

UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Economia e Statistica

Dipartimento di Statistica e Metodi Quantitativi

Corso di Laurea in Statistica e Gestione delle Informazioni



Notizie e tweet sulla pandemia in Italia: differenze tra comunicazione istituzionale e percezione della popolazione nel tempo

Relatore: Prof. Pescini Dario

Correlatore: Prof. Cesarini Mirko

Laureando:

Morganti Raffaele
Matricola 846595

Anno Accademico 2020-2021

Sommario

Obiettivi Lo scopo di questa tesi è quello di analizzare i tweet italiani relativi al coronavirus per comprendere come la percezione della situazione emergenziale si sia evoluta nel tempo, sia per quanto riguarda gli argomenti di maggiore interesse sia per l'opinione ad essi associata. Si vuole inoltre valutare l'esistenza di differenze comunicative tra canali istituzionali, esperti di settore e media tradizionali rispetto alle discussioni degli italiani.

Metodi I testi sono stati raccolti da 5 fonti: Articoli con commenti degli esperti, Notizie quotidiane, Tweet popolari, Tweet di account istituzionali, Tweet da parte delle regioni. Nelle analisi sono stati utilizzati il modello UMAP per la rappresentazione in dimensionalità ridotta, il GSDMM per il raggruppamento e il BERT per l'analisi del sentiment. Si introduce inoltre una tecnica bayesiana per il calcolo dei pesi delle parole associate al sentiment al fine di realizzare delle wordcloud più esplicative.

Risultati Si osserva una differenza tra gli stili comunicativi adottati dalle varie fonti analizzate, in particolare tra le istituzioni che mantengono una comunicazione più oggettiva da cui traspare principalmente paura, contro i tweet più popolari che presentano una polarizzazione negativa più marcata e per cui l'emozione predominante è la rabbia. In generale risultano meno apprezzati i commenti dei politici rispetto a quelli dei medici. Dall'analisi di casi specifici si evidenzia la tendenza alla diffusione di opinioni in contrasto con quelle ufficiali delle istituzioni.

Indice

Introduzione	3
COVID-19	4
Storia	4
Social Network	5
Twitter	5
Natural Language Processing	6
Evoluzione	6
Analisi e metodi	7
Obiettivi	8
Raccolta dati	12
Dataset: Esperti	12
Dataset: Notizie	13
Dataset: Tweet	14
Dataset: Istituzioni	15
Dataset: Regioni	16
Analisi	18
Preprocessing	18
Rappresentazione	20
Clustering	20
Sentiment Analysis	23
Primi risultati	24
Considerazioni finali	32
Risultati	33
Risposta alle misure contenitive	33
Opinioni sull'app Immuni	34
Fiducia in esperti e istituzioni	37
Efficacia della campagna vaccinale	37
Conclusioni	40
Differenze di comunicazione	40
Opinione espressa su Twitter	40
Criticità	41
Metodi	41
Censura	41
Appendice	44

Glossario

BERT il Bidirectional Encoder Representations from Transformers (Devlin et al. 2019) è un modello di machine learning per il Natural Language Processing. 6, 23

CCS Campione Casuale Semplice. 12–16

fastText fastText (Bojanowski et al. 2016) è una libreria per l'apprendimento non supervisionato dei word-embedding. 18

GSDMM Il collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (Yin e Wang 2014) è un modello sviluppato per clusterizzare collezioni di testi brevi. 20, 48

IC Intervallo di Confidenza. 12–16

NLP Natural Language Processing. 6

OMS Organizzazione Mondiale della Sanità. 4

scraping Lo scraping web è una tecnica che permette di estrarre dati da pagine o servizi presenti sul web. 12–14

UMAP L'Uniform Manifold Approximation and Projection (Sainburg, McInnes e Gentner 2021) è una tecnica di riduzione della dimensionalità in maniera non lineare, utile per la visualizzazione. 20, 23

word cloud La word cloud è una rappresentazione visiva di parole-chiave in cui ogni termine è rappresentato con una dimensione proporzionale alla sua importanza. 20, 24, 40

Introduzione

COVID-19

La COVID-19 è una malattia causata dal coronavirus SARS-CoV-2, il contagio avviene principalmente tramite contatto con droplet e aerosol di una persona infetta. In alcuni casi la malattia può causare sintomi gravi che possono comportare la necessità di ricovero o addirittura il decesso.

Storia

In seguito una breve descrizione dei momenti salienti nell'evoluzione della diffusione del virus:

Il nuovo coronavirus è stato individuato per la prima volta il 31 dicembre 2019 in Cina, nel corso del mese di gennaio del 2020 l'argomento inizia a entrare nell'opinione pubblica italiana a causa delle prime notizie sulla possibilità di trasmissione da uomo a uomo, il lockdown di Wuhan e il blocco dei voli verso la Cina.

Fino al 20 febbraio si ignora il fatto che il virus sia diffuso sul nostro territorio, il giorno successivo vengono scoperti i primi casi, ma inizialmente si crede che si tratti di eventi isolati. A causa della crescente diffusione del virus e dell'aumento del numero di ricoveri l'8 marzo la Lombardia entra in lockdown, misura estesa dopo 2 giorni al resto del paese. L'11 marzo l'OMS dichiara la malattia pandemica.

A inizio maggio, a seguito del miglioramento della situazione, inizia la cosiddetta "Fase 2" seguita a giugno dalla "Fase 3" caratterizzate dal progressivo allentamento delle misure restrittive durante il periodo estivo. Il 15 giugno viene attivata l'app per il contact-tracing "Immuni".

A settembre la situazione torna a peggiorare e a ottobre tornano le misure restrittive (seconda ondata). A novembre si inizia a parlare di vaccini per la prevenzione della malattia e vengono diffuse le prime notizie sulla loro efficacia.

Il 27 dicembre 2020 iniziano le prime vaccinazioni in Italia: viene data la precedenza a operatori sanitari e ultraottantenni. Con la disponibilità di nuove dosi nel corso del nuovo anno la possibilità viene estesa a fasce sempre più ampie della popolazione. Nei primi mesi del 2021 si iniziano a diffondere notizie sul tema delle varianti e sulla sicurezza dei vaccini (in particolare per il vaccino AstraZeneca a seguito della modifica delle raccomandazioni sull'età per la somministrazione).

Come un anno prima dal mese di maggio le restrizioni vengono via via allentate nelle varie regioni. Nel mese di giugno si inizia a parlare di "Green Pass": lo strumento inizialmente studiato per la mobilità tra stati europei e da luglio reso necessario per alcune attività all'interno del paese.

Social Network

I servizi di social networking nascono a cavallo del nuovo millennio segnando la transizione tra il “vecchio” web unidirezionale e la nuova struttura partecipativa del “Web 2.0”. L’evoluzione nel corso degli anni ha portato alla nascita e alla chiusura di vari social network, ognuno con caratteristiche e target differenti.

Secondo la definizione di Boyd e Ellison (2007) un social network ha la caratteristica di permettere all’interno del sistema la creazione di un profilo pubblico o semi-pubblico, l’articolazione di una lista di contatti e la possibilità di esplorare la lista di contatti degli amici.

Secondo il report di We Are Social e Hootsuite (2021) il 67.9% degli italiani sono utenti social attivi, percentuale che sale nella fascia d’età 16-64 con l’85.2% delle persone che dichiarano di aver partecipato attivamente su almeno un social network nell’ultimo mese.

Ad oggi in Italia i social legati a comunicazione tramite video o immagini sono tra i più affermati o in forte crescita (sono usati da più del 10% degli italiani: YouTube, Instagram, Pinterest, TikTok e Twitch), un’altra importante fetta è coperta dai social che offrono servizi di chat o voip (WhatsApp, Messenger, Skype, Telegram, Snapchat e WeChat). Gli altri social utilizzati attivamente da almeno il 10% della popolazione sono Facebook, Twitter, LinkedIn, Tumblr e Reddit (We Are Social e Hootsuite 2021).

Twitter

Twitter nasce nel 2006 come piattaforma di microblogging, ad oggi non è il social più diffuso nel nostro paese con solo il 32.8% degli italiani tra i 16 e i 64 anni attivi sulla piattaforma (We Are Social e Hootsuite 2021).

Rispetto ai competitor, su questa piattaforma ogni messaggio pubblicato (tweet) è pubblico di default e gli utenti possono interagire includendo nel tweet mentions (menzioni ad altri utenti), effettuando retweet (ricondivisione di un tweet già pubblicato), rispondendo o mettendo “Mi piace” a un tweet. Hanno inoltre la possibilità di seguire un utente per restare aggiornati sui suoi nuovi messaggi.

L’algoritmo di Twitter offre consigli agli utenti mostrando argomenti di tendenza e permettendo di ottenere contenuti di interesse tramite ricerca testuale o uso di hashtags (parole chiave precedute dal simbolo “#”).

Per le sue caratteristiche Twitter è visto come alternativa ai canali media classici ed è molto utilizzato per commentare in diretta eventi pubblici, con la possibilità di leggere in tempo reale i tweet più recenti o più popolari e di pubblicare le proprie opinioni.

Vista l’elevata diffusione di questa piattaforma anche le istituzioni nel corso degli ultimi anni hanno creato i propri account per comunicare in maniera efficace con gli utenti.

Natural Language Processing

Il Natural Language Processing (NLP) è la scienza che si occupa di trattare in modo automatico il linguaggio naturale, ovvero qualsiasi linguaggio utilizzato comunemente dagli esseri umani per comunicare, sia esso in forma scritta o parlata. La difficoltà principale deriva proprio dalle caratteristiche del linguaggio naturale tra cui la sua complessità sintattica e la presenza di ambiguità lessicali (parole che assumono più significati), sintattiche (la frase non è interpretabile in modo univoco) o pragmatiche (l'interpretazione dipende dal contesto).

Tra i vari campi di applicazione del NLP possiamo indicare, in maniera sicuramente non esaustiva:

- la traduzione automatica di testi;
- la classificazione;
- il question answering;
- il riconoscimento vocale;
- la sentiment analysis e l'opinion mining.

Evoluzione

Lo sviluppo del NLP procede parallelamente all'evoluzione dei calcolatori per motivi sia tecnici sia applicativi: macchine sempre più potenti permettono di analizzare moli di dati sempre più consistenti e con tecniche più avanzate; con la diffusione del World Wide Web inoltre la quantità di documenti testuali esplode e la necessità di estrarne informazioni acquisisce maggiore importanza.

I primi approcci di NLP avvengono con algoritmi “rule-based” definendo regole precise per la risoluzione di un determinato problema. Il principale limite di queste tecniche risiede nell'elevata complessità e la scarsa scalabilità di una soluzione per altre lingue o differenti obiettivi. Si tratta comunque di un approccio ancora utilizzato ed efficace, un esempio è il modello VADER (Hutto e Gilbert 2014) per la sentiment analysis di testi in lingua inglese.

Approcci di tipo statistico classici basati su Bag of Words (testo rappresentato come elenco non ordinato di parole) e N-grammi (a differenza dei bag of words considera sequenze di parole di lunghezza N) sono ampiamente utilizzati anche se recentemente sovrastati da uno degli sviluppi più interessanti degli ultimi anni che sfrutta il Machine Learning (ML): i word embeddings (rappresentazioni vettoriali di ogni termine nel dizionario) la cui prima implementazione è il modello word2vec (Mikolov et al. 2013).

Le tecniche di tipo ML possono prevedere l'apprendimento supervisionato (addestramento su testi annotati) o non supervisionato (non necessita la fase di annotazione). Con il transfer learning i modelli possono essere preaddestrati su dataset non supervisionati di grandi dimensioni, per poi subire una fase di “fine-tuning” su un dataset annotato di dimensioni ridotte. Un esempio di modello che sfrutta queste caratteristiche è BERT (Devlin et al. 2019), preaddestrato per il language modelling (predire una parola nascosta all'interno del testo), ma utilizzato anche per compiti di classificazione.

Analisi e metodi

Obiettivi

Lo scopo di questa tesi è quello di analizzare i tweet italiani relativi al coronavirus per comprendere come la percezione della situazione emergenziale si sia evoluta nel tempo, sia per quanto riguarda gli argomenti di maggiore interesse sia per l'opinione ad essi associata. Si vuole inoltre valutare l'esistenza di differenze comunicative tra canali istituzionali, esperti di settore e media tradizionali rispetto alle discussioni degli italiani.

Questa area di ricerca è stata parzialmente esplorata da:

- De Rosis et al. (2021): il paper si concentra su tweet in italiano durante i primi sviluppi (tra il 17 febbraio e il 22 marzo 2020). Cerca di confrontare l'opinione globale dei vari tweet mostrando come quella negativa sia costantemente maggiore rispetto a quella positiva (figura 1).

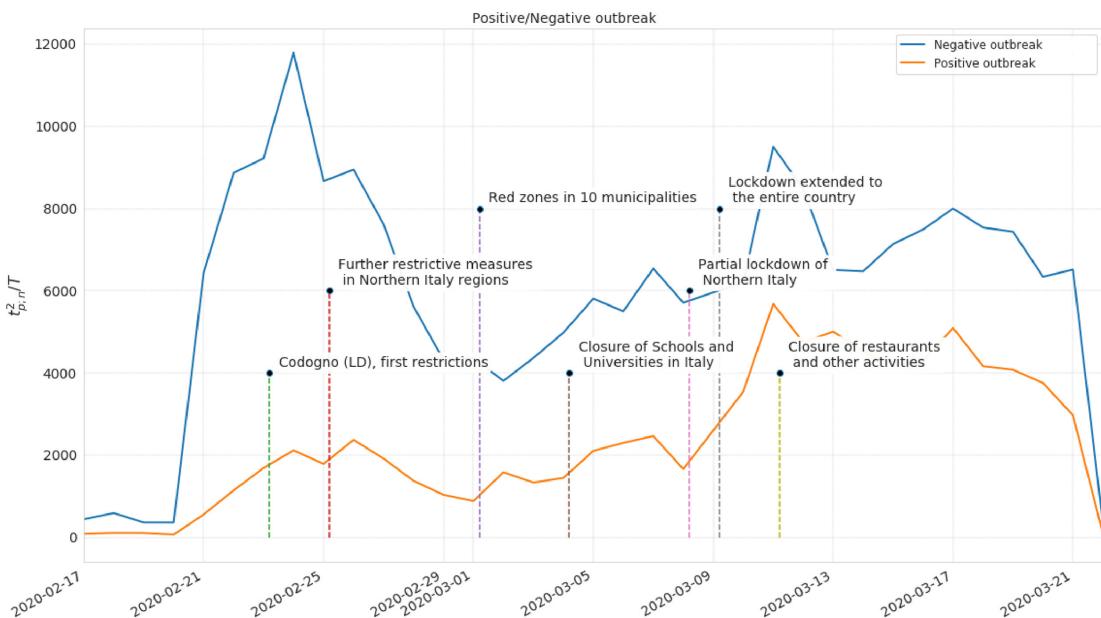


Figura 1: plots the national rescaled daily PO and NO indexes time series during the first four weeks of the Italian coronavirus outbreak. The dashed vertical lines indicate the main government announcements or events. (De Rosis et al. 2021)

- Melo e Figueiredo (2021): il paper si concentra sul Brasile e analizza notizie (dal portale Universo OnLine) e tweet durante i primi mesi di pandemia (tra gennaio e maggio 2020). Cerca di estrarre i topic più importanti per confrontare gli argomenti più trattati dai tweet e dai notiziari (figura 2). Inoltre confronta il sentiment generato dai media tradizionali (figura 3) con quello dei social media (figura 4) mostrando come, in particolare sugli aspetti economici e di prevenzione, il sentiment ottenuto dai tweet sia più negativo rispetto a quello ricavato dalle news.

