Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il management

Malattie autoimmuni gastrointestinali e sindrome del colon irritabile: un confronto infodemiologico, prima e durante il Covid-19

Relatore:
Chiar.mo Prof.
Roccetti Marco

Presentata da:
Dinelli Michele

Correlatore:
Dott. Casini Luca

Sessione III
Anno Accademico 2021/2022

Indice

Indice **2**

1 Background, obiettivi e ipotesi di ricerca: metodi e risultati attesi **4**
1.1 IBS e IBD 5
1.2 Il ruolo dello stress 6

2 Metodologie di raccolta dei dati **7**
2.1 Dati d'interesse 7
2.2 Praw - Python Reddit Api Wrapper 7
2.3 Estrazione dei dati 8
2.4 Dati raccolti 9

3 Data inisghts **10**

4 Interazioni correlate al virus SARS-CoV-2 **14**

5 Sentiment analysis **17**
5.1 Sentiment analysis e social media 18
5.2 Strumenti utilizzati: Vader e roBERTa 19

6 Sentiment Analysis: risultati ottenuti **22**
6.1 Submissions 22
6.2 Comments 25

7 Sentiment analysis su interazioni correlate al virus SARS-CoV-2: risultati ottenuti **29**

8 Distribuzioni di frequenza **31**

9 Risultati finali **33**

10 Conclusioni **42**

Elenco delle figure **44**

Bibliografia e Sitografia **47**

1 Background, obiettivi e ipotesi di ricerca: metodi e risultati attesi

Questa tesi ha lo scopo di analizzare l'impatto che ha avuto il virus SARS-CoV-2 sulla condizione psicologica dei pazienti affetti dalle malattie infiammatorie gastrointestinali (IBD - Inflammatory bowel disease) e dalla sindrome del colon irritabile (IBS - Irritable bowel syndrome) [1].
Per effettuare questa ricerca sono stati raccolti numerosi dati in formato testuale dal noto social media Reddit, concentrandosi su quattro subreddit che trattano le patologie d'interesse. I dati raccolti interessano un lasso temporale che va da Giugno 2019 e Settembre 2022, così da poter analizzare i cambiamenti avvenuti in seguito alla diffusione del virus.

Si è cercato quindi di determinare se esistono differenze d'interazione legate ai dati raccolti tra il periodo precedente e antecedente alla pandemia, in modo da poter affermare, eventualmente, se queste modifiche possono essere correlate all'avvento del virus SARS-CoV-2.
Per questo motivo i dati sono stati divisi in due gruppi delimitati da Febbraio 2019 e confrontati sotto tre aspetti: la variazione del numero di post pubblicati, l'andamento del sentimento negativo dei post utilizzando tecniche di sentiment analysis e un'eventuale differenza d'interazione nei singoli subreddit a seconda del tipo di patologia che tratta.

In relazione alle tre analisi effettuate si è giunti alla conclusione che il numero di post ha subito un incremento rispetto al periodo precedente alla pandemia, così come la negatività dei post ha subito variazioni significative in seguito alla diffusione del virus. Non è stato possibile invece determinare se i due gruppi di subreddit che rappresentano IBD e IBS hanno registrato differenze d'interazione significative tra di loro.

[1] Il termine virus SARS-CoV-2 fa riferimento al virus che causa la malattia da coronavirus (COVID-19). In questo documento verrà citato per nome completo o più semplicemente con il termine "virus"

1.1 IBS e IBD

IBS e IBD sono due tipologie di malattie croniche che interessano l'apparato gastrointestinale. Le due condizioni hanno sintomi simili, tuttavia mentre IBD è caratterizzata da infiammazione IBS non lo è, rendendo più difficile la sua diagnosi.

La sindrome dell'intestino irritabile (IBS) è un disturbo funzionale cronico del tratto gastrointestinale, caratterizzato da dolore addominale e alterazioni delle abitudini intestinali come costipazione, diarrea e crescita batterica anormale nell'intestino tenue.
Le cause sono molteplici e, nello stesso individuo, non è riconoscibile un singolo fattore scatenante. Da un lato vi sono fattori psico sociali, come il comportamento verso le malattie, aspetti cognitivi ed emotivi; dall'altro fattori biologici, come la predisposizione e la suscettibilità individuale, alterazioni della motilità del tratto digestivo, la sensibilità dei visceri, la percezione soggettiva del dolore, la flora batterica e infezioni intestinali [3] [2].

IBS è una malattia di gran lunga più comune di IBD, infatti IFFGD (International Foundation for Functional Gastrointestinal Disorders) stima che IBS affligga fino al 15% della popolazione mondiale [2].

IBD è un termine che raccoglie le due principali malattie infiammatorie gastrointestinali ovvero il Morbo di Crohn (CD - Crohn's Disease) e la Colite Ulcerosa (UC - Ulcerative Colitis). Un'altra forma di IBD, la Colite indeterminata, viene diagnosticata quando gli esami non riescono a distinguere inizialmente di quale forma di IBD si tratti. La maggior parte dei casi di colite indeterminata evolve verso la diagnosi del Morbo di Crohn o Colite Ulcerosa.
CD e UC sono malattie simili per molti aspetti e la diagnosi differenziale spesso risulta difficile. Tuttavia, un tratto che le differenzia è dato dall'interessamento di quasi tutto il tratto digerente da parte del Morbo di Crohn, mentre la colite ulcerosa interessa quasi sempre solo il colon.
I sintomi causati dalla IBD variano in funzione della parte dell'intestino coinvolta e dalla presenza di morbo di Crohn o colite ulcerosa. I pazienti con morbo di Crohn presentano generalmente diarrea cronica e dolore addominale. I pazienti con colite ulcerosa presentano di norma episodi intermittenti di dolore addominale e diarrea emorragica. In entrambe le malattie, i pazienti con diarrea di lunga anamnesi possono perdere peso e diventare denutrite. In alcuni casi l'IBD causa infiammazione in altre parti del corpo come articolazioni, occhi, bocca,

fegato, cistifellea e cute. L'IBD aumenta inoltre il rischio di cancro nelle aree dell'intestino affette.
La causa dell'IBD non è nota, ma le evidenze suggeriscono che i normali batteri intestinali attivano incorrettamente una reazione immune in soggetti con una predisposizione genetica. Normalmente, il sistema immunitario limita gli effetti di batteri, virus e tossine ambientali dannosi. Nelle persone affette da IBD, il sistema immunitario vacilla e si sviluppa un'infiammazione nel tratto gastrointestinale. Giocano un ruolo importante anche i fattori ambientali come lo stress. [2].

Dunque IBD e IBS, nonostante le similitudini sintomatiche, sono malattie molto diverse tra di loro. Un paziente a cui è stata diagnosticata una delle due può infatti manifestare i sintomi dell'altra e addirittura soffrire di entrambi contemporaneamente.

1.2 Il ruolo dello stress

Lo stress è una risposta psicofisica naturale e può avere la funzione benefica di attivare risorse e guidare alla risoluzione di problemi. Tuttavia, nella vita quotidiana sono numerose le fonti di stress e un'attivazione eccessiva per intensità e prolungata nel tempo può compromettere lo stato di salute di un individuo [8].

La ricerca pubblicata sull'NCBI (National Center for Biotechnology Information) [12] dimostra che lo stress può influire sia sull'insorgenza dell'IBD sia sul decorso della malattia. In particolare, può aggravare la malattia e contribuire alla comparsa di riacutizzazioni. Lo stress può anche intensificare i disturbi dell'umore che talvolta si manifestano nelle persone affette dalla malattia. I disturbi emotivi, come la depressione e l'ansia, sono comuni tra le persone affette da IBD e lo stress può intensificarne gli effetti [2].

È stato approfondito brevemente il ruolo dello stress per queste tipologie di malattie perché utile a ricordare lo scopo finale della ricerca. Lo stress è stata sicuramente una componente costante durante la Pandemia, quindi uno dei fattori ipoteticamente legato alle variazioni attese del sentiment estratto dai dati.

2 Metodologie di raccolta dei dati

2.1 Dati d'interesse

I dati utilizzati per questa ricerca sono informazioni in formato testuale in lingua prevalentemente Inglese estratte da Reddit.
Reddit è una rete di community in cui le persone interagiscono approfondendo i loro interessi, hobby e passioni. Le community prendono il nome di subreddit. Ogni subreddit tratta un argomento specifico e al suo interno la partecipazione è libera: ognuno può contribuire postando contenuti o commentando post già presenti [2].

La terminologia utilizzata da Reddit è submission, in riferimento a un post e comment in riferimento a un commento. Vedendola dal punto di vista informatico una submission rappresenta la radice dell'albero mentre i comment rappresentano i nodi, fino alle foglie.

Sono stati identificati quattro subreddit rilevanti che ospitano le discussioni su IBD e IBS e sono i seguenti: r\IBD, r\ibs, r\CrohnsDisease e r\UlcerativeColitis, tre di questi trattano di IBD mentre uno tratta di IBS.
I dati ottenuti da questi subreddit rappresentano il contributo personale di pazienti, medici o persone che hanno a che fare con IBD e IBS. Possono quindi conservare informazioni importanti riguardo allo stato psicofisico degli utenti. Questa tipologia di dati è perfetta per lo scopo della ricerca perché può essere sottoposta a natural language analysis e a verifiche statistiche.

2.2 Praw - Python Reddit Api Wrapper

L'accesso ai dati e la loro estrazione è stata effettuata utilizzando uno script scritto in Python. [3] Lo script fa uso di una libreria Python chiamata Praw [10].
Per poter utilizzare correttamente Praw è stato necessario registrare lo script al sito ufficiale di Reddit nella sezione riservata agli sviluppatori. Una volta registrato lo script vengono ottenuti un client id e un client secret. Il client id è l'identificativo pubblico di un'applicazione mentre il client secret è un segreto

[2] La notazione per descrivere un subreddit è la seguente: r\nomeSubreddit

[3] Il codice sorgente è disponibile su una repository pubblica a questo indirizzo https://github.com/micheledinelli/Crohns-Sentiment-Analysis

condiviso tra applicazione e authorization server, una sorta di password personale dell'applicazione.

Una volta in possesso di queste due informazioni è possibile istanziare un client Reddit messo a disposizione da Praw. Il client offre le api base di Reddit ed è molto vantaggioso in quanto astrae completamente l'utilizzo dalla logica sottostante.

Sono stati estratti in prima istanza gli id dei post. Gli id sono identificativi univoci interni al sistema Reddit, che sono messi a disposizione per avere un riferimento dell'entità submission a partire dal suo id. Successivamente grazie agli id sono state ottenute massivamente le submission, infine i relativi commenti [4] [5].

2.3 Estrazione dei dati

È stato determinato come periodo d'interesse il periodo tra Giugno 2019 e Settembre 2022. Fatto ciò sono stati recuperati gli id delle submission utilizzando Pushift Api [11]. Pushift espone due endpoint principali: /reddit/search/comment e /reddit/search/submissions dai quali possono essere ottenuti gli id di commenti e submissions. Le api permettono anche di filtrare, ordinare e raggruppare il payload di risposta attraverso query params nella richiesta.

Per un efficiente estrazione delle submissions è stato diviso il periodo temporale d'interesse in 27 mesi. Per ognuno dei quattro subreddit sono state effettuate richieste a Pushift per ottenere gli id delle submissions giorno per giorno. Per ogni mese e per ogni subreddit sono stati raggruppati gli id ottenuti ed esportati su file csv. Infine è stato creato a partire dai csv singoli di ogni mese un unico file csv per gli id di tutte le submissions di tutti i subreddit.
Iterando su tutti gli id è stato utilizzato Praw per ottenere e salvare le submissions sotto forma di dati strutturati.

L'estrazione dei commenti invece è stata più agile perché a partire da ogni submissions è possibile ottenere i commenti relativi accedendo al campo comments dell'oggetto submission questo è stato un dei vantaggi dell'utilizzo di Praw. Ottenuti i commenti sono stati anch'essi salvati su un file csv.

[4] La profondità dell'estrazione dei commenti è costante e uguale a 1. Per ogni submission vengono estratti i commenti di primo livello

[5] Reddit permette di commentare i commenti e non solo le submissions

2.4 Dati raccolti

Una volta estratti i dati in forma strutturata è stata prevista una pipeline di data cleaning, così da assicurarsi che quanti più dati possibili siano rilevanti per la ricerca. Per ogni entry è stato filtrato il contenuto testuale attraverso apposite espressioni regolari e sono stati eliminati i dati con testo nullo o vuoto.
La quantità di dati finale è modesta, in totale sono state ottenute **67 481** submissions e **480 165** commenti [6].

