Università degli studi di Milano-Bicocca

Social Media Analytics
Progetto finale

Super green pass: la reazione della comunità italiana su Twitter

Autori:
Armanini Justin, 830103, j.armanini@campus.unimib.it
Confalonieri Riccardo, 830404, r.confalonieri5@campus.unimib.it
Cormio Chiara, 868708, c.cormio@campus.unimib.it

A.A.: 2021/2022

Indice

1 Introduzione **2**
Note implementative 2

2 Dataset **2**
2.1 Preprocessing 4

3 Social network analysis **4**
3.1 Hub analysis 7
3.2 Community detection 8
3.2.1 Analisi semantica delle community 10

4 Social content analysis **11**
4.1 Sentiment analysis 12
4.2 Emotion Recognition 16

5 Conclusioni **18**

Sommario

L'introduzione del Super Green pass, avvenuta il 6 Dicembre 2021, è stata oggetto di un acceso dibattito in Italia. Questa nuova misura ha creato umori e idee molto diverse nella popolazione: da un lato i cittadini da tempo vaccinati hanno accettato con favore la misura, dall'altro invece i negazionisti del vaccino e del Green Pass si sono detti fin da subito contrari per le più svariate motivazioni. Il progetto si è posto dunque l'obiettivo di analizzare i pensieri e l'interazione degli utenti italiani attraverso i post social espressi tramite la piattaforma Twitter proprio nel giorno dell'entrata in vigore delle nuove misure. In particolare si vogliono ricercare le figure di riferimento presenti nella rete, ma anche il sentimento e le emozioni delle persone rispetto alle nuove limitazioni.

Keywords: twitter - sentiment - social network analysis - Super green pass

1 Introduzione

Dopo due anni di convivenza con il COVID-19, e ad un anno dall'inizio della campagna vaccinale, ogni cittadino ha sviluppato una propria opinione su questi temi, improvvisandosi all'occorrenza virologo, epidemiologo o medico.
I Social Network sono diventati quindi uno spazio dove poter esprimere le proprie opinioni, siano esse frutto di ricerche fondate o semplicemente delle supposizioni personali.
Uno dei temi più caldi in Italia negli ultimi mesi riguarda l'introduzione del Super Green Pass per accedere a locali al chiuso, eventi, palestre e luoghi di cultura. Questa misura di sicurezza è entrata in vigore il 6 dicembre 2021 e di fatto non permette più l'accesso a diversi servizi a coloro che non hanno completato il ciclo vaccinale.
Per questi motivi si è scelto, per lo sviluppo del progetto, di analizzare le impressioni della comunità italiana di Twitter riguardo il Super Green Pass, con l'obiettivo di capire chi sono gli utenti che scrivono e interagiscono sulla piattaforma e se esistono *community* specifiche tra gli utenti che hanno commentato l'introduzione di questa restrizione. Si vuole inoltre analizzare i possibili nodi influenti della rete e verificare il *Sentiment* attorno a questi.

Note implementative

Per svolgere il lavoro è stato utilizzato come linguaggio principale Python 3.6 [1] e in particolare le librerie Tweepy [2] per l'interfacciamento con Twitter, Networkx [3] e Pyvis [4] per la social network analysis, il lexicon NRC [5] e la libreria Sentita [6] per la emotion recognition e la sentiment analysis in Italiano.

2 Dataset

Per lo sviluppo del progetto si è utilizzata la piattaforma social Twitter dalla quale sono stati raccolti 10000 tweet pubblicati da diversi utenti il giorno 6 dicembre 2021, data dell'entrata in vigore del Super Green pass. Per la raccolta ci si è avvalsi della API proprietaria di Twitter e della libreria Tweepy che permette di interfacciarsi con la API. Tra le diverse librerie esistenti si è scelto di utilizzare proprio Tweepy in quanto è risultata essere la più aggiornata. Inoltre, consente di scaricare il testo completo di ogni post anche quando superiore a 140 caratteri, ovvero il limite inizialmente introdotto in Twitter ed oggi superato. Tale libreria permette di scaricare i dati secondo diversi parametri di ricerca, ma per questo progetto si è scelto di ricercare i tweet che rispettano i seguenti paramenti:

- **Lingua italiana**: dato l'interesse ad approfondire le reazione del popolo italiano si sono filtrati i tweet nella sola lingua nazionale. Questa scelta innanzitutto ha permesso di restringere la raccolta dei tweet ai soli utenti potenzialmente coinvolti nel fatto di interesse, quelli italiani per l'appunto, e in secondo luogo, dal punto di vista implementativo, ha permesso di adottare un'unica procedura per la Sentiment Analysis. Quest'ultimo, infatti, è un tipo di analisi fortemente dipendente dalla lingua utilizzata, e che quindi avrebbe richiesto una soluzione dedicata per ogni idioma rappresentato. In letteratura è risaputo che per la Sentiment Analysis multilingua sia molto difficile garantire la consistenza dei risultati per le diverse lingue.