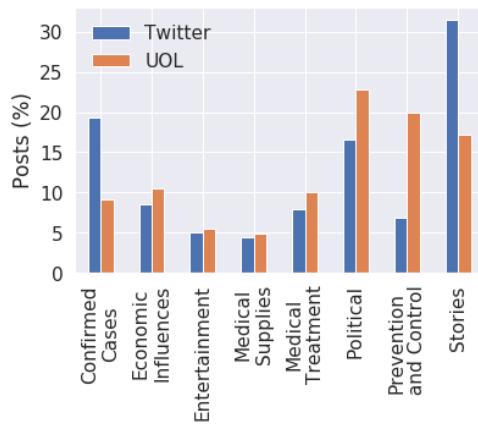


Figura 2: Distribution of themes. UOL: Universo Online. (Melo e Figueiredo 2021)

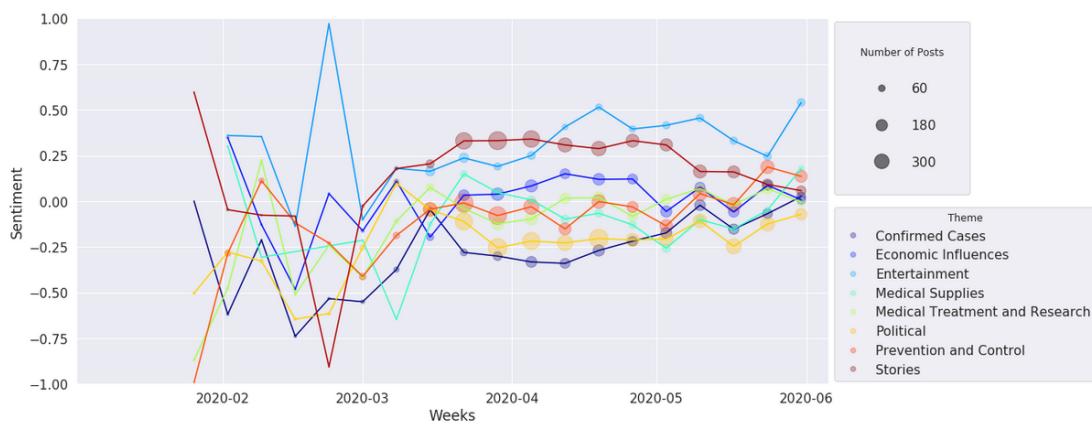


Figura 3: Universo Online sentiment analysis over time. (Melo e Figueiredo 2021)

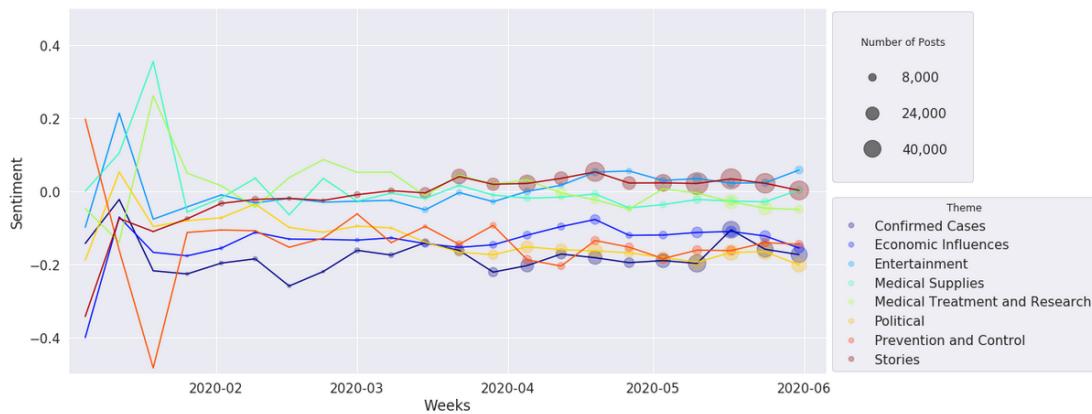


Figura 4: Twitter sentiment analysis over time. (Melo e Figueiredo 2021)

- Garcia e Berton (2021): il paper si concentra su tweet dal Brasile e dagli Stati Uniti durante i primi mesi di pandemia (tra il 17 aprile e l'8 agosto 2020). Cerca di estrarre i topic più importanti per valutare l'evoluzione della quantità di tweet relativi ai vari topic individuati e il rispettivo sentiment. Vengono mostrate le differenze tra i tweet brasiliani (figura 5) e quelli statunitensi (figura 6): quest'ultimi sono prevalentemente negativi per tutti i topic individuati, mentre nel caso brasiliano si nota un sentiment prevalentemente positivo sugli argomenti: politica, trattamenti e sport.

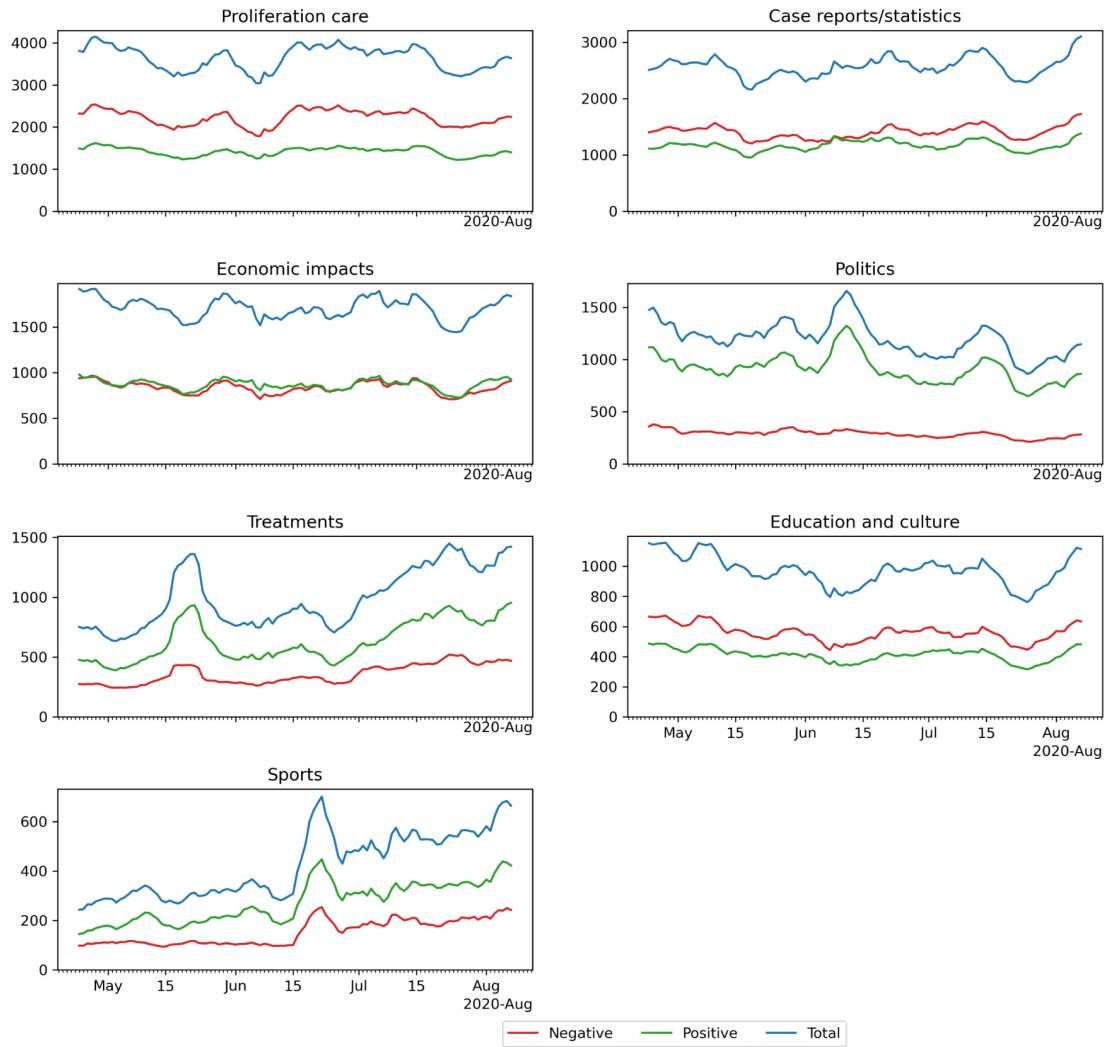


Figura 5: Portuguese volume variations for topic. The horizontal axis represents the posting dates where each point represents the sum of the messages over a week and the vertical axis the number of posts. The blue lines represent the total messages. The other lines are related to sentiment analysis where red are negative and green are positive. (Garcia e Berton 2021)

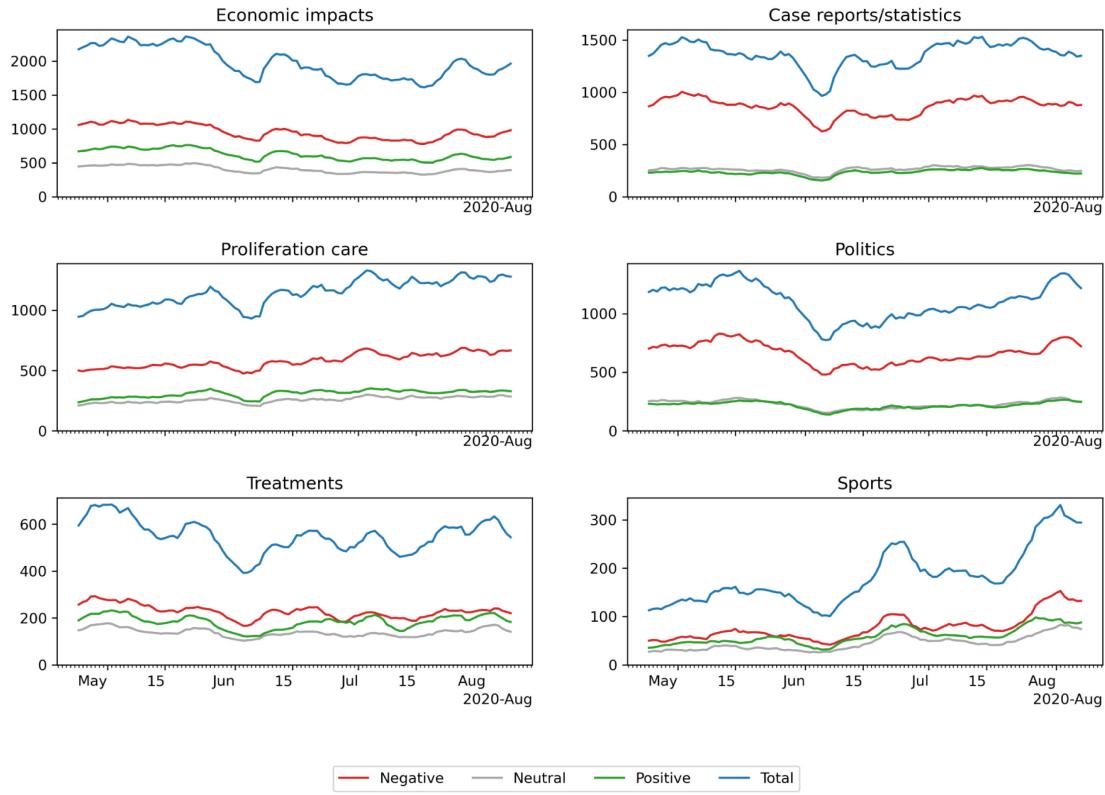


Figura 6: English volume variations for topic. The horizontal axis represents the posting dates where each point represents the sum of the messages over a week and the vertical axis the number of posts. The blue lines represent the total messages. The other lines are related to sentiment analysis where red are negative and green are positive. (Garcia e Berton 2021)

Non sono noti studi in lingua italiana né che abbiano considerato il fattore temporale, l'estrazione dei topic e l'opinion mining congiuntamente né che abbiano confrontato l'opinione sui social media rispetto alla comunicazione da altre fonti.

Raccolta dati

Per le analisi sono stati creati 5 dataset:

- Esperti: Opinioni e commenti dei medici mediaticamente più esposti;
- Notizie: Articoli pubblicati da una testata giornalistica;
- Tweet: Tweet a tema coronavirus in lingua italiana;
- Istituzioni: Tweet pubblicati da alcuni profili istituzionali;
- Regioni: Tweet pubblicati dai profili ufficiali delle regioni.

Dataset: Esperti

È stato scelto un gruppo di esperti tra quelli che più spesso hanno commentato eventi durante la pandemia, l'elenco è riportato nella tabella 6 riportata in appendice. I dati sono stati raccolti tramite scraping dal sito web di Adkronos¹.

Per ognuno dei medici è stata effettuata una ricerca degli articoli pubblicati tra il primo gennaio 2020 e il 31 luglio 2021 contententi nel titolo il nome dell'esperto. Da ogni risultato è stato utilizzato l'url per scaricare l'articolo completo e tramite espressione regolare è stato estratto il testo contenuto tra virgolette attribuendo le frasi selezionate all'esperto che appariva nel titolo dell'articolo.

Il dataset finale conta 2 969 record, con le variabili in tabella 1.

Tabella 1: Variabili del dataset degli esperti

Variabile	Descrizione
expert	nome dell'esperto
datetime	giorno e ora di pubblicazione
url	url dell'articolo
title	titolo della notizia
body	corpo dell'articolo
quoted	parti dell'articolo tra virgolette

Per verificare la validità del metodo utilizzato è stato estratto un CCS di 100 unità e si è osservato che nel 7% (IC al 90%: 3.3|12.7) dei casi il testo non è stato correttamente attribuito.

Durante la prima ondata l'esperto più attivo è stato *Roberto Burioni*, mentre durante la seconda *Andrea Crisanti* e *Massimo Galli* (figura 7).

¹<https://www.adnkronos.com/search>

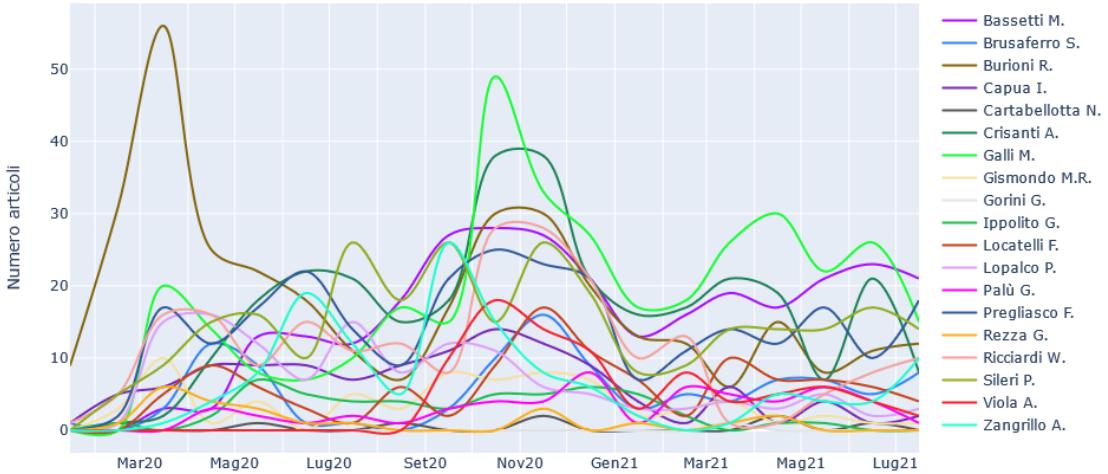


Figura 7: Distribuzione del numero di articoli pubblicati nel tempo per ogni esperto (aggregazione mensile)

Dataset: Notizie

I dati sono stati raccolti tramite scraping utilizzando la funzionalità di ricerca presente nell’archivio web di la Repubblica².

