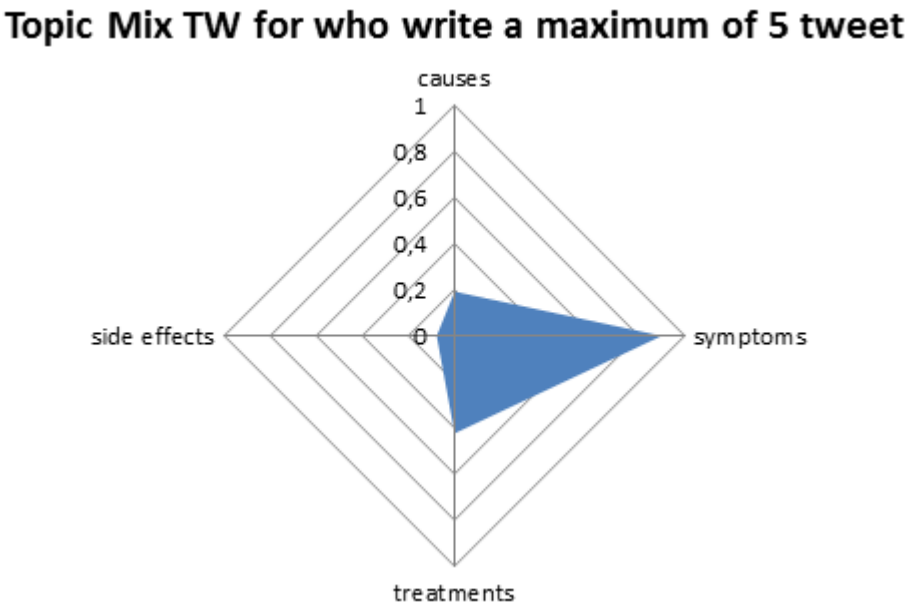


Figura 4.14: Distribuzione dei topic per gli autori che hanno scritto al massimo 5 tweet su Twitter

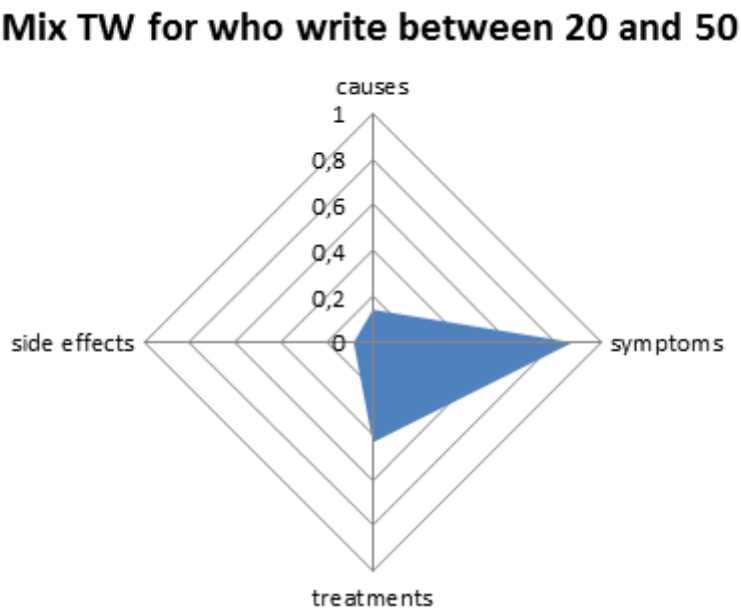


Figura 4.15: Distribuzione dei topic per gli autori che hanno scritto tra i 20 e i 50 tweet su Twitter

A differenza di Facebook, Twitter presenta quindi una differenza sostanziale, per gli argomenti trattati, tra gli utenti che scrivono almeno 100 tweets e tutti altri.

Un ultima conferma delle analisi fino ad ora presentate sulla popolarità dei topic trattati, si può visualizzare nelle immagini seguenti, dove, per i primi 100 autori più prolifici, viene mostrata la distribuzione di quanti post hanno scritto, con indicazione del rapporto tra post che trattano di cause, sintomi, trattamenti ed effetti collaterali.
In figura 4.16 è mostrata la distribuzione per Facebook, che fa notare subito sia come il numero dei post scritti dagli autori crolli velocissimamente e come le proporzioni tra le categorie trattate siano più eque e meno sbilanciate.
In figura 4.17 si vede questa distribuzione per Twitter, che fa notare come anche in questo caso il numero di tweets scritti cali velocemente e mostra chiaramente come siano sbilanciati i discorsi verso un unico argomento.

Tutte queste analisi che riguardano i topic, sono fatte ovviamente analizzando i termini indicati nei vocabolari di ogni categoria, ma quanta percentuale di discorso coprono questi termini?

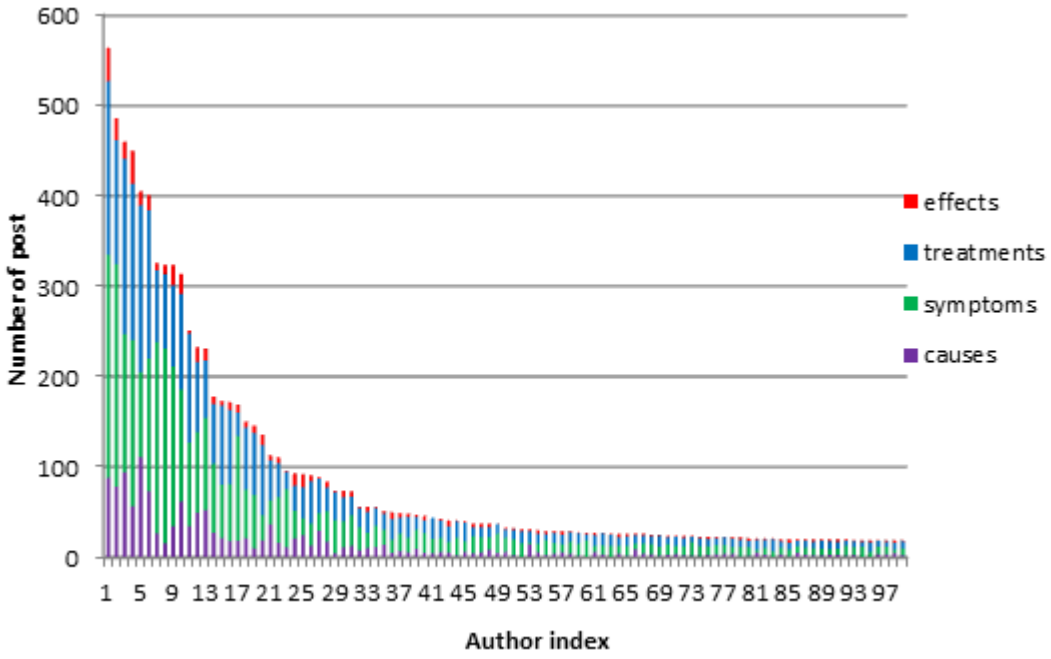


Figura 4.16: Distribuzione dei topic per i 100 autori più attivi tweet su Facebook

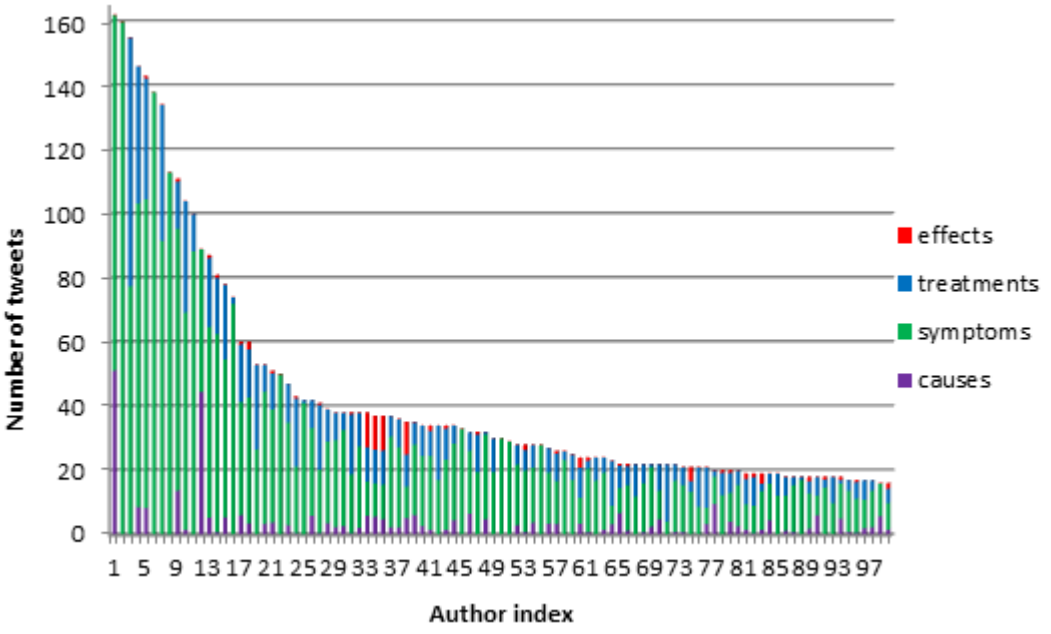


Figura 4.17: Distribuzione dei topic per i 100 autori più attivi tweet su Twitter

In figura 4.18 e in figura 4.19 è mostrato il rapporto tra i termini del nostro vocabolario utilizzati dagli utenti e le restanti parole scritte ma non analizzate, per i post e tweet che presentano più parole.
Si vede chiaramente come i termini analizzati siano in realtà solo una piccolissima parte degli interi discorsi fatti.

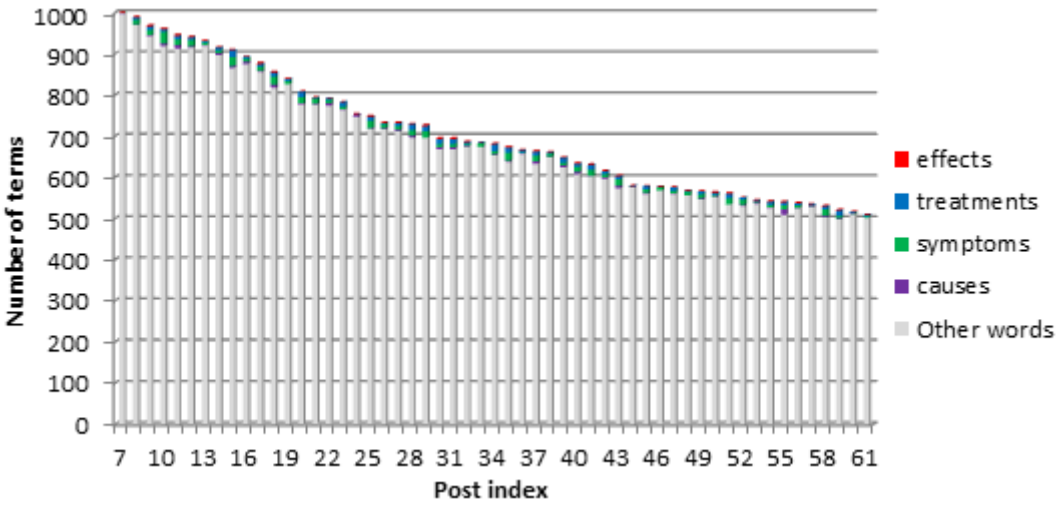


Figura 4.18: Relazione tra termini analizzati ed altre parole su Facebook

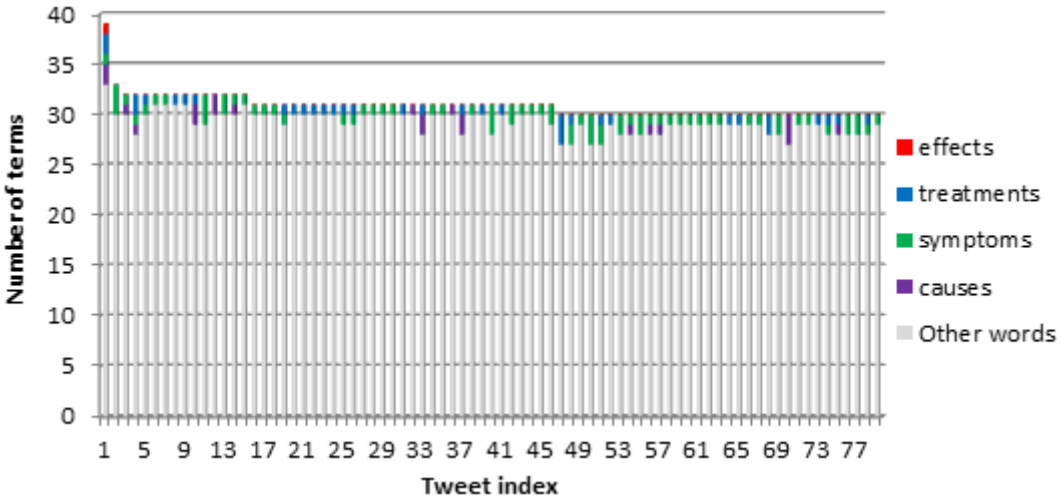


Figura 4.19: Relazione tra termini analizzati ed altre parole su Twitter

Procedendo con le analisi, di particolare interesse sono le relazioni di "distanza" tra i termini dei vari topic.
Utilizzando la nuova formula di distanza entropica indicata nella metodologia, che sfrutta il concetto di "cosine similarity" e di "sparsità", sono state calcolate le distanze tra tutti i termini delle categorie che presentano una relazione logica, ipotizzata con il nostro approccio bayesiano.
Le varie coppie di termini, presentano relazioni di distanza che vanno da un massimo di 2.17 ad un minimo di 1.09.
Studiando i valori è stata scelta la soglia per considerare due termini in relazione tra loro e il valore stimato è 1.55. Nei grafici successivi vengono quindi mostrate le relazioni tra i termini che presentano una distanza minore di 1.55.
Per mantenere una coerenza visiva, in tutti i grafici, fino ad ora mostrati e prossimi, vengono utilizzati colori che indicano la categoria di appartenenza, in particolare:

- Viola → "causes"
- Verde → "symptoms"
- Blu → "treatments"
- Rosso → "side effects"

In figura 4.20 sono mostrate le relazioni trovate tra i termini che rappresentano le cause del morbo e i sintomi su Facebook.
Le relazioni trovate sono state mostrate ai medici specializzati nel morbo di Crohn e tutte sono state confermate come veritiere, questo dimostra la validità della metodologia utilizzata che calcola quindi in modo opportuno le distanze tra i termini che determinano se sono presenti relazioni o no tra questi.
Oltre a validare il metodo usato, ricordiamo che questi dati sono il risultato dei discorsi fatti dagli utenti sui social network, questo significa che gli utenti parlano effettivamente del morbo in modo specifico e preciso dicendo cose reali e perfettamente conosciute in ambito medico.
Alcune relazioni sono molto generali, come il fatto che fumando cresca l'appetito, che i virus provochino la febbre o che il burro sia difficile da digerire e il latte dia problemi di stomaco. Altre sono più specifiche come il fatto che i batteri causano le infezioni e i disordini all'intestino e tra questi un batterio in particolare "Mycobacterium avium paratuberculosis" (map). Altra causa del morbo discussa è il sistema immunitario e i geni.

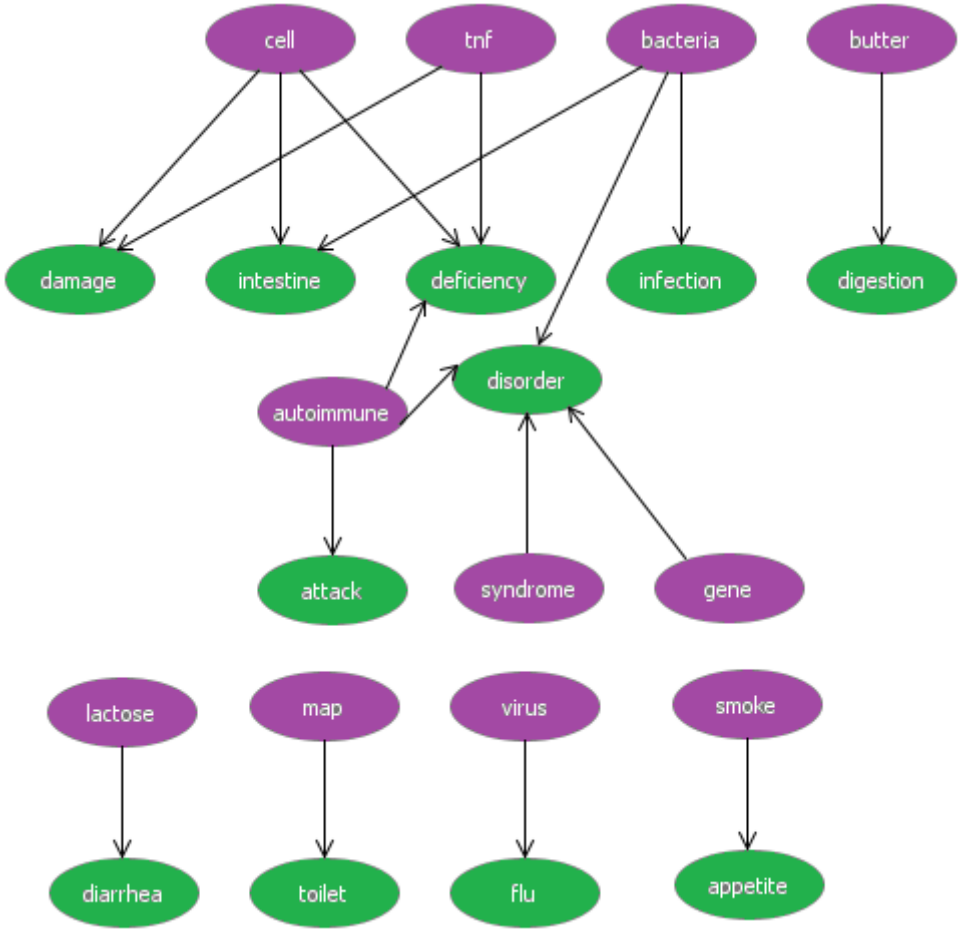


Figura 4.20: Rete delle relazioni Causes - Symptoms per Facebook

In figura 4.21 sono invece indicate le relazioni tra i termini che rappresentano i sintomi e quelli che rappresentano i trattamenti.
E' chiaramente visibile come questa immagine sia molto più ricca di relazioni e questo è conseguenza del fatto che, come visto nei risultati delle analisi precedenti, gli argomenti più discussi su Facebook sono appunto i sintomi e i trattamenti.
Anche queste relazioni, come le precedenti, sono perfettamente credibili e confermate dai medici. Per esempio, si vede come chi soffre di problemi intestinali si reca da uno specialista o da un dottore, quando ha dolori fisici esegue delle terapie e se ha perdite di sengue deve sottoporsi ad una colonscopia. Entrando più nello specifico, troviamo che chi soffre di artrite prende il medicinale "methotrexate", chi ha ascessi prende antibiotici e fa risonanze magnetiche (mri) e drenaggi. Chi presenta problemi alle ossa deve assumere

più calcio e può arrivare a fare trapianti e chi come sintomo ha problemi al retto deve sottoporsi ad interventi chirurgici come la colostomia.