La struttura delle submissions è leggermente differente dalla struttura dei commenti ma in generale entrambi condividono i seguenti campi: id, subreddit, author, score, url, body, created.

title	author	score	id	subreddit	url	num_comments	body	created
Align Probiotic Gas	Complete_Lack4089	6	k4x6e3	ibs	https://www.reddit.com/r/ibs/comments/k4x6e3/a...	9	Does Align Prebiotic + Probiotic give anyone e...	2020-12-02 00:41:35

Figura 1: Esempio della struttura di una submission

Il campo title si riferisce al titolo della submission, così come author si riferisce al suo autore. Lo score è invece il numero di upvotes ricevuto. Si ha anche informazione in merito al subreddit trattato e al numero di commenti. I campi di maggior interesse sono il body e il campo created che definisce il momento di creazione della submission.
I commenti presentano inoltre il campo parent_id che è la combinazione del prefisso t3 (indicante un riferimento a una submissions) e l'id della submissions stessa. L'utilizzo dei prefissi è documentata nelle api di Reddit [4].

parent_id	author	score	id	subreddit	body	created
t3_k4xgh0	biljardbal	6	gebmqwv	CrohnsDisease	I tried to make one like this, but it seems th...	2020-12-02 02:32:38

Figura 2: Esempio della struttura di un comment

[6] I numeri riportati fanno riferimento ai dati leggibili, quindi con corpo non vuoto e informazioni non inconsistenti

3 Data inisghts

L'analisi dei dati e la loro visualizzazione sono state effettuate utilizzando Python e alcune delle librerie maggiormente utilizzare a tali scopi. Per l'accesso e la manipolazione dei dati è stata utilizzata la libreria pandas che offre combinata alla libreria numpy accesso rapido ed efficiente ai dati raccolti, anche per datasets abbastanza grandi.
Per la visualizzazione dei dati è stata utilizzata la libreria matplotlib e nello specifico si è sfruttata l'interfaccia offerta da seaborn per poter produrre grafici agilmente e quanto più possibile chiari.
Il codice Python è stato eseguito utilizzando Jupyter Notebook, una web application che permette l'esecuzione di svariati linguaggi ed è molto utilizzato con Python in quanto crea un ambiente virtuale che può essere gestito attraverso il package manager classico di Python, importando facilmente le librerie necessarie.

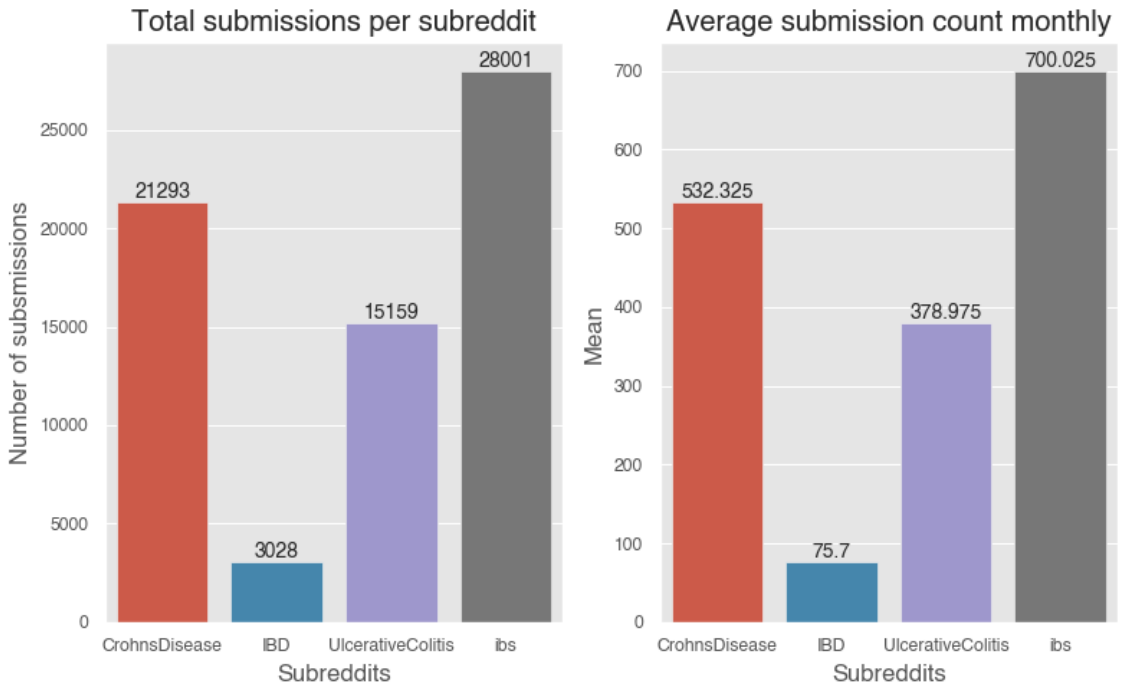

Figura 3: Numero totale di submissions e media mensile per subreddit

CrohnsDisease Dal subreddit r\CrohnsDisease sono state estratte **21 293** submission e **172 959** commenti. Il mese che ha contato più submission è stato Agosto 2022 con **821** submissions. Mediamente sono state postate circa **532** sub-

missions al mese. Marzo 2022 è stato il mese con più commenti ben **6 878** a fronte di una media mensile di circa **4 324**.

IBD Dal subreddit r\IBD sono state estratte **3 028** submission e **21 334** commenti. Il mese che ha contato più submission è stato Agosto 2022 con **135** submissions. Mediamente sono state postate circa **75** submissions al mese. Maggio 2021 è stato il mese con più commenti **983** a fronte di una media mensile di circa **533**.

ibs Dal subreddit r\ibs sono state estratte **28 001** submission e **176 659** commenti. Il mese che ha contato più submission è stato anche in questo caso Agosto 2022 con **1 203** submissions. Mediamente sono state postate circa **700** submissions al mese. Luglio 2022 è stato il mese con più commenti **7 542** a fronte di una media mensile di circa **4 416**.

UlcerativeColitis Dal subreddit r\UlcerativeColitis sono state estratte **15 159** submission e **109 178** commenti. Il mese che ha contato più submission è stato Agosto 2022 con **737** submissions. Mediamente sono state postate circa **378** submissions al mese. Gennaio 2022 è stato il mese con più commenti **4 898** a fronte di una media mensile di circa **2 729**.

In figura 4 viene mostrato l'andamento del numero di submissions mensili per tutti e quattro i subreddit. Sull'asse delle ascisse è posto il riferimento al mese mentre l'asse delle ordinate fa riferimento al numero di submission.
Dal grafico spicca l'alto numero di submissions pubblicate nel subreddit r\ibs, infatti in nessun mese si è verificato che un subreddit registrasse più submissions. Va però considerato che r\ibs è l'unico subreddit analizzato che tratta della sindrome del colon irritabile. I subreddit r\CrohnsDisease, r\IBD e r\UlcerativeColitis trattano tutti di malattie infiammatorie gastrointestinali. Quindi considerando questo è accettabile che r\ibs abbia molte più submission in quanto probabilmente è il subreddit di riferimento per la sindrome del colon irritabile.

Per tutti i subreddit a Marzo 2021 si è verificato un decremento del numero di interazioni che è probabilmente dovuto a un disservizio delle api e può essere considerato come mese outlier . Si può inoltre notare un incremento comune ai subreddit del numero di submissions dopo Marzo 2020.

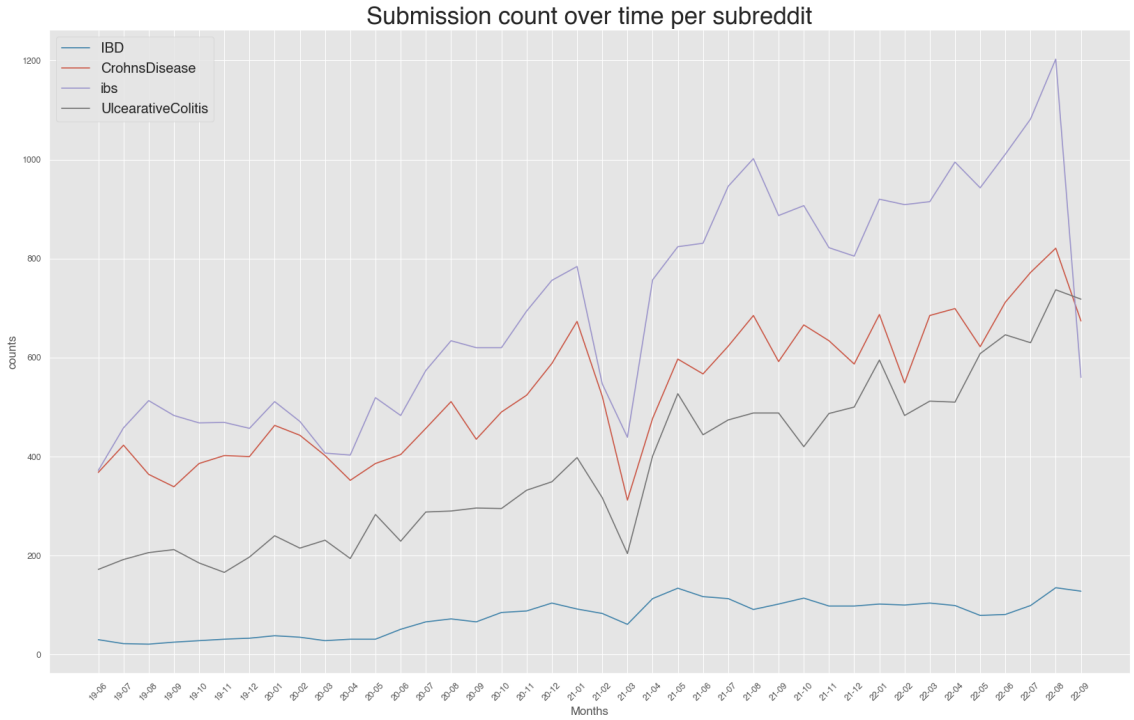

Figura 4: Andamento temporale del numero di submissions mensili per ogni subreddit analizzato

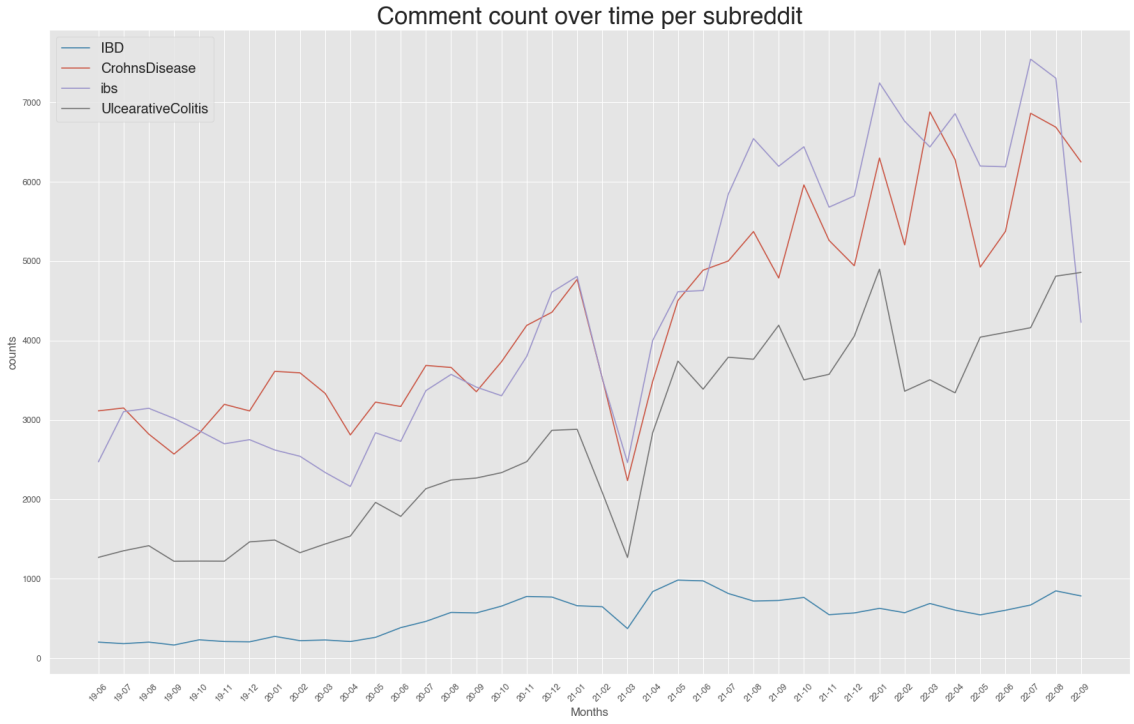

Figura 5: Andamento temporale del numero di commenti mensili per ogni subreddit analizzato

In figura 5 viene mostrato l'andamento del numero di commenti mensili per tutti e quattro i subreddit. Sull'asse delle ascisse è posto il riferimento al mese mentre l'asse delle ordinate fa riferimento al numero di commenti.
Si può notare che per quanto riguarda i commenti le interazioni sono distribuite in maniera più equa tra i subreddit analizzati. Anche in questo caso il subreddit r\IBD non conta numerose interazioni, mentre r\CrohnsDisease nonostante abbia meno submissions rispetto a r\ibs per quanto riguarda i commenti ne pareggia il numero.

Anche in questo caso per tutti i subreddit a Marzo 2021 si è verificato un decremento. Si può inoltre notare un incremento costante del numero di commenti dall'inizio del periodo di analisi fino alla fine, a eccezione del subreddit r\IBD che rimane pressoché piatto.

4 Interazioni correlate al virus SARS-CoV-2

Avendo accesso ai contenuti testuali delle submissions e dei commenti è stata approfondita l'analisi sulla distribuzione delle interazioni mensili cercando d'isolare i contenuti correlati al virus SARS-CoV-2. Per farlo è stato predisposto un filtro sul campo body di submissions e commenti. Il filtro consiste in un'espressione regolare che ricerca i match con una lista di parole chiave inerenti al virus. La lista di parole chiave è la seguente: Pandemic, pandemic, pandemics, coronavirus, CORONAVIRUS, Coronavirus, COVID-19, COVID19, COVID, COVID 19, covid-19, covid19, covid, covid 19, Covid-19, Covid19, Covid, Covid 19, Vaccine, vaccine, SARS-COV-2, SARS COV-2, SARS-COV2, SARS-COV 2, SARS COV.