Per ogni giorno tra l’11 gennaio 2020 e il 31 luglio 2021 è stata effettuata una richiesta per scaricare le pagine contenenti l’elenco degli articoli pubblicati. Sono state utilizzate espressioni regolari per estrarre le informazioni e memorizzarle in formato CSV in un unico file contenente 87 943 record, con le variabili in tabella 2.

Tabella 2: Variabili del dataset delle notizie

Variabile	Descrizione
url	url della notizia
title	titolo della notizia
body	sommario o inizio dell’articolo
date	giorno di pubblicazione

Per determinare gli articoli che parlavano di argomenti collegati al coronavirus gli articoli sono stati filtrati utilizzando l’espressione regolare (codice 1) riportata in appendice.

Per verificare la validità del metodo utilizzato è stato estratto un CCS di 100 unità dai 18 923 articoli filtrati dal quale si è osservato che solo nel 4% (IC al 90%: 1.4|8.9) dei casi l’articolo non trattava gli argomenti di interesse.

Il periodo di maggiore interesse mediatico è stato il mese di marzo 2020 con circa il 55% degli articoli che trattavano l’argomento in studio (figura 8).

Sono stati effettuati tentativi di recupero dei dati da altre testate giornalistiche che a causa di paywall, rate limits, formattazione delle pagine html non omogenee non hanno portato alla possibilità di ottenere dataset di qualità paragonabile a quello utilizzato.

²<https://ricerca.repubblica.it/ricerca/repubblica-it>

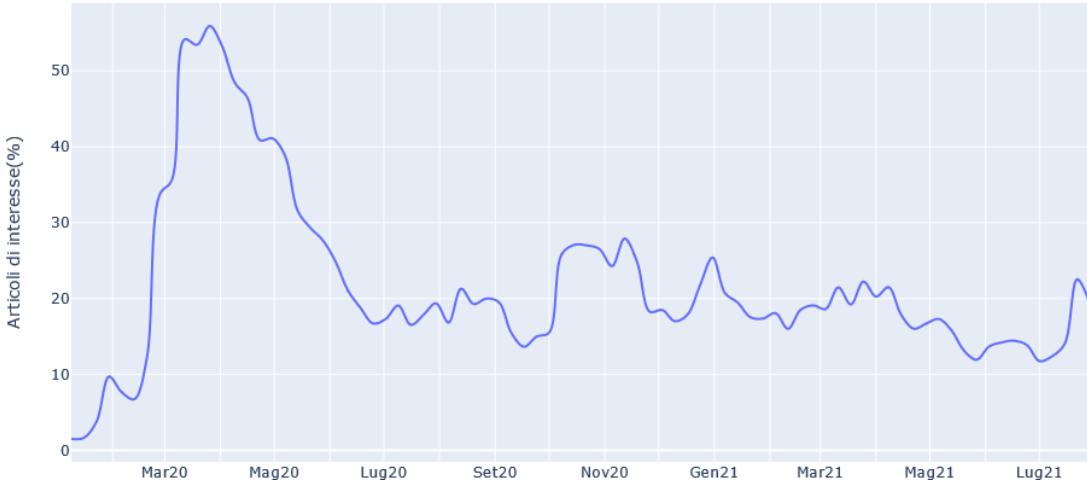


Figura 8: Percentuale degli articoli pubblicati relativi al covid rispetto al totale nel tempo (aggregazione settimanale)

Dataset: Tweet

I dati sono stati raccolti tramite scraping sfruttando il pacchetto snscreape³.

Per ogni giorno tra il 21/01/2020 e il 25/07/2021 è stata effettuata una richiesta dei tweet contenenti almeno uno dei termini elencati in testo 1 riportato in appendice. Con i parametri `-filter:links -filter:replies lang:it` sono stati esclusi retweet e link e considerati solo tweet in italiano.

Per ogni giornata è stata ottenuta risposta in formato JSON contenente circa 480 tweet più popolari (secondo l'algoritmo di Twitter). I risultati sono stati convertiti e uniti in un unico file CSV contenente 265 959 record, con le variabili in tabella 3.

Tabella 3: Variabili del dataset creato

Variabile	Descrizione
date	Data e ora del tweet
text	Corpo del tweet
source	Strumento utilizzato per twittare
hashtags	Lista degli hashtag presenti nel tweet
mentions	Lista degli utenti menzionati nel tweet

La maggior parte dei tweet è creata tramite app o client web (figura 9), mentre l'1.45% sono tweet automatici che sono stati esclusi dalle analisi.

Per verificare la validità del metodo utilizzato è stato estratto un CCS di 100 unità dai 262 107 tweet rimasti osservando che solo nel 2% (IC al 90%: 0.0|6.2) dei casi non venivano trattati argomenti collegati al coronavirus.

³<https://github.com/JustAnotherArchivist/snscreape>

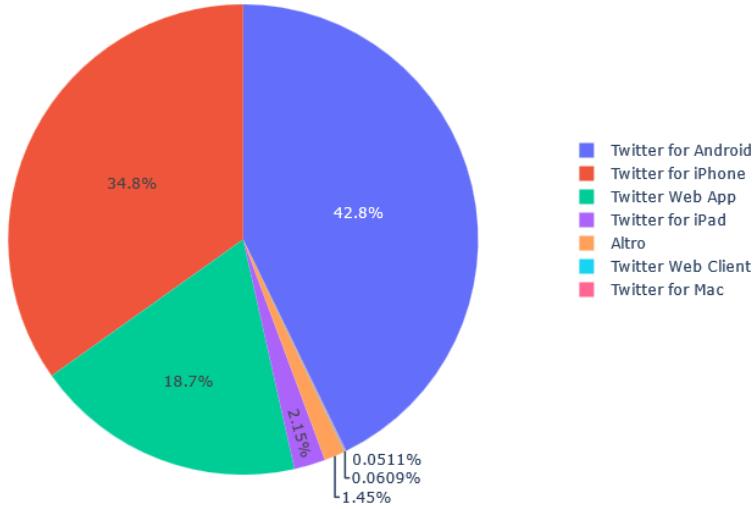


Figura 9: Distribuzione delle applicazioni utilizzate per la pubblicazione di tweet

Dataset: Istituzioni

I dati sono stati raccolti usando snscreape. Sono stati scaricati tutti i tweet fino al 31/07/2021 dagli account ufficiali di:

- **Ministero della Salute**⁴ a partire dal 20/01/2020
- **Presidenza del Consiglio dei Ministri**⁵ a partire dal 30/01/2020
- **Agenzia Italiana del Farmaco**⁶ a partire dal 03/01/2020
- **Dipartimento della Protezione Civile**⁷ a partire dal 31/01/2020
- **Istituto Superiore di Sanità**⁸ a partire dal 15/01/2020

Ottenendo 4 812 tweet con le variabili in tabella 4 che sono stati ulteriormente filtrati per mantenere solo quelli di interesse utilizzando l'espressione regolare (codice 2) riportata in appendice.

Tabella 4: Variabili dei dataset istituzioni e regioni

Variabile	Descrizione
user	Nome dell'account che ha twittato
date	Data e ora del tweet
text	Corpo del tweet
hashtags	Lista degli hashtag presenti nel tweet
mentions	Lista degli utenti menzionati nel tweet

Per verificare la validità del metodo utilizzato è stato estratto un CCS di 100 unità dai 2 405 tweet osservando che nel 6% (IC al 90%: 2.6|11.54) dei casi non venivano trattati argomenti collegati al coronavirus.

⁴<https://twitter.com/MinisteroSalute>

⁵https://twitter.com/Palazzo_Chigi

⁶https://twitter.com/Aifa_ufficiale

⁷<https://twitter.com/DPCgov>

⁸<https://twitter.com/istsupsan>

Nel primo periodo di emergenza @MinisteroSalute e @DPCgov sono risultati i canali più attivi. @Palazzo_Chigi è invece più attivo dall'insediamento del Governo Draghi (figura 10).

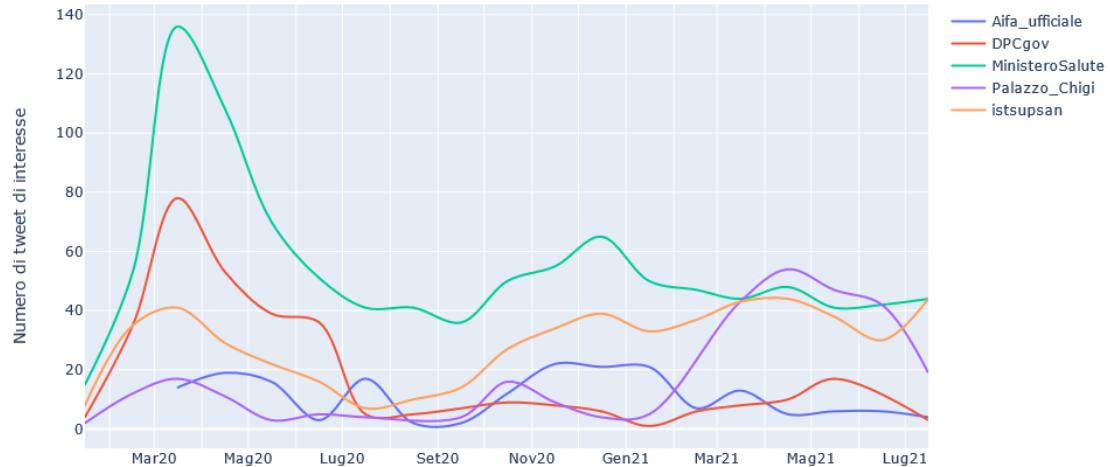


Figura 10: Numero di tweet relativi al covid pubblicati per canale istituzionale (aggregazione mensile)

Dataset: Regioni

I dati sono stati raccolti usando snscreape. Sono stati scaricati tutti i tweet dal 30 gennaio 2020 al 31 luglio 2021 dagli account twitter delle regioni (elenco in tabella 7 in appendice) ottenendo 37112 tweet con le variabili in tabella 4 che sono stati ulteriormente filtrati utilizzando un espressione regolare (codice 2 in appendice) per ricavare quelli di interesse.

Per verificare la validità del metodo utilizzato è stato estratto un CCS di 100 unità dai 18 994 tweet osservando che solo nell'1% (IC al 90%: 0.0|4.7) dei casi non venivano trattati argomenti collegati al coronavirus.

Come mostrato in figura 11 l'account @RegLiguria risulta generalmente molto attivo, con @RegLombardia che è molto attiva nei periodi di massima emergenza (marzo 2020, periodo del lockdown e marzo 2021 con la regione che passa in zona rossa).

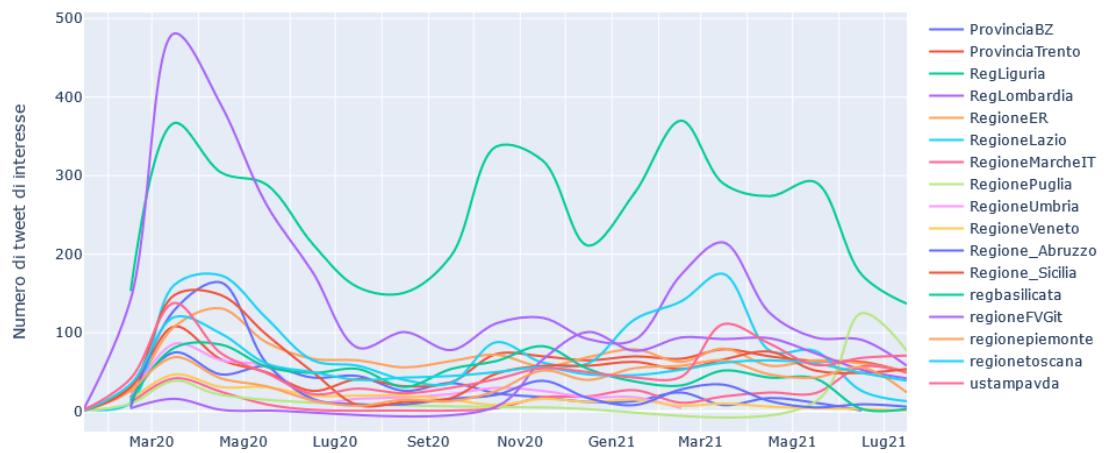


Figura 11: Numero di tweet relativi al covid pubblicati per account regionale (aggregazione mensile)

Analisi

Preprocessing

Ogni documento è stato sottoposto a:

- Tokenizzazione: le frasi vengono suddivise in parole (token)
- Lowercasing: tutti i caratteri vengono convertiti in minuscolo
- Part of Speech (PoS) tagging: a ogni token è associata la relativa categoria lessicale
- Lemmatizzazione: ogni token è ridotto alla forma canonica
- Rimozione stopword: token molto frequenti e/o non significativi vengono rimossi
- Vettorizzazione: ad ogni token viene associato un vettore fastText di 300 componenti

Per il preprocessing è stata utilizzata la libreria SpaCy (Honnibal et al. 2020) e il rispettivo modello preaddestrato per la lingua italiana `it_core_news_lg`⁹

Tabella 5: Esempio delle varie fasi di preprocessing

Fase	Esempio
Testo originale	Oggi sono andato a prendere un caffè in una pasticceria e si sentiva più l'odore dell'Amuchina che dei dolci. #COVID19 #Coronavirus
Tokenizzazione	Oggi, sono, andato, a, prendere, un, caffè, ... , COVID, 19, #, Coronavirus
Rimozione stopword	andato, prendere, caffè, pasticceria, sentiva, odore, Amuchina, dolci, COVID, Coronavirus
Lowercasing	andato, prendere, caffè, pasticceria, sentiva, odore, amuchina, dolci, covid, coronavirus
PoS Tagging	andato:VERB, prendere:VERB, ... , amuchina:PROPN, dolci:NOUN, covid:PROPN, coronavirus:PROPN
Lematizzazione	andare, prendere, caffè, pasticceria, sentire, odore, amuchina, dolci, covid, coronavirus
Word vector	andare:[-1.24, ..., 0.15], prendere:[1.59, ..., -3.86], caffè:[3.71, ..., -0.09], ... , coronavirus:[0.54, ..., 0.80]