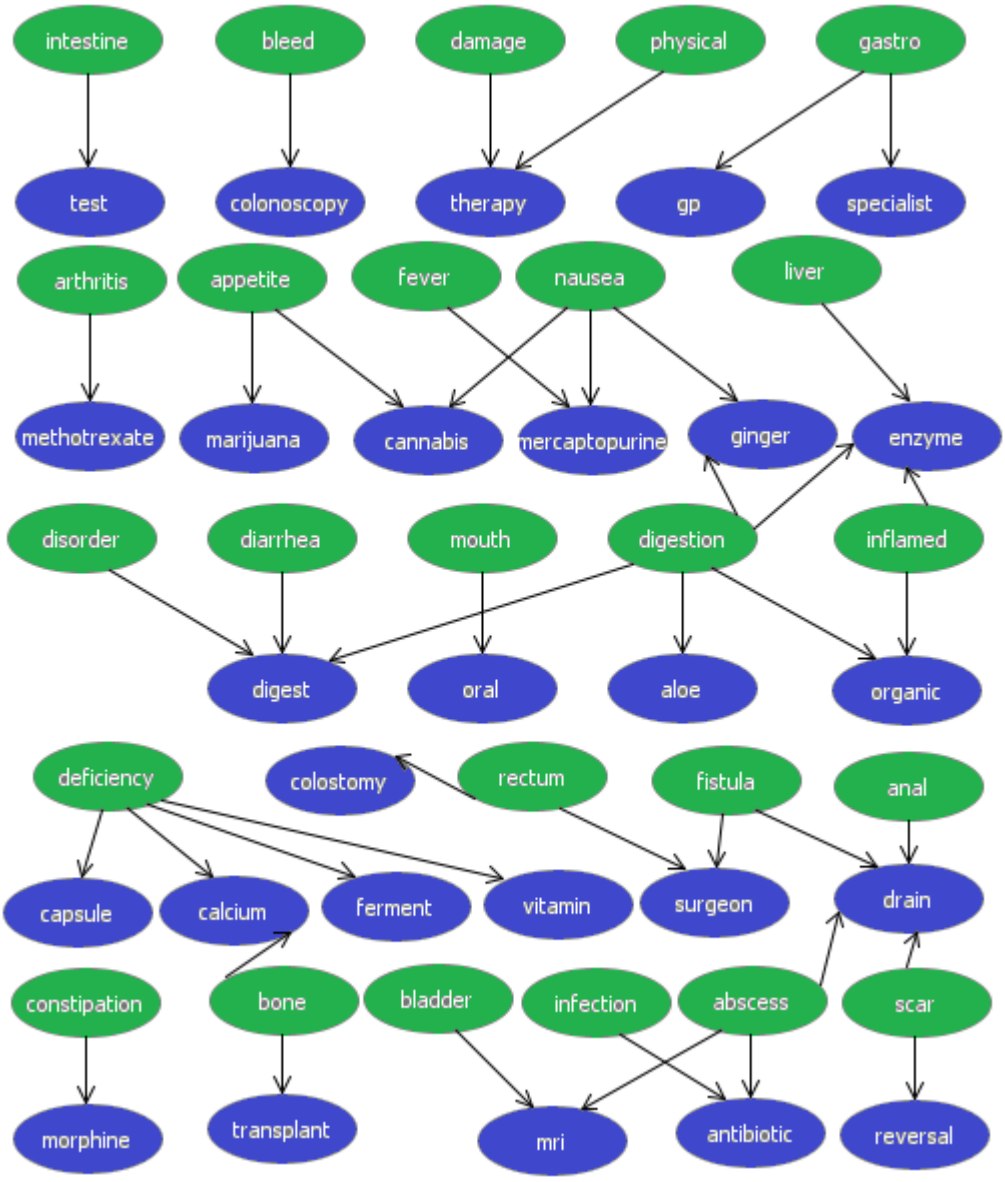


Figura 4.21: Rete delle relazioni Symptoms - Treatments per Facebook

Per quanto riguarda le relazioni possibili tra i trattamenti e gli effetti collaterali, non sono state trovare distanze minori della soglia prefissata, non

ci sono quindi relazioni particolarmente significative tra i termini.

Procedendo con l'analisi dei risultati per Twitter, anche questi nel complesso rispecchiano relazioni reali e confermate dai medici.
In figura 4.22 sono mostrate le relazioni tra le cause e i sintomi del morbo e molto interessante è leggere come le persone riportano che virus, batteri, ma soprattutto cibo ed in particolare carne di mucca provochi la leucemia. Virus come quello di epstein e una malattia cronica di natura autoimmune che è il lupus portano ad avere addirittura il cancro.
Questa ultima malattia cronica (lupus) viene molto discussa e gli si collegano diversi sintomi spiacevoli, come le artriti, sclerosi multipla, problemi al sistema immunitario e fibromialgia. Infine gli utenti parlano anche del morbo di parkinson e gli associano sintomi come depressione, epilessie, disordini e bipolarismo.

In figura 4.23 si parla delle relazioni tra sintomi e trattamenti e la quantità di dati riportati è sicuramente molto minore rispetto ai risultati di Facebook. Una particolarità che troviamo su Twitter è che gli utenti parlano davvero molto dell'utilizzo di droghe per alleviare i sintomi del morbo e addirittura come possibile cura. Come rimedio infatti per i sintomi di depressione, disordini e bipolarismo vengono indicati trattamenti con cannabis e canapa (hemp).
Chiaramente per la febbre è consigliato il vaccino e per infiammazioni all'intestino che presenta un effetto "cobblestone" vengono eseguite endoscopie. Due farmaci specifici sono presenti in queste relazioni e sono Infliximab ed Humira che vengono usati in presenza di artriti o di sangue (nelle feci) associato anche a Fioricet.

In figura 4.24 si trovano relazioni tra trattamenti ed effetti collaterali che per Twitter sono presenti. Come già accennato molto discusse sono le droghe e non può mancare la discussione che riguarda gli effetti che queste possono avere, si vede infatti come marijuana e canapa siano collegate alla termine effetti ed in generale delle droghe e/o farmaci si discute delle possibili reazioni allergiche.

Per concludere l'analisi di distanza tra i termini è possibile affermare che gli argomenti discussi nei due social sono diversi e si concentrano su elementi differenti della malattia che infatti non si trovano riportati in entrambi i social.

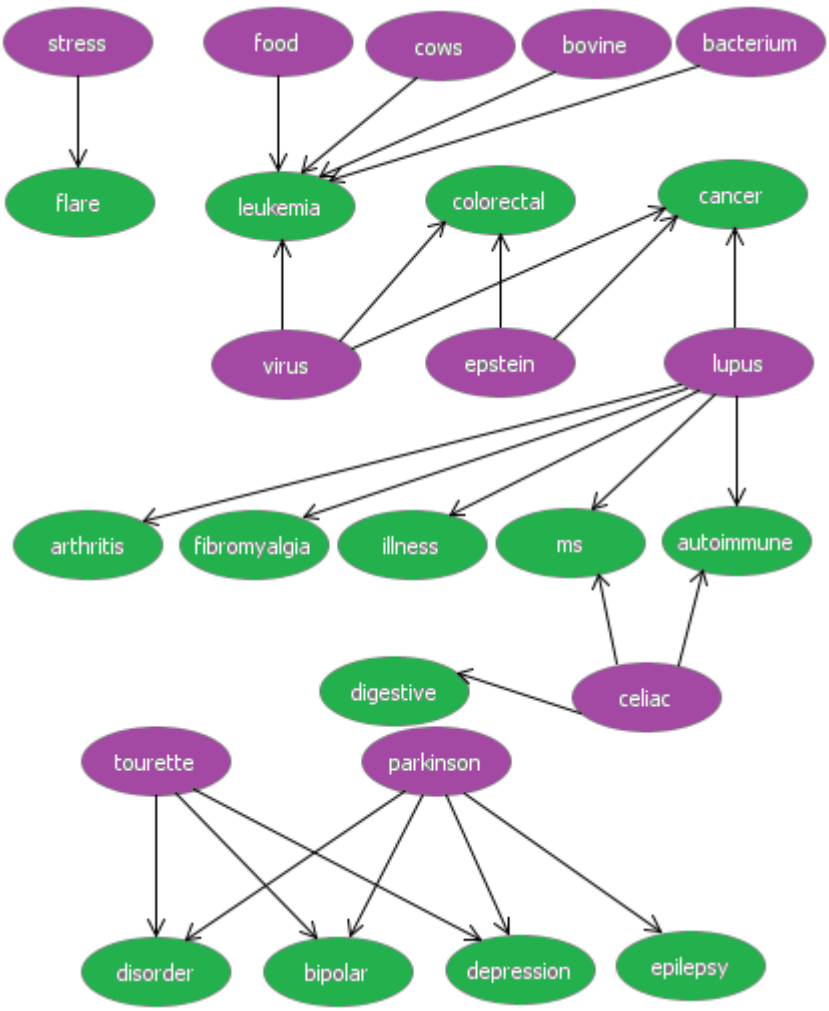


Figura 4.22: Rete delle relazioni Causes - Symptoms per Twitter

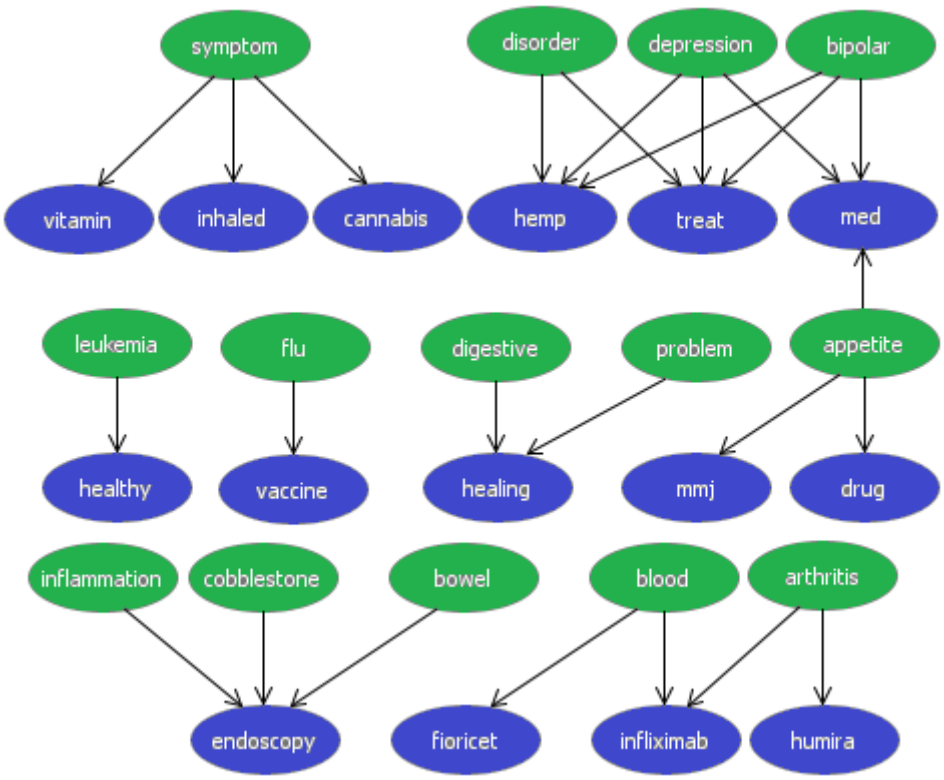


Figura 4.23: Rete delle relazioni Symptoms - Treatments per Twitter

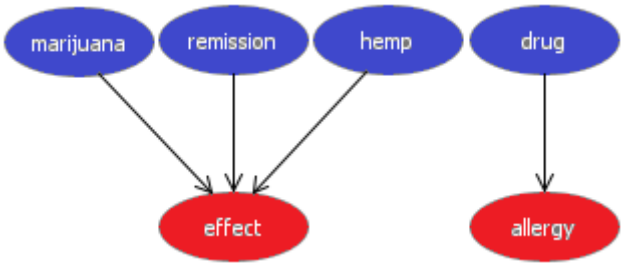


Figura 4.24: Rete delle relazioni Treatments - Side effects per Twitter

Procediamo con un ulteriore approccio bayesiano che continua ad indagare le relazioni tra i vari termini attraverso il calcolo delle probabilità condizionate.
In altre parole, si è andata a calcolare la frequenza con cui un dato termine compare, sapendo che un altro termine (di un'altra categoria) è comparso.
In figura 4.25, 4.26 e 4.27, sono riportati i risultati delle probabilità condizionate tra i termini che presentano un valore di almeno 0.25 per Facebook.
I risultati sono infine divisi in ulteriori categorie in base al loro argomento secondario che ne facilita la lettura.
Per esempio, nel gruppo del sistema immunitario, troviamo che ogni volta che appare il termine "autoimmune" il 65% delle volte appare anche il termine "disease".
Interessanti sono i dati osservati, come il fatto che la parola intestino compaia il 51% delle volte che compare la parola sindrome, o il fatto che la chirurgia ("surgery") venga menzionata con alta probabilità quando compaiono i termini ascesso, cicatrice, fistola, retto, vescica e addome. Quando si indica una carenza di qualcosa si consigliano il 53% delle volte delle vitamine e per moltissimi sintomi del morbo vengono discussi i relativi test.
Una sola relazione è presente tra i trattamenti e gli effetti e pare che il 30% delle volte che si parla del medicinale "entocort" vengano discussi gli effetti.

Facebook	P(symptom|cause)	Val	P(symptom|cause)	Val
Immune system	P(disease|autoimmune) P(disease|bacteria)	0.65 0.35		
Genetic	P(disease|gene)	0.43	P(disease|cell)	0.31
Syndrome	P(bowel|syndrome)	0.51	P(disease|syndrome)	0.39

Figura 4.25: Probabilità condizionate per Facebook

Facebook	P(treatment|symptom)	Val	P(treatment|symptom)	Val
Vitamin	P(vitamin|deficiency)	0.53		
Exams	P(test|inflamed)	0.36	P(test|constipation)	0.31
	P(test|damage)	0.36	P(test|liver)	0.29
	P(test|blood)	0.34	P(test|abdomen)	0.28
	P(test|diarrhea)	0.31	P(test|disorder)	0.26
	P(test|deficiency)	0.31	P(test|rectum)	0.25
Surgery	P(surgery|abscess)	0.36		
	P(surgery|scar)	0.33	P(surgery|bladder)	0.27
	P(surgery|fistula)	0.32	P(surgery|abdomen)	0.26
	P(surgery|rectum)	0.27		

Figura 4.26: Probabilità condizionate per Facebook 2

Facebook	P(effect|treatment)	Values
Medicines	P([effect]|[entocort])	0.30

Figura 4.27: Probabilità condizionate per Facebook 3

Stessa analisi è stata eseguita anche per Twitter ed i risultati a prima vista sembrano davvero molto particolari, come si può vedere in figura 4.28, 4.29, 4.30 e 4.31.

Twitter	P(symptom|cause)	Val	P(symptom|cause)	Val
Smoke	P([disease]|[smoking])	0.98		
Nutrition	P([disease]|[bovine]) P([disease]|[cows]) P([leukemia]|[bovine]) P([leukemia]|[cows]) P([colitis]|[meat])	1 0.99 0.99 0.90 0.73	P([disease]|[food]) P([disease]|[milk]) P([disease]|[eat]) P([disease]|[gluten])	0.70 0.67 0.35 0.34
Immune system	P([disease]|[bacterium]) P([disease]|[virus])	0.97 0.95	P([leukemia]|[bacterium]) P([leukemia]|[virus])	0.79 0.25

Figura 4.28: Probabilità condizionate per Twitter

Risultano molto particolari questi dati, perchè presentano dei valori veramente altissimi di probabilità condizionata. Sembrerebbe quindi che ogni volta che si parla di carne bovina e di mucche si parla sempre di malattia e di leucemia, così come il fumo, i batteri e virus sono legati strattamente alla malattia. Troviamo anche che ogni volta che si nomina il virus di epstein viene parlato di malattia, cancro e problemi al colon-retto ed altre relazioni molto strette tra cause e sintomi.
Anche le relazioni tra sintomi e trattamenti presentano questi valori altissimi e tutti questi risultati sarebbero sorprendenti in quanto rappresentano una dipendenza quasi totale di più termini, ma proprio per criticità a questo fatto si è approfondito meglio il motivo di queste relazioni.
Il motivo scoperto è che su Twitter è possibile un meccanismo di retweet, che permette di pubblicare molte volte uno stesso tweet, scritto anche da altre persone. Questo fatto porta ad avere molti tweet ripubblicati ma che

Twitter	P(symptom|cause)	Val	P(symptom|cause)	Val
Syndrome and chronic disease	P([disorder]|[tourette])	1	P([disorder]|[parkinson])	0.72
	P([disease]|[epstein])	0.99	P([bipolar]|[parkinson])	0.72
	P([cancer]|[epstein])	0.99	P([depression]|[parkinson])	0.72
	P([colorectal]|[epstein])	0.99	P([disease]|[lupus])	0.47
	P([disease]|[tourette])	0.98	P([disease]|[celiac])	0.46
	P([bipolar]|[tourette])	0.98	P([autoimmune]|[lupus])	0.32
	P([depression]|[tourette])	0.98	P([ms]|[lupus])	0.30
	P([disease]|[parkinson])	0.93	P([arthritis]|[lupus])	0.27
General	P([disease]|[cause])	0.80	P([disease]|[stress])	0.38

Figura 4.29: Probabilità condizionate per Twitter 2

in effetti sono un unico messaggio che non cambia.
E' stato scoperto quindi che, nel nostro database di tweet estratti, sono presenti solamente tre tweet differenti che contengono i termini "cows" e "bovine" ed uno di questi è stato retweettato 125 volte. Allo stesso modo un tweet che contiene i termini "smoke" e "marijuana" e che descrive l'uso di queste sostenze in modo benefico per arrivare alla remissione della malattia ha avuto addirittura 580 condivisioni.
Una lista di cibi che influenzano il morbo di Crohn è stata retweettata 80 volte, ed un altro tweet che metteva in evidenza l'influenza del latte sulla malattia è stato pubblicato 44 volte.
Il termine "tourette" che dalle analisi sembra molto frequente e con alte relazioni è stato in realtà trovato in soli due tweet distinti, uguale per "epstein" che è contenuto in un tweet ripubblicato 75 volte.
Continuando con l'analisi dei motivi di queste alte relazioni è stato trovato che "hemp" viene considerato un possibile trattamento per il morbo in un tweet retweettato 98 volte, come il sintomo "cobblestone" è nominato in due soli tweet, ricondivisi 72 e 30 volte ciascuno. I termini febbre e vaccino sono stati discussi insieme una sola volta anche questa ripubblicata 40 volte.