Prima di applicarla è stato effettuato un join tra il campo title e il campo body delle submissions così da non escludere eventualmente le submissions con un titolo correlato al virus ma che per qualche ragione senza match nel body.

Dopo aver applicato il filtro sono state ottenute **2 986** submissions e **7 672** commenti.

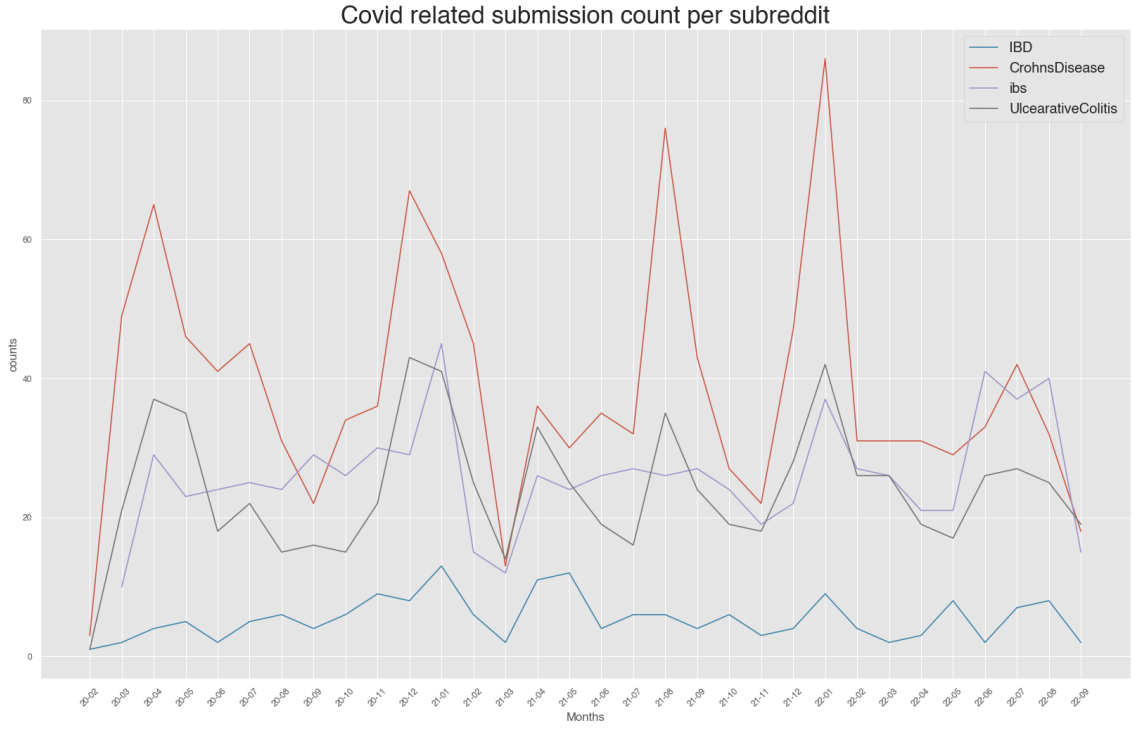

Figura 6: Andamento temporale del numero di submissions mensili contenenti parole chiave legate al virus

In figura 6 è riportato il numero di submissions mensili inerenti al virus SARS-CoV-2 per i quattro diversi subreddit. Sull'asse delle ascisse si trovano i

mesi mentre sull'asse delle ordinate il numero di submission.
Le prime interazioni esplicitamente correlate al virus appaiono a Febbraio 2020 con un picco significativo per il subreddit r\CrohnsDisease rispetto agli altri tre. Infatti r\CrohnsDisease è proprio il subreddit che ha verificato più submissions inerenti al virus.

In generale si possono notare quattro picchi distinti che coincidono esattamente con le cosiddette "ondate" della pandemia, ovvero quei periodi dove il virus, forte di nuove varianti, si è diffuso più velocemente. Sulla dashboard aggiornata settimanalmente da WHO (World Health Organization) è possibile verificare la coincidenza dei picchi riscontrati [1]. Questa è stata una buona prova del successo del filtraggio avvenuto sui dati.

Sicuramente l'ordine di grandezza delle interazioni correlate al virus è molto modesto rispetto all'ordine di grandezza dei dati raccolti ma va considerato ancora una volta che sono state catturate solo le interazioni esplicitamente correlate al virus.

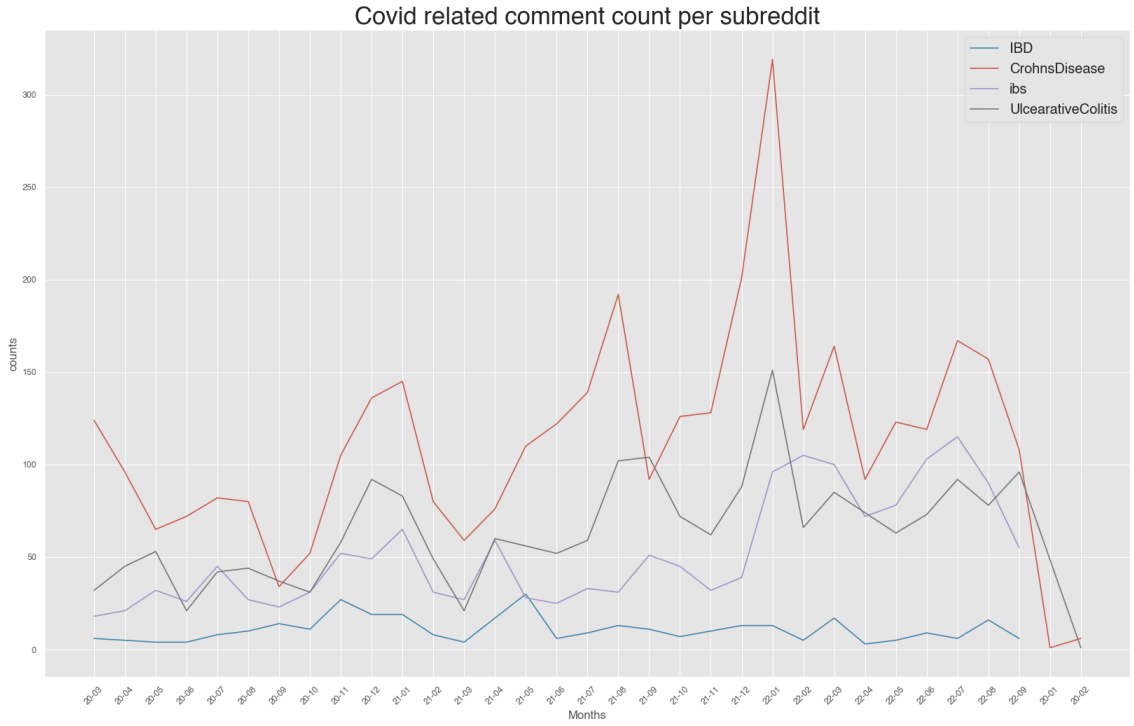

Figura 7: Andamento temporale del numero di commenti mensili contenenti parole chiave legate al virus

In figura 7 viene mostrato il numero di commenti inerenti al virus, differenziando per subreddit. Sull'asse delle ascisse si trovano i mesi mentre sull'asse

delle ordinate il numero di commenti.
In questo caso appaiono meno marcati i picchi definiti dalle ondate della pandemia e i subreddit hanno un andamento differente tra di essi.

Chiaro il motivo dal picco del subreddit r\CrohnsDisease a Gennaio 2022, in quanto tra Gennaio e Febbraio 2022 sono stati registrati numerosissimi casi di COVID-19 e prima di Dicembre 2022, Gennaio 2022 è stato un mese con uno dei massimi globali di casi registrati.

5 Sentiment analysis

La sentiment analysis (nota anche come opinion mining) è il processo di estrazione del sentimento da un corpo di testo. Nel dominio della Data Science per sentimento si intende un insieme discreto di valori la cui dimensione determina l'efficienza ed il risultato finale del modello. Possiamo vedere il sentimento come la classificazione di un corpo di testo in base alle parole che lo compongono, incorniciate da un contesto. Si può avere classificazione a due classi (positiva o negativa) oppure a tre classi (positiva, negativa, neutra). Nella realtà per sentimento legato a un corpo di testo si intendono le sensazioni ed emozioni che esso suscita; quello che si cerca di ottenere è la riproduzione questo processo naturale.

La sentiment Analysis (SA in breve) utilizza un ramo specifico dell'Intelligenza Artificiale nota come Natural Language Processing (NLP). La NLP viene utilizzata per comprendere la struttura e il significato del linguaggio umano analizzandone diversi aspetti come la sintassi, la semantica, e la morfologia. Utilizza tecniche di apprendimento automatico, come reti neurali e algoritmi di classificazione, per elaborare il linguaggio naturale e fornire risultati precisi e accurati. L'obiettivo della NLP è quello di creare sistemi intelligenti che possano comprendere, interpretare e generare il linguaggio naturale come farebbe un essere umano
Esistono due tipologie di approcci alla Sentiment Analysis: lexicon based [7] o machine learning based. L'approccio lexicon based è caratterizzato da apprendimento non supervisionato mentre gli approcci machine learning based rientrano nell'apprendimento supervisionato.

L'approccio lexicon based si fonda sull'utilizzo di dizionari di parole noti come valence dictionary. Un valence dictionary è un elenco di parole con assegnate loro polarità, che può essere utilizzato per determinare appunto la polarità di frasi o documenti completi. Questo dizionario assegna una valenza positiva, negativa o neutra a ciascuna parola in base al suo significato e al contesto in cui viene utilizzata. Questo punteggio di polarità viene quindi utilizzato per determinare la polarità complessiva del testo o del documento. Gli approcci lexicon based possono presentare alcune limitazioni, come la difficoltà nel gestire il sarcasmo

[7] Usato spesso in concomitanza del termine rule based approach, ovvero la classificazione ottenuta attraverso regole, queste regole sono solitamente chiamate lexicons

o le figure retoriche, la mancanza di considerazione del contesto e la difficoltà di valutare la polarità di frasi complesse o di documenti lunghi. Per questa ricerca è stato utilizzato VADER (Valence Aware Dictionary and Sentiment Reasoner) [9] un tool per la sentiment analysis totalmente open source. VADER è un tool lexicon e rule based affinato appositamente per captare il sentiment dei contenuti testuali sui social media e può essere applicato direttamente a dati non classificati.

Gli approcci basati su machine learning si basano sull'utilizzo di algoritmi di apprendimento automatico che analizzano il testo e identificano i pattern e le relazioni tra le parole e le loro polarità. Per farlo è necessario avere a disposizione una grande quantità di dati con associata una label, ovvero dati etichettati già associati a una classe (positiva, negativa o neutra). Con questo approccio è possibile allenare dei modelli di apprendimento artificiale sulla base di dati già classificati, affinarne le performance e studiarne il comportamento su nuovi dati.

5.1 Sentiment analysis e social media

L'utilizzo della sentiment analysis sui dati estratti dai social media è diventato sempre più diffuso negli ultimi anni, grazie alla grande quantità di informazioni disponibili sui vari canali social. Questo tipo di analisi consente di comprendere le opinioni, le emozioni e le percezioni degli utenti riguardo a specifici argomenti.

La sentiment analysis applicata ai social media è praticabile dati gli ottimi strumenti attualmente a disposizione ma va sempre considerato che i dati trattati sono prodotti da persone reali all'interno dei social media pertanto bisogna considerare la presenza di limitazioni e sfide.

L'analisi lessico sintattica dei dati provenienti dai social media può essere alterata da fattori come ad esempio l'utilizzo dello slang, eventuali errori di battitura, la presenza di emoji, corpi di testo monosillabici, o al contrario corpi di testo molto lunghi, il linguaggio figurativo, la negazione e l'ironia. Inoltre, la sentiment analysis potrebbe essere influenzata da bias, come la lingua o la cultura. Ad esempio, un'analisi di sentimenti in lingua inglese potrebbe non essere accurata quando applicata a testi in altre lingue. Allo stesso modo, la comprensione delle emozioni e delle percezioni potrebbe variare in base alle differenze culturali e sociali. La sentiment analysis non è una scienza esatta, e l'analisi delle emozioni umane rimane sempre complessa e multidimensionale.

Diventa molto importante assicurarsi che quanti più dati possibili non abbiano caratteristiche che possono influenzare nettamente i risultati. Per questo motivo, (soprattutto per dati provenienti dai social media) è buona pratica attuare una fase preliminare all'analisi chiamata data preparation. La fase di data preparation consiste nella pulizia dei dati grezzi trasformandoli e preparandoli per l'analisi successiva. Può essere personalizzata in base alle finalità dell'utilizzo dei dati e soprattutto in base alla loro tipologia. Per dati che contengono contenuti testuali sono previste tecniche efficaci che sono comunemente utilizzate come la pulizia del testo, l'uniformazione, lo stemming o derivazione e la gestione delle stopwords.

5.2 Strumenti utilizzati: Vader e roBERTa

In questa ricerca è stato utilizzato un approccio lexicon based sfruttando Vader e in misura minore roBERTa.

Vader (Valence Aware Dictionary and sEntiment Reasoner) è un tool di sentiment analysis lexicon-based e rule-based. È stato sviluppato dalla Georgia Institute of Technology. Utilizza un approccio valence-aware per l'analisi del sentiment, ovvero tiene conto del fatto che le parole possono avere valenze positive o negative e che la loro intensità può variare a seconda del contesto in cui vengono utilizzate. Fa uso di un dizionario di parole valutate manualmente, con punteggi di polarità e intensità associati a ogni parola. In questo modo, Vader è in grado di gestire meglio il linguaggio informale, il sarcasmo, l'ironia e altri aspetti del linguaggio naturale che possono influire sulla comprensione del sentiment, per questo motivo si adatta meglio di altri tool al contesto dei social media.