⁹https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.1.0

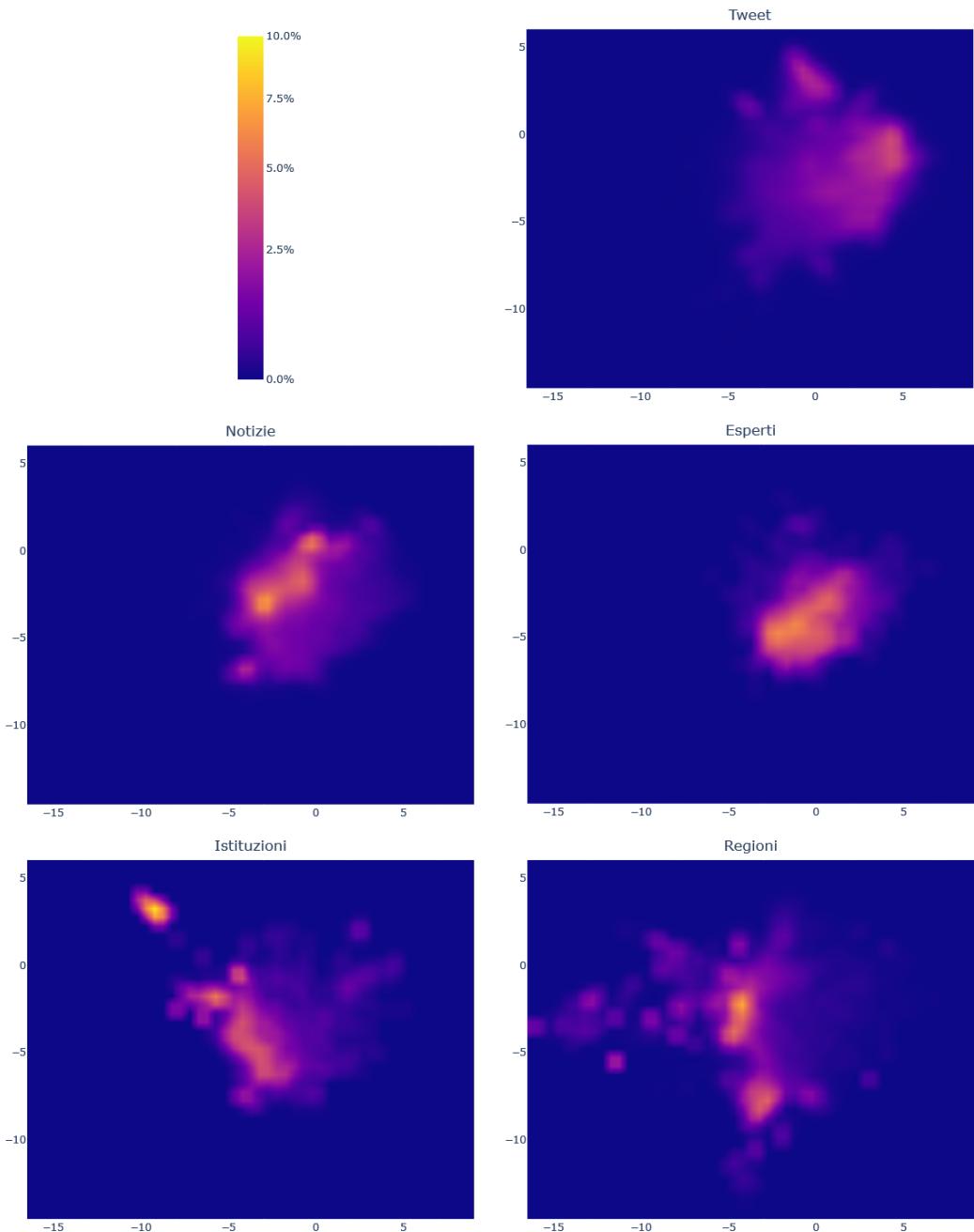


Figura 12: Distribuzione dei testi nello spazio bidimensionale per ogni fonte

Rappresentazione

A ogni documento, in seguito al preprocessing, è stato associato un vettore reale 300-dimensionale calcolato come media aritmetica semplice dei word-vector. Per rappresentare bidimensionalmente i vari testi è stato stimato un modello utilizzando l'approccio UMAP parametrico, utilizzando come metrica la dissimilità del coseno e come numero di vicini il valore 5 (figura 12). Si può notare come in base alle varie fonti i testi si distribuiscano in maniera differente, indice di possibili differenze tra le varie sorgenti analizzate.

Clustering

Per interpretare più facilmente i testi ne è stato effettuato un raggruppamento utilizzando il GSDMM sui testi preprocessati con parametri di α e β pari a 0.1, 200 cluster iniziali e 30 epoche. Avendo ottenuto 180 cluster, numero troppo elevato per essere facilmente interpretabile, i cluster sono stati ulteriormente raggruppati tramite clustering gerarchico aggregativo (figura 13) utilizzando come metrica una versione modificata dell'importanza della parola entro cluster ($\phi_{c,w}$) proposta da Yin e Wang (2014):

$$\phi_{c,w}^* = \frac{\phi_{c,w}}{\sum_{i \in C} \phi_{i,w}} \quad (1)$$

Così da correggere l'importanza proporzionalmente alla frequenza di apparizione della parola nei testi. Per il clustering sono state utilizzate distanza di Manhattan e il metodo del legame di Ward. Si è scelto di mantenere 24 cluster, estraendo l'argomento trattato osservando le word cloud costruite utilizzando l'equazione (1) (figura 14).

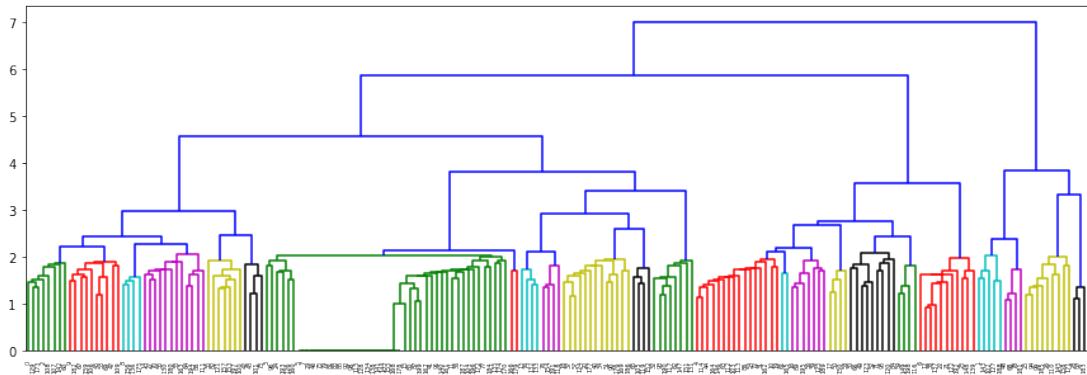


Figura 13: Aggregazione dei cluster ottenuti tramite GSDMM

In figura 15 si nota come, rappresentando nel sottospazio UMAP la distribuzione dei testi sotto l'assunzione che si distribuiscono condizionatamente al cluster di appartenenza secondo una normale bivariata, i cluster relativi a aggiornamenti (#23, #21, #20), dati su tamponi, positivi e decessi (#22), dirette e conferenze stampa (#11) siano prevalentemente trattati da *istituzioni* o *regioni*. Al contrario il tema dei colori delle zone (#18) è più diffuso tra i *tweet*.



Figura 14: Word cloud dei 24 cluster

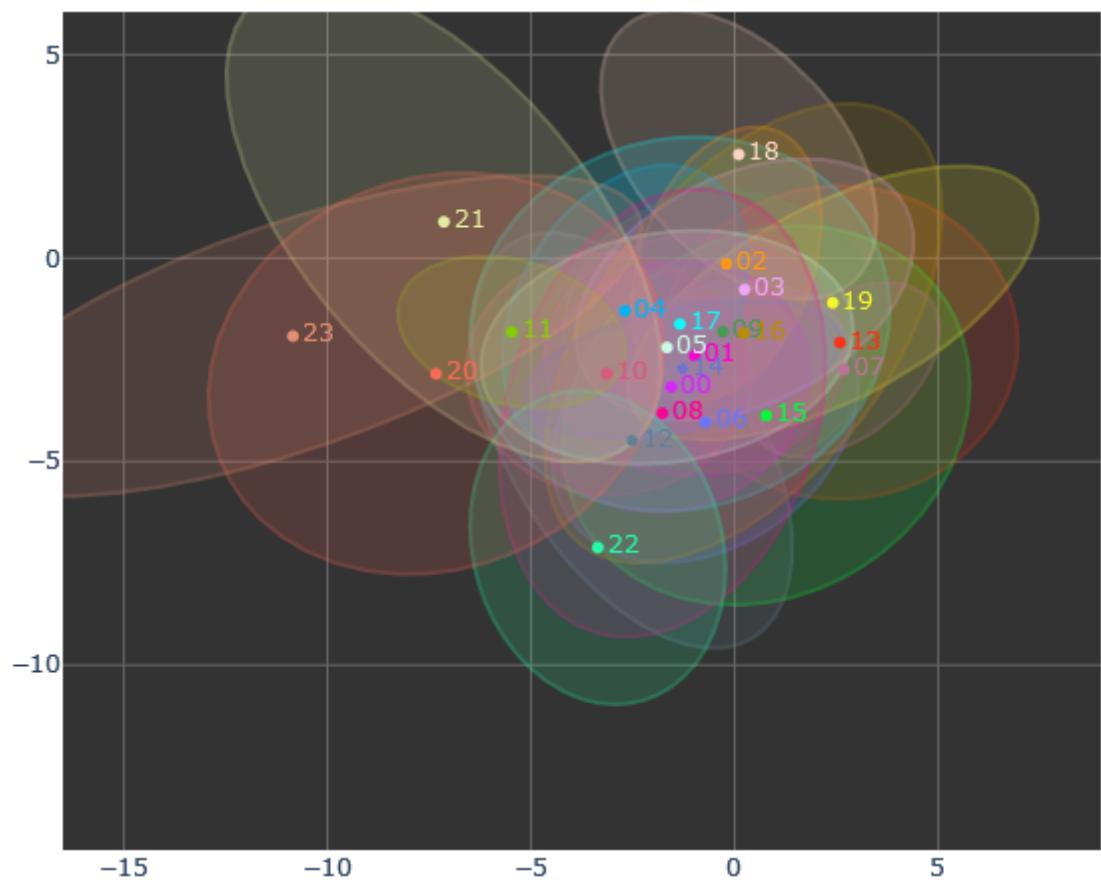


Figura 15: Distribuzione dei cluster nello spazio 2d con regione di confidenza al 90%

Sentiment Analysis

Per analizzare l'opinione dei testi si è optato per modelli classificativi basati su BERT applicati sui testi preprocessati esclusivamente con tokenizzazione. Sono stati utilizzati due modelli classificativi preaddestrati sulla lingua italiana per individuare:

- Sentiment (Positivo, Neutrale, Negativo): modello preaddestrato da Neuraly (2021)
- Emotion (Gioia, Paura, Rabbia, Tristezza): modello preaddestrato da Bianchi, Nozza e Hovy (2021)

Rappresentando nel sottospazio UMAP la distribuzione dei sentiment dei testi sotto l'assunzione che si distribuiscano secondo una normale bivariata (figura 16) si nota come i testi neutrali siano distribuiti in tutto lo spazio mentre quelli polarizzati (positivamente o negativamente) occupino lo spazio dei tweet e delle opinioni degli esperti. Per quanto riguarda la paura sembra essere più diffusa nei testi di *istituzioni* e *regioni* rispetto alla rabbia che è prevalente in *tweet* e *esperti*.

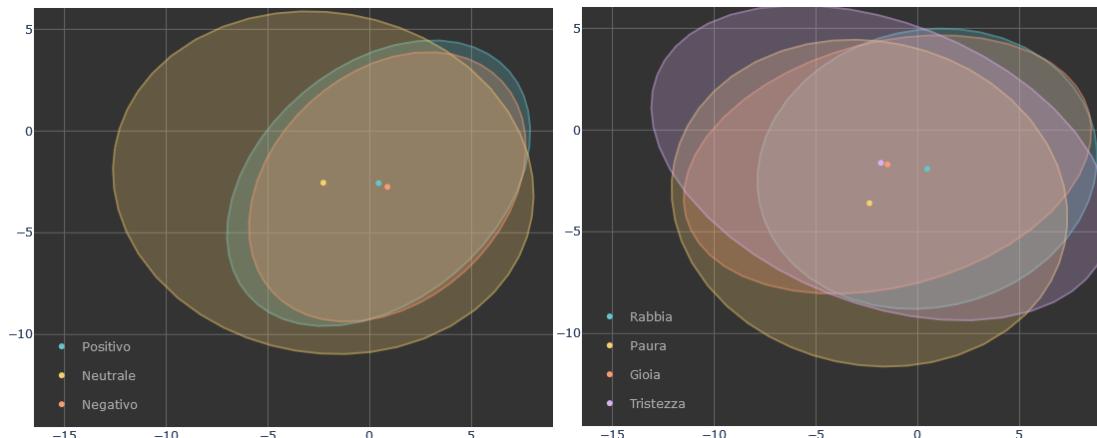


Figura 16: Distribuzione dei sentiment e emozioni nello spazio 2d con regione di confidenza al 99%

Per analizzare più nel dettaglio le motivazioni e gli argomenti correlati al sentimento è stato utilizzato un approccio bayesiano, sfruttando gli output della funzione softmax associati al singolo testo per ognuna delle categorie:

1. Assumendo che ogni testo contenga una sola opinione (plausibile poiché stiamo trattando testi brevi) assegnamo ad ogni parola nel testo il valore del sentimento globale del testo;
2. Per ogni parola consideriamo la collezione dei punteggi dei testi che la includono come estrazioni da una distribuzione Beta;
3. Stimiamo i parametri α e β della distribuzione utilizzando il metodo della massima verosimiglianza;

4. Consideriamo l'indice di asimmetria γ della distribuzione Beta come importanza della parola associata al sentiment;
5. Calcoliamo i pesi come:

$$\left(\frac{1}{n_d} \sum_{i=1}^{n_d} \gamma_i - \frac{1}{n_w} \sum_{w \in d} \gamma_w \right) \cdot \log(n_w)$$

con n_d numero dei documenti e n_w numero dei documenti contenenti la parola considerata.

6. Creiamo la word cloud delle parole più importanti escludendo le parole con pesi negativi. Rimuovendo anche gli aggettivi non esplicativi si possono estrarre dei topic più significativi.

Primi risultati

Notizie

Da figura 17 si nota come temi diversi vengano trattati con costanza nel tempo, in particolare per tutto il periodo: la situazione economica (#00), le questioni politiche (#01), le notizie sportive (#02) ed eventi e cultura (#03) ricoprono un'elevata quota degli articoli pubblicati. Nel mese di febbraio 2020 il cluster #09 (relativo alla situazione cinese) è molto trattato. Appaiono da novembre 2020 il tema delle zone (#18) e dal mese successivo #12 (campagna vaccinale) e #15 (vaccini).