Twitter	P(treatment|symptom)	Val	P(treatment|symptom)	Val
Exam	P([endoscopy]|[cobblestone])	0.96		
General	P([healthy]|[leukemia]) P([treat]|[bipolar]) P([remission]|[appetite]) P([treat]|[depression])	0.91 0.89 0.83 0.63	P([treat]|[disorder]) P([inhaled]|[symptom]) P([cure]|[ms]) P([cure]|[autoimmune])	0.51 0.35 0.26 0.25
Medicines	P([med]|[bipolar]) P([med]|[appetite]) P([vaccine]|[flu]) P([med]|[depression])	0.90 0.83 0.75 0.64	P([med]|[disorder]) P([med]|[epilepsy]) P([fioricet]|[blood])	0.53 0.35 0.31
Drugs	P([hemp]|[bipolar]) P([marijuana]|[appetite]) P([drug]|[appetite]) P([hemp]|[depression])	0.89 0.84 0.75 0.62	P([hemp]|[disorder]) P([cannabis]|[symptom]) P([marijuana]|[epilepsy])	0.50 0.36 0.33

Figura 4.30: Probabilità condizionate per Twitter 3

Infine, per quanto riguarda gli effetti dei trattamenti, il termine “remission” compare ben in 1780 tweet ed insieme a questo termine, i termini “marijuana”, “hemp” ed “effect” compaiono in un solo tweet retweettato addirittura 700 volte. Il termine “hemp” compare poi anche in qualche altro tweet, ma anche questi in generale vengono ricondivisi parecchie volte.

Twitter	P(effect|treatment)	Val	P(effect|treatment)	Val
General	P([effect]|[remission])	0.61	P([effect]|[diet])	0.28
	P([effect]|[hemp])	0.60	P([effect]|[mmj])	0.26
	P([effect]|[marijuana])	0.52		

Figura 4.31: Probabilità condizionate per Twitter 4

Avendo ora analizzato le relazioni tra i diversi termini estratti dai social è opportuno verificare che queste relazioni mantengano la legge di potenza o power law, tipica del linguaggio naturale.
Come si vede in figura 4.32 e in figura 4.33 la distribuzione di frequenza della co-orrelazione tra i termini di Facebook sembra chiaramente seguire una power law, mentre per Twitter l'andamento è lo stesso ma presenta un comportamento insolito durante l'inizio della coda.
Non basta ovviamente osservare semplicemente questi grafici ma è necessario analizzare anche l'istogramma su scala logartmica della distribuzione.
La figura 4.34 mostra il grafico log-log per Facebook e vediamo che i dati seguono bene la linea retta in scala logaritmica, mentre la figura 4.35 mostra il grafico log-log per Twitter e come già preannunciato dall'immagine precedente i dati non seguono la linea retta in scala logaritmica.
L'ultima conferma della presenza o meno di una power law si ha effettuando una stima di curve, grazie al software di statistica SPSS e verificando se i dati osservati sono approssimati bene da una curva di potenza.
La figura 4.36 mostra il grafico dei dati osservati e della regressione con la cuva di potenza per Facebook e sono indicati anche i valori del modello per una sua analisi.
Dal valore R-quadrato si può dedurre se il modello utilizzato è buono o meno, il valore può oscillare tra 0 ed 1 e se il valore è basso il modello non è buono, se il valore è alto il modello utilizzato è buono.
Il valore F indica un confronto di due varianze e più questo valore è alto più i due fenomeni hanno la stessa varianza, ma ancora più importante è il valore "Sig." che corrisponde alla significabilità della statistica F conosciuto anche come p-value che deve essere inferiore a 0.05 per poter rigettare l'ipotesi nulla.

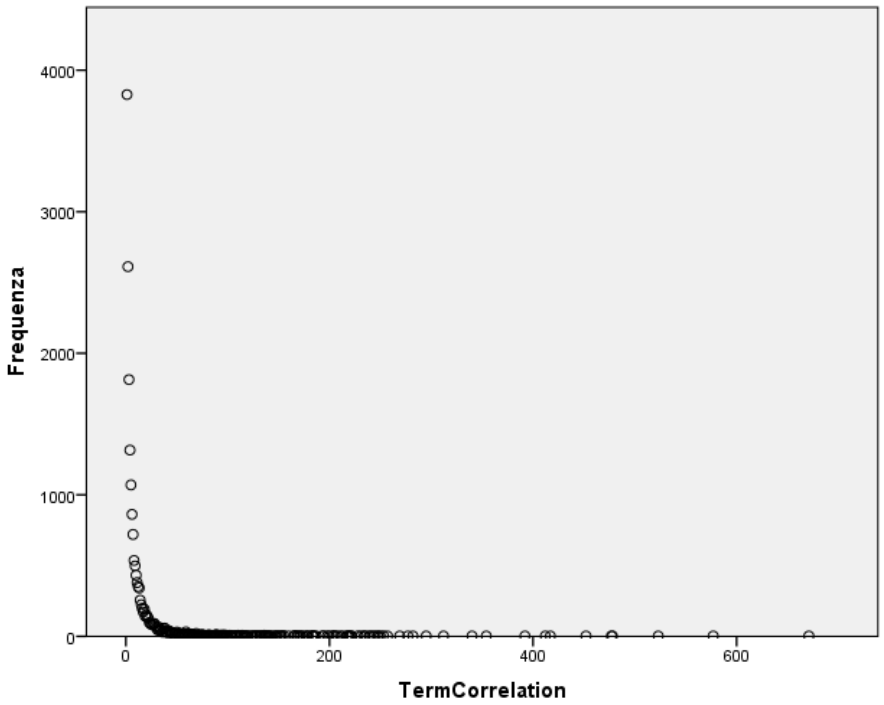


Figura 4.32: Dist. di frequenza della correlazione tra i termini di Facebook

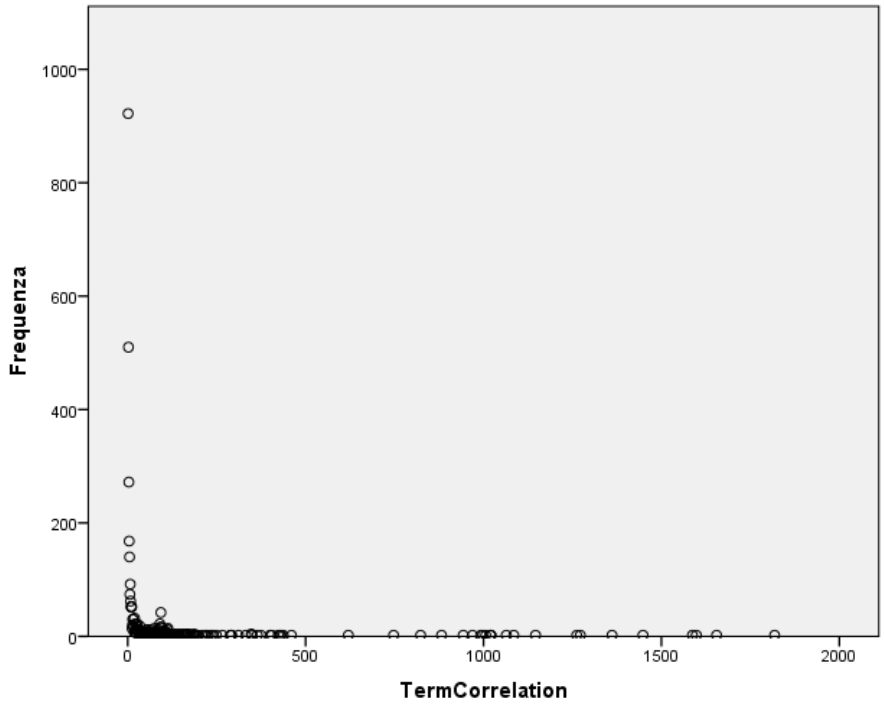


Figura 4.33: Dist. di frequenza della correlazione tra i termini di Twitter

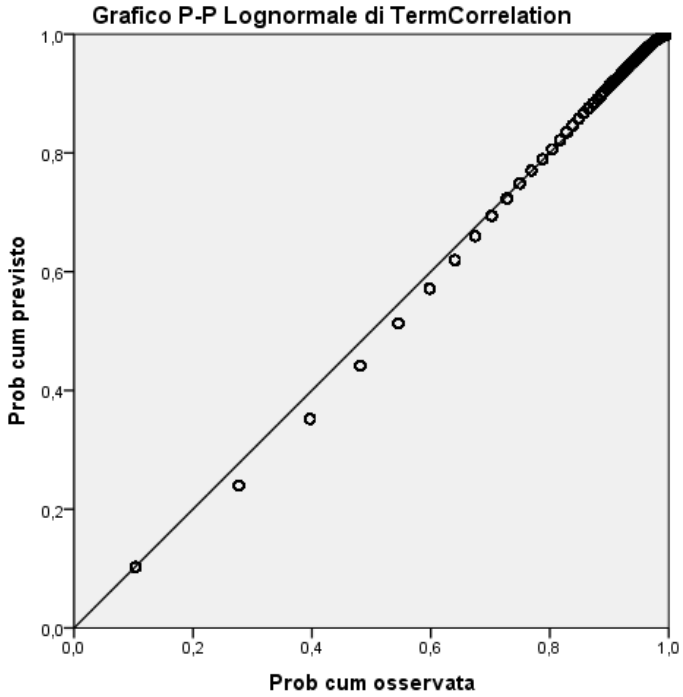


Figura 4.34: Dist. log-log della correlazione tra i termini di Facebook

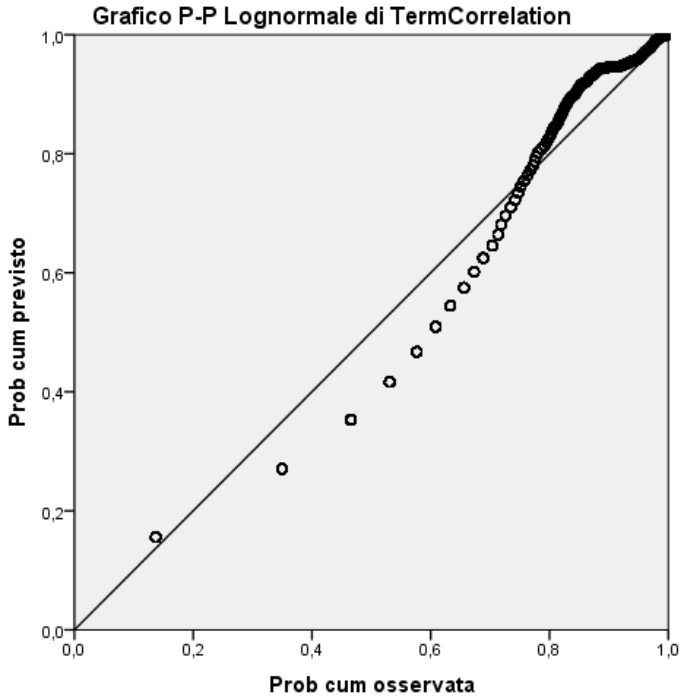
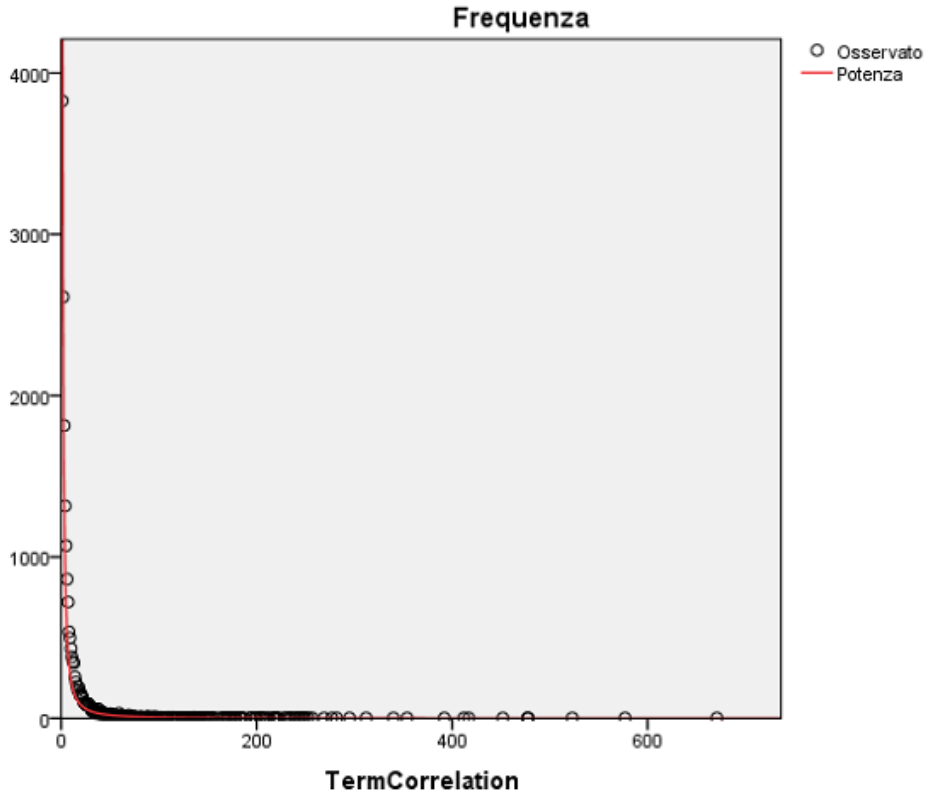


Figura 4.35: Dist. log-log della correlazione tra i termini di Twitter



Riepilogo del modello e stime dei parametri

Variabile dipendente: Frequenza

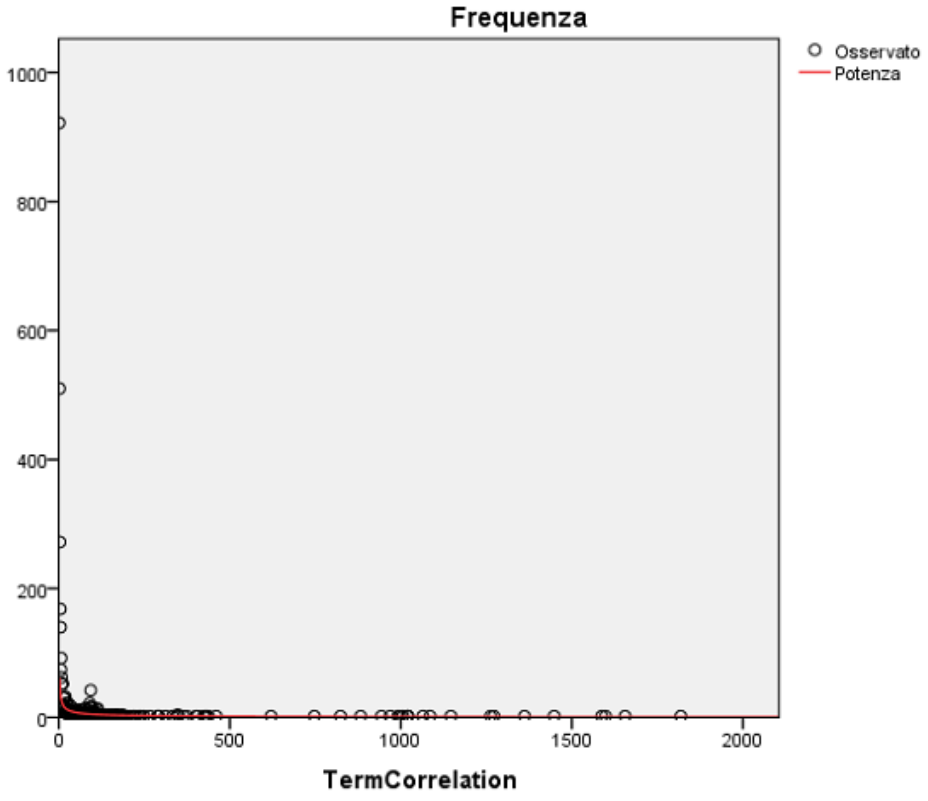
Equazione	Riepilogo del modello					Stime di parametri	
	R-quadrato	F	df1	df2	Sig.	Costante	b1
Potenza	,894	1670,839	1	199	,000	8719,339	-1,551

La variabile indipendente è TermCorrelation.