- **Data di pubblicazione**: si è scelto di scaricare soltanto i tweet del 6 dicembre 2021. Questo per poter raccogliere le impressioni a caldo sull'entrata in vigore del Super Green Pass. Per motivi computazionali si è posto un limite di tweet da scaricare (10000), anche se da una successiva verifica è emerso che si è riusciti a scaricare tutti i tweet della giornata scelta.

- **Hashtag**: si è scelto di filtrare tra tutti i tweet esistenti quelli che contenessero come hashtag uno o più dei seguenti:
  - #supergreenpass
  - #greenpass
  - #novax
  - #nogreenpass
  - #greenpassrafforzato
  - #vaccino

Inoltre si è scelto di includere non soltanto i tweet originali, ma anche i retweet Questo è stato fondamentale per analizzare le interazioni tra i diversi utenti. Al termine dello scaricamento si è dunque ottenuto un dataset con 10000 tweet complessivi ognuno con le seguenti feature:

- date: giorno e ora della pubblicazione del tweet.
- id: identificativo univoco del tweet.
- text: contenuto testuale del tweet.
- n_rt: numero di retweet ricevuti dal tweet.
- n_like: numero di like ricevuti dal tweet.
- author_name: username dell'utente che ha postato il tweet. Nel caso sia un retweet riporta il nome dell'utente che sta retweettando il post e non il nome dell'utente originale.
- author_id: identificativo univoco dell'utente.
- location: se disponibile, posizione geografica di dove è stato pubblicato il tweet.
- rt_authors: contiene una lista in cui vengono salvati tutti i nomi degli utenti retweettati e/o menzionati nel post in questione. Nel caso in cui il post sia originale e non contenga menzioni questo campo sarà una lista vuota. La feature è modellata con un dizionario contenente per ciascun utente le seguenti informazioni:
  - **screen_name**: lo username del proprio profilo twitter, quello che compare in caso di menzione o retweet.
  - **name**: il nome che compare sul proprio profilo twitter.
  - **id**: identificativo unico dell'utente.
  - **id_str**: identificativo unico dell'utente in formato stringa.
- hashtag: lista degli hashtag presenti nel tweet.
- author_follower: numero di follower dell'autore del post.
- author_friends: numero di persone seguite dall'autore del post.

2.1 Preprocessing

I tweet sono documenti testuali e in quanto tali necessitano una fase di preprocessing al fine di eliminare tutti gli elementi di rumore che potrebbero interferire con i task successivi. Inizialmente si sono convertiti tutti i testi dei tweet in *lowercase* per facilitare gli step successivi. Si sono dunque individuati e rimossi i seguenti elementi tramite espressioni regolari:

- emoji
- menzioni: le menzioni presenti nel tweet di riferimento sono contenute nel dataset di partenza.
- hashtag: gli hashtag presenti nel tweet sono contenuti nel dataset di partenza.
- url
- spazi bianchi extra

Si è poi proseguito a pulire la feature *rt_authors* dalle informazioni non utili ai fini del progetto. Si è scelto di mantenere come informazione solo lo *screen_name*.
Si evidenzia che non sono state rimosse la punteggiatura e le stopword (le parole più comuni) nonostante sia solitamente buona pratica farlo. Questo perché Sentita è una rete neurale che è stata addestrata su testi italiani che non hanno subito importanti operazioni di preprocessing. Seguendo lo stesso ragionamento non sono state svolte operazioni di *stemming* o *lemmatization*.