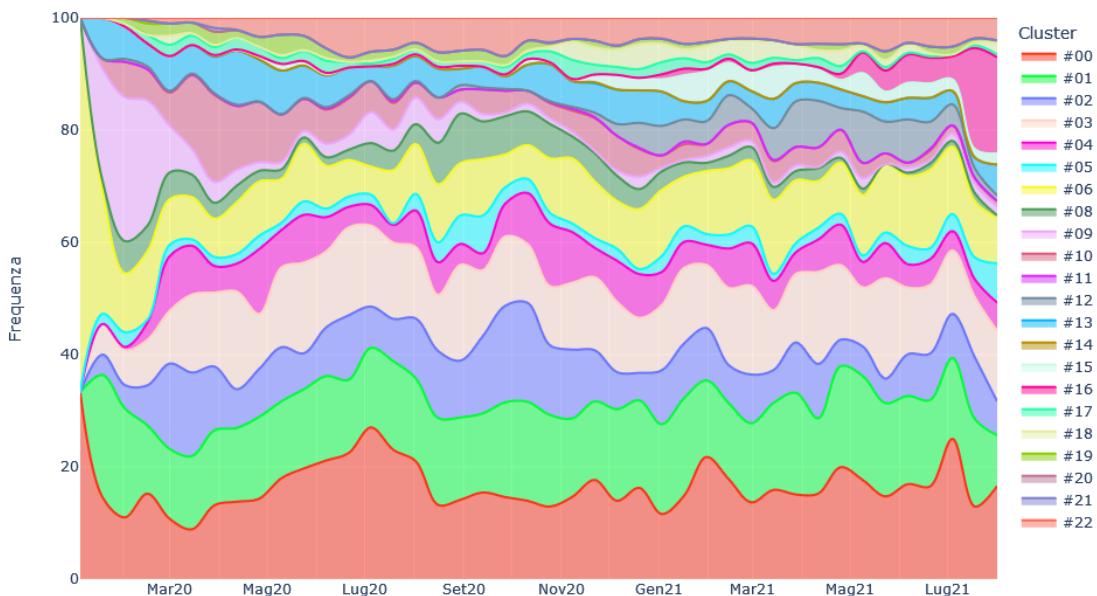
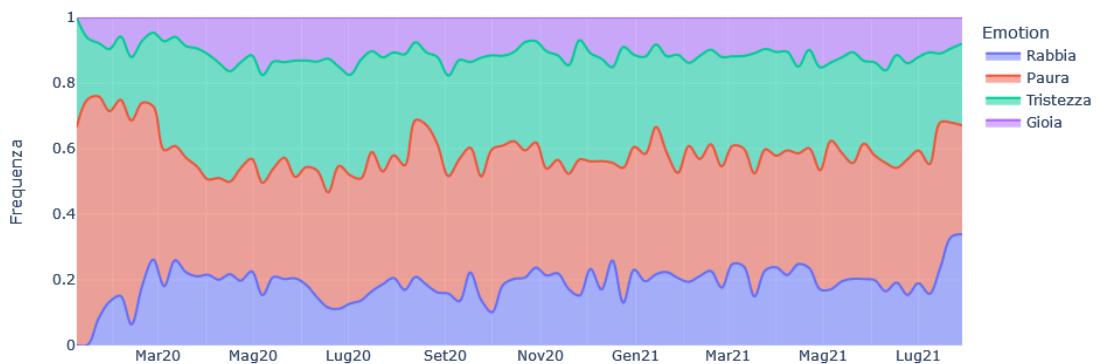
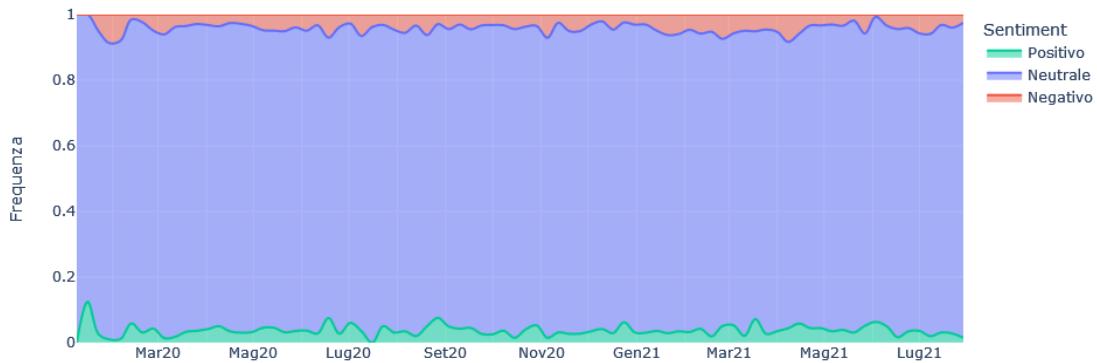


Figura 17: Evoluzione dei topic sul dataset notizie nel tempo (aggregazione quindicinale)

In figura 18 si osserva come il sentiment sia prevalentemente neutrale, mentre da figura 19 si rileva che la paura è l'emotion più diffusa delle altre e che per quanto

riguarda la rabbia ci sono 2 periodi di aumento: il primo a inizio pandemia (marzo 2020) e il secondo nell'ultimo periodo di osservazione (luglio 2021).



Tweet

Da figura 20 si nota come il tema #13 (pandemia: quarantena e lockdown) sia stato il più trattato nel tempo, seguito dagli argomenti collegati all'utilizzo della mascherina (#19) soprattutto nel periodo tra maggio e ottobre 2020. Nel mese di febbraio 2020 il cluster #09 (relativo alla situazione cinese) è più diffuso per poi perdere interesse con il peggioramento della situazione italiana. Il tema delle zone a colori (#18) è di forte interesse dall'introduzione della misura a novembre 2020 fino a maggio 2021 con l'allentamento delle misure in tutta Italia (da notare anche il leggero picco nel mese di marzo 2020 relativo alla discussione sulla zona rossa nell'area bergamasca). Il tema delle vaccinazioni (#15) acquisisce interesse nel periodo tra maggio e giugno 2021 per poi lasciare spazio al tema del green pass (#16) che dal mese di luglio esplode diventando il tema più trattato al termine del periodo osservato.

In figura 21 si osserva come il sentiment sia omogeneo nel tempo e in prevalenza neutrale (60% circa), con una forte presenza di opinioni negative (25%

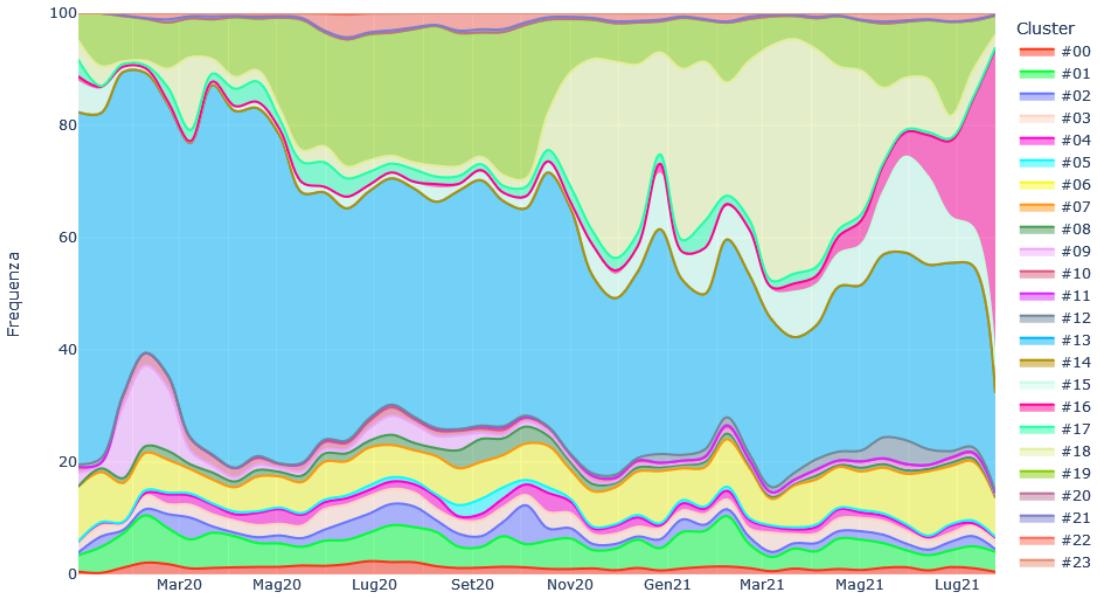


Figura 20: Evoluzione dei topic sul dataset tweet nel tempo (aggregazione quindicinale)

circa) e il rimanente 15% positivo. A differenza delle altre fonti, in figura 22 si rileva che l'emotion principale è la rabbia. Nel periodo tra febbraio e marzo 2020 si nota un picco di paura, mentre dalla seconda settimana di luglio 2021 un forte aumento della rabbia.

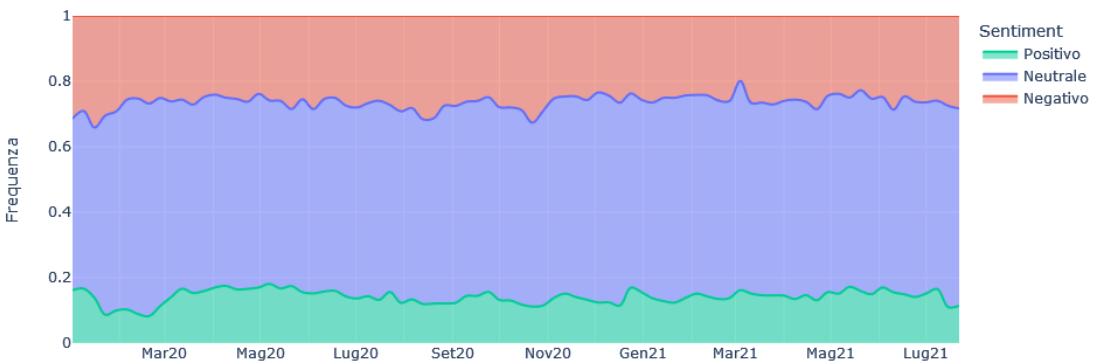


Figura 21: Evoluzione del sentiment sul dataset tweet nel tempo (aggregazione settimanale)

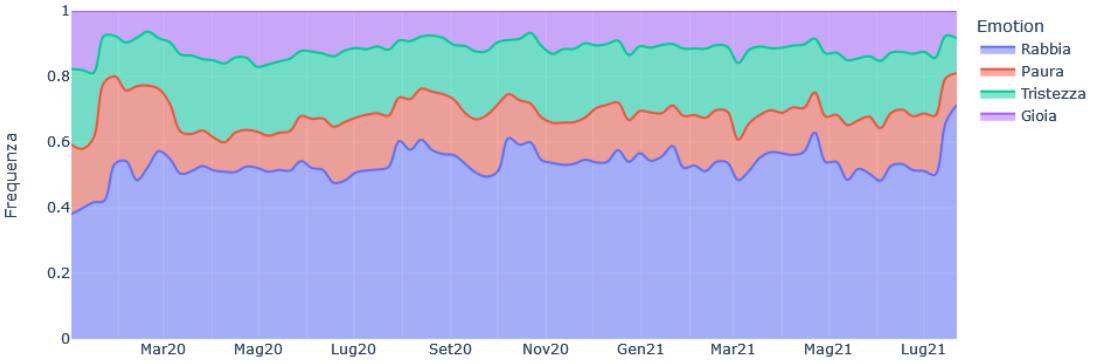


Figura 22: Evoluzione dell'emotion sul dataset tweet nel tempo (aggregazione settimanale)

Esperti

Da figura 23 si nota come #06 (numeri su virus e vaccini) sia il tema prevalente (circa il 60% degli articoli), seguito a distanza dagli aspetti economici (#00) che hanno acquisito interesse con l'arrivo del nuovo anno. Per la situazione cinese (#09) il comportamento è analogo a quello dei tweet delle persone comuni. Il tema #13 (pandemia: quarantena e lockdown) è stato fortemente discusso soprattutto durante la prima ondata A cavallo tra settembre e ottobre 2020, con l'inizio dell'anno scolastico, c'è stato un leggero picco sul tema delle scuole (#05).

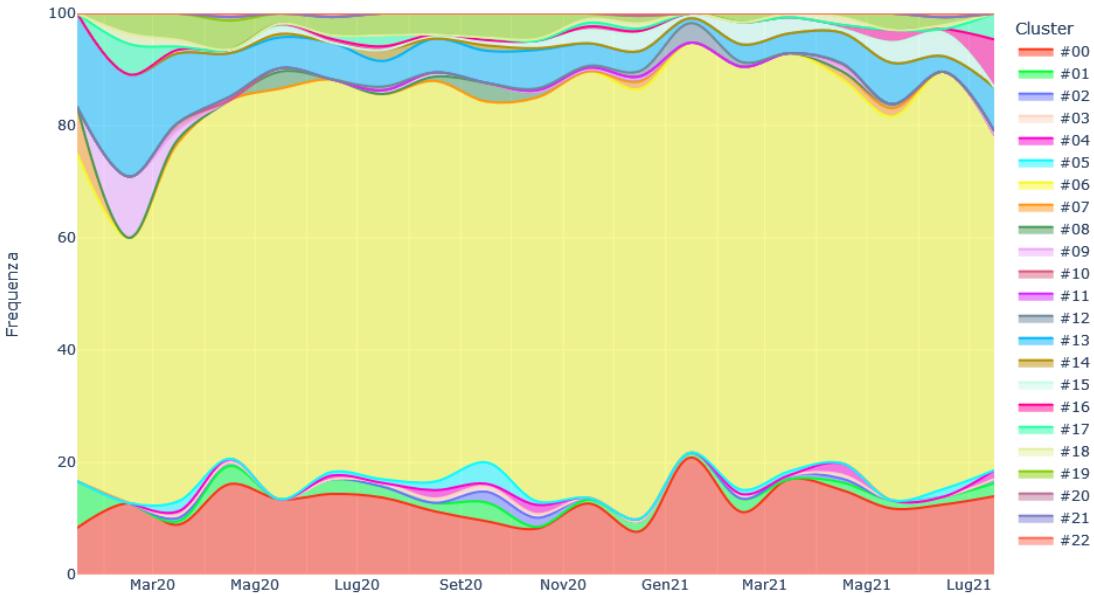


Figura 23: Evoluzione dei topic sul dataset esperti nel tempo (aggregazione mensile)

In figura 24 si osserva come il sentiment sia più negativo che positivo, con un lieve miglioramento da giugno 2021. Da figura 25 si rileva che la paura è l'emotion più diffusa, seguita da rabbia e tristezza.