Figura 4.36: Stima della curva di potenza per Facebook

Per Facebook vediamo che i dati sono tutti positivi, il valore R-quadrato è alto e la significabilità è ottima.
Nella figura 4.37 viene mostrato lo stesso grafico di regressione per Twitter e dai dati vediamo che anche in questo caso si ha un'ottima significabilità con p-value pari a 0, mentre il valore R-quadrato non è altissimo e questo rispecchia il fatto che l'andamento sarebbe anche per Twitter quello di una power law ma per colpa di alcune perturbazioni i dati non sono precisi.
Ma cosa è che disturba i dati di Twitter? Abbiamo già detto che per Twitter è possibile un meccanismo di retweet cioè di ripubblicazione di altri tweet

molto utilizzato e abbiamo già visto come questo meccanismo invalidi molte statistiche effettuate.
Anche in questo caso, questo meccanismo fa si che diverse coppie di termini con co-occorenza media vengano ripresentate più volte con questi retweet e questo provoca quel picco mostrato nel primo grafico all'inizio della coda.



Riepilogo del modello e stime dei parametri

Variabile dipendente: Frequenza

Equazione	Riepilogo del modello					Stime di parametri	
	R-quadrato	F	df1	df2	Sig.	Costante	b1
Potenza	,618	289,645	1	179	,000	100,674	-,660

La variabile indipendente è TermCorrelation.

Figura 4.37: Stima della curva di potenza per Twitter

Proseguono le analisi con un ulteriore passo avanti, cercando di studiare non solo il comportamento delle persone online, ma anche come le persone si sentono e qual'è il loro sentimento sugli argomenti di interesse.

4.3 Qual'è il sentimento emergente?

Le reti sociali possono essere luoghi dove valutare non soltanto i temi a cui le persone sono più interessate, ma anche come (in che termini) la gente parla di questi argomenti.
Nel nostro caso specifico è sicuramente interessante scoprire quanto positivamente o negativamente le persone considerano un certo trattamento, o se l'umore dei pazienti dipende dal farmaco che stanno assumendo.
Queste informazioni possono essere molto utili ai medici per la prescrizione del trattamento migliore e più gradito dagli utenti e per lo sviluppo futuro di nuovi trattamenti.
Nella metodologia sono già stati elencati quali sono i trattamenti persi in considerazione da questa analisi e prima di analizzare i risultati sul sentimento di questi farmaci è utile capire quanto si parla di questi farmaci e quindi quanto sono discussi e quali sono i farmaci più popolari (sia in positivo che in negativo).

Concentrandoci ora solo su Facebook, in figura 4.38, 4.39 e 4.40 sono riportati gli andamenti settimanali di discussione dei vari trattamenti. I trattamenti più discussi sembrano quindi essere Prednisone, Infliximab ed Adalimumab che vengono nominati anche 40, 50 e 80 volte in un unica settimana. A seguire un altro farmaco molto discusso è Azatioprina seguito da Mesalazina, Certulizumab e 6-Mercanturopine. Altri farmaci sono discussi invece solo poche volte e sono Budesonide, Metotrexate e Natalizumab, infine due farmaci non sono quasi mai nominati e sono Beclotemasone dioprinato e Metilprednisolone.

Proseguendo con l'analisi del sentimento, sono state estratte le polarità del sentimento per ogni post ed in figura 4.41 è mostrato l'andamento dell'umore generale per Facebook. Per graficare l'andamento e l'evoluzione del sentimento in modo immediatamente comprensibile e con un singolo grafico è stata calcolata la differenza tra i sentimenti positivi e negativi espressi.
Da questo primo grafico di evoluzione generale del sentimento per Facebook si evince che l'umore degli utenti sia molto negativo, soprattutto nell'ultimo periodo, dove addirittura non si vede mai un segno positivo nel grafico. Ovviamente questo risultato ci sembra più che normale, dal momento che si parla di un morbo cronico e con sintomi molto dolorosi, difficilmente ci si può aspettare un sentimento positivo a riguardo.

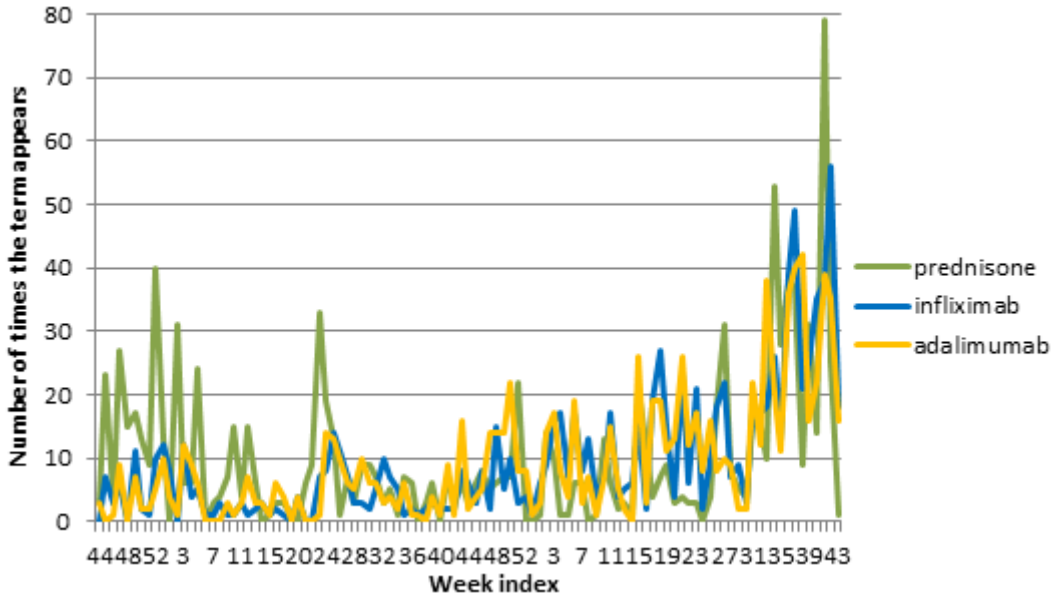


Figura 4.38: Distribuzione dei farmaci su Facebook

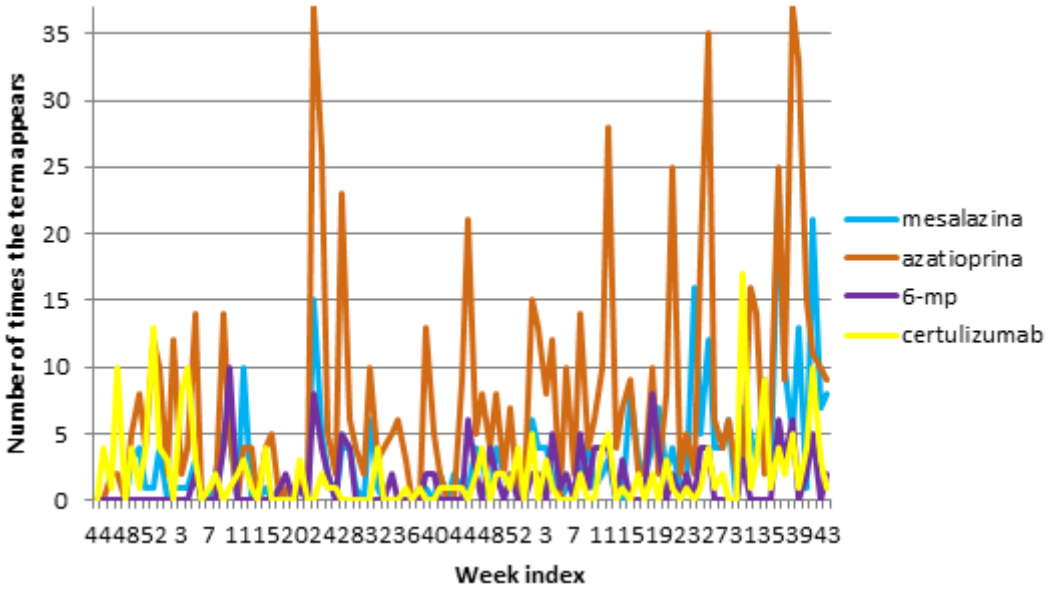


Figura 4.39: Distribuzione dei farmaci su Facebook 2

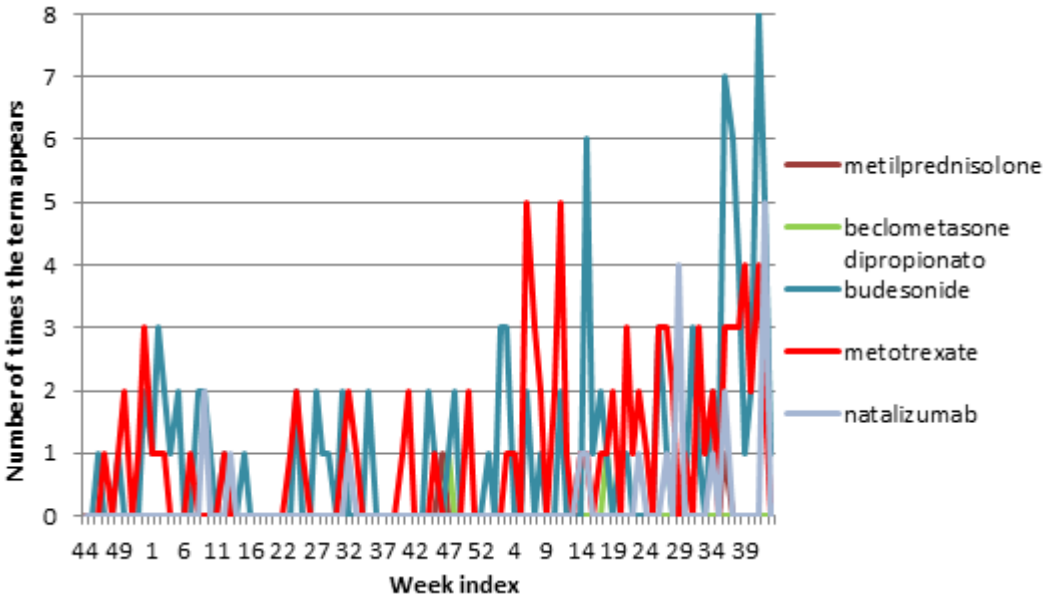


Figura 4.40: Distribuzione dei farmaci su Facebook 3

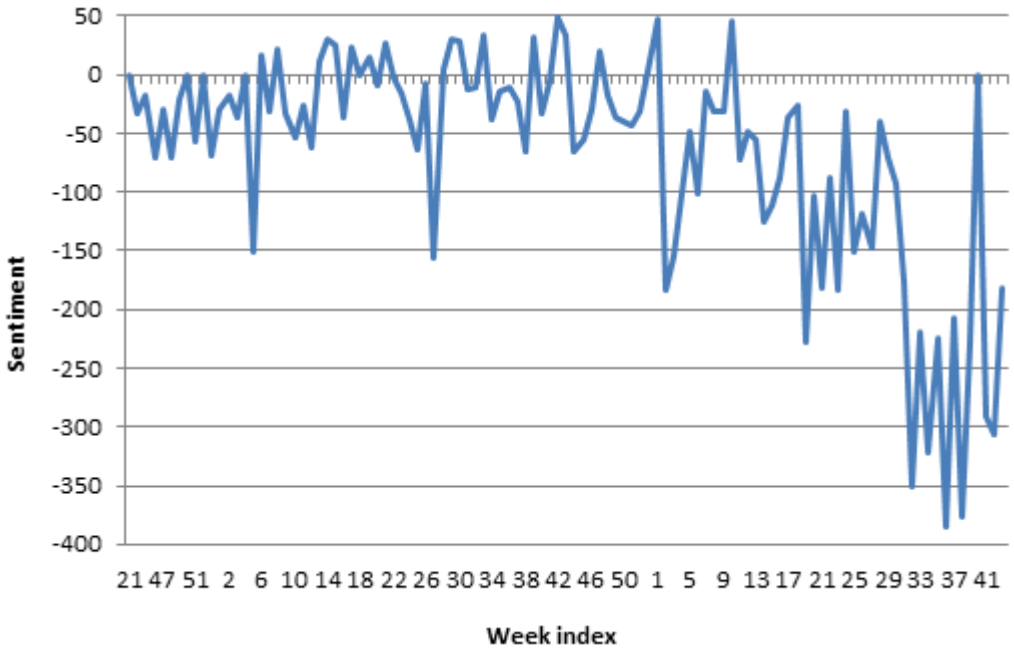


Figura 4.41: Sentimento settimanale generale su Facebook

Volendoci concentrare sul sentimento specifico per ogni trattamento, per ogni farmaco (nominato almeno 40 volte) sono stati estratti i sentimenti dei post dove il farmaco compare ed è stato graficato l'andamento e l'evoluzione del sentimento giornaliero associato ad esso.
In figura 4.42 è espressa l'evoluzione dell'umore sul farmaco "6-mercaptopurine". Si nota che a parte qualche piccolo picco in positivo, l'andamento prevalente è negativo o neutro.
In figura 4.43 l'andamento per "Adalimumab" è più altalenante, mostrando sia picchi positivi che negativi e senza una netta prevalenza per uno dei due versi.
In figura 4.44 il sentimento per "Azatioprina" sembra avere una chiara tendenza ad essere negativo.
In figura 4.45 l'evoluzione per "Budesonide" è chiaramente molto neutra, con pochissimi picchi positivi e leggermente di più negativi.
In figura 4.46 l'andamento per "Certulizumab" sembra essere più positivo che negativo, con qualche picco negativo molto rilevante ma raro e sentimento neutro molto più frequente.
In figura 4.47 è mostrato il farmaco "Infliximab", che sicuramente non lascia spazio a sentimenti neutri, ma altera molto l'umore degli utenti, sia in positivo che in negativo, con una leggera tendenza al positivo nell'ultimo periodo.