3 Social network analysis

Il primo task prevede la costruzione di una rete sociale costruita attraverso le menzioni all'interno dei tweet scaricati. In questo caso con la parola menzione ci si sta riferendo sia all'utente retweettato sia ad eventuali utenti menzionati a parte attraverso l'utilizzo del simbolo chiave @ seguito da uno screen_name. A partire dal dataset scaricato si è dunque creata una rete in cui ogni utente (u1) avrà un arco diretto verso un altro utente (u2) se il primo ha menzionato in almeno un post il secondo (u2). Il peso degli archi corrisponderà invece al numero di interazioni totali tra i due nodi. Si è così ottenuto un grafo *pesato e diretto* con 5272 nodi e 9681 archi avente le seguenti caratteristiche:

Metrica	**Valore**
Diametro	9
Degree media	1.97
Densità del grafo	0.00034
Node connectivity	0
Edge connectivity	0
Assortativity Degree	–0.1347
Overall_reciprocity	0.008

Tabella 1: Metriche del grafo

Il valore di assortativity negativo indica che i singoli hub tendono a connettersi ai nodi di grado minore senza connettersi tra loro, creando così una rete *disassortativa*. Inoltre il valore di reciprocità indica che in generale tutti gli utenti tendono a retweettarsi poco tra loro, questo influisce anche sulla densità del grafo che risulta essere molto bassa. Infine il valore di connectivity indica che il *grafo è disconnesso*. Tutto questo porta a pensare che la

rete sia di tipo ego-centrico con la presenza di diversi hub che agglomerano la maggior parte dei retweet degli altri utenti, generando cosi il fenomeno noto come 'preferential attachment'. Queste metriche trovano conferma anche dai grafici riportati in Figura 1 e 2, che riportano rispettivamente il numero di retweet ricevuti ed effettuati dagli utenti. Da queste immagini si evince che la maggior parte degli utenti ha effettuato un solo retweet durante tutta la giornata ricevendone zero.
A partire da queste analisi si è cercato di individuare i possibili hub della rete, per farlo è stato esteso il range nel grafico riportante il numero di retweet ricevuti, e analizzando la figura 3 è emerso che pochissimi utenti hanno ricevuto più di 50 retweet. Per questa motivazione è stata utilizzata la metrica dell'*in_degree* per evidenziare graficamente i possibili hub della rete.