Vader utilizza una metrica di valutazione chiamata **compound**.
Il compound di una singola frase è calcolato come la somma dei punteggi di ogni parola nel testo aggiustati secondo le regole definite dal tool e poi normalizzati tra -1 (estremo negativo) e 1 (estremo positivo).

Il compound è la metrica di riferimento se si desidera una singola misura unidimensionale del sentimento per una data frase. Risulta utile ed efficace se si vogliono definire soglie standardizzate per classificare le frasi come positive, neutre o negative. La documentazione di Vader [6] fornisce dei valori soglia default per classificare i corpi di testo. Possono essere modificati e ritoccati a

piacimento ma per questa ricerca sono stati mantenuti i valori soglia tipici che sono:

1. positive sentiment: compound score >= 0.05
2. neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
3. negative sentiment: compound score <= -0.05

Vader è stato utilizzato per la maggior parte dell'analisi per due ragioni: perché offre le metrica di compound, e per la sua velocità (considerando circa 500 000 entry è ragionevole pensare alle performance)

roBERTa (Robustly Optimized BERT Pretraining Approach) è un modello di linguaggio naturale basato su trasformatori (Transformer-based) sviluppato da Facebook AI Research (FAIR). Nasce come un modello generico di Rete Neurale applicato principalmente nell'elaborazione del linguaggio naturale, ma è in grado di raggiungere risultati ottimi in molti altri task. roBERTa utilizza una tecnica di pre training completamente unsupervised.
Si tratta di un modello molto più raffinato rispetto a VADER perché basato sui Transformer, un tipo di architettura di rete neurale applicabile al linguaggio naturale. I Transformer sono stati introdotti nel 2017 e hanno rivoluzionato il NLP introducendo una nuova tecnica di attenzione chiamata Self-Attention (espansa poi a Self-Attention-Multihead [8]). Le tecniche di attenzione permettono ai modelli di concentrarsi su parti specifiche del testo in modo da comprendere meglio il significato dei dati di input e generare risposte più accurate. Invece di elaborare il testo in sequenza, il Transformer utilizza un meccanismo di attenzione per elaborare tutte le parole del testo contemporaneamente. Inoltre, i Transformer utilizzano una tecnica di attenzione posizionale, chiamata positional encoding. Questa tecnica permette al modello di comprendere l'ordine e la posizione delle parole all'interno della frase. Modelli come roBERTa hanno contribuito in modo significativo allo sviluppo di modelli di linguaggio naturale avanzati e all'avanzamento del NLP. Hanno permesso di raggiungere risultati estremamente positivi in molte applicazioni, compresa la sentiment analysis

[8] La Self-Attention-Multihead è una versione avanzata della Self-Attention, in cui il modello divide l'input in diverse parti (chiamate testate) e calcola l'attenzione su ognuna di esse in modo indipendente. In pratica, il modello utilizza più meccanismi di Self-Attention in parallelo, ciascuno dei quali concentrandosi su sottoinsiemi diversi delle parole dell'input

Per questa ricerca roBERTa non è stato utilizzato come tool principale di sentiment analysis per limiti legati alla computazione. Infatti il tempo impiegato da roBERTa per classificare un insieme di dati è nettamente superiore a quello impiegato da Vader. Per questa ragione roBERTa è stato utilizzato solo in determinati scenari, e in tutto ciò sempre a supporto di Vader, così da confermarne o smentirne le intuizioni.

6 Sentiment Analysis: risultati ottenuti

Vader è stato lo strumento principale per effettuare sentiment analysis sui dati estratti. Sono state analizzati separatamente le submission e i commenti per ognuno dei quattro subreddit. Inoltre sono stati prodotti grafici comuni per analizzare il sentiment generico presente su reddit riguardo all'argomento IBD e IBS.

Utilizzando Python risulta molto semplice interfacciarsi con Vader e consultarne i risultati. La cosa interessante però è che per grandi quantità di dati non si ha riscontro immediato dei risultati perché Vader ritorna la polarità di un dato e non lo classifica automaticamente. Per questo motivo è stata prevista una funzione da applicare ai risultati dopo essere stati processati da Vader, la quale in base al compound produce la label. Questa funzione può essere vista come una funzione di configurazione in quanto definisce quante classi utilizzare e quando classificare un dato in una classe o in un'altra. Come visto in precedenza Vader definisce dei "threshold values" default per classificare un dato in base al compound ottenuto (5.2). Questa configurazione può essere eventualmente modificata, ad esempio è possibile rimuovere l'intera classe neutrale, o spostare leggermente i valori soglia ma per questa ricerca sono stati utilizzati i valori default.

I dati sono stati sottoposti a una fase di data preparation. Sono stati infatti filtrati utilizzando varie espressioni regolari per rimuovere quelli con contenuto testuale nullo o irrilevante per sostituire gli url e i caratteri non codificati.

6.1 Submissions

In figura 8 sono mostrati i risultati ottenuti utilizzando Vader per effettuare sentiment analysis sulle submissions. I due grafici presentano sull'asse delle ascisse la label prodotta (la classificazione finale) mentre sull'asse delle ordinate è riportato il numero di dati classificati.

Si può notare come in generale il numero di submissions classificate come positive sia minore del numero di negative ma con una differenza non elevata. Circa il 7% dei dati è stato classificato come neutro.
Quindi prevalgono le submission con sentimento generale negativo ma di poco rispetto alle positive mentre una modesta componente dei dati è stata classificata come neutra.

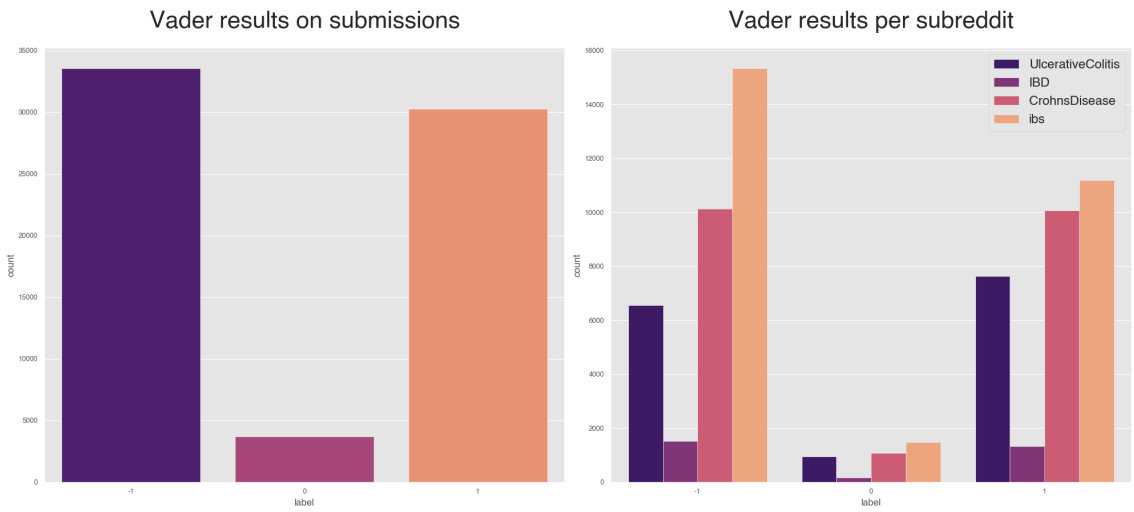

Figura 8: Risultati ottenuti effettuando sentiment analysis sulle submissions

Per quanto riguarda i singoli subreddit è molto interessante notare che r\UlcerativeColitis presenta più dati classificati come positivi rispetto ai dati classificati come negativi. Questo certamente non vale per r\ibs che il subreddit con più distanza tra numero di negativi e positivi e prevalgono appunto i negativi mentre r\CrohnsDisease e r\IBD sono bilanciati tra positivi e negativi.

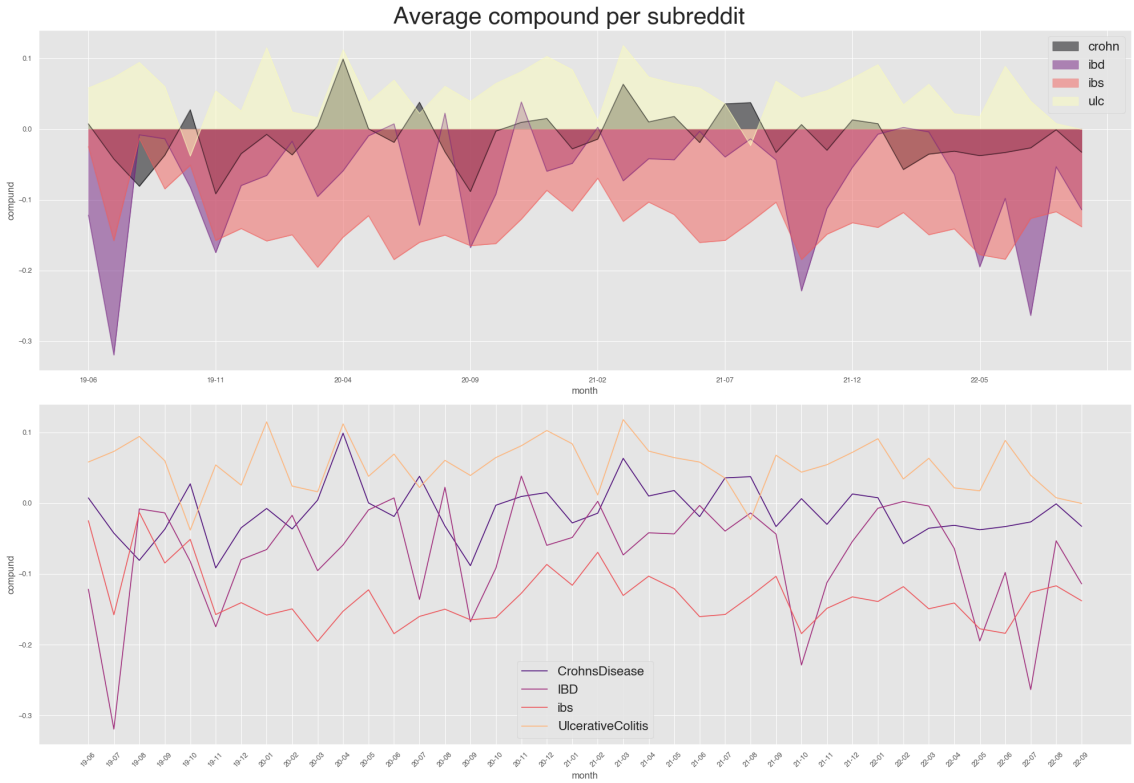

Figura 9: Compound medio mensile dei quattro subreddit analizzati

In figura 9 viene mostrato il compound medio mensile dei singoli subreddit. Sull'asse delle ascisse sono presenti i mesi di osservazione mentre sull'asse delle ordinate viene riportato il compound. Per ottenere questo grafico sono state raggruppate le submissions per ogni subreddit, è stato calcolato il compound totale per ogni mese ed è stato diviso per il numero di submissions nel dato mese.
Si può notare come il subreddit r\ibs giace sempre al di sotto della soglia di compound positivo (5.2), questo indica che mediamente le submissions pubblicate all'interno del subreddit hanno un carattere negativo. \UlcerativeColitis invece è il subreddit con il compound medio più alto.
Il comportamento dei subreddit è tutt'altro che uniforme. Va considerata anche la differenza di scala tra il numero di submissions tra i vari subreddit. Quindi la metrica del compound medio risulta utile per monitorare l'andamento del sentiment per i singoli subreddit ma non per confrontarli.