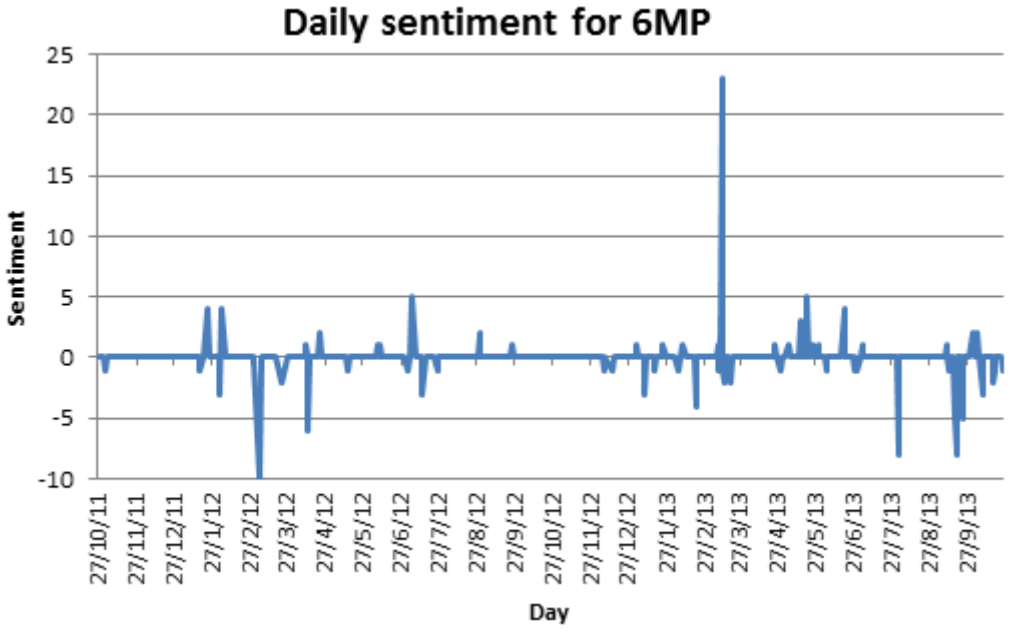


Figura 4.42: Sentimento giornaliero su Facebook per 6-Mercaptopurine

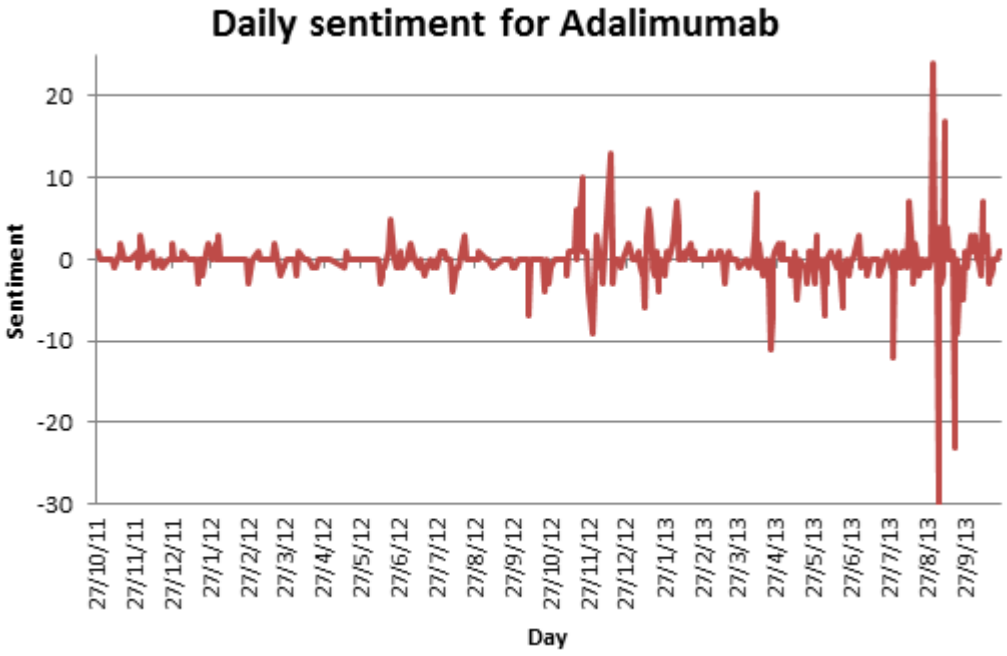


Figura 4.43: Sentimento giornaliero su Facebook per Adalimumab

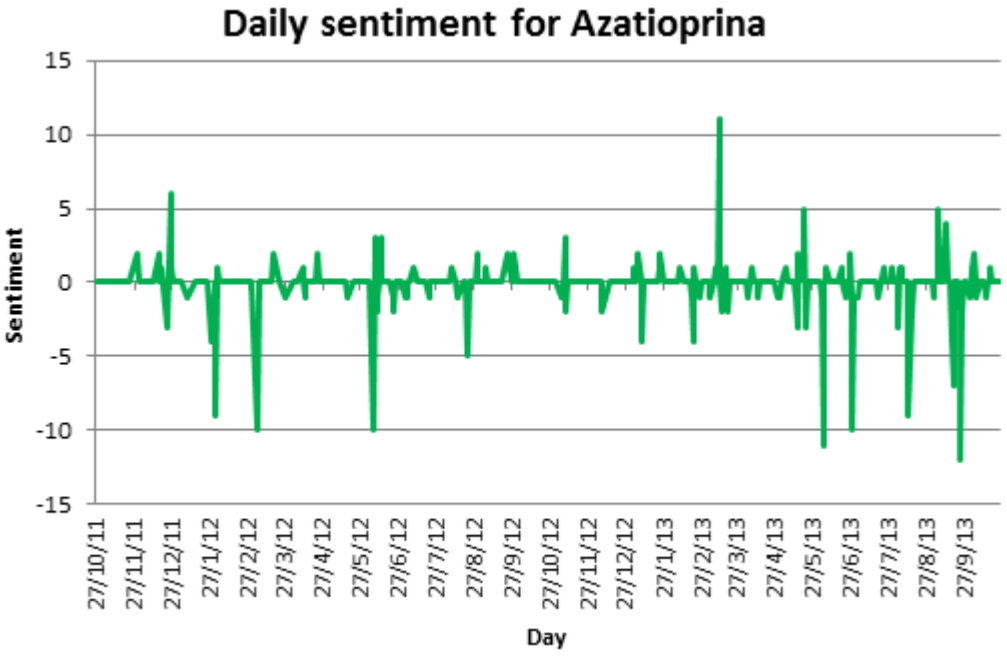


Figura 4.44: Sentimento giornaliero su Facebook per Azatioprina

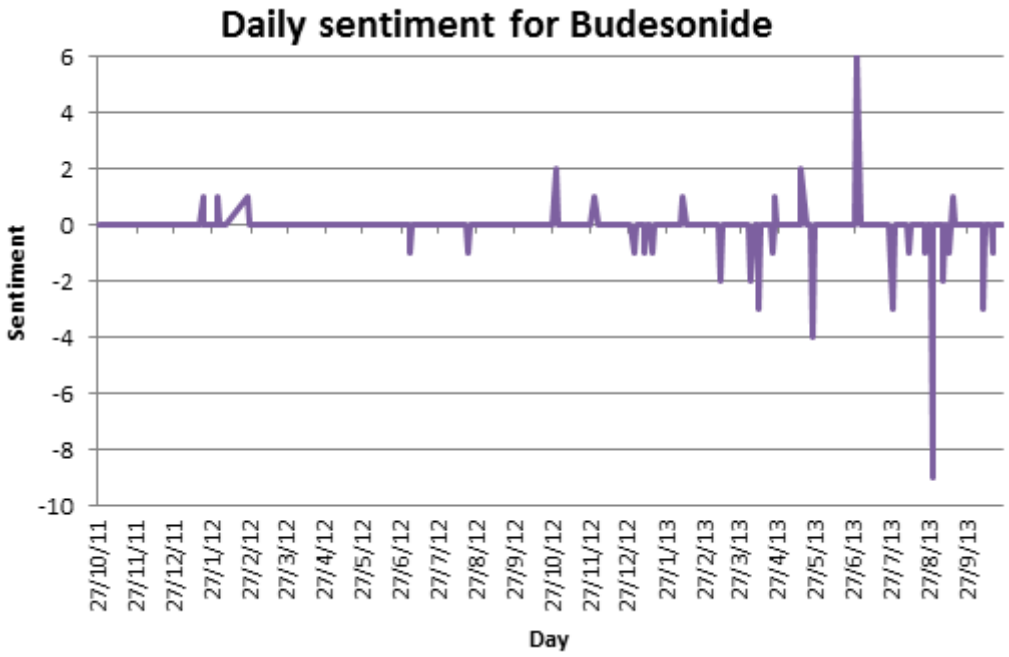


Figura 4.45: Sentimento giornaliero su Facebook per Budesonide

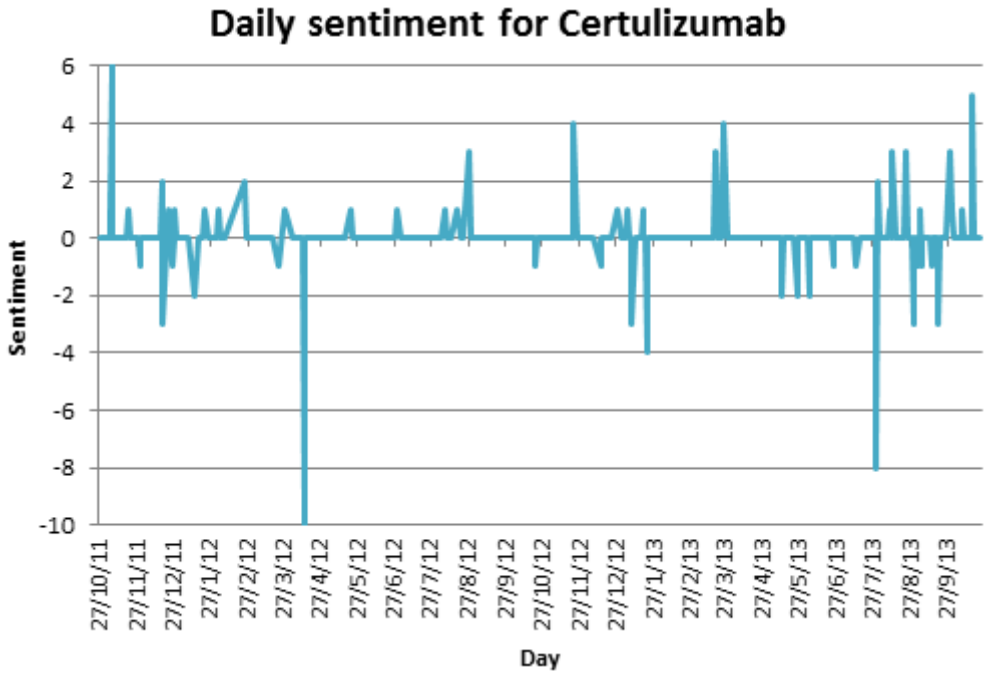


Figura 4.46: Sentimento giornaliero su Facebook per Certulizumab

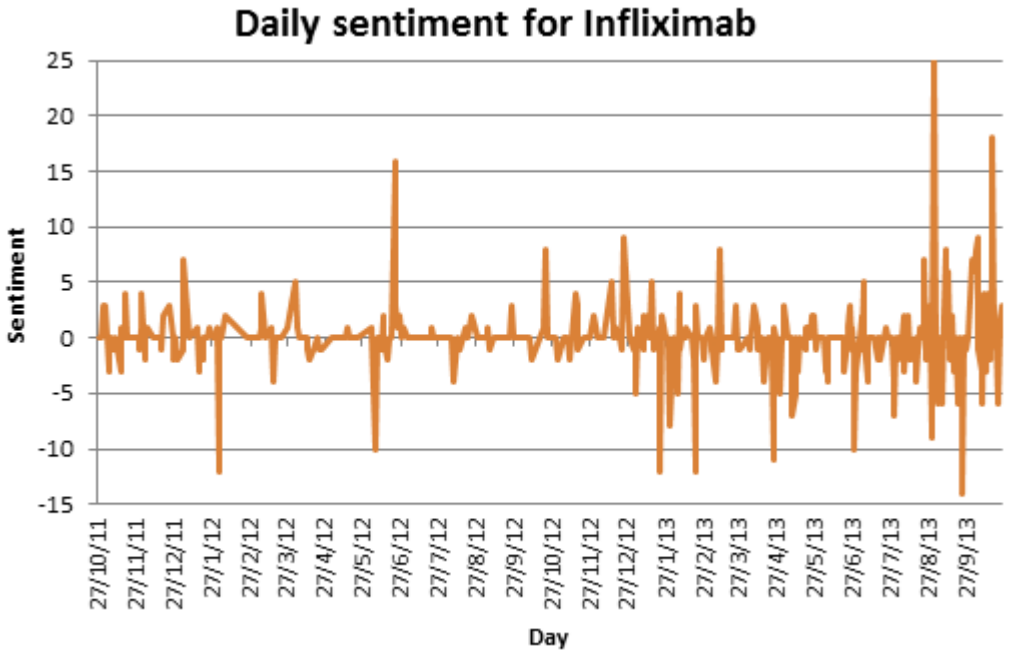


Figura 4.47: Sentimento giornaliero su Facebook per Infliximab

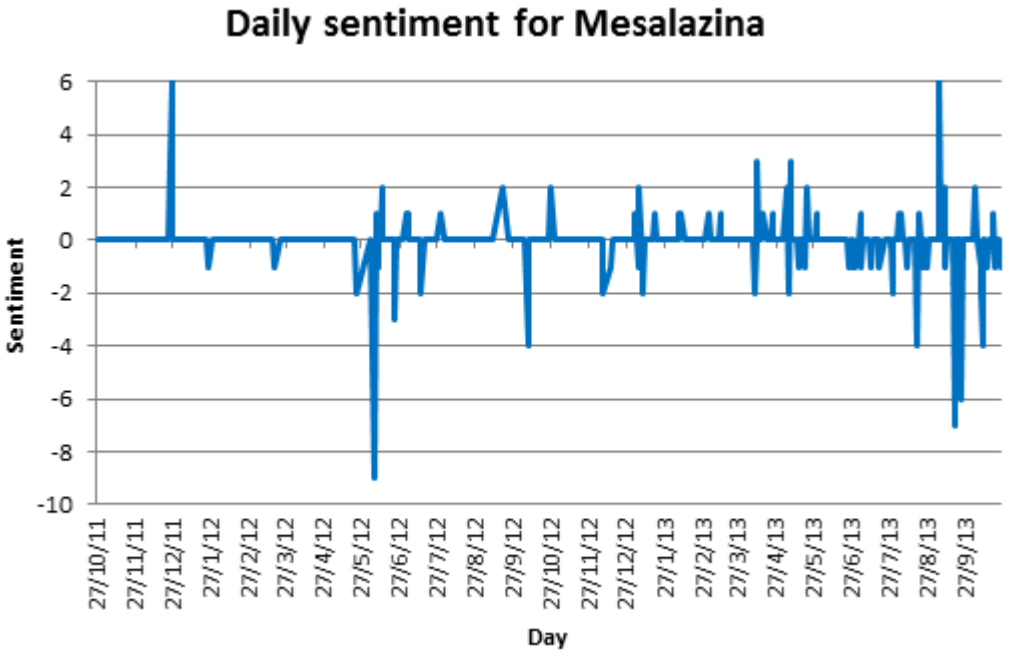


Figura 4.48: Sentimento giornaliero su Facebook per Mesalazina

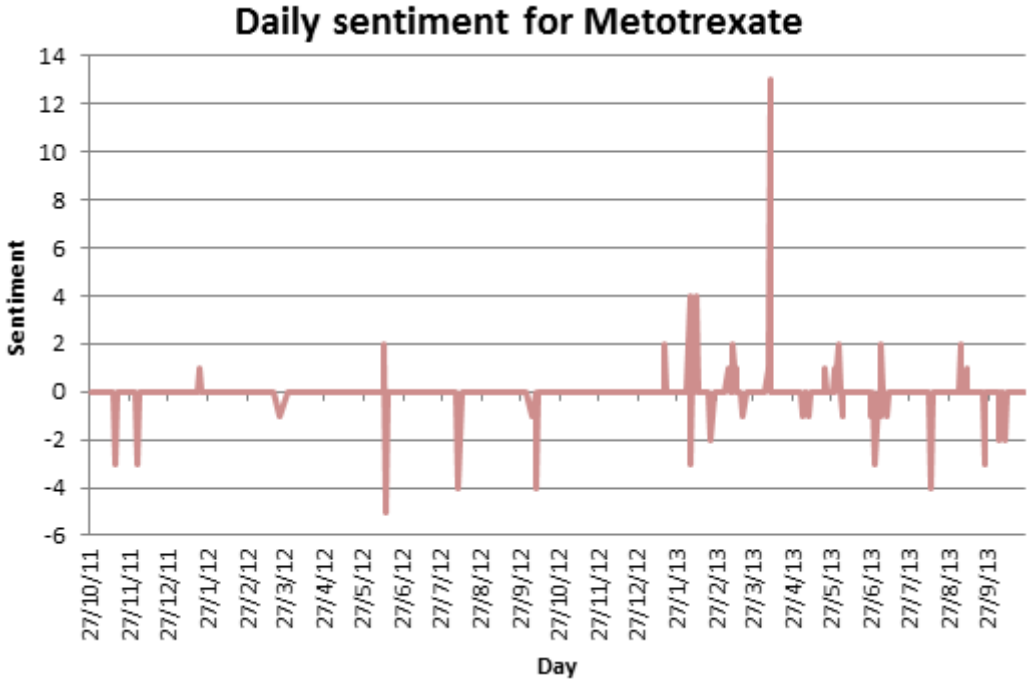


Figura 4.49: Sentimento giornaliero su Facebook per Metotrexate

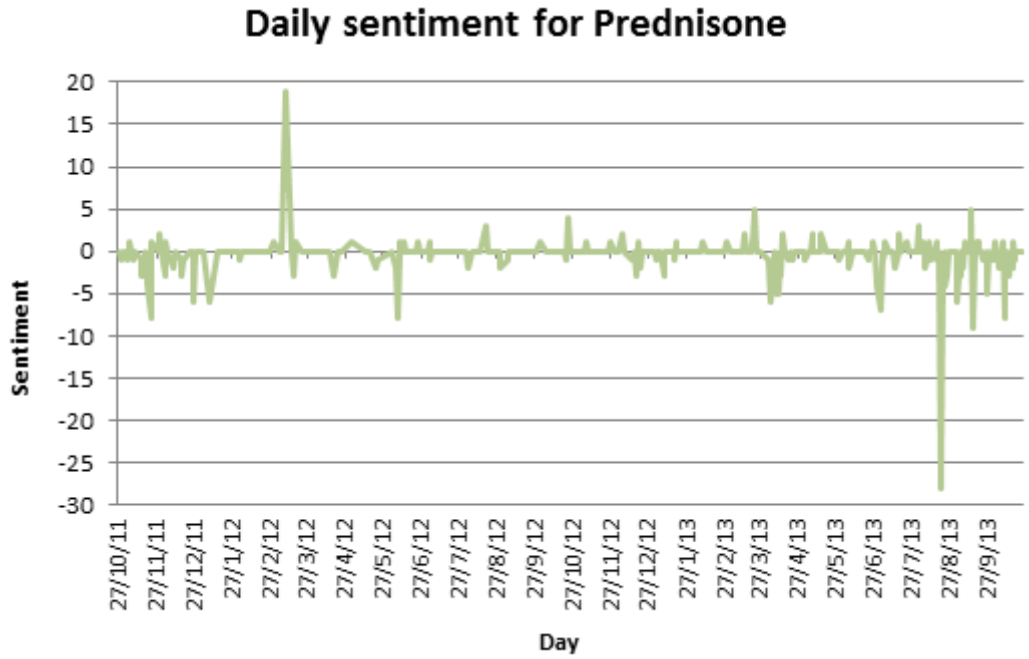


Figura 4.50: Sentimento giornaliero su Facebook per Prednisone

In figura 4.48 per "Mesalazina" si trova invece un andamento più neutro all'inizio e meno più recentemente, dove sembra invece prevalere il sentimento negativo.
In figura 4.49 l'evoluzione di "Metotrexate" indica un sentimento molto neutro e maggiormente negativo all'inizio, con un unico periodo fortemente positivo a metà, per tornare ad un sentimento prevalentemente negativo attuale.
In figura 4.50 è riportato infine "Prednisone" che è il farmaco più discusso come abbiamo visto dall'analisi precedente e che presenta sia sentimenti positivi che negativi, senza rendere chiara da questa prima analisi quale dei due prevalga.

Tra tutti i termini, positivi e negativi, che vengono utilizzati quando si parla di questi trattamenti, quali sono quelli più frequenti?
In Tabella 4.5 sono riportati i vocaboli più frequenti utilizzati con ogni farmaco analizzato.

Treatment	**Three Most Frequent Positive Words**	**Three Most Frequent Negative Words**
Adalimumab	good (46), better (43), remission (36)	bad (39), severe (23), worse (18)
Azatioprina	good (29), better (26), remission (19)	bad (12), sick (11), problems (11)
Budesonide	relief (2), good (2), remission (1)	worse (3), bad (3), problems (3)
Certulizumab	good (11), better (8), best (5)	bad (9), sick (5), problems (3)
Infliximab	better (54), good (51), remission (47)	bad (45), severe (22), sick (20)
Prednisone	better (16), good (15), remission (11)	bad (19), hard (9), severe (9)
Mesalazina	good (21), remission (14), better (13)	bad (17), problems (7), severe (7)
Metotrexate	better (6), good (6), remission (4)	problems (5), bad (4), sick (3)
6MP	good (19), better (11), remission (10)	bad (11), risk (10), sick (8)

Tabella 4.5: Termini più utilizzati positivi e negativi insieme ai trattamenti su Facebook

Da questa tabella è possibile notare un termine che risulta sicuramente ovvio, ma che è stato giustamente estratto dalle nostre metodologie, ed è il fatto che l'argomento più di interesse per i pazienti affetti dal morbo di Crohn e che viene usato frequentemente come termine positivo è "remission". Per remissione si intende l'assenza dei sintomi della malattia, per pazienti con malattie croniche, e questa fase temporanea potrebbe quindi riportare alla comparsa di questi sintomi in un secondo momento.

Dopo aver visto l'andamento del sentimento dei vari trattamenti è possibile dedurre una qualche relazione tra questi? E tra le relazioni possibili, ci sono relazioni di causalità?
Partendo dai dati relativi alle "distanze" tra i trattamenti ed i sentimenti, questa nuova metrica è stata applicata sia alle distribuzioni settimanali dei dati, sia a quelle giornaliere, sia alle distribuzioni che considerano tutti i post analizzati, sia a quelle che scremano i dati ai soli post dove compare almeno un trattamento.
Concentradoci ora sulla sola distribuzione settimanale, in tabella 4.6 sono mostrati i dati relativi alle distribuzioni settimanali considerando tutti i post di Facebook. Si nota da questi risultati che le relazioni tra i farmaci e i sentimenti, sono presenti sempre per entrambi i sentimenti, questo è molto interessante, non trattandosi di relazioni di causalità, ma di relazioni generali sembra logico trovare che un certo trattamento ha relazioni con il sentimento in generale, positivo o negativo che sia. Quando si parla di certi trattamenti quello che risulta è che quindi si può identificare un'umore, qualunque esso sia.