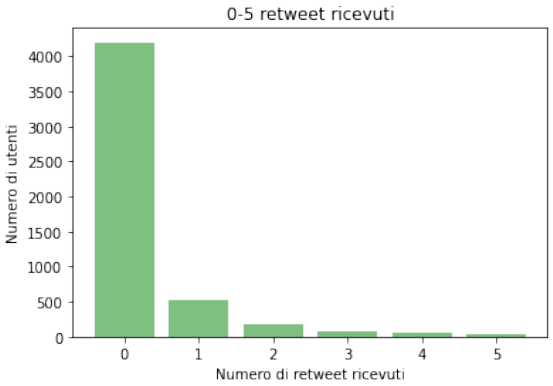

Figura 1: Numero di retweet ricevuti dagli utenti nel range $[0, 5]$

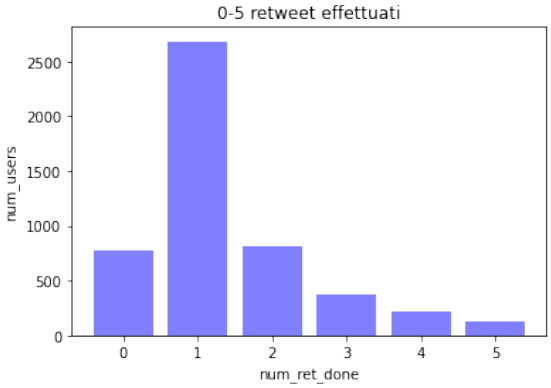

Figura 2: Numero di retweet effettuati dagli utenti nel range $[0, 5]$

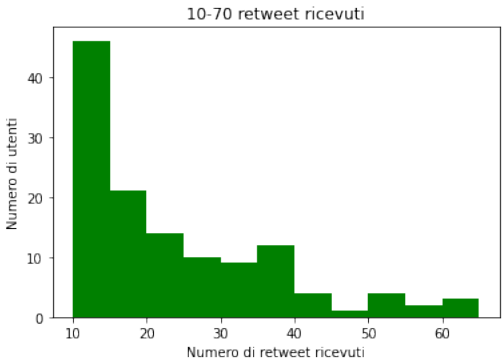

Figura 3: Numero di retweet ricevuti dagli utenti nel range $[10, 70]$

Successivamente a questa prima analisi si è proceduto con la visualizzazione del grafo attraverso la libreria Pyvis che permette di ottenere un grafo interattivo in formato .html. Sfruttando quanto osservato dalle prime analisi è stata aumentata la dimensione degli hub, ovvero i nodi con in_degree ≥ 50. Inoltre, per semplificare la visualizzazione, tutte le label dei nodi sono state nascoste e vengono visualizzate soltanto se il mouse viene posizionato sopra uno specifico nodo, le uniche label visualizzate sono quelle degli hub anche se dalla figura 4 non si riesce ad apprezzare dato il forte zoom out. Infine il layout scelto per la visualizzazione è il 'forceAtlas2Based' che respinge i nodi tra loro mentre gli archi attraggono i nodi, questo permette di individuare delle possibili community facilitando i passi successivi.

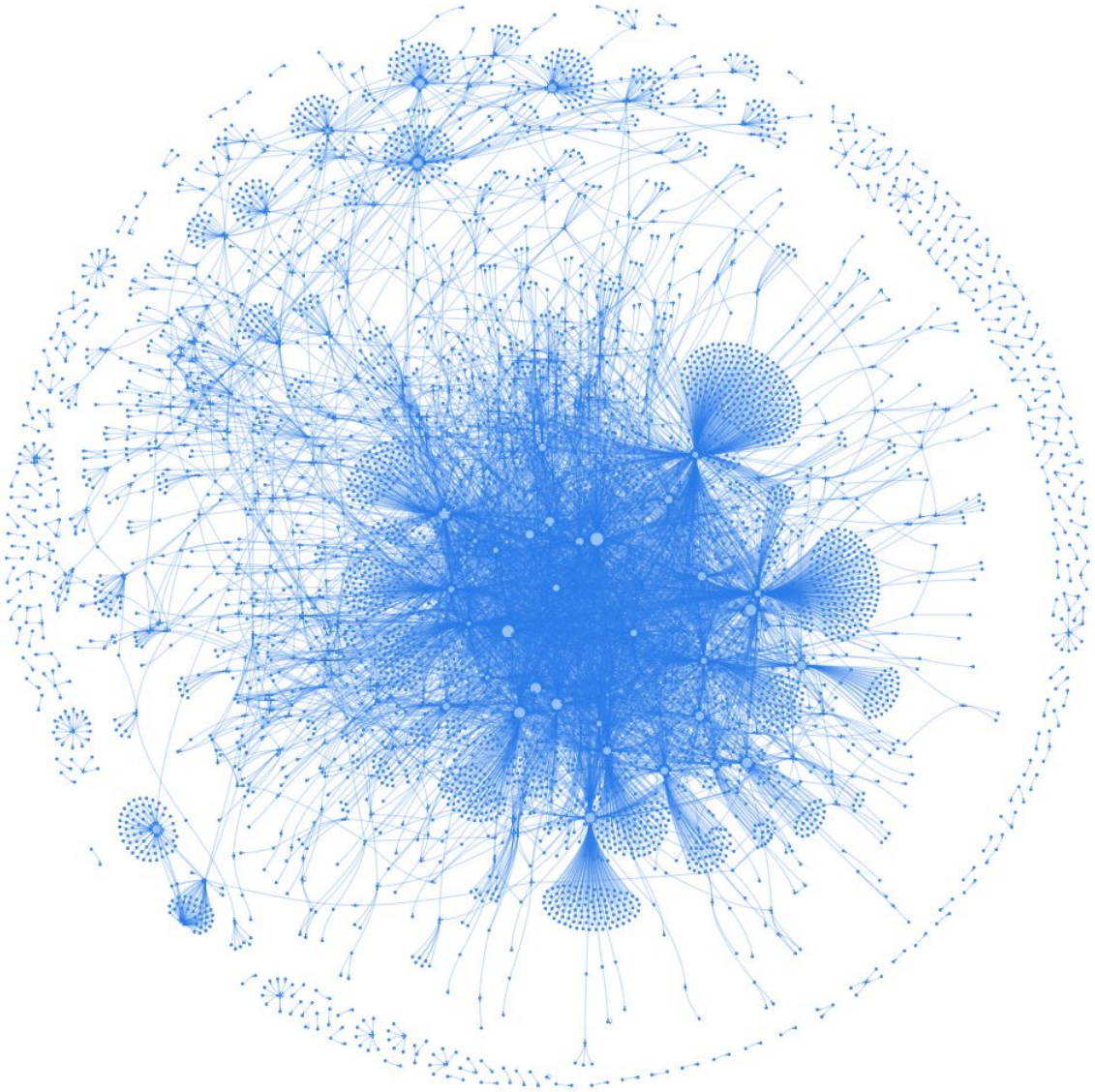

Figura 4: Visualizzazione del grafo

Osservando più nel dettaglio la figura si è inoltre giunti a diverse considerazioni:

- conferma visiva del fatto che il grafo sia disconnesso. Si notano infatti numerosi nodi nel cerchio più esterno che non sono collegati agli altri nodi.
- sembrano essere presenti diverse community di dimensioni diverse. La maggior parte di esse sembrano comunque attaccarsi ad uno degli hub individuati come accade nelle strutture a mongolfiera che vengono a crearsi al centro del grafo.
- gli hub non sembrano far riferimento a personaggi noti e/o famosi.

3.1 Hub analysis

Come anticipato nessuno dei nomi riportati come hub del grafo sembra essere famoso o comunque conosciuto ai più. In tal senso è stata condotta una verifica manuale sulle pagine twitter di questi utenti dalla quale è emersa la presenza, tra tutte, delle seguenti persone:

Nome utente	**Descrizione**
Corpodelledonne	Account di Lorella Zanardo, attivista e scrittrice.
fratotolo2	Scrittrice e collaboratrice de 'Il Primato Nazionale'
robertalerici	Scrittrice ed attrice
ladyonorato	Account di Francesca Donato, membro del parlamento europeo.
TgrRaiLazio	Account ufficiale della Testata Giornalistica Regionale Rai Lazio.
AlexBazzaro	Deputato della Lega e consigliere comunale a Venezia.
chiaralucetw	Account di Carlotta ChiaraLuce, politica.
amodeomatrix	Account di Francesco Amodeo, giornalista d'inchiesta.
NicolaPorro	Account del vicedirettore del ilgiornale.
CislNazionale	Account del sindacato dei lavoratori

Tabella 2: Personaggi famosi presenti come hub del grafo

Oltre a questi account è emersa la presenza di 'GiocatricePerTE', appartenente ad un medico chirurgo no-vax, e di 'RomaNoGreenPass', appartenente ad un gruppo di studenti universitari contro il Green pass.
Durante questa verifica si è notato come la maggior parte di questi utenti abbia un numero abbastanza esiguo di follower. Da ciò si può concludere che per questa particolare tematica non conta tanto essere famosi, piuttosto i contenuti e le ideologie che si cerca di portare avanti con i propri post. A tal proposito, osservando le Tabelle 3 e 4, le quali riportano rispettivamente gli utenti con più follower e gli hub, è possibile notare come nella prima tabella, ad eccezione di 'NicolaPorro' e 'ByoBlu', non siano presenti hub benché gli utenti siano molto conosciuti e con tanti follower.

Nome utente	Num. follower
sole24ore	1595551
MediasetTgcom24	1198635
RaiNews	1132252
rtl1025	878548
lercionotizie	848316
CorSport	728477
TgLa7	701123
ilgiornale	547311
Adnkronos	542252
La7tv	541627
NicolaPorro	431457
fanpage	369278
Libero_official	300032
Capezzone	135725
borghi_claudio	122162
OttoemezzoTW	112446
byoblu	101944
Tg1Rai	98169
paoloigna1	93424
MassimGiannini	93302
a_meluzzi	84599
tempoweb	84597
SusannaCamusso	81692
GiovanniToti	76851

Tabella 3: Nodi del grafo con più followers

Nome utente	Num. follower
NicolaPorro	431457
byoblu	101944
mgmaglie	69536
ProfCampagna	63862
ladyonorato	52726
fdragoni	44052
FmMosca	43781
valy_s	40079
CislNazionale	28858
BarbaraRaval	20745
chiaralucetw	10579
Solocarmen1	10358
fratotolo2	9864
siriomerenda	9441
RomaNoGreenPass	7945
strange_days_82	7832
AntonelloZedda	6719
OrtigiaP	6461
viaggrego	6301
Lorenzo62752880	4596
giopge	3594
Zippo88lrr	2738
AlexBazzaro	2625
CianPdc	2610