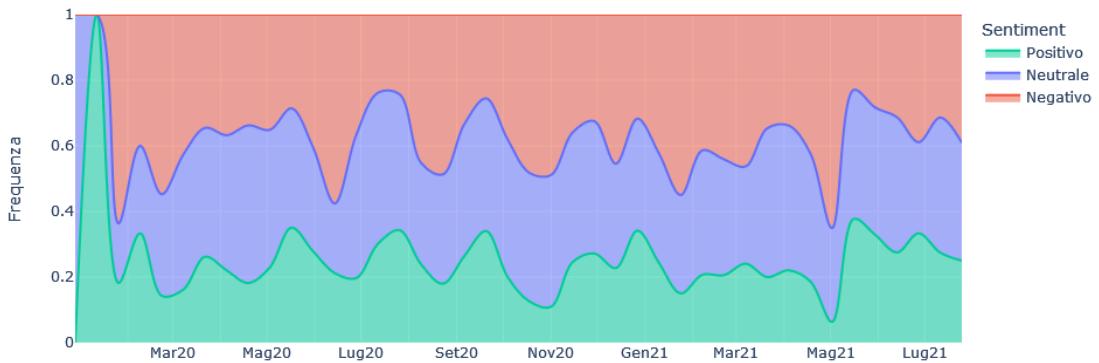


Figura 24: Evoluzione del sentiment sul dataset esperti nel tempo (aggregazione quindicinale)

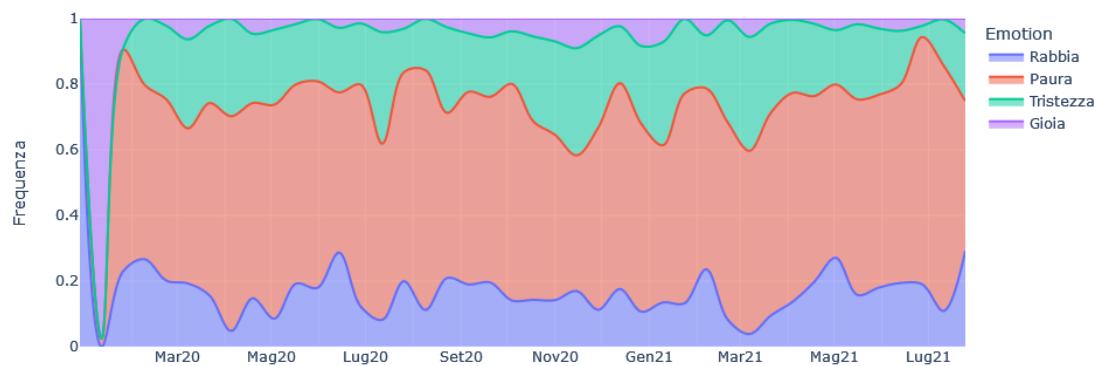


Figura 25: Evoluzione dell'emotion sul dataset esperti nel tempo (aggregazione quindicinale)

Istituzioni

In figura figura 26 si nota come aggiornamenti relativi alla situazione (#21), Numeri su virus e vaccini (#06) siano fortemente trattati. Altri temi molto trattati soprattutto durante la prima fase sono sullo stato del servizio sanitario e di medici e infermieri (#10) e su conferenze stampa e dirette (#11). Da dicembre 2020 anche i cluster #12 (campagna vaccinale) e #15 (vaccini) acquisiscono interesse. da marzo 2021 il cluster #01 (temi politici) è molto più presente.

In figura 27 si osserva come il sentiment sia decisamente neutrale, mentre da figura 28 si rileva che la paura è l'emotion più diffusa delle altre, con un forte calo nel periodo compreso tra giugno e ottobre 2020 e massimi nel periodo iniziale (marzo 2020) e durante la seconda ondata (novembre 2020). Sostanzialmente assente la rabbia.

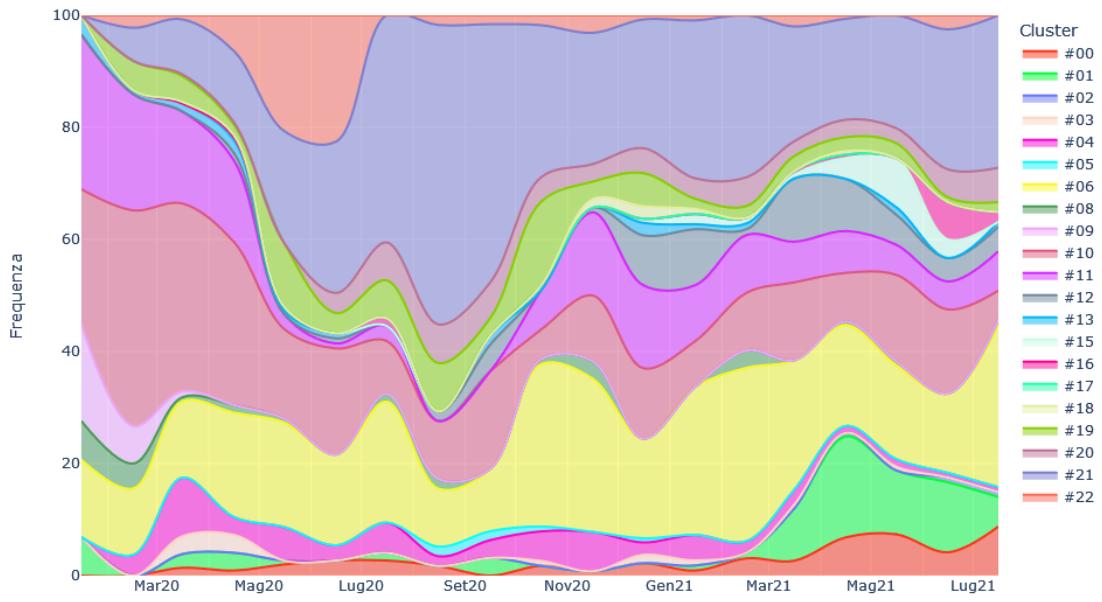


Figura 26: Evoluzione dei topic sul dataset istituzioni nel tempo (aggregazione mensile)

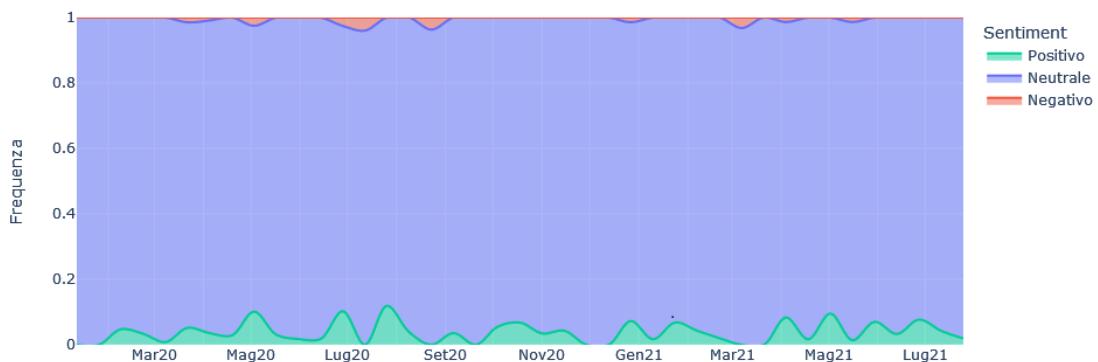


Figura 27: Evoluzione del sentimento sul dataset istituzioni nel tempo (aggregazione quindicinale)

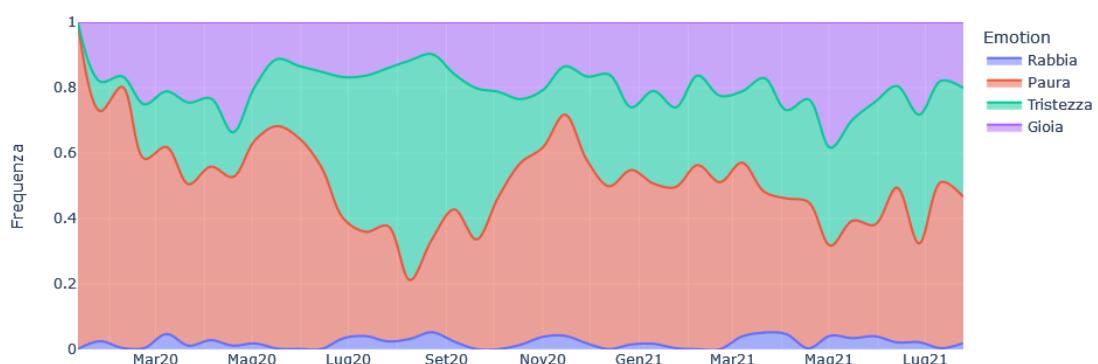


Figura 28: Evoluzione dell'emotion sul dataset istituzioni nel tempo (aggregazione quindicinale)

Regioni

In figura 29 si nota come aggiornamenti relativi alla situazione (cluster #20, #21, #23) e dati su tamponi, positivi e decessi (#22) siano gli argomenti principali. Altro tema importante quello relativo al sostegno del servizio sanitario e ringraziamento a medici e infermieri (#10) che perde interesse nel tempo passando dal 30% dei tweet fino a maggio 2020 al 10% nel nuovo anno. Informazioni su ordinanze e decreti (#04) è diffuso con picchi a marzo 2020 (inizio prima ondata), maggio 2020 (fine prima ondata) e ottobre 2020 (inizio seconda ondata). Da gennaio 2021 c'è una forte spinta sul tema delle vaccinazioni (#12).

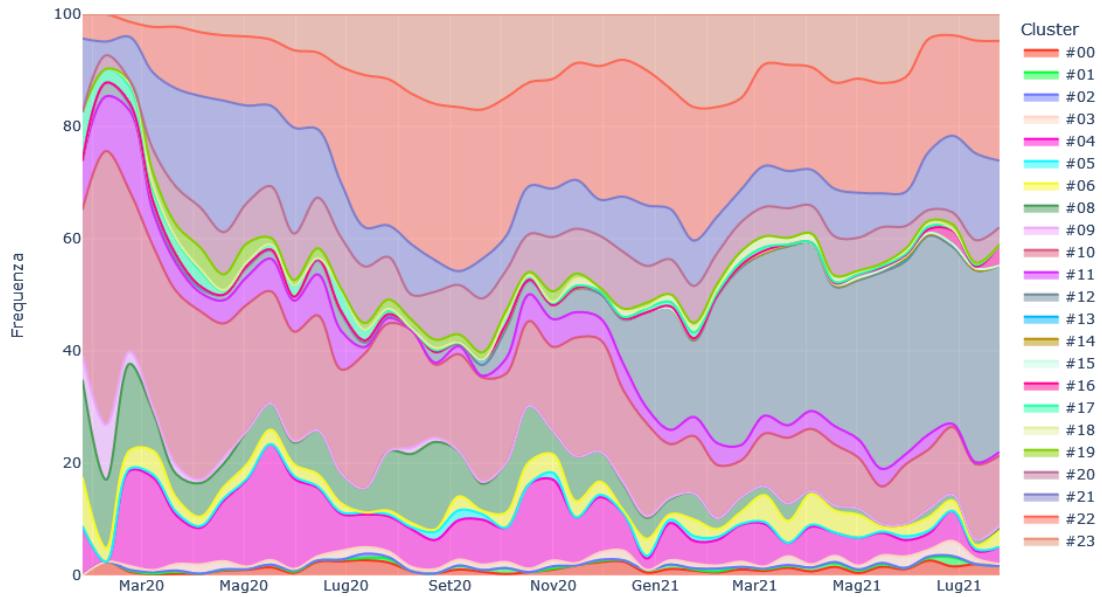
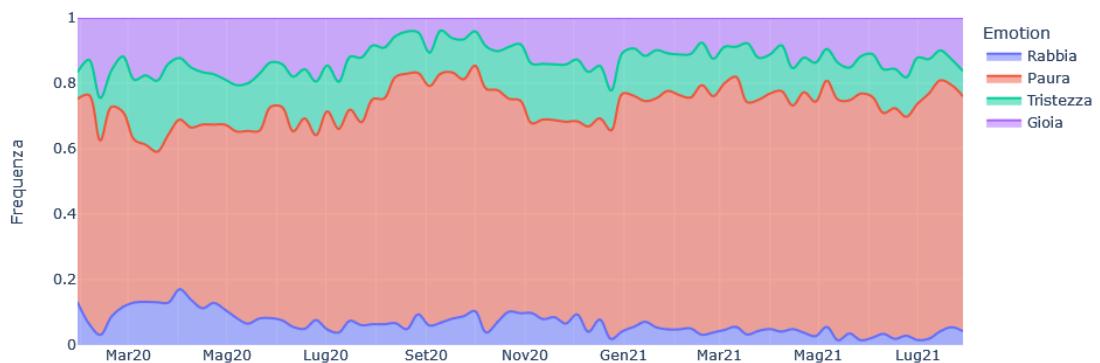


Figura 29: Evoluzione dei topic sul dataset regioni nel tempo (aggregazione quindicinale)

In figura 30 si osserva come il sentiment sia decisamente neutrale, mentre da figura 31 si rileva che la paura è costantemente l'emotion più diffusa e registra un aumento nel periodo agosto-novembre 2020, assestandosi su valori più elevati nel corso del 2021 rispetto al 2020. La rabbia è più elevata durante la prima ondata rispetto al resto del periodo osservato.



Considerazioni finali

Risultati

Risposta alle misure contenitive

Come osservabile dai dati sui ricoveri in figura 32 l'evoluzione pandemica è stata caratterizzata da varie ondate.

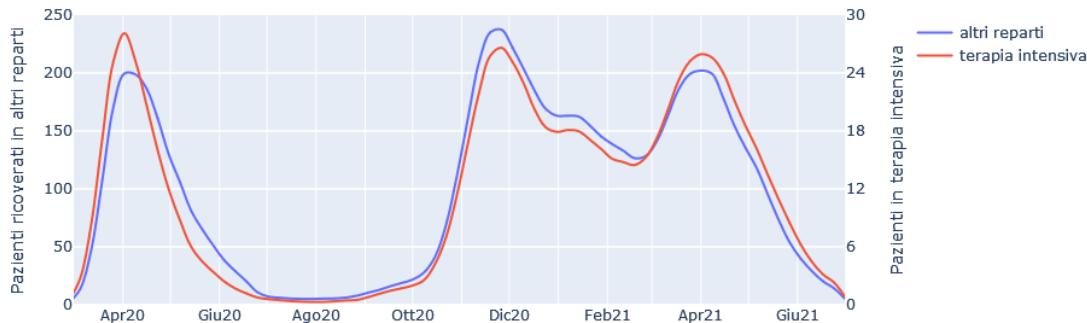


Figura 32: Numero di ricoveri ordinari e in terapia intensiva nel tempo (migliaia di persone per settimana)

Ogni ondata è stata divisa in due fasi: la prima con il trend dei ricoveri in aumento e la seconda con il calo. Valutando le cause dell'emotion nelle varie fasi (figura 33) osserviamo alcune caratteristiche dell'evoluzione in ognuno dei temi osservati:

- Rabbia: Durante l'inizio della seconda ondata sono “multe” e “negazionisti” le principali cause di rabbia, mentre dal termine della seconda aumenta il disappunto verso chi si lamenta. Salvini risulta il politico meno apprezzato sul social sia a inizio pandemia sia nell'ultimo periodo di osservazione.
- Paura: Le preoccupazioni iniziali relative al covid sono legate alle notizie dei contagi sulla “Diamond Princess” e sul panico nella prima fase, per poi passare alla paura per il rischio di essere contagiati e la presenza di focolai sul finire della prima ondata. All'inizio della seconda ondata si registra una maggiore attenzione verso i sintomi e verso le ordinanze e le altre misure adottate.
- Gioia: Durante la fase iniziale della prima ondata il termine “lockdown” è spesso utilizzato con positività. Nella fase finale della terza ondata il termine “rivedere” è correlato alla gioia nonostante le maggiori restrizioni nel periodo natalizio.
- Tristezza: Resta piuttosto omogenea nel tempo anche se sul finire della seconda ondata aumenta la voglia di un ritorno alla normalità, seguita dall'emergere di termini legati alla stanchezza durante l'inizio della terza ondata.



Figura 33: Word cloud per emotion durante le varie fasi pandemiche

Opinioni sull'app Immuni

L'applicazione per il tracciamento dei contatti “Immuni” disponibile in Italia dal 15 giugno 2020 non ha riscosso il successo sperato, con la proporzione di casi tracciati costantemente intorno al 5% rispetto ai casi totali (figura 34).

In figura 35 mostriamo quanto si sia parlato di questa applicazione nel tempo (clustering con GSDMM) e appare evidente come non sia mai entrato nell'interesse collettivo, con solo i canali istituzionali che hanno dedicato a questo tema una percentuale superiore al 2% del totale dei tweet pubblicati.