In figura 10 sono presentati i risultati della classificazione dei dati monitorando l'andamento temporale per i singoli subreddit. Nella parte superiore della figura sono presenti quattro istogrammi con hue sulla classificazione [9]. Sull'asse delle ascisse sono posti i mesi osservati, mentre sull'asse delle ordinate viene riportato il numero di dati classificati.
La prima parte del grafico riassume molte informazioni come l'andamento della classificazione così come il numero di submission pubblicate per subreddit. Emerge anche il sentiment dei singoli subreddit e il suo andamento temporale. Si nota che le curve tendono a crescere al crescere dei mesi, questo sicuramente è correlato a un'ipotesi della tesi, ovvero cercare di capire se le interazioni nei subreddit trattati hanno subito variazioni dopo l'avvento del virus. Questa crescita sarà approfondita con opportuni test statistici.
Il grafico conferma ancora una volta la contrapposizione tra r\ibs e \UlcerativeColitis, il quale con vari picchi gialli conferma ancora una volta la classificazione positiva di molte delle submissions al suo interno. Al contrario r\ibs ha solamente picchi negativi.
Nella seconda parte del grafico viene mostrato l'andamento della classificazione dei dati per singoli subreddit. Sull'asse delle ascisse si trova il periodo di osser-

[9] In analisi dei dati, il termine **hue** ci si riferisce generalmente a una variabile categorica che viene utilizzata per raggruppare e differenziare i dati o le osservazioni in una visualizzazione. In strumenti come seaborn o matplotlib in Python, dove è possibile utilizzare il parametro hue per mappare diversi valori di una variabile categorica a colori diversi nel grafico

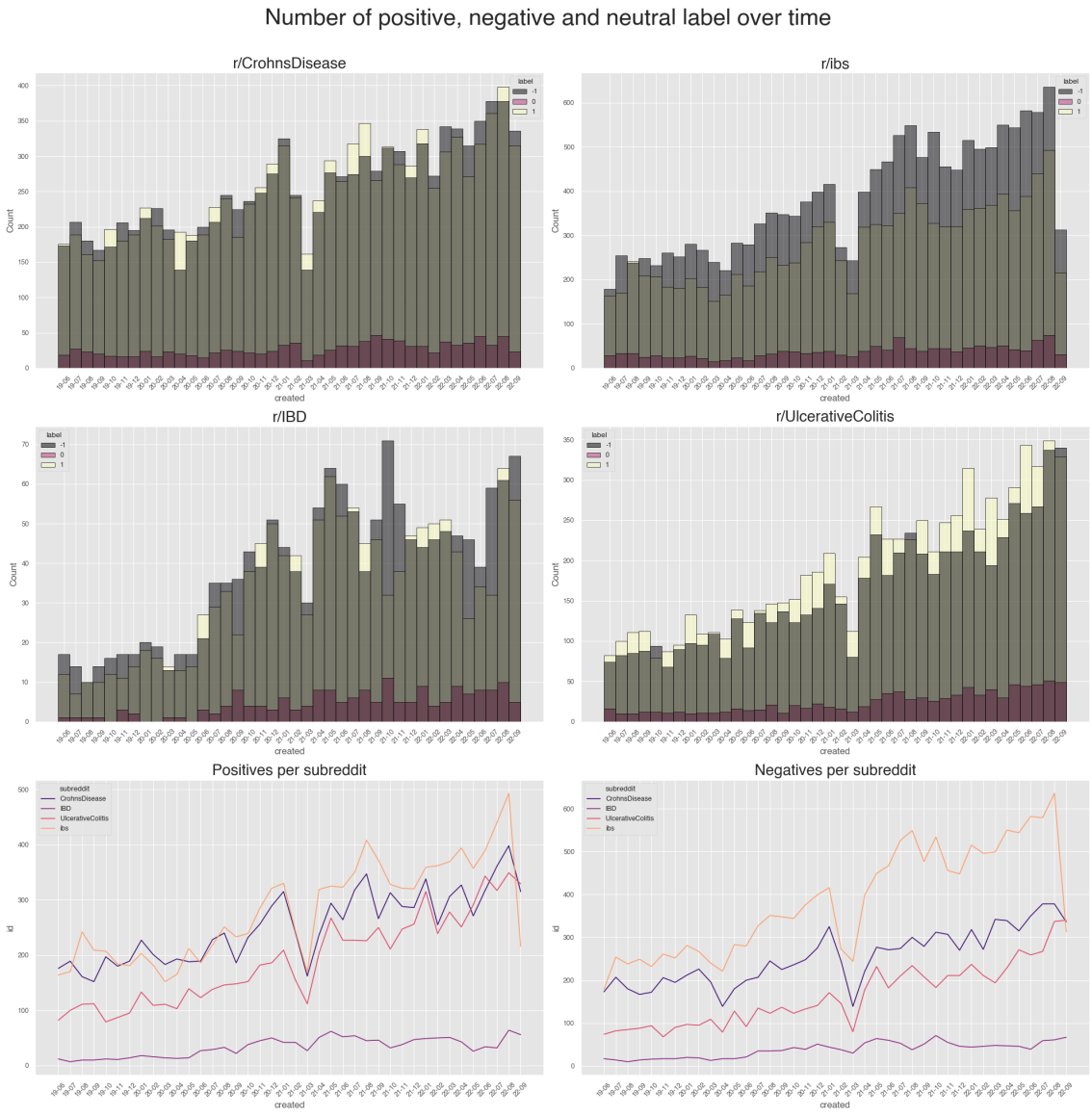

Figura 10: Risultati ottenuti effettuando sentiment analysis sulle submissions

vazione e sull'asse delle ordinate il numero di classificazioni.
Anche in questo caso non i subreddit hanno un andamento abbastanza unico fatta eccezione per Marzo 2021 dove si è riscontrato un decremento generale nelle interazioni.

6.2 Comments

Per quanto riguardi l'analisi del sentimento effettuata sui commenti vengono mostrati i risultati ottenuti in figura 11.

Come si può notare moltissimi dati sono stati classificati come positivi da Vader, addirittura più del 50 %. Molti dati sono stati classificati come neutri, a

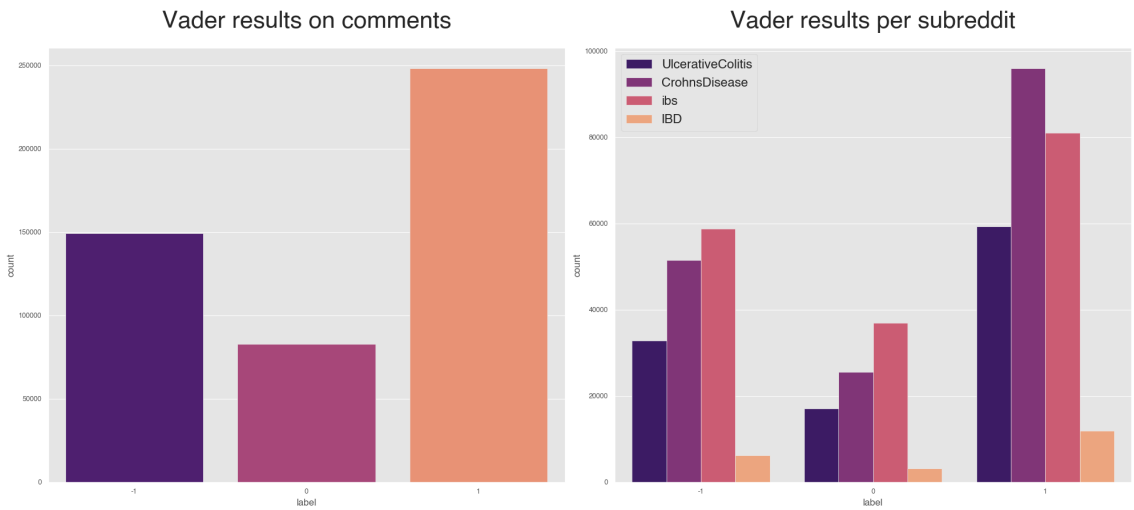

Figura 11: Risultati ottenuti effettuando sentiment analysis sui commenti

livello percentuale molti di più rispetto ai dati neutri relativi alle submissions.

In figura 11 nel secondo grafico sono esposti i risultati di classificazione per i differenti subreddit. In questo caso per tutti e quattro si è verificato un numero di dati classificati come positivi maggiore rispetto a quelli negativi. Da notare il subreddit r\CrohnsDisease che ha un elevatissimo numero di commenti positivi. Questo risultato può dare indicazioni su come gli utenti esprimano supporto reciproco.

Si suppone che a fronte di un contenuto chiaramente negativo la risposta delle persone sia solitamente positiva, cercando quindi di supportare. Per verificare questa ipotesi sono stati analizzati a campione alcuni subtree. Un subtree può essere visto come un "albero d'interazione", quindi un albero che ha come radice la submission e ha come nodi i commenti, fino alle foglie. Si ribadisce che in questa ricerca la profondità dell'albero d'interazione è costante e uguale a 1 quindi si è posto limite ai commenti di primo livello. Una volta raccolte a campione circa 20 submission con un compound nettamente negativo (compound ¡ 0.9 quindi classificate come negative), sono stati raggruppati i commenti che riferiscono alle submissions raccolte ed è stato calcolato il compound medio nei subtree. Il grafico in figura 12 riassume questo processo. Nella parte superiore del grafico è presente l'istogramma dei compound medi per i differenti subtree. Sulle ascisse sono mostrati gli identificativi univoci del "parent" ovvero la submissions, mentre sulle ordinate è posto il compound medio. Nella parte inferiore del grafico sono presentate le submission in corrispondenza dei relativi subtree. Sulle viene riportato il compound relativo alle singole submission e

sulle ordinate è presente l'identificativo univoco affiancato dal subreddit.
Si può notare come in generale nonostante una submission abbia un sentiment strettamente negativo, il compound medio all'interno del subtree non viene influenzato, anzi tende spesso a essere positivo e quasi mai negativo. Quindi come si ipotizzava, c'è la probabilità che la tendenza degli utenti sia di supportare coloro che manifestano un sentimento negativo, per questa ragione l'analisi del sentimento dei commenti ha prodotto risultati molto più positivi.

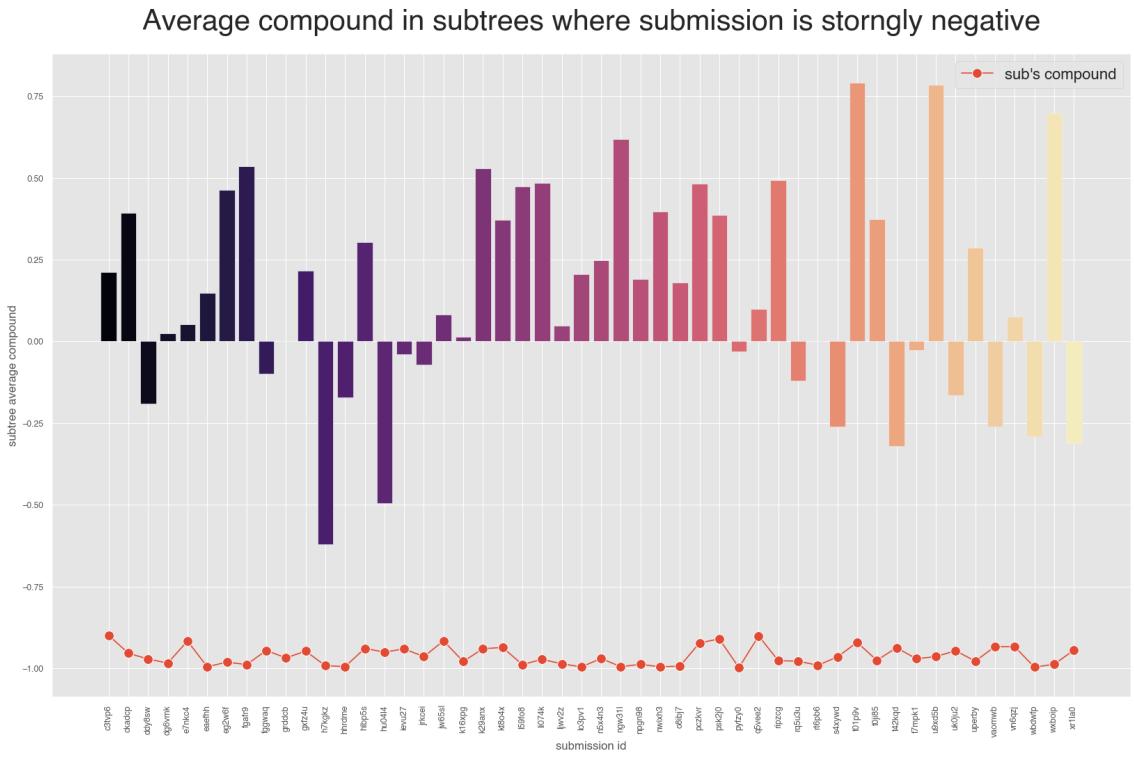

Figura 12: Analisi del compound nei subtree legati a submissions strettamente negative

In figura 13 viene proposto il resoconto dei risultati della classificazione dei commenti. Per quanto riguarda la parte superiore del grafico gli istogrammi presentano sulle ascisse i mesi osservati e sulle ordinate il numero di dati classificati. Si nota sicuramente il colore giallo molto più presente rispetto al grafico analogo rispetto alle submissions, sottolineando ancora una volta la positività dei commenti.