Trattamento	**Sentimento**	**Distanza**
All treatments	negative	0.477947
All treatments	positive	0.50719
Infliximab	negative	0.661796
Infliximab	positive	0.683258
Adalimumab	negative	0.72496
Adalimumab	positive	0.755222
Prednisone	negative	1.146788
Prednisone	positive	1.173216
Azatioprina	negative	1.323242
Azatioprina	positive	1.333466
Mesalazina	negative	1.391368
Mesalazina	positive	1.44751

Tabella 4.6: Distanza con la distribuzione settimanale sulla totalità dei post su Facebook

In tabella 4.7 vediamo gli stessi risultati per le distribuzioni scremate però ai soli post dove compare un trattamento e notiamo la stessa tendenza ad avere relazioni con entrambi i sentimenti per i farmaci elencati.

Trattamento	**Sentimento**	**Distanza**
All treatments	negative	0.325055358
All treatments	positive	0.332793309
Infliximab	positive	0.540705086
Infliximab	negative	0.624028649
Adalimumab	positive	0.657503893
Adalimumab	negative	0.668282584
Prednisone	positive	0.945697601
Prednisone	negative	0.997588098
Azatioprina	negative	1.222452581
Azatioprina	positive	1.249887724
Mesalazina	negative	1.309758234
Mesalazina	positive	1.32983372
Budesonide	negative	1.360887249
Budesonide	positive	1.36264562
Metotrexate	positivi	1.398879567
Metotrexate	negativi	1.480689943

Tabella 4.7: Distanza con la distribuzione settimanale sui soli post dove compaiono trattamenti su Facebook

Ma è possibile dire che qualche farmaco provoca un sentimento negativo o positivo negli utenti?
Dalla semplice visualizzazione dei grafici e dalle relazioni di distanza questo non è assolutamente possibile, è per questo che si è fatto ricorso all'analisi di causalità di Granger, così come specificato nella metodologia. Questa analisi ci restituisce come risultato se è possibile affermare che una distribuzione causa secondo Granger un'altra distribuzione e le nostre due distribuzioni sono ovviamente quella dei trattamenti e quella dei sentimenti.
In tabella 4.8 sono riportati i risultati che hanno rivelato una relazione di causalità secondo Granger e per ottenere questi risultati si è analizzata la distribuzione settimanale dell'intero insieme di post estratti, divisa ovviamente in trattamenti menzionati e sentimenti trovati.
Nella tabella troviamo indicata la relazione estratta, il numero dei passi di esecuzione di Granger a cui si è trovata questa relazione, il valore di statistica F, il valore P e l'R^2 della regressione. Le relazioni di causalità per essere considerate tali e per rientrare in questa tabella devono presentare un valore

$P \leq 0.05$, questa soglia stabilisce il minimo livello di significatività per il quale l'ipotesi nulla viene rifiutata.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Negative cause Infliximab	3	3.548683	0.01727422	1
Mesalazina cause positive	4	5.01364	0.001038633	1
Prednisone casue positive	4	10.2414	6.15186e-07	1
Prednisone cause negative	4	3.010074	0.02191128	1
Adalimumab cause positive	4	2.610279	0.04027593	1
Certulizumab cause positive	4	4.888949	0.001253192	1
Certulizumab cause negative	5	2.617699	0.02930866	1
Budesonide cause positive	5	2.696842	0.02548282	1
All treatments cause positive	4	5.375741	0.0006034341	1

Tabella 4.8: Causalità di Granger, distribuzione settimanale sulla totalità dei post su Facebook

Osservando questi risultati è interessante notare come sembra che il farmaco Mesalazina causi sentimenti positivi, cosa che dal grafico non si sarebbe affermato, ma il grafico rappresenta una distribuzione giornaliera del sentimento, per questa analisi di granger la distribuzione considerata è quella settimanale.
Prednisone e Certulizumab sembrano causare sia sentimenti positivi che negativi e dal grafico già si notava questa presenza di periodi altamenti positivi e periodi negativi.
Adalimumab, Budesonide e la distribuzione con tutti i sentimenti indicati causano sentimenti positivi, mentre caso interessante è Infliximab, che non causa direttamente sentimenti negativi, ma sono i sentimenti negativi a causare il fatto che si stia parlando di Infliximab.
Confrontando questi dati con le relazioni di distanza precedentemente mostrate, si vede che tutti i farmaci che presentavano relazioni con i sentimenti, tranne Azatioprina, presentano anche relazioni di causalità significative con almeno una polarità specifica del sentimento.
Questo fatto è importante per validare entrambe le metodologie applicate.

In tabella 4.9 sono mostrati i risultati dell'analisi di causalità di Granger per le distribuzioni settimanali dei soli post dove compare almeno un trattamento.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Infliximab cause positive	1	5.767553	0.0181348	1
Negative cause infliximab	1	7.986094	0.005672471	1
Infliximab cause positive	4	3.675809	0.007979758	1
Methotrexate cause positive	1	4.485374	0.03661639	1
Negative cause prednisone	2	3.231391	0.04370496	1
Positive cause azatioprina	2	4.640384	0.01184818	1
Azatioprina cause positive	5	2.333483	0.04847505	1
Certulizumab cause positive	3	3.333811	0.02267097	1
Certulizumab cause negative	4	2.548318	0.0443944	1
Budesonide cause positive	4	3.133465	0.01824568	1
Mesalazina cause positive	4	3.022639	0.02160248	1
Mesalazina cause positive	5	4.009302	0.002492051	1
Negative cause all treatment	2	4.669235	0.01153981	1
All treatments cause positive	4	2.77388	0.03153938	1

Tabella 4.9: Causalità di Granger, distribuzione settimanale sui soli post dove compaiono trattamenti su Facebook

In questo caso, considerando i soli post dove si parla di trattamenti, i risultati sono maggiori in quanto più accurati e centrati nel problema.
Quasi tutti i risultati precedenti vengono riportati anche in questo caso, tranne Adalimumab che sparisce dai dati non presentando una chiara preferenza tra sentimenti positvi e negativi e Prednisone che non causa più entrambi gli umori, ma viene causato da sentimenti negativi.
In più da questa analisi troviamo che Infliximab causa anche sentimenti positivi, compaiono altri farmaci che causano reazioni positive, come Azatioprina o Methotrexate.
Interessante è notare che tutti i farmaci insieme causano prevalentemente sentimenti positivi, ma se si sta analizzando un sentimento negativo allora si sta parlando di qualche trattamento.
Ancora una volta, confrontando questi dati con quelli relativi alla distanza tra le distribuzioni, si nota che, tranne per Adalimumab, tutti i trattamenti che presentano qualche relazione con l'umore mostrano anche una relazione specifica di causalità con questo.

Spostando ora le analisi sulle distribuzioni giornaliere e non più settimanali dei dati vengono mostrati risultati sicuramente più scremati e ridotti che mettono in evidenza le relazioni più forti effettivamente presenti.

Le distribuzioni che considerano tutti i post di Facebook presentano relazioni di distanza con solo valori molto alti, per questo non vengono considerati, analizzando invece le distribuzioni scremate ai soli post dove compare almeno un trattamento vengono mostrati più risultati interessanti, mostrati in tabella 4.10.
Solo i primi due valori rispettano il limine di 1.55, ma i restanti valori indicati seppure sforano questo limite possono essere ancora valutati.

Trattamento	**Sentimento**	**Distanza**
All treatments	negative	0.968138878
All treatments	positive	1.012965448
Adalimumab	negative	1.973710883
Infliximab	positive	1.996939321
Infliximab	negative	2.01973775
Adalimumab	positive	2.043675607

Tabella 4.10: Distanza con la distribuzione giornaliera sui soli post dove compaiono trattamenti su Facebook

Anche i risultati di Granger dove vengono considerati tutti i post, non soddisfano il vincolo del valore P, ma due risultati soltanti, che riportiamo in Tabella 4.11, presentano un valore P almeno inferiore a 0.1.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Mesalazina cause positive	3	2.189464	0.087975	1
Prednisone cause positive	5	1.94094	0.085475	1

Tabella 4.11: Causalità di Granger, distribuzione giornaliera sulla totalità dei post su Facebook

In Tabella 4.12 sono presenti invece i dati scremati ai soli post dove compare un trattamento e notiamo invece la presenza di risultati con valore P ottimo. I risultati trovati indicano che Prednisone, Infliximab, Mercaptopurine e Adalimumab causano sentimenti negativi.
Questi ultimi valori per le distribuzioni giornaliere scremate, presentano sempre delle coerenze con le relazioni di distanza, Infliximab ed Adalimumab infatti compaiono in entrambe le tabelle.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Prednisone cause positive	2	3.40503	0.009215	1
Prednisone cause negative	4	4.42166	0.012482	1
Infliximab cause negative	2	4.23446	0.015004	1
Mercaptopurine cause negative	2	3.542576	0.029659	1
Adalimumab cause negative	2	3.469487	0.031876	1

Tabella 4.12: Causalità di Granger, distribuzione giornaliera sui soli post dove compaiono trattamenti su Facebook

Presentiamo ora tutte queste analisi su Facebook anche per Twitter.
In figura 4.51, 4.52, 4.53 sono riportati gli andamenti settimanali di discussione dei vari trattamenti. Si vede subito come la discussione dei trattamenti su Twitter sia meno presente rispetto a Facebook, i trattamenti più discussi sono Azatioprina, Infliximab ed Adalimumab che sono stati menzionati fino a 20 o 30 volte in una settimana, a seguire ci sono Prednisone, Metotrexate, Budesonide e Natalizumab, mentre nell'ultimo grafico sono mostrati farmaci che non vengono quasi mai o mai menzionati.
Nel proseguimento delle analisi successive non verranno più considerati i farmaci che sono stati menzionati meno di 10 volte.
Entrando nel vivo del sentimento per Twitter, anche per questi tweet sono state estratte le polarità del sentimento e la sua evoluzione è mostrata in figura 4.54. L'umore è quasi constantemente negativo tranne per un picco pesantemente positivo tra il 15 maggio 2013 e il 30 maggio 2013.
Volendoci concentrare sul sentimento specifico per ogni trattamento (nominato almeno 10 volte), sono stati estratti i sentimenti dei tweet dove il farmaco compare e nelle prossime figure è visibile l'andamento e l'evoluzione del sentimento giornaliero associato ad esso.

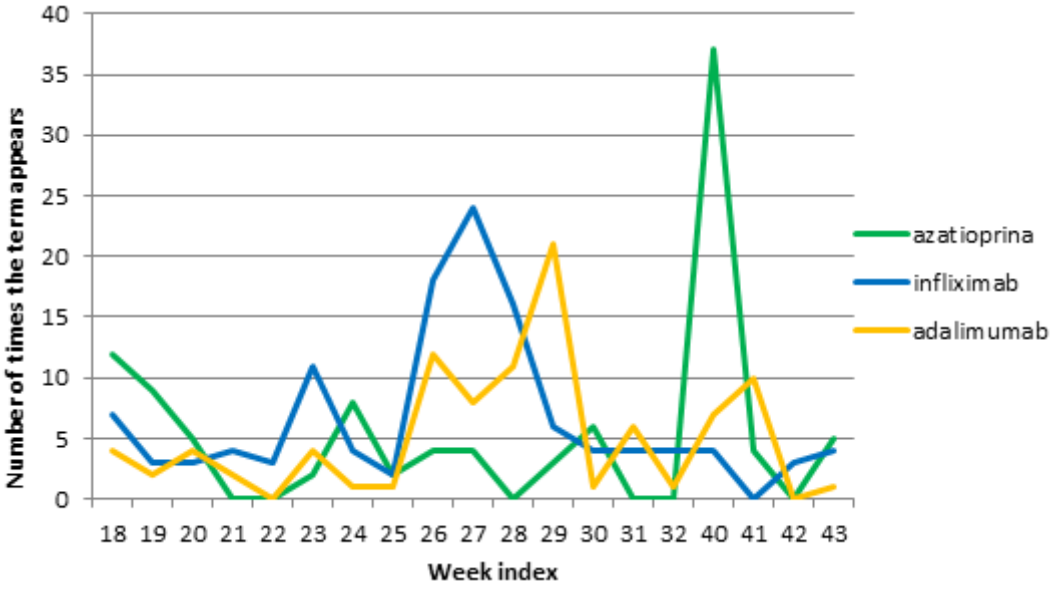


Figura 4.51: Distribuzione dei farmaci su Twitter

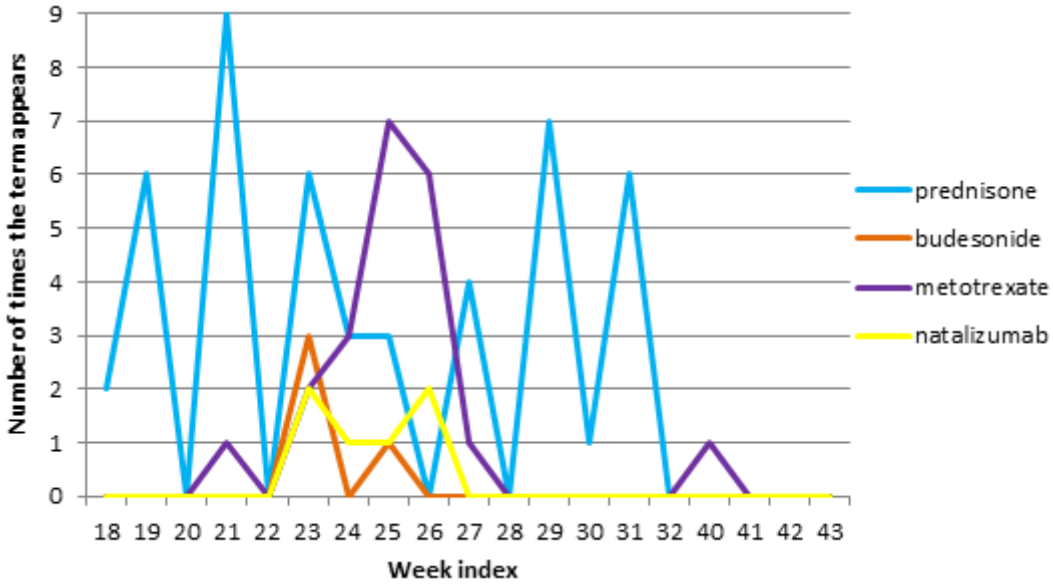


Figura 4.52: Distribuzione dei farmaci su Twitter 2

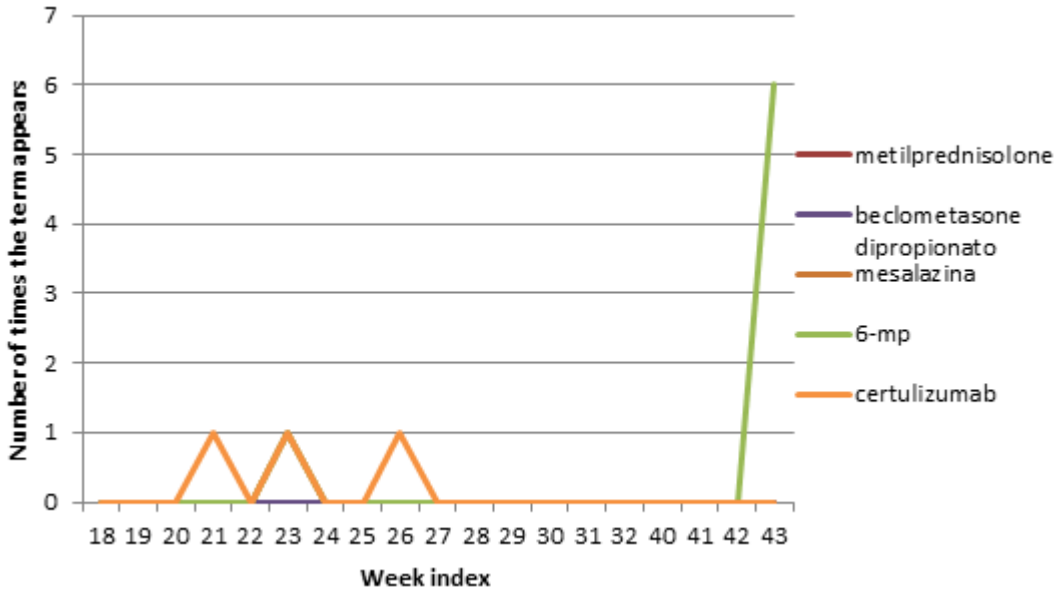


Figura 4.53: Distribuzione dei farmaci su Twitter 3

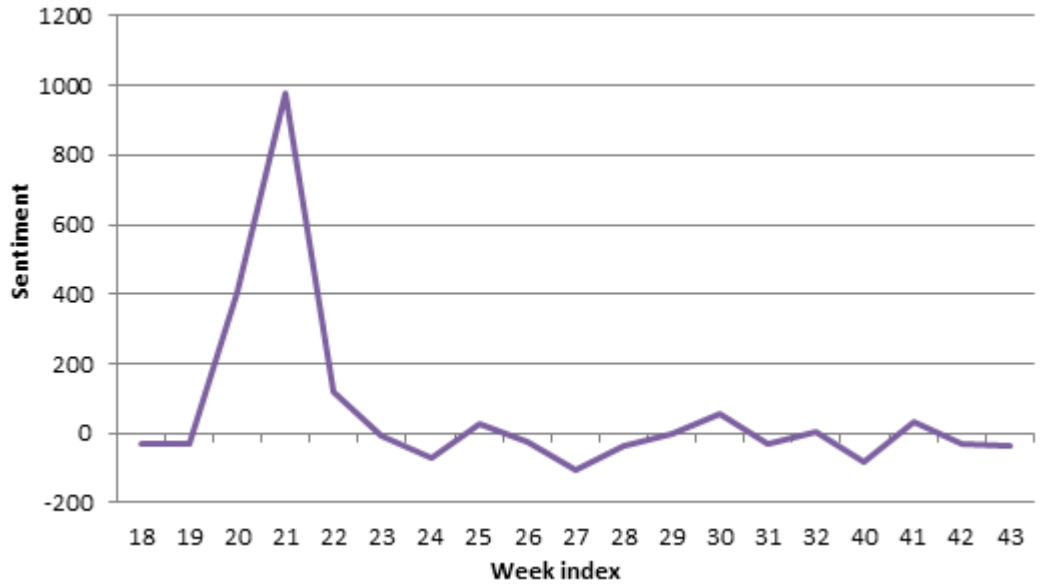


Figura 4.54: Sentimento settimanale generale su Twitter

Dal momento che i trattamenti sono poco discussi su Twitter i grafici successivi che mostrano l'andamento giornaliero dei sentimenti nei tweet in cui i farmaci sono menzionati presenteranno sicuramente molti dati neutri e pochissimi valori.
In figura 4.55 l'andamento per "Adalimumab" è negativo nel primo periodo e più positivo negli ultimi giorni.
In figura 4.56 il sentimento per "Azatioprina" sembra essere leggermente più negativo, ma i dati non nulli sono veramente pochi per poter definire un sentimento associato a questo farmaco.
In figura 4.57 è mostrato il farmaco "Infliximab", che vive diversi momenti, inizia con un sentimento negativo, segue un periodo positivo immediatamente seguito da uno altamente negativo. Solo nell'ultimo periodo sembra diventare neutro e positivo il sentimento.
In figura 4.58 l'evoluzione di "Metotrexate" indica un sentimento quasi sempre nullo, con due unici sentimenti negativi associati.
In figura 4.59 è riportato infine "Prednisone" che mostra solo sentimenti negativi, per quanto siano sporadici.