Tabella 4: Followers degli hub del grafo

3.2 Community detection

I social, come Twitter, hanno consentito alle persone di tutto il mondo di interagire tra loro e costruire relazioni con altri individui con cui condividono interessi comuni. Questo può essere osservato anche nella vita reale: naturalmente, tendiamo a sviluppare e mantenere relazioni con altre persone simili a noi. Le persone con interessi simili tendono infatti a gravitare l'una verso l'altra e ad associarsi in comunità o gruppi di persone che condividono tratti simili tra loro. Questo concetto può essere applicato anche alle reti complesse nelle quali si cercano quegli insiemi di nodi, detti comunità, che rispettino delle determinate caratteristiche a livello di connessioni o struttura.
In tal senso, in questo progetto è stato implementato l'algoritmo di *massimizzazione della modularità* al fine di individuare le community presenti. Questo tipo di algoritmo fa parte della famiglia dei network-centric che cercano di trovare delle comunità considerando l'insieme globale di connessioni e dividendo la rete in sottoinsiemi disgiunti, quindi senza overlapping.
Il concetto di modularità si basa sulle interazioni dei nodi della rete: dato un gruppo la sua modularità Q è la differenza tra il numero di archi effettivi all'interno del gruppo e il numero previsto di archi all'interno del gruppo. Questo valore viene poi normalizzato per restituire un valore compreso tra $[-1, 1]$. Quanto più positivo è il valore di Q tanto più significativo è il raggruppamento, ne segue che considerando l'intero grafico come un'unica comunità si ottiene $Q = 0$. Il numero previsto di archi all'interno di un gruppo è calcolato sulla base dell'ipotesi nulla che gli archi siano formati casualmente, cioè non esista alcuna struttura.

L'obiettivo è quindi massimizzare la distanza dalla rete randomica.
Dall'applicazione di questo algoritmo sono risultate 204 community con uno score medio di modularità pari a 0.63. La Figura 5 mostra la rappresentazione grafica della community detection. Sono state evidenziate le community che al loro interno contengono almeno un hub di quelli citati nel corso dell'analisi. Come ci si poteva aspettare molti cluster fanno proprio riferimento agli hub individuati. Alcuni hub, come 'CislNazionale' (riportato in alto a sinistra in fucsia) creano delle community quasi completamente isolate dal resto della rete. Mentre molti hub tendono a creare delle community condivise con almeno un altro hub. Si può notare nella parte centrale del grafo, ad esempio per la community riportata in verde che include al suo interno due hub.