Si nota però anche come la classificazione di dati negativa sia in crescita per i subreddit escluso r\IBD. La crescita in realtà è distribuita perché come evidenziato precedentemente l'andamento del numero di interazioni è crescente.
La seconda parte del grafico mostra l'andamento del numero di classificazioni positive e negative per i singoli subreddit. Anche in questo caso r\ibs conta

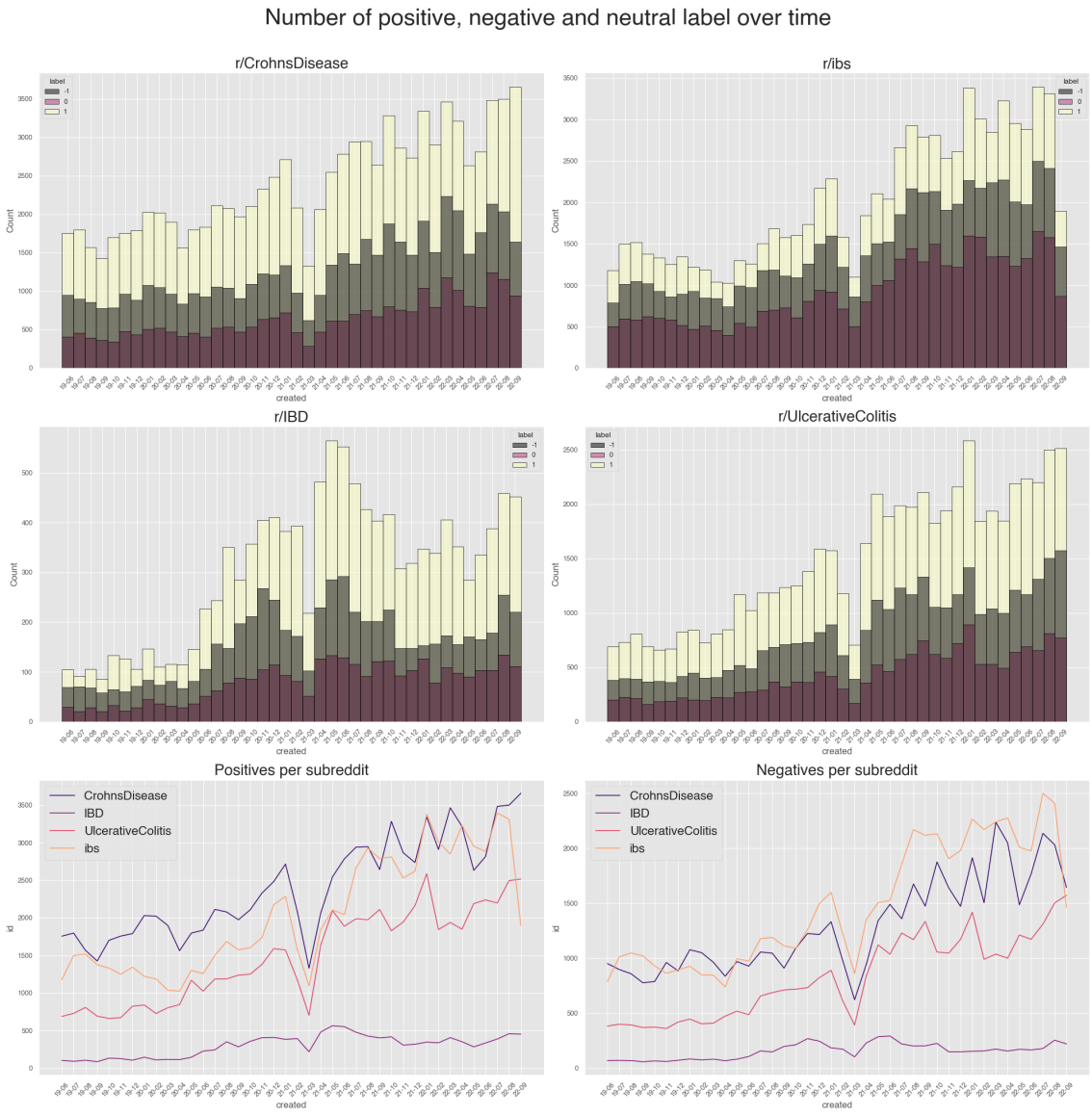

Figura 13: Risultati ottenuti effettuando sentiment analysis sui commenti

il maggior numero di interazioni negative. Non sono presenti picchi comuni rilevanti.

7 Sentiment analysis su interazioni correlate al virus SARS-CoV-2: risultati ottenuti

Si è ritenuto interessante approfondire l'andamento dei dati inerenti al virus anche per quanto riguarda l'analisi del sentimento. Sono state utilizzate le stesse parole chiave utilizzate precedentemente (4) per filtrare i dati isolando quelli correlati al virus.

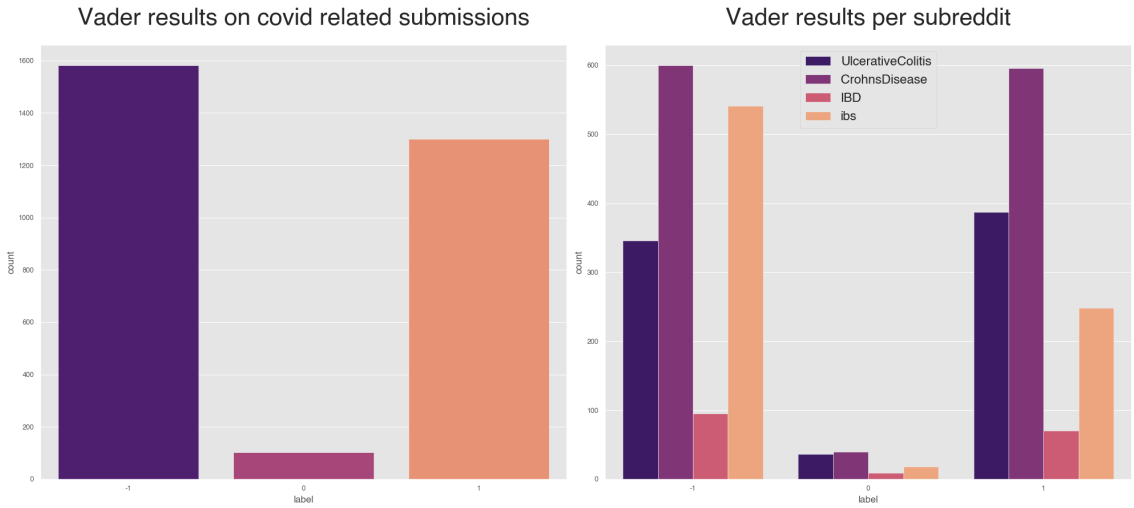

Figura 14: Risultati ottenuti effettuando sentiment analysis sulle submission correlate al virus

In figura 14 vengono mostrati i risultati della classificazione di Vader. Sulle ascisse è posta la classe mentre sulle ordinare il numero di classificazioni.
Si può notare come il numero di classificazioni neutre sia molto basso. Questo potrebbe essere determinato dal fatto che trattare di un argomento come il virus SARS-CoV-2 in concomitanza di malattie come IBD e IBS lasci poco spazio alla neutralità.
Il subreddit r\ibs presenta molte classificazioni negative anche per submissions inerenti al virus mentre gli altri tre subreddit sono più bilanciati.

Si riportano inoltre i risultati ottenuti dalla classificazione dei commenti correlati al virus. In figura 15 viene mostrata sulle ascisse la label scelta mentre sull'asse delle ordinate il numero di classificazioni effettuate.
In questo caso prevalgono ancora una volta le interazioni di carattere positivo. Infatti trattandosi di commenti si può far riferimento all'argomentazione sostenuta in precedenza (12).

Non sarà approfondita ulteriormente l'analisi del sentimento effettuata per in-

terazioni legate al virus SARS-CoV-2 in quanto l'andamento del compound così come del sentiment generico ricalca i risultati già mostrati.

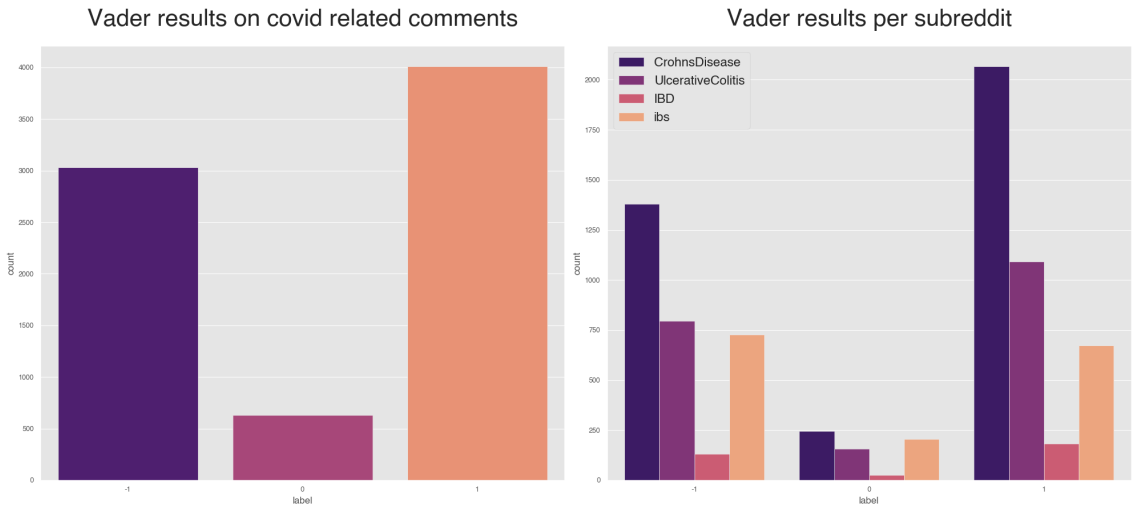

Figura 15: Risultati ottenuti effettuando sentiment analysis sui commenti correlati al virus

8 Distribuzioni di frequenza

Con distribuzione di frequenza si fa riferimento al processo di conteggio del numero di volte in cui ogni parola appare in un testo o in una collezione di testi, e quindi alla visualizzazione di queste informazioni in una tabella di distribuzione di frequenza o in un istogramma. Consultare la distribuzione di frequenza di un testo può inoltre dare indicazioni sul suo carattere generale.

Sono state calcolate le distribuzioni di frequenza dei quattro subreddit e per farlo è stata utilizzata la libreria NLTK [7] di Python. Sono state raccolte tutte le parole di ogni subreddit in un testo per ognuno dei quattro, unendo submissions e comments.
Prima di ottenere la distribuzione di frequenza dei testi sono stati previsti due step preliminari: la tokenizzazione e la rimozione delle stop words.
La tokenizzazione è il processo di suddivisione un testo in unità più piccole, chiamate token. In NLP (Natural Language Processing), la tokenizzazione viene utilizzata per preparare un testo per l'elaborazione automatica e può essere eseguita in diversi modi, a seconda delle esigenze. Per ottenere le distribuzione di frequenza dei testi è stato utilizzato un tokenizer basato su espressioni regolari.
Le stop words sono parole molto comuni che vengono spesso ignorate in un testo durante l'elaborazione automatica. Queste parole sono considerate "senza significato" o "meno significative" rispetto ad altre parole che trasmettono informazioni più rilevanti sul contenuto del testo. La rimozione delle stop words è una tecnica comune nell'elaborazione del linguaggio naturale per ridurre la complessità del testo e migliorare le prestazioni di alcune analisi, come la classificazione o la ricerca di informazioni. Tuttavia, la rimozione delle stop words può anche comportare la perdita di alcune informazioni importanti, ad esempio in contesti in cui le stop words hanno un significato specifico o indicano relazioni tra le parole.

Si possono ottenere ottime visualizzazioni sfruttando i WordCloud. Un WordCloud è una rappresentazione visiva di un insieme di parole, in cui la grandezza di ogni parola è proporzionale alla sua frequenza nell'insieme. Un WordCloud viene creato a partire da un testo e le parole presenti nel testo vengono estratte e rappresentate graficamente. Le parole che compaiono più frequentemente nel testo vengono rappresentate con una dimensione maggiore rispetto alle parole che compaiono meno frequentemente.
In figura 16 vengono mostrati due WordCloud: il primo riguarda tutte le parole

presenti nei subreddit che trattano la malattia IBD (r\CrohnsDisease, r\IBD e r\UlcerativeColitis) e il secondo si riferisce al subreddit r\ibs che tratta IBS.

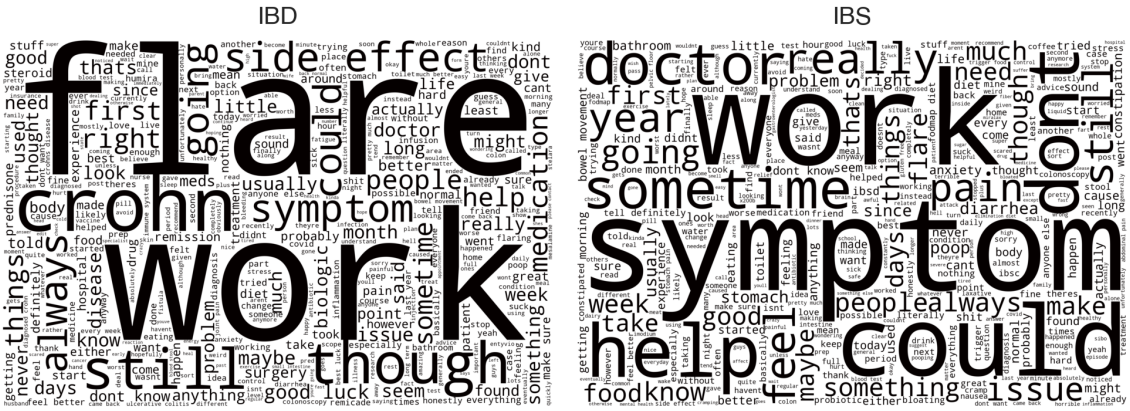

Figura 16: WordCloud dei due gruppi di malattie trattati

9 Risultati finali

In questa sezione vengono illustrati i risultati finali, le risposte ai quesiti della tesi. I risultati ottenuti sono supportati da evidenza statistica, infatti sono prodotti utilizzando uno strumento statistico: il test di Kolmogorov-Smirnov.

Il test di Kolmogorov-Smirnov (KS) è un test non parametrico che verifica la forma delle distribuzioni campionarie. Permette di confrontare tra loro un campione di dati e una distribuzione teorica (oppure due campioni di dati) allo scopo di verificare l'ipotesi statistica che la popolazione da cui i dati provengono sia quella in esame (oppure l'ipotesi che entrambi i campioni provengano dalla stessa popolazione). Il test può essere di due tipi a una coda o a due code. Il test a una coda viene utilizzato per verificare se una distribuzione di dati segue una distribuzione di probabilità specifica in una sola direzione (per esempio, se la distribuzione dei dati è maggiore o minore di una distribuzione di probabilità specifica). Questo tipo di test è utile quando si vuole sapere se i dati seguono una distribuzione di probabilità specifica solo in una direzione. Il test di Kolmogorov-Smirnov a due code, d'altra parte, viene utilizzato per verificare se una distribuzione di dati segue una distribuzione di probabilità specifica in entrambe le direzioni (ad esempio, se la distribuzione dei dati è significativamente diversa sia in eccesso che in difetto rispetto a una distribuzione di probabilità specifica) È stato utilizzato per confrontare i dati precedenti a Febbraio 2020 e successivi in modo da verificare statisticamente se i campioni hanno due distribuzioni diverse e quindi se c'è una differenza statisticamente rilevante tra i due gruppi. Verificare la differenza tra i due gruppi suggerirebbe che in effetti le interazioni dei pazienti sono variate dopo il virus.