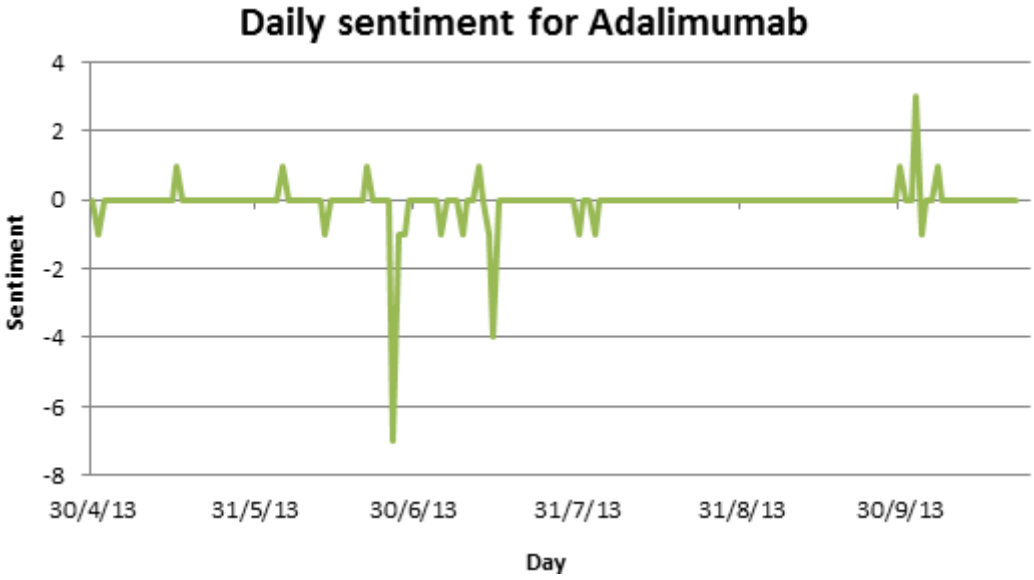


Figura 4.55: Sentimento giornaliero su Twitter per Adalimumab

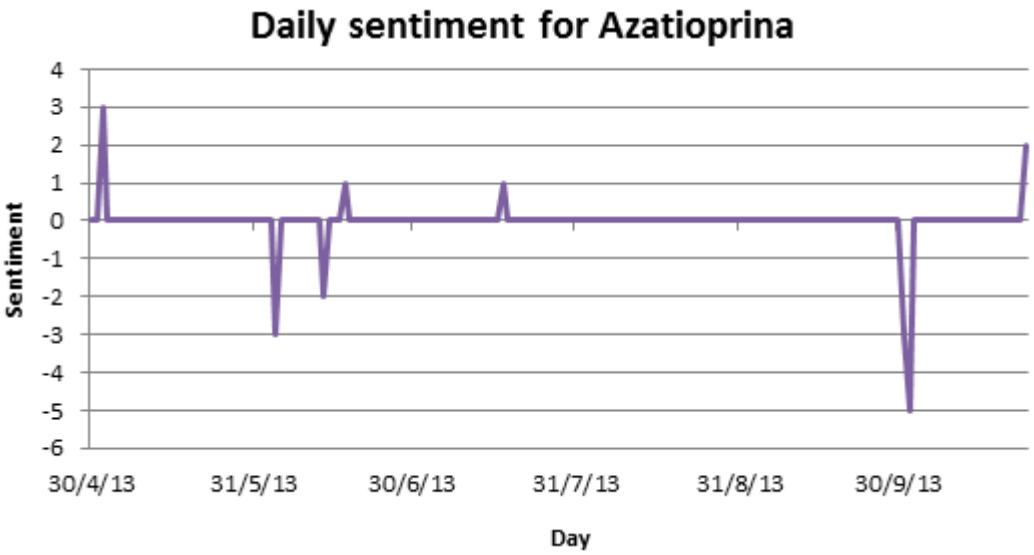


Figura 4.56: Sentimento giornaliero su Twitter per Azatioprina

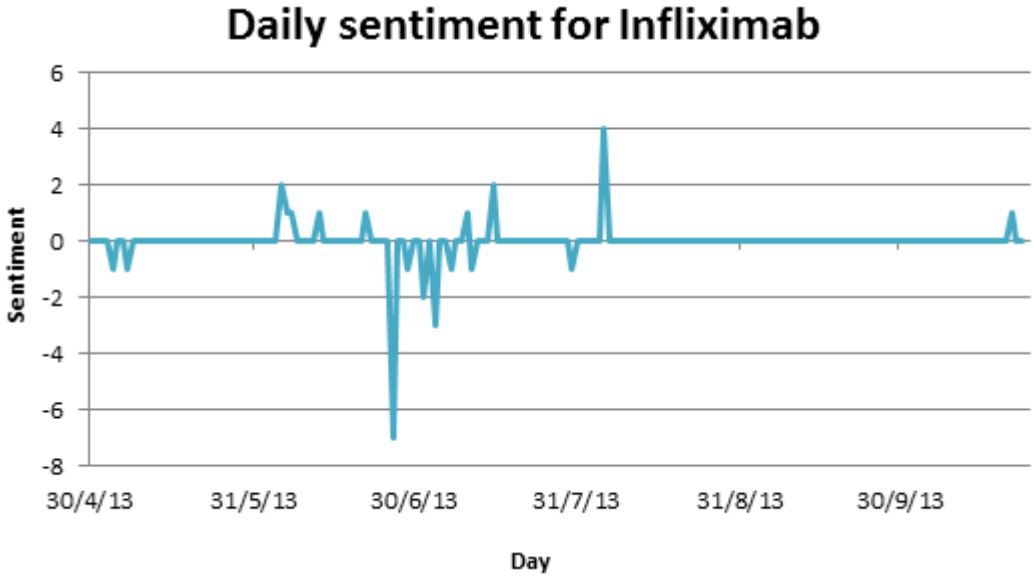


Figura 4.57: Sentimento giornaliero su Twitter per Infliximab

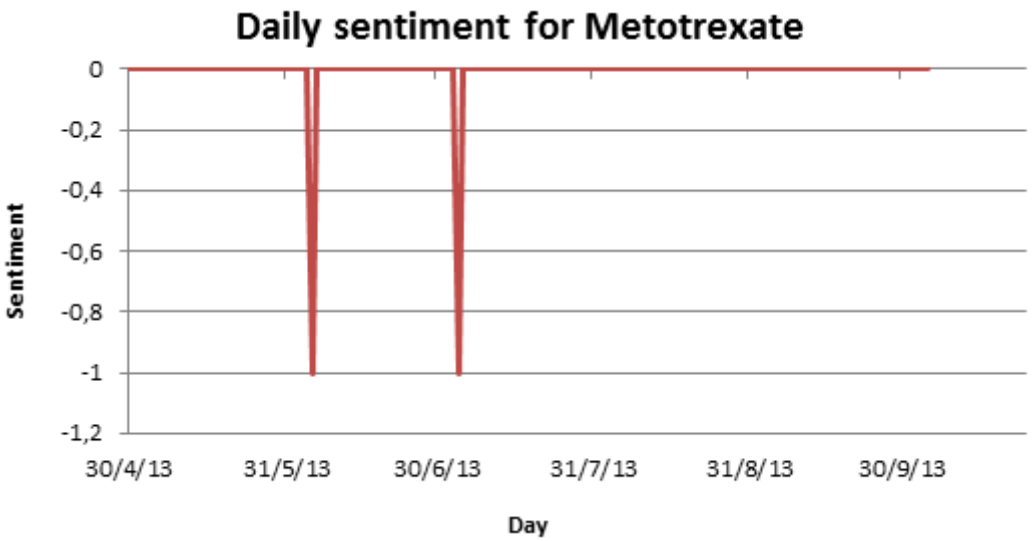


Figura 4.58: Sentimento giornaliero su Twitter per Metotrexate

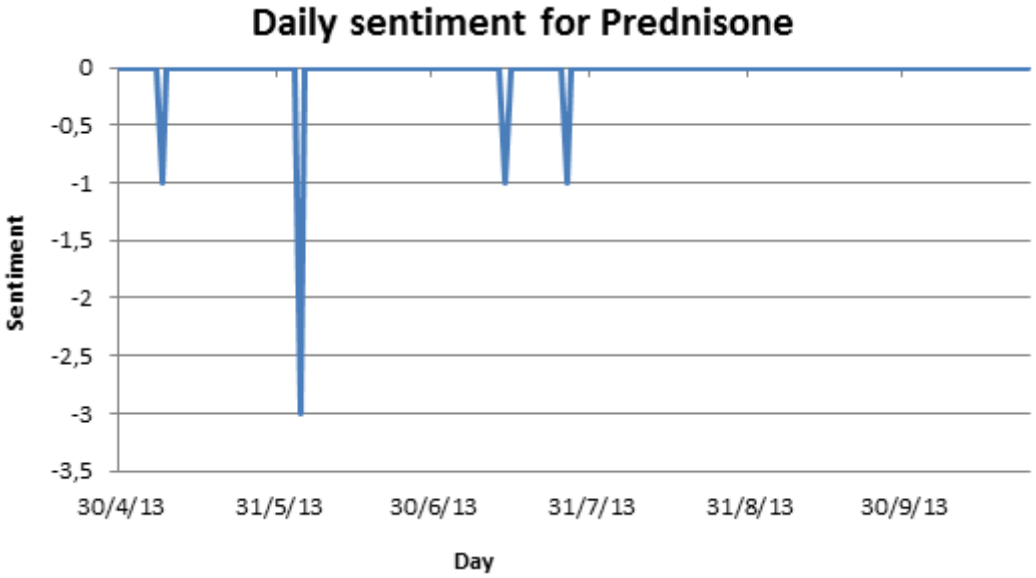


Figura 4.59: Sentimento giornaliero su Twitter per Prednisone

Anche per Twitter sono stati estratti i termini più frequenti (positivi e negativi) usati insieme ad ogni trattamento ed è possibile osservarli in tabella 4.13.
Il centro delle discussioni positive sembra quindi riguardare l'effettivo aumento dei benefici di certi trattamenti, mentre in negativo ci si concentra di più sul tema del fallimento.

Treatment	**Three Most Frequent Positive Words**	**Three Most Frequent Negative Words**
Adalimumab	effective (9), advanced (3), effectiveness (3)	failure (3), refusing (2), severely (2)
Azatioprina	effective (7), benefit (4), surprising (2)	adverse (2), severe (1), pain (1)
Infliximab	effective (13), popular (8), best (4)	failure (4), lose (3), pain (2)
Prednisone	enhanced (1)	shocked (1), complications (1)
Metotrexate	safe (8), benefits (5)	adverse (2)

Tabella 4.13: Termini più utilizzati positivi e negativi insieme ai trattamenti su Twitter

Per quanto riguarda le distribuzioni settimanali su Twitter, le distanze tra trattamenti e sentimenti presentano valori molto più alti rispetto a Facebook, sembrano esserci quindi relazioni meno forti su questo social e questo si vede già bene anche dai grafici appena mostrati per ogni farmaco.
In tabella 4.14 sono indicati i risultati per le distribuzioni che considerano tutti i tweet analizzati. Solamente Infliximab e l'insieme di tutti i farmaci sembrano presentare relazioni significative.
In tabella 4.15 questi stessi risultati sono mostrati per le distribuzioni scremate. Nessun valore rispetta la soglia di 1.55 fissata, ma i risultati con distanza minore vediamo che sono gli stessi trovati anche per tutti i tweet.

Sempre per queste distribuzioni settimanali su Twitter sono analizzate le vere e proprie relazioni di causalità, che rivelano se è possibile affermare che un farmaco provoca un certo umore positivo o negativo. Come fatto in precedenza si utilizza come metodologia l'analisi di causalità di Granger e si considerano i soli valori che presentano un valore $P \leq 0.05$.
In tabella 4.16 sono mostrati i risultati dell'analisi di Granger per la distribuzione settimanale dei sentimenti e dei trattamenti considerando tutti i tweet.

Trattamento	Sentimento	Distanza
All treatments	negative	1.1762962492130975
All treatments	positive	1.1817236142172427
Infliximab	negative	1.3426016702628183
Infliximab	positive	1.351086671108888

Tabella 4.14: Distanza con la distribuzione settimanale sulla totalità dei post su Twitter

Trattamento	Sentimento	Distanza
All treatments	negative	1.8186832110403501
All treatments	positive	1.8371592273009074
Infliximab	negative	1.9820856480981448
Infliximab	positive	2.0131510690763808

Tabella 4.15: Distanza con la distribuzione settimanale sui soli post dove compaiono trattamenti su Twitter

Sembra immediatamente strano vedere che Prednisone causa i sentimenti positivi, in quanto dal grafico vediamo che questo farmaco non ha mai un valore di sentimento positivo, questo può essere spiegato perchè Granger trova che tra le due distribuzioni ci sia una relazione di causalità, ma non dice se è positiva o negativa, l'interpretazione quindi corretta è che Prednisone causa il fatto di non avere sicuramente umori positivi.
Altro dato interessante è che sia sentimenti negativi che positivi causano il fatto che si sta parlando di trattamenti, anche questo significa che entrambi i sentimenti vengono usati per i farmaci e che sicuramente quando viene espresso un umore (positivo o negativo) allora si sta parlando di trattamenti.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Prednisone cause positive	1	6.1022	0.02513	1
Infliximab cause negative	1	6.0441	0.02573	1
Negative cause all treatments	2	5.805414	0.015787	1
Positive cause all treatments	2	4.41144	0.034491	1
Prednisone cause positive	2	3.92376	0.046428	1

Tabella 4.16: Causalità di Granger, distribuzione settimanale sulla totalità dei post su Twitter

Per quanto riguarda la distribuzione settimanale dei tweet però scremati ai soli che contengono trattamenti, questa non presenta causalità con un valore P adeguato.
Confrontando questi dati con quelli precedenti sulle distanze sono confermate le relazioni trovate con Infliximab e tutti i trattamenti in generale che compaiono in entrambe le tabelle.

Analizzando invece la distribuzione giornaliera del sentimento e dei trattamenti, in tabella 4.17 e 4.18 sono riportate le relazioni di distanza, rispettivamente per le distribuzioni totali di tutti i tweet e quelle scremate ai soli tweet dove compare un farmaco.
In entrambi i casi la sola relazione che presenta un valore di distanza adeguato si trova considerando l'insieme di tutti i trattamenti.

Trattamento	**Sentimento**	**Distanza**
All treatments	negative	1.4202032101632547

Tabella 4.17: Distanza con la distribuzione giornaliera sulla totalità dei post su Twitter

Trattamento	**Sentimento**	**Distanza**
All treatments	negative	1.5505646950225067

Tabella 4.18: Distanza con la distribuzione giornaliera sui soli post dove compaiono trattamenti su Twitter

Di queste stesse distribuzioni sono analizzati anche i rapporti di causalità e per quanto riguarda l'intero insieme dei tweet i risultati sono mostrati in tabella 4.19 e mostrano che i trattamenti in generale causano sentimenti negativi e il singolo caso del farmaco Azatioprina li causa entrambi.
Questo stesso risultato è confermato anche dall'analisi di causalità di Granger per la distribuzione giornaliera del sentimento e dei trattamenti sui soli tweet dove compare un farmaco, che come si vede in tabella 4.20 mostra gli stessi dati.
I risultati per queste distribuzioni giornaliere confermano le relazioni di distanza trovate con tutti i trattamenti in generale.