Oltre alla scarsa penetrazione anche la percezione degli utenti è rimasta negativa per il sentimento (figura 36) e per l'emotion con la prevalenza di rabbia e paura (figura 37). In entrambi i casi i grafici mostrano una percezione peggiore rispetto a quella non filtrata per argomento nello stesso periodo temporale.

In figura 38 si nota come al sentimento positivo siano associate l'utilità per combattere il covid, soprattutto nell'individuazione dei focolai. A quello negativo invece i dubbi sul funzionamento, con paura per cosa fare in caso di possibile contatto con infetti e rabbia per quanto riguarda le tematiche legate alla privacy.

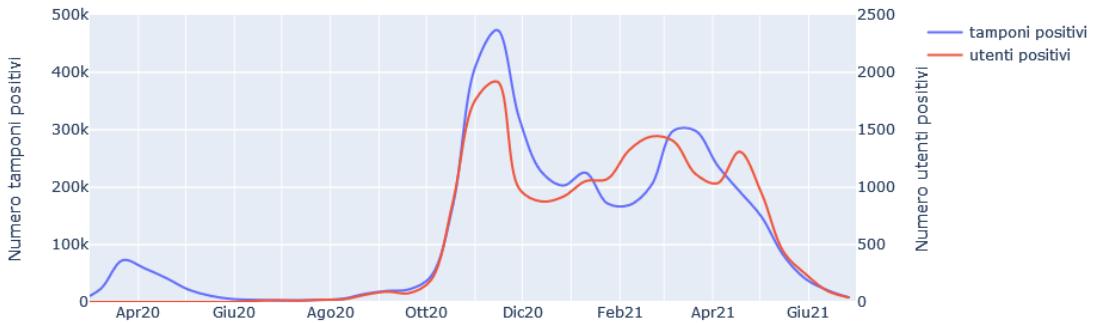


Figura 34: Confronto tra numero di persone positive in Italia e positivi registrati sull'app nel tempo (aggregazione quindicinale)

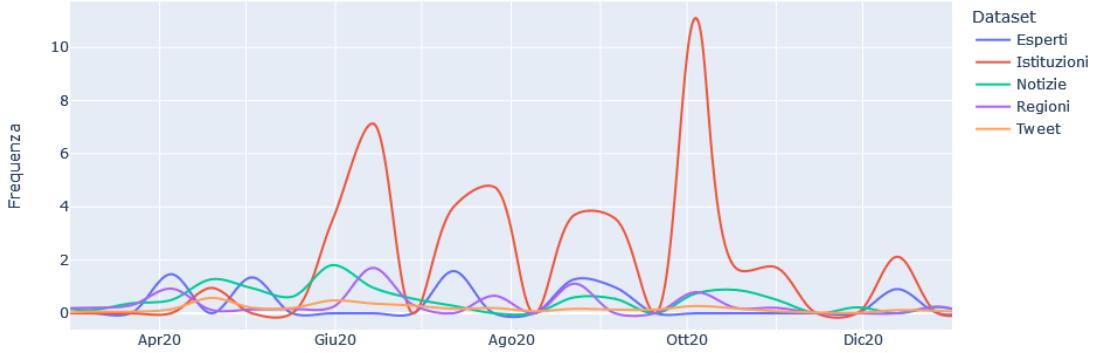


Figura 35: Percentuale di testi relativi l'app Immuni rispetto al totale (aggregazione quindicinale)

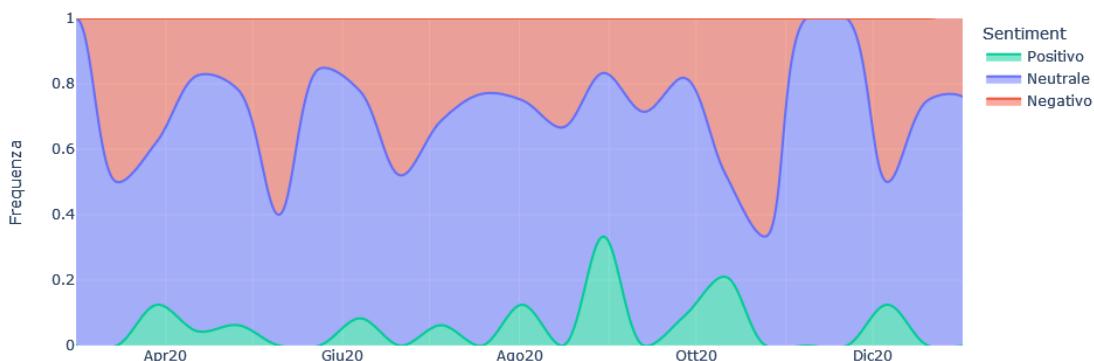


Figura 36: Evoluzione del sentimento verso l'app Immuni nel tempo (aggregazione quindicinale)

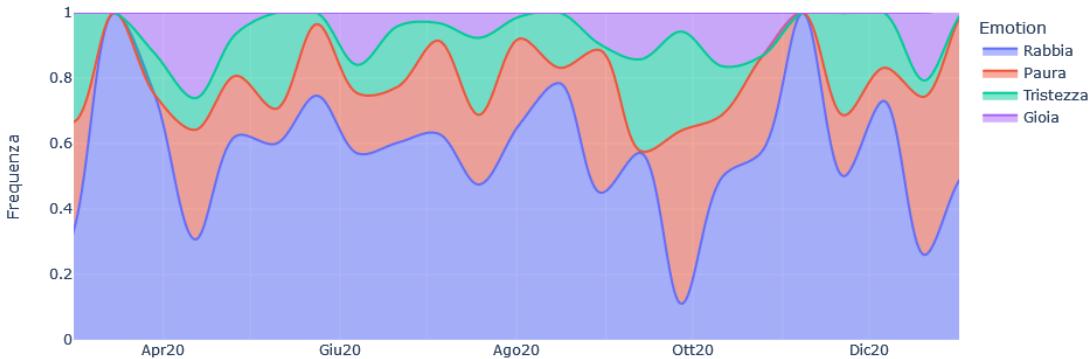


Figura 37: Evoluzione dell'emotion verso l'app Immuni nel tempo (aggregazione quindicinale)



Figura 38: Word cloud per sentimento ed emotion relative all'app Immuni

Fiducia in esperti e istituzioni

In figura 39 possiamo confrontare il sentimento espresso dagli utenti nei confronti degli esperti con quello verso le istituzioni. Si nota come i primi risultino generalmente meno apprezzati, con un livello di sentimento positivo costantemente più basso. Dall'analisi dell'emotion (figura 40) risulta che le istituzioni tendano a generare più rabbia negli utenti rispetto agli esperti, mentre quest'ultimi causino un livello di paura maggiore.

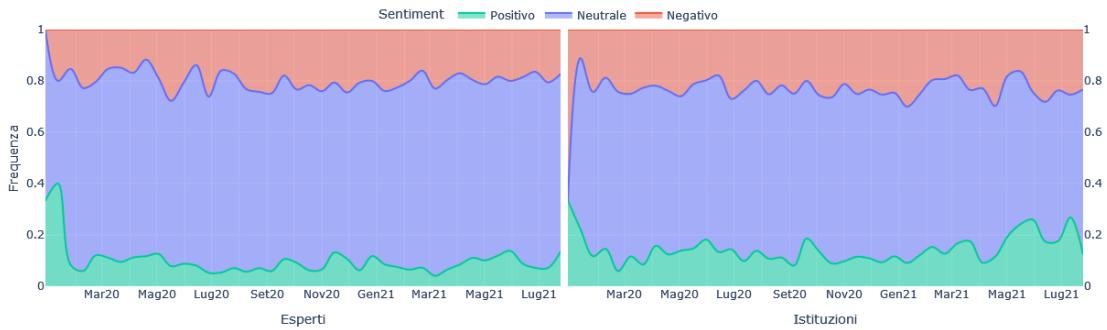


Figura 39: Evoluzione del sentimento delle opinioni nei confronti delle istituzioni e degli esperti (aggregazione quindicinale)

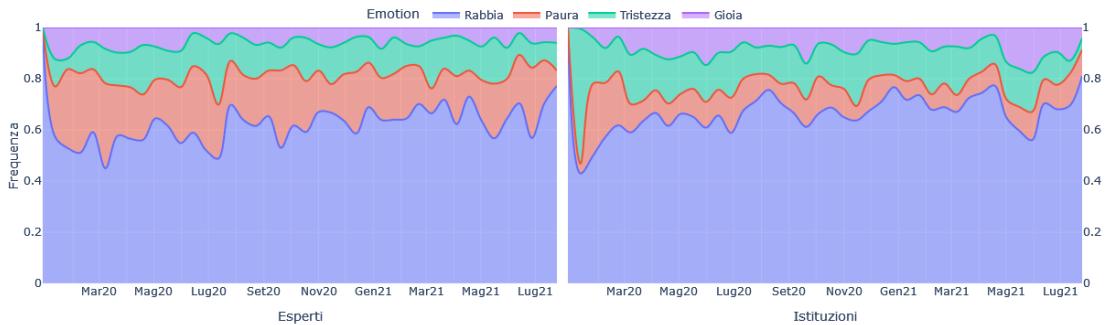


Figura 40: Evoluzione dell'emotion delle opinioni nei confronti delle istituzioni e degli esperti (aggregazione quindicinale)

Efficacia della campagna vaccinale

In figura 41 vediamo l'evoluzione dell'interesse sull'argomento. Il tema delle vaccinazioni è iniziato a essere trattato dagli esperti nel periodo di agosto/settembre 2020 (soprattutto per via delle vaccinazioni antinfluenzali), mentre da dicembre ne hanno parlato con costanza (22% del totale degli articoli). Da parte delle istituzioni si nota una forte spinta nel periodo gennaio/febbraio 2021, subito seguito dal picco delle regioni tra marzo e maggio. L'interesse degli utenti è aumentato progressivamente nel tempo, in linea con la proporzione di notizie relative all'argomento fino a maggio 2021, poi l'interesse degli utenti è esploso raggiungendo circa il 30% di tweet a giugno 2021 e superando il 60% a fine luglio.

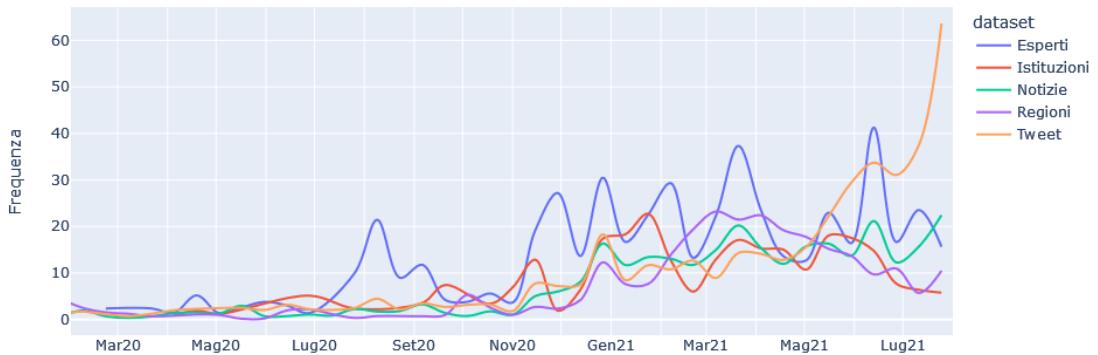


Figura 41: Percentuale di testi relativi le vaccinazioni rispetto al totale (aggregazione quindicinale)

In figura 42 si nota un sentiment maggiormente neutrale per il tema trattato rispetto a quello non filtrato per argomento. Per quanto riguarda l'emotion (figura 43) il livello di rabbia è leggermente superiore, mentre la tristezza è costantemente inferiore a quella globale.

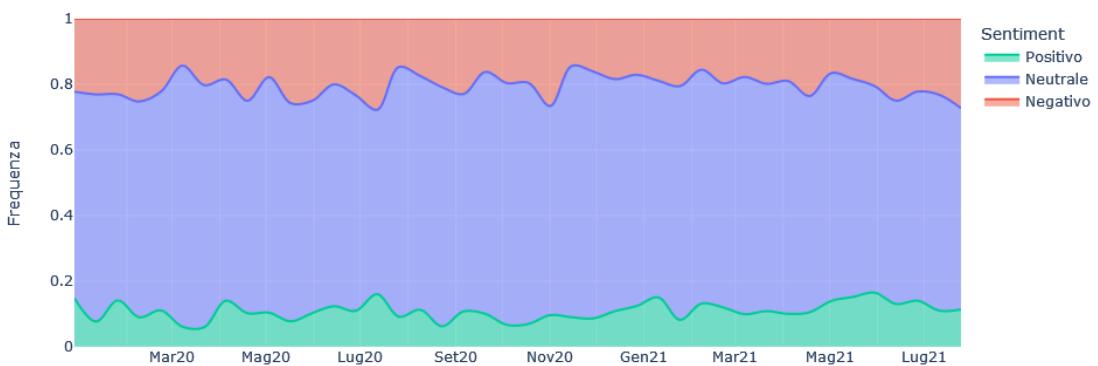


Figura 42: Evoluzione del sentiment verso le vaccinazioni nel tempo (aggregazione quindicinale)

In figura 44 si nota come al sentiment negativo siano associati i termini “inutile” e “pericoloso” relativi al vaccino. Sono presenti opinioni negative anche per termini come “nogreenpass” ma in misura decisamente minore. La paura deriva prevalentemente dal rischio di reazioni avverse al vaccino mentre la rabbia è presente in testi contenenti termini come “dittatura”, “imporre” e “costituzione”.

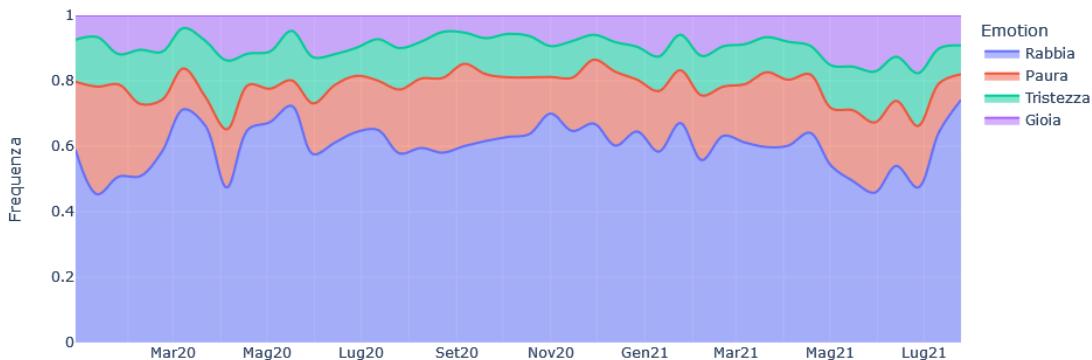


Figura 43: Evoluzione dell'emotion verso le vaccinazioni nel tempo (aggregazione quindicinale)



Figura 44: Word cloud per sentimento ed emotion relative alle vaccinazioni

Conclusioni

Differenze di comunicazione

I risultati riportati mostrano come le diverse fonti abbiano adottato stili comunicativi differenti e li abbiano mantenuti pressoché costanti nel periodo di osservazione.