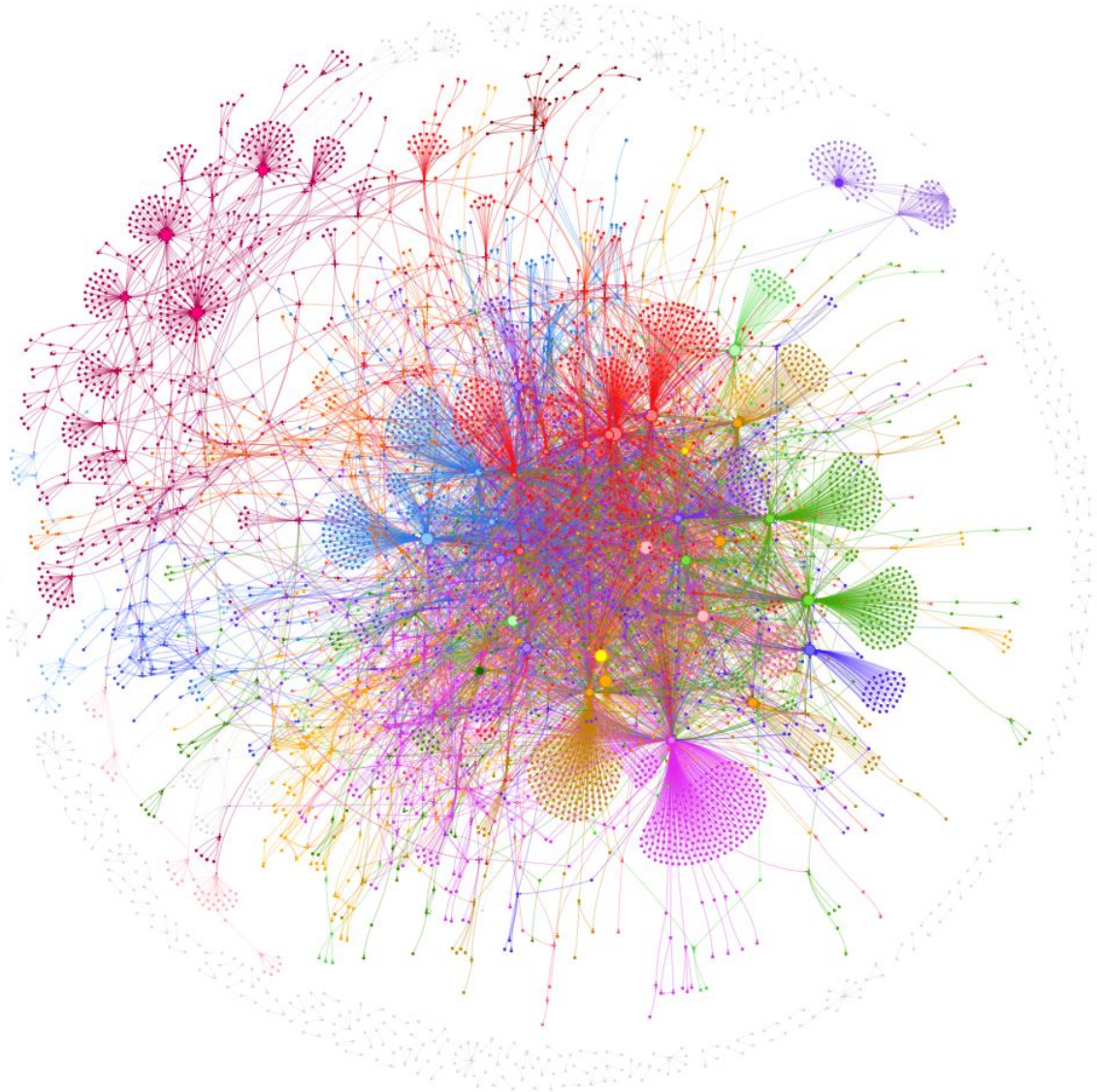

Figura 5: Visualizzazione delle community

3.2.1 Analisi semantica delle community

Analizzare le community graficamente risulta essere abbastanza complesso. Si è quindi provato ad analizzare il contenuto semantico dei tweet delle diverse community attraverso delle wordcloud al fine di individuare eventuali differenze nei contenuti. Durante la creazione delle wordcloud si è deciso di eliminare le stopwords per migliorare la visualizzazione ed evitare di stampare parole poco significative (ad esempio gli articoli). Si ricorda in tal senso che le stopwords saranno importanti per le successive analisi sulla sentiment e per questo non erano già state eliminate in fase di preprocessing.
La figura 6 riporta le wordcloud estratte per le community degli hub: si può notare che le tematiche variano leggermente nelle diverse community, ma esistono molti aspetti comuni. É possibile però individuare delle differenze, ad esempio le community $24, 26$ e 28 si concentrano sul vaccino ai bambini, la community 8 è chiaramente quella del medico chirurgo novax mentre altre community si concentrano su problematiche diverse come la richiesta di esibire il Green pass sui mezzi pubblici. Infine si segnala la community 36 che discute sul numero di Green pass scaricati e si differenzia dalle altre in quanto contiene tweet di approvazione rispetto al Super green pass, è quindi un gruppo pro-vax.

Figura 6: Wordclouds delle community contenenti gli hub

Infine si sono analizzate anche le community formate da pochi utenti. In questo caso è emerso che i contenuti siano principalmente no-vax con argomenti più specifici che probabilmente interessano poche persone e per questo sono stati isolati dagli altri utenti come ad esempio:

```
Ma davvero la "nostra storia" deve finire in questa maniera?
↪ #NoAlPaseSanitario #nogreenpass

Sembra che, in Italia,  un 85% di super mega iper immunizzati, abbia
↪ una paura folle di un 15% (inclusi i bambini...) di non vaccinati.
↪ Ditemi che è un film di fantascienza di serie B.
```