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Azatioprina cause negative	2	10.24185	0.0000877	1
Azatioprina cause positive	2	8.04187	0.000568	1
All treatments cause negative	2	3.29519	0.04101	1
Azatioprina cause positive	3	6.67084	0.0003757	1
Azatioprina cause negative	3	6.28988	0.0005929	1
Azatioprina cause positive	4	4.86498	0.00128	1
Azatioprina cause negative	4	4.8608	0.00129	1
Azatioprina cause negative	5	4.16467	0.001839	1
Azatioprina cause positive	5	3.9678	0.002614	1

Tabella 4.19: Causalità di Granger, distribuzione giornaliera sulla totalità dei post su Twitter

Risultati	**L = n^o di passi**	**Statistica F**	**Valore P**	**R^2 della regressione**
Azatioprina cause negative	2	9.4557	0.000202	1
Azatioprina cause positive	2	8.6166	0.000401	1
All treatments cause negative	2	3.9419	0.023196	1
Azatioprina cause positive	3	7.24958	0.0002336	1
All treatments cause negative	3	4.56975	0.005283	1
Azatioprina cause negative	3	4.20291	0.00821	1
Azatioprina cause positive	4	5.50683	0.00060148	1
Azatioprina cause negative	4	3.52991	0.01071	1
Azatioprina cause positive	5	4.08231	0.00253	1
Azatioprina cause negative	5	3.65733	0.005224	1

Tabella 4.20: Causalità di Granger, distribuzione giornaliera sui soli post dove compaiono trattamenti su Twitter

Capitolo 5

Architettura del sistema software

Questa tesi non si è conclusa con la sola analisi delle metodologie necessarie per analizzare il comportamento e lo stato d'animo dei pazienti malati del Morbo di Crohn tramite i social networks e l'analisi dei relativi risultati, ma si è concentrata anche sullo sviluppo di un sistema software in grado di riprodurre tutte queste analisi in modo automatico e per di più per qualsiasi tema che si voglia analizzare con queste metodologie.
Il software è stato realizzato in Java, con l'utilizzo di Eclipse e la seguente figura 5.1 ne mostra l'architettura.

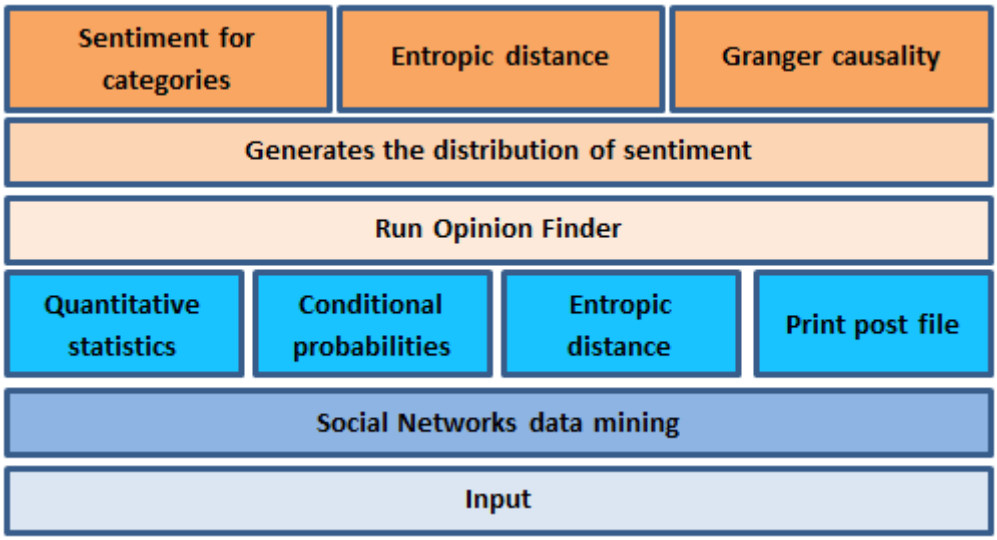


Figura 5.1: Architettura del sistema software

Per poter utilizzare questo sistema è richiesto un solo ulteriore software installato che è R e in questo deve essere installato il pacchetto "Runiversal".

Nel **primo livello** sono presenti gli input da dare al software per permettergli di eseguire le analisi, nella cartella input devono essere quindi presenti i seguenti file:

- **inputTerm.txt** - file che contiene nella prima riga il termine chiave per cui si vuole fare questa ricerca, nelle righe successive contiene per ogni riga il nome della categoria del topic di interesse e l'elenco di tutti i termini che ne fanno parte separati da un tab;
- **inputTermForSentiment.txt** - file che contiene per ogni riga il nome della categoria di cui si vuole analizzare il sentimento e l'elenco di tutti i termini che ne fanno parte separati da tab. Queste categorie potrebbero ovviamente combaciare con quelle precedenti;
- **Rfolder.txt** - file che contiene il percorso assoluto di dove si trova il file Rscript.exe per poter utilizzare il programma R;
- **GrangerR.txt** - file che contiene il codice della funzione da dare ad R per calcolare la causalità di Granger.

Di seguito sono mostrati degli esempi dei primi tre file indicati, che possono e devono essere modificati dagli utenti che utilizzano il programma. Viene indicato anche cosa contiene l'ultimo file "GrangerR" che invece non deve essere mai modificato.

Esempio di "inputTerm.txt":

```
crohn
CAUSES smoking cause virus food eat cows bovine celiac parkinson..
SYMPTOMS disease colitis ulcerative suffer symptom bowel ...
TREATMENTS marijuana remission cannabis vitamin cure diet ...
EFFECTS effect allergy
```

Esempio di "Rfolder.txt":

```
C:/R/R-3.1.0/bin/x64/Rscript.exe
```

Esempio di “inputTermForSentiment.txt”:

PREDNISONE adasone ancortone apo-prednisone bicortone ...
BUDESONIDE bidien budeson budesonide budesonido budesonidum..
MESALAZINA mesalazina azulfidine lialda apriso delzicol ...
AZATIOPRINA azamun azanin azasan azathioprin azathioprine ...
METOTREXATE a-methopterin a-methpterin abitrexate ...
MERCAPTOPURINE 6mp 6 mp 6-mercaptopurin 6mercaptopurine..
INFLIXIMAB infliximab avakine remicade remsima inflectra
ADALIMUMAB adalimumab humira trudexa
CERTULIZUMAB certulizumab cimzia
METILPREDNISOLONE methylprednisolone a-methapred artisone..
BECLOMETASONE aerobec alanase aldecin aldecina aldecine ...
NATALIZUMAB natalizumab tysabri

Cosa contiene “GrangerR.txt”:

```
granger<-function(d, L, k = 1) {
names.d <- dimnames(d)[[2]]
D<-d
for(i in 1:L)
{
D<-ts.intersect(D, lag(d,  - i))
}
dimnames(D)[[2]] <- paste(rep(names.d, L + 1), "_",
rep(0:L, times = rep(2, L + 1)), sep = "")
y  <- D[, k]
n  <- length(y)
x1 <- D[,  - (1:2)]
x0 <- x1[, ((1:L) * 2) - (k %% 2)]
z1 <- lm(y ~ x1)
z0 <- lm(y ~ x0)
S1 <- sum(z1$resid^2)
S0 <- sum(z0$resid^2)
ftest <- ((S0 - S1)/L)/(S1/(n - 2 * L - 1))
pval <- 1 - pf(ftest, L, n - 2 * L - 1)
R2 <- summary(z1)$r.squared
list(ftest = ftest, pval = pval, R2 = R2)
}
```

Il **secondo livello** dell'architettura contiene la componente per estrarre i dati dai social networks.
Ogni volta che si esegue una nuova ricerca e si utilizza il software, viene creata una cartella che ha come nome la chiave di ricerca data in input e il timestamp del momento dell'avvio, per distinguere stesse ricerche fatte in momenti diversi.
Su Facebook vengono cercate le pagine pubbliche che contengono il termine chiave nel titolo e vengono estratti tutti i post presenti all'interno di queste, su Twitter vengono estratti i tweet che contengono il termine chiave nel testo, sia come hashtag che come testo normale.
Oltre al testo dei post e tweet sono salvate anche le informazioni di ognuno, come identificativo, autore, data e tutti gli altri campi già indicati nella metodologia e vengono create tante matrici quante sono le categorie indicate nel file di input dove ogni matrice presenta nelle colonne i termini che fanno parte della categoria e nelle righe l'informazione se il termine è presente o no nei vari post. La lunghezza di queste matrici è quindi pari al numero di post o tweet estratti.
Tutte queste informazioni, i testi e le matrici sono salvate su file e verranno usate da tutte le componenti dei livelli successivi.
Nel solo caso che la ricerca sia fatta sul tema del morbo di Crohn, i post e tweet estratti vengono sommati a quelli già analizzati fino a quel momento, in modo da procedere con analisi più accurate e complete.

Il **terzo livello** dell'architettura contiene più componenti che utilizzano gli output del livello precedente e che sono indipendenti tra di loro.
La prima componente "Quantitative statistics" si occupa di analizzare i dati estratti dai social e di creare le prime statistiche quantitative su questi dati, restituendo in output i risultati di queste statistiche sia in formato testuale che attraverso grafici.
Il tipo di statistiche quantitative effettuate è già stato menzionato nella metodologia e tratta tutte le informazioni sulla distribuzione temporale dei post o tweet scritti, sui diversi autori che scrivono, sulla distribuzione temporale con cui gli utenti riscrivono, sul rapporto tra i vari argomenti trattati e sulla popolarità dei topic affrontati.
La seconda componente "Conditional probabilites" studia i dati con l'approccio bayesiano per cui alcuni argomenti possono condizionare la comparsa di altri. Essendo il software generico ed applicabile a qualsiasi caso e non sapendo quindi a priopri quale argomento causi un altro in senso logico, sono provati tutti i possibili accoppiamenti tra i topic e vengono tutti riportati in vari file, dove è possibile scoprire qual'è la probabilità condizionata tra i vari termini.

La terza componente "Entropic distance" calcola la distanza tra tutti i possibili accoppiamenti di termini di categorie diverse, per lo stesso motivo indicato precedentemente. Il calcolo della distanza è quello spiegato nelle metodologie e i risultati sono salvati su file per cui è possibile osservarli interamente, ma viene anche generato in automatico un grafico a "rete" che mostra le relazioni tra i termini che hanno distanza minore di una certa soglia.
Quarta ed ultima componente di questo livello è "Print post file" e si occupa si preparare i dati che serviranno al livello successivo per l'analisi del sentimento.
La preparazione dei dati consiste nel creare un singolo file per ogni post o tweet perchè Opinion Finder richiede un elenco di file testuali da analizzare. Oltre a tanti file quanti sono i post e tweet, i nomi dati a questi file sono tutti indicati in un unico altro file che servirà sempre per l'esecuzione successiva. Tutti questi file creati sono convertiti in formato UTF-8.

Il **quarto livello** dell'architettura segna il passaggio definitivo all'analisi del sentimento espresso sui social, che sarà esaminato da ogni futura componente.
Questa componente è esterna e si tratta del sistema Opinion Finder che processa documenti e identifica in modo automatico il sentimento delle varie frasi. Questo servizio viene richiamato dando in input il file che contiene destinazione e nome dell'elenco dei file in cui sono salvati i testi dei post e tweet. L'output di Opinion Finder viene salvato nella stessa destinazione dove si trovano i file di input e contiene il sentimento espresso dagli utenti in ogni post o tweet.

Il **quinto livello** dell'architettura analizza i risultati trovati da Opinion Finder, mappando quindi il sentimento espresso in ogni post e ordina questi risultati in modo da generare la distribuzione temporale (giornaliera) del sentimento e la sua evoluzione.
Questa distribuzione rappresenta il punto di partenza per poi derivare quella associata ad ogni singola categoria per la quale si vuole analizzare l'umore. Ricordo che queste categorie sono indicate in un file di input apposito.

Il **sesto livello** dell'architettura utilizza le distribuzioni generate al livello precedente e comprende tre diverse componenti.
La componente "Sentiment for categories" si occupa di generare tutti i grafici che mostrano l'evoluzione del sentimento associato ad ogni categoria, nonchè l'evoluzione della discussione delle categorie stesse.
La componente "Entropic distance" analizza le relazioni di distanza presenti tra le varie categorie e i sentimenti, per tutti i post e tweet, e per le distri-

buzioni scremate ai soli testi dove compaiono le categorie. Viene utilizzata la formula di "distanza" indicata nelle metodologie.
La componente "Granger causality" studia le relazioni di causalità tra le distribuzioni delle categorie e quella del sentimento, per tutti i post e tweet, e per le distribuzioni scremate. Questa componente utilizza il file di input "GrangerR" che contiene la funzione necessaria per eseguire questo studio utilizzando il software R. Da questa analisi vengono restituiti in output 5 file per ogni social e per ogni tipo di analisi (tutti i post o post scremati), ogni file corrisponde all'esecuzione di Granger ad un certo passo e contiene i risultati di tutti gli accoppiamenti fatti a quel passo di analisi.

Utilizzare questo programma è molto semplice, si tratta di due file JAR eseguibili, dal nome "AnalisiSocialParte1.jar" e "AnalisiSocialParte2.jar" e due ulteriori file eseguibili "runOF.bat" e "runOF.sh".
Nella stessa cartella dove si trovano questi file, devono trovarsi anche il file "OpinionFinder.jar", la cartella "models" con i modelli utilizzati da OpinionFinder per la sua esecuzione, la cartella "lib" con al suo interno i file "stanford-postagger.jar" e "weka.jar" e la cartella "input" con i file già discussi precedentemente.
Il primo file da eseguire è "AnalisiSocialParte1.jar" e questo eseguirà in automatico tutte le funzionalità dei primi tre livelli dell'architettura.
Al suo termine dovrà essere eseguito il file "runOF.bat" o "runOF.sh", in base al sistema operativo che si sta utilizzando e questo provvederà a richiamare le funzionalità offerte da OpinionFinder ed a coprire il quarto livello dell'architettura.
Terminata l'esecuzione di OpinionFinder sarà allora possibile eseguire l'ultimo file neccesario "AnalisiSocialParte2.jar" che porterà a termine l'analisi con le funzionalità del quinto e sesto livello dell'architettura.

La necessità di separare il software in due parti, una da eseguire prima e una dopo l'analisi di OpinionFinder, è nata dal fatto che questo programma non espone delle API direttamente utilizzabili da Java, ma deve essere lanciato come programma a parte.
Inizialmente si era tentato di integrare questa chiamata all'interno del resto del codice, generano un nuovo processo e monitorandolo, ma OpinionFinder richiede molta memoria per portare a termine le sue analisi e fare partire questo processo all'interno di un altro portava via molto spazio che non permetta a questo di terminare.
Per questo il sistema è stato diviso in tre parti, per permettere a OpinionFinder di trovare libera tutta la memoria necessaria e terminare senza problemi.

L'unico software che necessita di essere installato è R e come già precisato è necessario installargli il pacchetto "Runiversal" che permette di utilizzare R da Java, grazie alla libreria "RCaller".
R viene usato in questo progetto sia per la generazione di tutti i grafici, sia per il calcolo di causalità secondo Granger.
Se R non è presente o non ne viene indicato correttamente il percorso nel file di input "Rfolder.txt", non verranno restituiti in output i grafici e la Granger causality, ma questo non comprometterà il resto delle operazioni, che verranno lo stesso portate a termine.

Capitolo 6

Conclusioni

In questo lavoro ci si è concentrati sulla comprensione del comportamento online di pazienti affetti dal morbo di Crohn.
I risultati mostrati dimostrano che è presente ed è tutt'ora in corso una ricca discussione intorno a tale tema, e questo è emerso dai messaggi che si trovano sui social networks e come ci si poteva aspettare la maggior parte delle discussioni sono dedicate ai trattamenti e ai sintomi del morbo.
In conclusione il lavoro qui svolto è stato quello di:

- analizzare il livello di discussione del morbo di Crohn sui social networks e il comportamento degli utenti;
- trovare le relazioni che collegano gli argomenti principali legati con la malattia, come cause, sintomi, trattamenti ed effetti collaterali;
- cercare eventuali rapporti tra i trattamenti discussi e i diversi stati d'animo espressi dai pazienti online;
- rendere tutte le analisi automatiche attraverso la creazione di un sistema software e permettere l'applicazione delle metodologie a qualsiasi ambito, creando un sofware quindi il più generale possibile.

Chiaramente i risultati qui presentati, indicano un punto di partenza per futuri approfondimenti ed analisi sulla malattia, sono state indicate le informazioni più rilevanti che emergono quando si cercano le opinioni dei pazienti e che possono dare una direzione per studi clinici approfonditi, ma non si vuole certamente affermare, con questo lavoro, che un certo trattamento sia migliore o peggiore rispetto ad altri.

Bibliografia

[1] Charu C Aggarwal. *An introduction to social network data analytics*. Springer, 2011.

[2] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] AA VV Universtità di Bologna (Dipartimento di Informatica policlinico S. Orsola). Analyzing Crohn's disease patients online behavior and sentiment via social networks.

[5] Andrea Freyer Dugas, Yu-Hsiang Hsieh, Scott R Levin, Jesse M Pines, Darren P Mareiniss, Amir Mohareb, Charlotte A Gaydos, Trish M Perl, and Richard E Rothman. Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clinical infectious diseases*, 54(4):463–469, 2012.

[6] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010.

[7] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[8] Clive WJ Granger. Causality, cointegration, and control. *Journal of Economic Dynamics and Control*, 12(2):551–559, 1988.

[9] Jeremy A Greene, Niteesh K Choudhry, Elaine Kilabuk, and William H Shrank. Online social networking by patients with diabetes: a quali-

tative evaluation of communication with facebook. *Journal of general internal medicine*, 26(3):287–292, 2011.

[10] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

[11] Tim Harford. Big data: are we making a big mistake?, 2014. available online.

[12] David Jensen and Jennifer Neville. Data mining in social networks. In *Dynamic Social Network Modeling and Analysis: workshop summary and papers*, pages 287–302. National Academies Press, 2003.

[13] Jon M Kleinberg. Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 4–5. ACM, 2007.

[14] EV Loftus, P Schoenfeld, and WJ Sandborn. The epidemiology and natural history of crohn's disease in population-based patient cohorts from north america: a systematic review. *Alimentary pharmacology & therapeutics*, 16(1):51–60, 2002.

[15] Danilo Montesi, Matteo Magnani, and Marco Gemelli. Ricostruzione del profilo mediale tramite dati di social network sites. 2010.

[16] Lucila Ohno-Machado. Health surveillance using the internet and other sources of information. *Journal of the American Medical Informatics Association*, 20(3):403–403, 2013.

[17] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 2011.

[18] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.

[19] Elissa R Weitzman, Emily Cole, Liljana Kaci, and Kenneth D Mandl. Social but safe? quality and safety of diabetes-related online social networks. *Journal of the American Medical Informatics Association*, 18(3):292–297, 2011.

[20] Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3):404–408, 2013.

[21] Wikipedia. Crohn's disease — wikipedia, the free encyclopedia, 2014.

[22] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

Elenco delle figure

Elenco delle tabelle