Sono stati svolti 3 test in base alle 3 ipotesi formulate:

1. Test sul numero di submissions mensili per ogni subreddit
2. Test sul sentiment negativo per ogni subreddit
3. Test sul sentiment negativo confrontando le due malattie (IBD e IBS)

Metodi utilizzati Per effettuare i test è stato utilizzato il package stats di scipy [5] per Python che mette a disposizione il test di Kolmogorov-Smirnov. È stato effettuato il test a una coda con alternativa "maggiore". Se l'alternativa è impostata a "maggiore" significa che l'ipotesi nulla è

$$F(x) <= G(X) \forall x$$

dove F e G sono le CDF (cumulative distribution function) [10]. Quindi scegliendo l'intervallo di confidenza al 95% l'ipotesi nulla sarà rifiutata se il p-value sarà minore di 0,05.

Per l'analisi del sentimento è stato utilizzato anche il modello roBERTa per confermare le classificazioni di Vader.

[10]La funzione di distribuzione cumulativa è una funzione matematica che fornisce la probabilità che una variabile casuale X assuma un valore minore o uguale a un certo valore x. In altre parole, la CDF descrive la distribuzione dei dati in termini di probabilità

Test 1: Numero di submission prima e dopo il virus Sono stati divisi i dati per ogni subreddit in due gruppi spariti da Febbraio 2020. Sono stati raggruppati i numeri di submissions mensili per ogni subreddit ed è stato effettuato il test confrontando i due gruppi per ciascun subreddit.

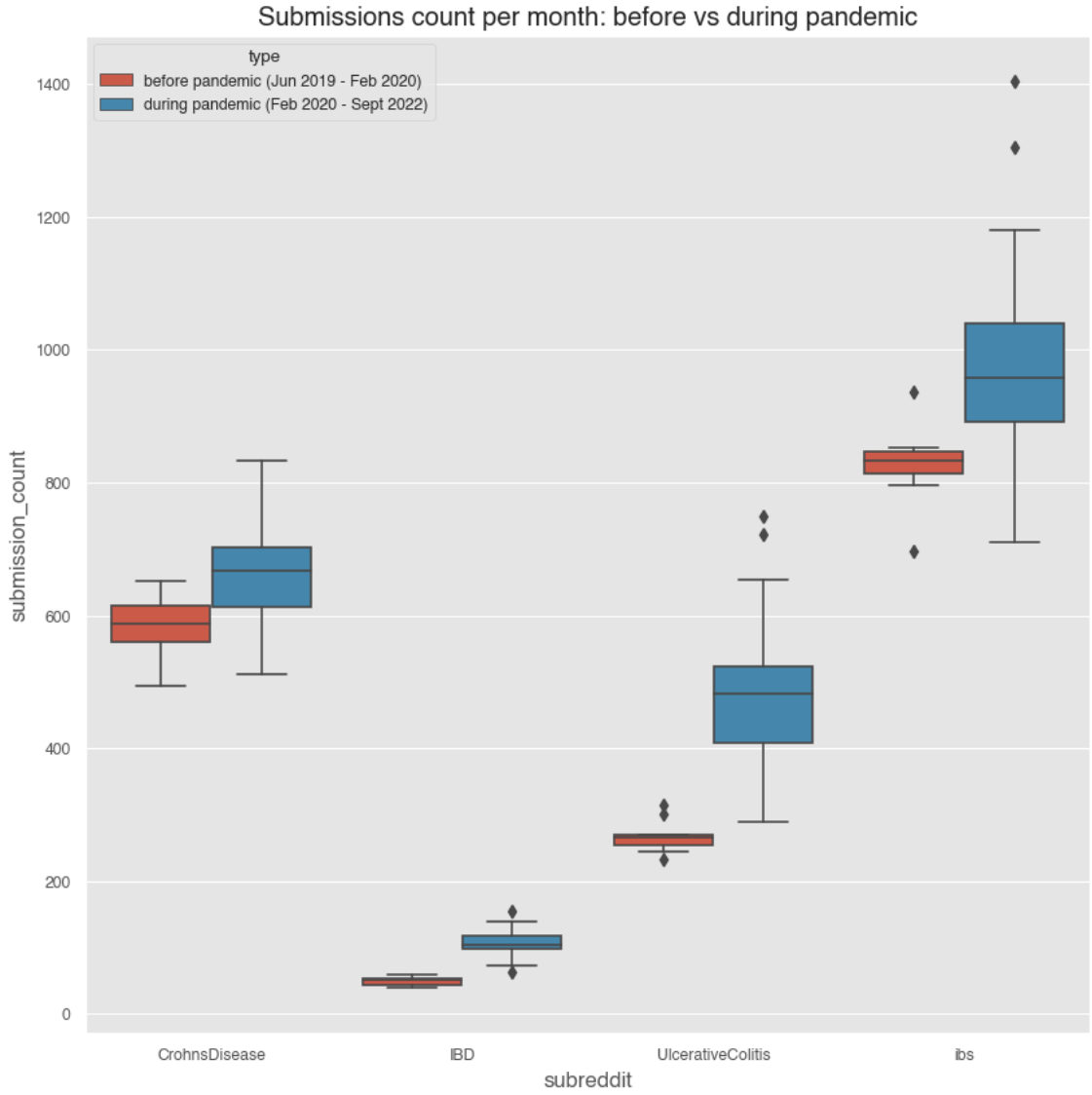

Figura 17: Boxplot rappresentante il primo test: numero di submission mensili prima vs dopo il virus

I risultati ottenuti sono riassunti in figura 17 rappresentati da un boxplot. [11]. Sulle ascisse è posto il subreddit mentre sulle ordinate il numero di submission

[11]Il boxplot è composto da un rettangolo, che rappresenta la mediana e il 50% dei dati, e da due linee verticali, chiamate baffi, che rappresentano i valori massimo e minimo dei dati, escludendo eventuali valori anomali (rappresentati da rombi in nero)

mensile. Ogni box rappresenta un insieme di mesi. A ogni mese è associato il numero di submissions pubblicate.

Per tutti i subreddit il p-value è minore di 0.05 quindi viene rifiutata l'ipotesi nulla e si può affermare che i dati non provengono dalla stessa popolazione.

```
Subreddit: CrohnsDisease
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.5357142857142857, pvalue=0.010805071427985776)

Subreddit: IBD
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=1.0, pvalue=8.038351295565194e-09)

Subreddit: UlcerativeColitis
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.9285714285714286, pvalue=4.4210932125608565e-07)

Subreddit: ibs
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.6746031746031746, pvalue=0.0009209458695816086)
```

Test 2: Sentiment negativo prima e dopo il virus Anche in questo caso sono stati divisi i dati per ogni subreddit in due gruppi spariti da Febbraio 2020. Sono stati classificati i dati utilizzando Vader e sono stati raggruppati i dati classificati come negativi per ogni mese e per ogni subreddit.

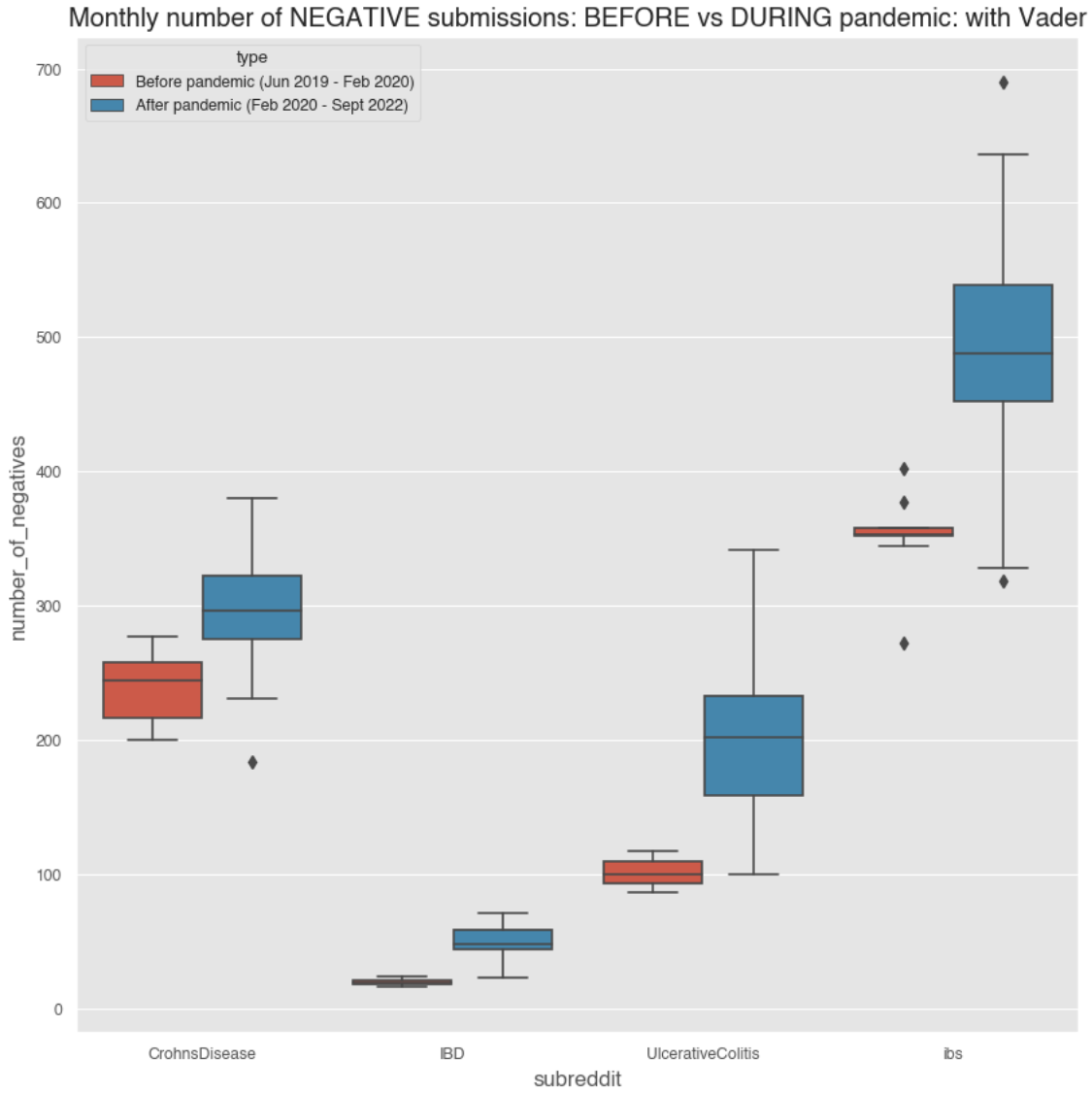

Figura 18: Boxplot rappresentante il secondo test: sentiment negativo prima e dopo il virus

I risultati ottenuti sono presentati in figura 18. Sulle ascisse sono riportati i subreddit e sulle ordinate il numero di submission classificate come negative da Vader.

Per tutti i subreddit pare vi sia una differenza netta tra i due periodi analizzati.