In particolare è stato evidenziato come i testi provenienti da *Istituzioni*, *Regioni* e *Notizie* fossero sostanzialmente non polarizzati, in netto contrasto con *Esperti* e *Tweet* che hanno espresso opinioni rimaste in prevalenza negative su tutto l'intervallo temporale.

È stata osservata una differenza nelle emozioni rilevate con *Istituzioni*, *Regioni* e *Esperti* dove è prevalse la paura, a differenza di *Tweet* che è stato fortemente caratterizzato dalla rabbia, infine *Notizie* con le distribuzioni all'interno delle quattro categorie studiate più omogenee tra loro. In tutti i casi le emozioni negative sono state preponderanti rispetto a quelle positive.

Un'altra importante differenza risiede negli argomenti trattati: se per *Esperti* e *Notizie* i temi trattati nel tempo siano rimasti piuttosto costanti, invece *Istituzioni*, *Regioni* e in particolare *Tweet* hanno mostrato una maggiore tendenza a cambiare gli argomenti di interesse nel tempo.

Opinione espressa su Twitter

Confrontando i casi studiati dell'app Immuni e della campagna vaccinale si può osservare con chiarezza come i due argomenti siano stati percepiti differentemente, non è stato comunque possibile stabilire la presenza di nessi causali tra l'opinione sul social e l'influenza nella vita reale delle misure adottate.

In generale i commenti polemici verso le decisioni istituzionali si sono rivelati essere più popolari di quelli positivi, ciò potrebbe dipendere dall'effettiva presenza di più utenti contrari alle decisioni prese ma non essendo noto l'algoritmo con cui Twitter classifica i tweet per popolarità non è possibile trarre conclusioni più dettagliate da questo studio. Il risultato ottenuto non è pertanto generalizzabile sull'intera popolazione italiana senza raccogliere ulteriori dati o effettuare altre assunzioni.

Per quanto riguarda le tecniche lo studio effettuato suggerisce che la sola analisi del sentiment non sia sufficiente per analizzare il fenomeno nella sua interezza. Come è stato osservato nell'analisi sulla rabbia generatasi per temi legati alla campagna vaccinale dove erano presenti sia i testi in cui l'emozione era nei confronti delle misure adottate, sia quelli contrari a chi ha protestato contro le stesse misure. Solo osservando più nel dettaglio tramite le word cloud realizzate con il metodo proposto è stato possibile distinguere quale pensiero fosse maggiormente condiviso.

Criticità

Metodi

Per ottenere risultati più accurati su un topic in particolare potrebbe essere necessario utilizzare filtri diversi e più restrittivi già in fase di raccolta dei testi così da comporre un corpus meno generico.

Essendo la diffusione del coronavirus relativamente recente e vista l'assenza di precedenti paragonabili nell'epoca del Web 2.0, potrebbe essere opportuno riaddestrare i modelli NLP utilizzati con nuovi dataset per ottenere risultati più accurati.

Censura

Twitter ha adottato un campagna contro la disinformazione su contenuti relativi al COVID-19, in particolare¹⁰:

Perché un contenuto correlato al COVID-19 possa essere etichettato o rimosso, in base a queste norme, deve:

- presentare un'affermazione di fatto, espressa in termini perentori;
- essere indiscutibilmente falso o ingannevole, sulla base di fonti autorevoli e ampiamente disponibili; e
- essere suscettibile di influire sulla sicurezza pubblica o di causare gravi danni.

Sono comunque consentite opinioni forti o satira, aneddoti personali, dibattiti sulla ricerca relativa al COVID-19.

Non è possibile determinare a priori se alcune opinioni diffuse nella popolazione non siano state evidenziate nelle analisi a causa della rimozione di alcuni tweet secondo queste politiche.

¹⁰<https://help.twitter.com/it/rules-and-policies/medical-misinformation-policy>

Bibliografia

- Bianchi, Federico, Debora Nozza e Dirk Hovy (2021). “FEEL-IT: Emotion and Sentiment Classification for the Italian Language”. In: *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.
- Bojanowski, Piotr et al. (2016). “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606*.
- Boyd, Danah M. e Nicole B. Ellison (ott. 2007). “Social Network Sites: Definition, History, and Scholarship”. In: *Journal of Computer-Mediated Communication* 13.1, pp. 210–230. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2007.00393.x. eprint: <https://academic.oup.com/jcmc/article-pdf/13/1/210/22316979/jjcmcom0210.pdf>.
- De Rosis, Sabina et al. (2021). “The early weeks of the Italian Covid-19 outbreak: sentiment insights from a Twitter analysis”. In: *Health Policy* 125.8, pp. 987–994. ISSN: 0168-8510. DOI: 10.1016/j.healthpol.2021.06.006. URL: <https://www.sciencedirect.com/science/article/pii/S0168851021001627>.
- Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv: 1810.04805 [cs.CL]*.
- Garcia, Klaifer e Lilian Berton (2021). “Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA”. In: *Applied Soft Computing* 101, p. 107057. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2020.107057. URL: <https://www.sciencedirect.com/science/article/pii/S1568494620309959>.
- Honnibal, Matthew et al. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. DOI: 10.5281/zenodo.1212303.
- Hutto, C. e Eric Gilbert (mag. 2014). “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 216–225. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>.
- Melo, Tiago de e Carlos M. S. Figueiredo (2021). “Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach.” In: *JMIR PUBLIC HEALTH AND SURVEILLANCE* 7.2. ISSN: 23692960. DOI: 10.2196/24585. URL: <https://publichealth.jmir.org/2021/2/e24585>.
- Mikolov, Tomas et al. (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *arXiv: 1301.3781 [cs.CL]*.
- Neuraly (2021). *Italian BERT Sentiment model*. URL: <https://huggingface.co/neuraly/bert-base-italian-cased-sentiment>.

- Sainburg, Tim, Leland McInnes e Timothy Q Gentner (2021). “Parametric UMAP embeddings for representation and semi-supervised learning”. In: arXiv: 2009. 12981 [cs.LG].
- We Are Social e Hootsuite (2021). *Digital in Italy: All the Statistics You Need in 2021*. URL: <https://datareportal.com/reports/digital-2021-italy> (visitato il 30/06/2021).
- Yin, Jianhua e Jianyong Wang (2014). “A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering”. In: KDD ’14, pp. 233–242. DOI: 10.1145/2623330.2623715.

Appendice

Tabella 6: Elenco degli esperti considerati nello studio

Nome	Ruoli
Matteo Bassetti	direttore del reparto di Malattie Infettive dell'ospedale San Martino di Genova
Silvio Brusaferro	presidente dell'Istituto Superiore di Sanità
Roberto Burioni	virologo e professore ordinario all'Università San Raffaele
Ilaria Capua	ricercatrice in virologia presso l'Università della Florida
Nino Cartabellotta	presidente della fondazione GIMBE
Andrea Crisanti	microbiologo all'Università di Padova
Massimo Galli	direttore reparto di Malattie Infettive dell'ospedale Luigi Sacco di Milano
Maria Rita Gismondo	virologa presso l'ospedale Luigi Sacco di Milano
Giacomo Gorini	immunologo presso il Jenner Institute dell'Università di Oxford
Giuseppe Ippolito	direttore scientifico dell'Istituto Nazionale per le Malattie Infettive Lazzaro Spallanzani di Roma
Franco Locatelli	Presidente del Consiglio superiore di sanità
Pierluigi Lopalco	professore ordinario di Igiene presso l'Università di Pisa
Giorgio Palù	presidente dell'Agenzia Italiana del Farmaco
Fabrizio Pregliasco	direttore sanitario dell'IRCCS Istituto Ortopedico Galeazzi di Milano
Giovanni Rezza	direttore generale della Prevenzione del Ministero della Salute
Walter Ricciardi	medico igienista e rappresentante italiano presso il consiglio dell'OMS
Pierpaolo Sileri	sottosegretario di Stato al Ministero della salute, viceministro della salute
Antonella Viola	immunologa e docente di Patologia generale presso l'Università di Padova
Alberto Zangrillo	primario e professore ordinario di Anestesiologia e Rianimazione dell'IRCCS Ospedale San Raffaele di Milano

Testo 1: Elenco dei termini utilizzati per cercare i tweet inerenti al coronavirus

prima ondata, seconda ondata, terza ondata,
fase 1, fase 2, fase 3, fase uno, fase due, fase tre,
zona rossa, zona arancione, zona gialla, zona bianca,
covid, coronavirus, virus, pandemia, tamponi,
lockdown, coprifuoco, quarantena, mascherina, mascherine,
variante, varianti, vaccino, vaccini, greenpass, green pass

Codice 1: Espressione regolare case insensitive per filtrare le notizie d'interesse

(prima|seconda|terza|nuova) ondata|tampon|coronavirus|virus|
covid|cts|pandemia|lockdown|coprifuoco|quarantena|mascherin|
variant|vaccin|contagi|zona (rossa|arancione|gialla|bianca)|
green ?pass|fase (1|2|3|uno|due|tre)

Codice 2: Espressione regolare case insensitive per filtrare i tweet d'interesse da parte di regioni e istituzioni

(prima|seconda|terza|nuova) ?ondata|fase ?(1|2|3|uno|due|tre)|
distanziament|virus|coronavirus|covid|pandemi|tampon|lockdown|
coprifuoco|quarantena|pnrr|cts|dpcm|sostegn|epidemi|certifica|
immun|mascherin|variant|zona ?(rossa|arancione|gialla|bianca)|
vaccin|contagi|sintom|green ?pass

Tabella 7: Elenco dei profili Twitter delle regioni italiane (Molise e Sardegna non hanno nessun account)

Regione	Account
Abruzzo	@Regione_Abruzzo
Basilicata	@regbasilicata
Calabria	@La_Calabria
Campania	@Reg_Campania
Emilia Romagna	@RegioneER
Friuli-Venezia Giulia	@regioneFVGit
Lazio	@RegioneLazio
Liguria	@RegLiguria
Lombardia	@RegLombardia
Marche	@RegioneMarcheIT
Molise	
Piemonte	@regionepiemonte
Puglia	@RegionePuglia
Sardegna	
Sicilia	@Regione_Sicilia
Toscana	@regionetoscana
Trentino-Alto Adige	@ProvinciaBZ e @ProvinciaTrento
Umbria	@RegioneUmbria
Valle d'Aosta	@Stampavda
Veneto	@RegioneVeneto

Elenco delle tabelle

1	Variabili del dataset degli esperti	12
2	Variabili del dataset delle notizie	13
3	Variabili del dataset creato	14
4	Variabili dei dataset istituzioni e regioni	15
5	Esempio delle varie fasi di preprocessing	18
6	Elenco degli esperti considerati nello studio	44
7	Elenco dei profili Twitter delle regioni italiane (Molise e Sardegna non hanno nessun account)	46

Elenco delle figure

1	plots the national rescaled daily PO and NO indexes time series during the first four weeks of the Italian coronavirus outbreak. The dashed vertical lines indicate the main government announcements or events. (De Rosis et al. 2021)	8
2	Distribution of themes. UOL: Universo Online. (Melo e Figueiredo 2021)	9
3	Universo Online sentiment analysis over time. (Melo e Figueiredo 2021)	9
4	Twitter sentiment analysis over time. (Melo e Figueiredo 2021) . .	9
5	Portuguese volume variations for topic. The horizontal axis represents the posting dates where each point represents the sum of the messages over a week and the vertical axis the number of posts. The blue lines represent the total messages. The other lines are related to sentiment analysis where red are negative and green are positive. (Garcia e Berton 2021)	10
6	English volume variations for topic. The horizontal axis represents the posting dates where each point represents the sum of the messages over a week and the vertical axis the number of posts. The blue lines represent the total messages. The other lines are related to sentiment analysis where red are negative and green are positive. (Garcia e Berton 2021)	11
7	Distribuzione del numero di articoli pubblicati nel tempo per ogni esperto (aggregazione mensile)	13
8	Percentuale degli articoli pubblicati relativi al covid rispetto al totale nel tempo (aggregazione settimanale)	14
9	Distribuzione delle applicazioni utilizzate per la pubblicazione di tweet	15
10	Numero di tweet relativi al covid pubblicati per canale istituzionale (aggregazione mensile)	16
11	Numero di tweet relativi al covid pubblicati per account regionale (aggregazione mensile)	17
12	Distribuzione dei testi nello spazio bidimensionale per ogni fonte .	19
13	Aggregazione dei cluster ottenuti tramite GSDMM	20
14	Word cloud dei 24 cluster	21
15	Distribuzione dei cluster nello spazio 2d con regione di confidenza al 90%	22

16	Distribuzione dei sentiment e emozioni nello spazio 2d con regione di confidenza al 99%	23
17	Evoluzione dei topic sul dataset notizie nel tempo (aggregazione quindicinale)	24
18	Evoluzione del sentiment sul dataset notizie nel tempo (aggregazione settimanale)	25
19	Evoluzione dell'emotion sul dataset notizie nel tempo (aggregazione settimanale)	25
20	Evoluzione dei topic sul dataset tweet nel tempo (aggregazione quindicinale)	26
21	Evoluzione del sentiment sul dataset tweet nel tempo (aggregazione settimanale)	26
22	Evoluzione dell'emotion sul dataset tweet nel tempo (aggregazione settimanale)	27
23	Evoluzione dei topic sul dataset esperti nel tempo (aggregazione mensile)	27
24	Evoluzione del sentiment sul dataset esperti nel tempo (aggregazione quindicinale)	28
25	Evoluzione dell'emotion sul dataset esperti nel tempo (aggregazione quindicinale)	28
26	Evoluzione dei topic sul dataset istituzioni nel tempo (aggregazione mensile)	29
27	Evoluzione del sentiment sul dataset istituzioni nel tempo (aggregazione quindicinale)	29
28	Evoluzione dell'emotion sul dataset istituzioni nel tempo (aggregazione quindicinale)	29
29	Evoluzione dei topic sul dataset regioni nel tempo (aggregazione quindicinale)	30
30	Evoluzione del sentiment sul dataset regioni nel tempo (aggregazione settimanale)	31
31	Evoluzione dell'emotion sul dataset regioni nel tempo (aggregazione settimanale)	31
32	Numero di ricoveri ordinari e in terapia intensiva nel tempo (migliaia di persone per settimana)	33
33	Word cloud per emotion durante le varie fasi pandemiche	34
34	Confronto tra numero di persone positive in Italia e positivi registrati sull'app nel tempo (aggregazione quindicinale)	35
35	Percentuale di testi relativi l'app Immuni rispetto al totale (aggregazione quindicinale)	35
36	Evoluzione del sentiment verso l'app Immuni nel tempo (aggregazione quindicinale)	35
37	Evoluzione dell'emotion verso l'app Immuni nel tempo (aggregazione quindicinale)	36
38	Word cloud per sentiment ed emotion relative all'app Immuni	36

39	Evoluzione del sentimento delle opinioni nei confronti delle istituzioni e degli esperti (aggregazione quindicinale)	37
40	Evoluzione dell'emotion delle opinioni nei confronti delle istituzioni e degli esperti (aggregazione quindicinale)	37
41	Percentuale di testi relativi le vaccinazioni rispetto al totale (aggregazione quindicinale)	38
42	Evoluzione del sentimento verso le vaccinazioni nel tempo (aggregazione quindicinale)	38
43	Evoluzione dell'emotion verso le vaccinazioni nel tempo (aggregazione quindicinale)	39
44	Word cloud per sentimento ed emotion relative alle vaccinazioni . .	39