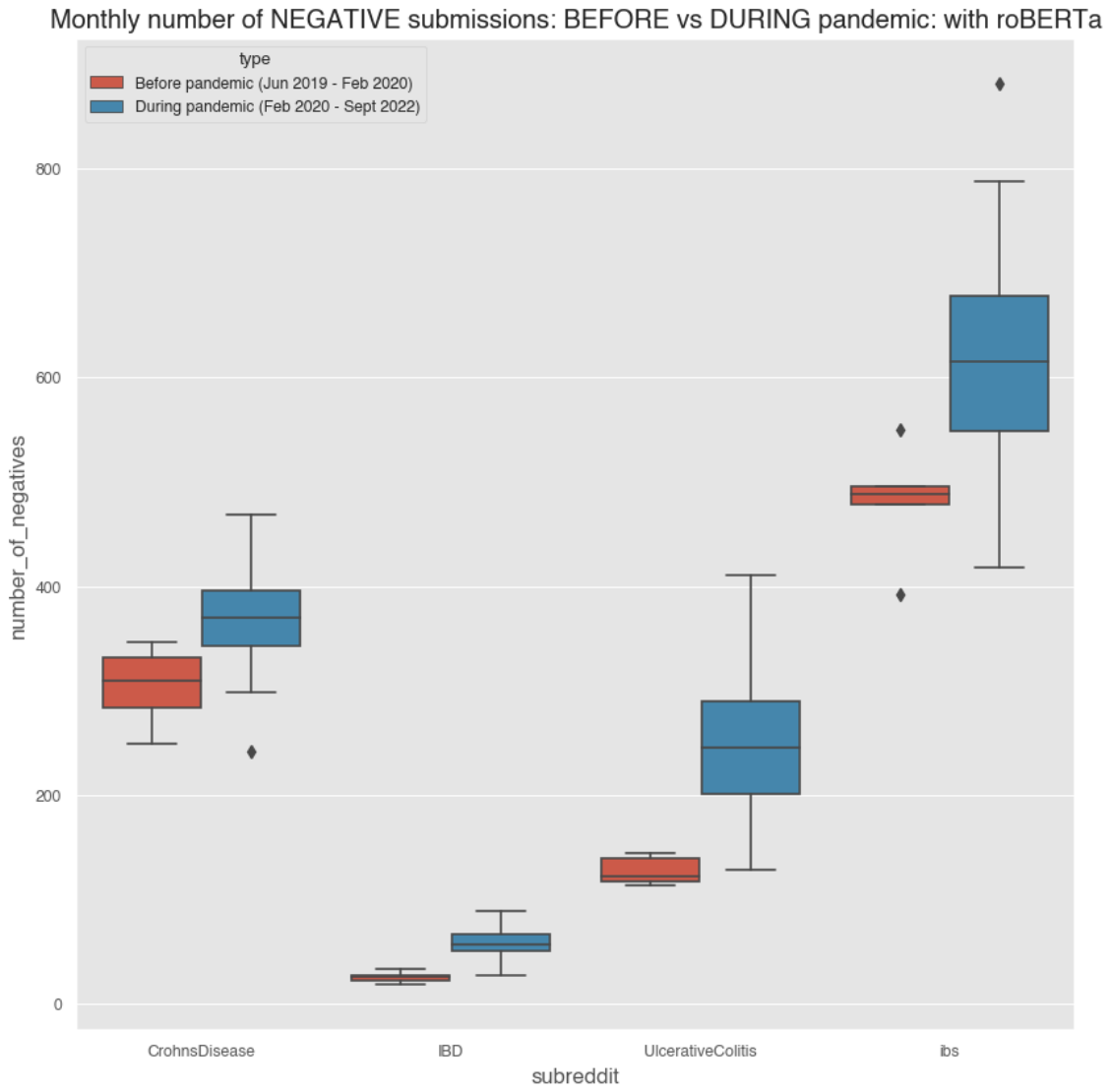

Figura 19: Boxplot rappresentante il secondo test: sentiment negativo prima e dopo il virus

Lo stesso test è stato ripetuto utilizzando roBERTa in modo da confermare i risultati ottenuti con Vader.

In figura 19 è riportato il boxplot ottenuto utilizzando la classificazione di roBERTa. Sulle ascisse sono riportati i subreddit e sulle ordinate il numero di submission classificate come negative da roBERTa.
I risultati sono molto simili a quelli ottenuto con Vader: le distribuzioni dei due periodi analizzati appaiono differenti per ogni subreddit.
Per verificarlo statisticamente è stato effettuato il test KS per entrambi le classificazioni dei modelli (figure 20 e 21).

Per tutti i subreddit risulta ancora una volta un p-value al di sotto della soglia

0,05 quindi si è portati anche in questo caso a rifiutare l'ipotesi nulla.
Quindi si è verificata una variazione del sentimento statisticamente rilevante.

```
Subreddit: CrohnsDisease
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.7103174603174603, pvalue=0.00041968232114145875)

Subreddit: IBD
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.9642857142857143, pvalue=8.038351295565193e-08)

Subreddit: UlcerativeColitis
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.9642857142857143, pvalue=8.038351295565193e-08)

Subreddit: ibs
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.8214285714285714, pvalue=1.775671801190351e-05)
```

Figura 20: Risultati del ks test utilizzando le label di Vader

```
Subreddit: CrohnsDisease
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.6785714285714286, pvalue=0.0006643616962271676)

Subreddit: IBD
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.9285714285714286, pvalue=4.4210932125608565e-07)

Subreddit: UlcerativeColitis
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.9642857142857143, pvalue=8.038351295565193e-08)

Subreddit: ibs
Period: before pandemic (Jun 2019 - Feb 2020) vs during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.7817460317460317, pvalue=6.941116343720545e-05)
```

Figura 21: Risultati del ks test utilizzando le label di roBERTa

Test 3: Confronto tra il sentiment negativo delle due malattie Questo test è stato effettuato per verificare se vi fosse differenza tra le due malattie (IBD e IBS) per quanto riguarda la variazione del sentiment negativo per dati antecedenti a Febbraio 2020 (dopo il virus).

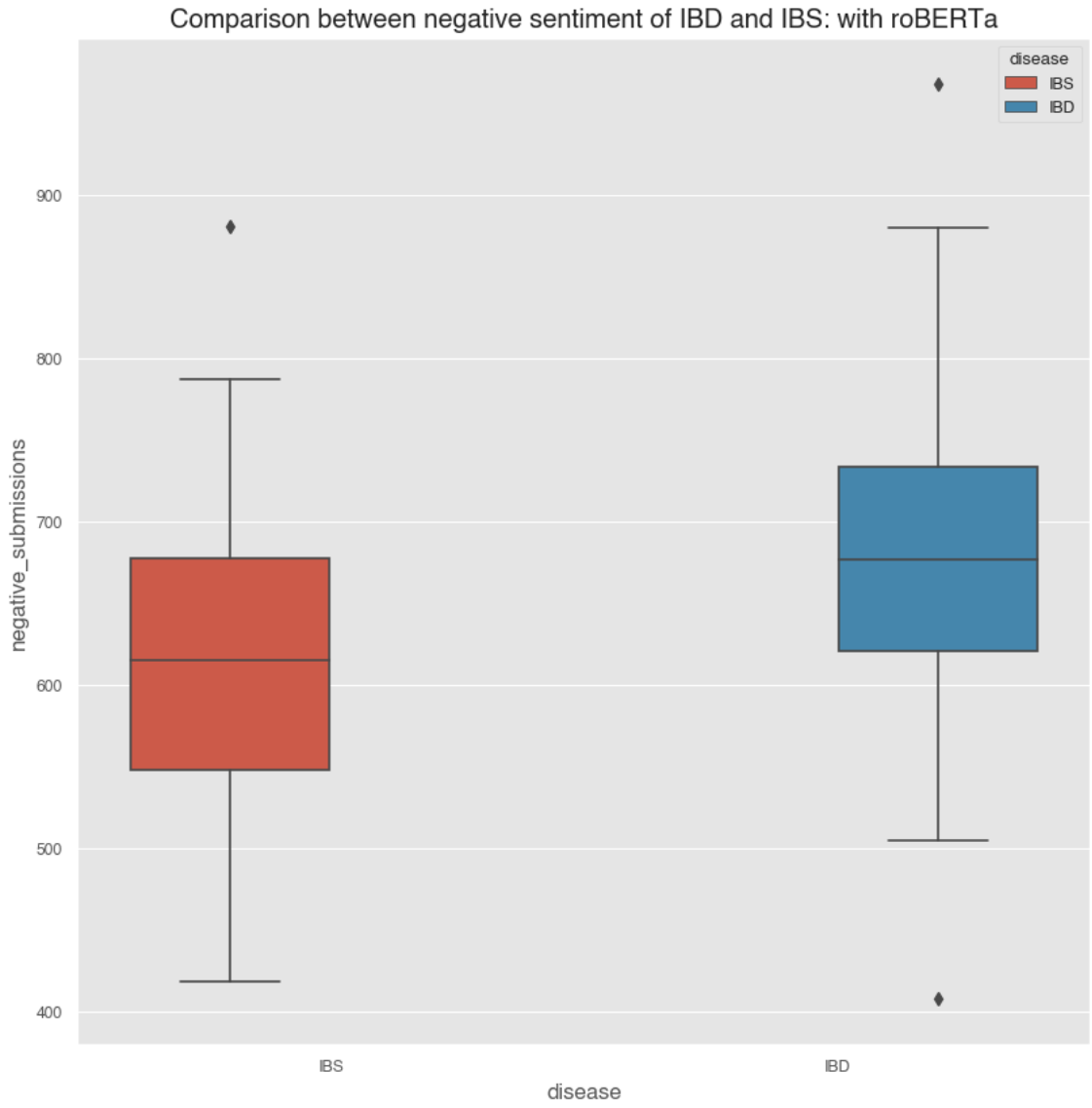

Figura 22: Boxplot rappresentante il terzo test: sentiment negativo dopo il virus confrontando IBD e IBS

Avendo già ottenuto la classificazione con roBERTa sono state utilizzate le label prodotte dal modello per ottenere il boxplot in figura 22. Sulle ascisse viene mostrato il tipo di malattia (IBD o IBS) mentre sulle ordinate viene riportato il numero di submission negative per ogni mese dopo Febbraio 2020.

Per produrre questo grafico sono stati raggruppati i tre subreddit che trattano di IBD (r\CrohnsDisease, r\IBD e r\UlcerativeColitis) e sono stati confrontati con il subreddit r\ibs. I risultati ottenuti in questo caso non permettono di rifiutare l'ipotesi nulla (come mostrato in figura 23).

```
Comparison of IBD and IBS
Period: during pandemic (Feb 2020 - Sept 2022)
KstestResult(statistic=0.03571428571428571, pvalue=0.9655172413793104)
```

Figura 23: Risultati del test ks sul sentiment negativo dopo il virus confrontando le due malattie

Il p-value è nettamente al di sopra del valore soglia 0,05 quindi non si può rifiutare l'ipotesi nulla. Questo significa che le distribuzioni dei due campioni sono simili e provengono probabilmente dalla stessa popolazione.

10 Conclusioni

Date le tre ipotesi della tesi si può affermare che il numero di submission mensili ha subito una variazione netta confrontando i periodi prima e dopo il virus, così come il sentiment negativo, infatti i dati classificati come negativi antecedenti al virus hanno una distribuzione differente rispetto ai dati classificati come negativi precedenti al virus. Per quanto riguarda l'ultima ipotesi non è stato possibile determinare se si è verificata una variazione d'interazione statisticamente rilevante differenziando per tipo di patologia (IBD o IBS).

Elenco delle figure

1 Esempio della struttura di una submission 9
2 Esempio della struttura di un comment 9
3 Numero totale di submissions e media mensile per subreddit . . . 10
4 Andamento temporale del numero di submissions mensili per ogni subreddit analizzato 12
5 Andamento temporale del numero di commenti mensili per ogni subreddit analizzato 12
6 Andamento temporale del numero di submissions mensili contenenti parole chiave legate al virus 14
7 Andamento temporale del numero di commenti mensili contenenti parole chiave legate al virus 15
8 Risultati ottenuti effettuando sentiment analysis sulle submissions 23
9 Compound medio mensile dei quattro subreddit analizzati 23
10 Risultati ottenuti effettuando sentiment analysis sulle submissions 25
11 Risultati ottenuti effettuando sentiment analysis sui commenti . . 26
12 Analisi del compound nei subtree legati a submissions strettamente negative 27
13 Risultati ottenuti effettuando sentiment analysis sui commenti . . 28
14 Risultati ottenuti effettuando sentiment analysis sulle submission correlate al virus 29
15 Risultati ottenuti effettuando sentiment analysis sui commenti correlati al virus 30
16 WordCloud dei due gruppi di malattie trattati 32
17 Boxplot rappresentante il primo test: numero di submission mensili prima vs dopo il virus 35
18 Boxplot rappresentante il secondo test: sentiment negativo prima e dopo il virus 37
19 Boxplot rappresentante il secondo test: sentiment negativo prima e dopo il virus 38
20 Risultati del ks test utilizzando le label di Vader 39
21 Risultati del ks test utilizzando le label di roBERTa 39
22 Boxplot rappresentante il terzo test: sentiment negativo dopo il virus confrontando IBD e IBS 40
23 Risultati del test ks sul sentiment negativo dopo il virus confrontando le due malattie 41

Ringraziamenti

Ringrazio innanzitutto il Prof. Marco Roccetti e il Dott. Luca Casini, per avermi accettato come tesista, per la disponibilità e per i preziosi consigli durante la realizzazione della tesi.

Ringrazio Filippo e Youssef, due colleghi speciali con cui ho condiviso tutto il percorso di studi. Li ringrazio per i meravigliosi anni trascorsi insieme.

Ringrazio mia sorella, mia madre e mio padre, per il supporto infinito e incondizionato nei miei confronti che posso ricambiare solamente rendendovi orgogliosi.

Bibliografia e Sitografia

[1] *Dashboard virus SARS-CoV-2 del WHO*. https://covid19.who.int. Ultimo accesso 14-11-2022.

[2] *Descrizione di IBD*. https://www.healthline.com/health/crohns-disease/ibs-vs-ibd. Ultimo accesso 16-02-2023.

[3] *Descrizione di IBS*. https://www.humanitas.it/malattie/sindrome-dell-intestino-irritabile/. Ultimo accesso 16-02-2023.

[4] *Documentazione delle api di Reddit*. https://www.reddit.com/dev/api/. Ultimo accesso 16-02-2023.

[5] *Documentazione di scipy.stats*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html. Ultimo accesso 22-02-2023.

[6] *Documentazione di Vader*. https://vadersentiment.readthedocs.io/en/latest/index.html. Ultimo accesso 27-02-2023.

[7] *Documentazione nltk*. https://www.nltk.org. Ultimo accesso 16-02-2023.

[8] *Effetti dello stress sull'organismo*. https://www.healthline.com/health/stress/effects-on-body. Ultimo accesso 16-02-2023.

[9] C.J. & Gilbert Hutto. «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text». In: *The Computer Journal* (2014).

[10] *PRAW api wrapper*. https://praw.readthedocs.io/en/stable/. Ultimo accesso 14-11-2022.

[11] *Pushift Api*. https://github.com/pushshift/api. Ultimo accesso 14-11-2022.

[12] *Stress Triggers Flare of Inflammatory Bowel Disease in Children and Adults*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6821654/. Ultimo accesso 16-02-2023.