Si è inoltre notata una piccola presenza di community pro-vax, ad esempio nel gruppo 98 e 122, che trattano tematiche più generali come:

```
Oggi ho incontrato un #novax in coda in farmacia. Abbiamo provato
↪ civilmente a confrontarci. Poi lui ha cominciato a motivare le sue
↪ scelte in greco antico. Ho lasciato perdere.

non ne posso più di chi rifiuta il vaccino: andare in farmacia è
↪ diventato impossibile. Code per chi ha diritto a curarsi a causa
↪ di chi se ne frega degli altri. Per non parlare degli ospedali.

Mi stavo domandando, ma tutti quelli #novax, cosa vorrebbero che si
↪ facesse per uscire da questa pandemia? Fingere che non esista e
↪ dare il via libera a tutti? Sono curioso.
```

4 Social content analysis

Una volta studiato il grafo delle interazioni si è voluto approfondire e migliorare la comprensione del fenomeno sul piano semantico. Per questa ragione il secondo task prevede l'estrazione e l'analisi del contenuto dei tweet distribuiti nei diversi cluster.
Sin dalla loro nascita, infatti, i Social Network hanno rappresentato un mezzo attraverso cui gli utenti possono esprimere le proprie emozioni e opinioni in diversi formati: testo, video o audio. Tali dimensioni sono particolarmente informative perché sono il risultato dell'interpretazione soggettiva che l'individuo trae dalla propria esperienza del mondo esterno, dagli input che da esso riceve. In questo modo egli comunica in maniera implicita, e in alcuni casi inconsapevole, informazioni sul suo benessere, stato d'animo, addirittura qualità della vita, così come anche il grado di apprezzamento di un determinato bene o servizio, oppure ancora l'avversione o meno nei confronti di una determinata persona (un personaggio famoso, un politico per esempio) o più in generale rispetto ad un certo evento.
Tuttavia, data la natura del dato, intrinsecamente non strutturato, ecco emergere il problema di come riuscire ad estrarre conoscenza che possa essere sfruttata nel processo di decision-making. Proprio in questa cornice si colloca la Social content analysis.
In questo progetto ci si è quindi occupati di interpretare il linguaggio naturale di cui sono costituti i tweet rispetto a due dimensioni:

- Polarizzazione (detta anche sentiment), che può essere positiva, negativa o neutrale;
- Emozioni, secondo il modello di Plutchik [7]

Non è stato effettuato nessun task di Named Entity Recognition perché, data la natura del dato, non ci si aspettavano entità rilevanti o che dovessero essere disambiguate.

4.1 Sentiment analysis

L'attività di Sentiment analysis consiste nell'inferire il valore di polarizzazione di un contenuto testuale (che sia un tweet come nel caso specifico, post o recensione) soggettivo, ovvero che esprime un'opinione personale, un pensiero o una convinzione. Tale polarizzazione può essere positiva, negativa o neutrale, dove quest'ultima, essendo difficile da definire, viene interpretata come "ciò che non è né positivo né negativo".
Per la realizzazione di questo task, l'approccio supervisionato è stato scartato per due ragioni principali: la prima è che non era disponibile nessun dataset etichettato compatibile con i dati a disposizione, la seconda è che la costruzione di un dataset tramite etichettatura manuale, da poter dividere in train e test, è stata reputata poco praticabile in termini di tempo e risorse umane. Alla luce di queste considerazioni si è optato per un approccio non supervisionato. Sono state comparate diverse soluzioni, sia neurali, tra cui Sentita [6] e FEEL-IT [8], sia lexicon-based, come Sentix [9], Vader [10] e NRC [5]. Fissato un campione di 100 tweet estratti dal dataset, per ciascuno di essi sono state applicate tutte le soluzioni appena menzionate. La scelta finale è ricaduta su Sentita, ovvero il metodo che tra tutti ha prodotto risultati migliori rispetto alla valutazione di un revisore umano, in altri termini, quello che nel maggior numero di casi, ha prodotti valori di sentiment coerenti con quelli che il revisore stesso avrebbe attribuito.
Sentita è una rete LSTM-CNN che, a partire dalla frase che si desidera valutare, riceve in input ciascuna parola rappresentata in forma di word embeddings e restituisce due valori compresi tra $[0, +1]$, uno che rappresenta il punteggio di polarità positivo e uno quello negativo. Questo modello è stato addestrato su circa 100mila esempi ed è possibile utilizzarlo attraverso l'omonima libreria Python che si occupa di preprocessare opportunamente il testo, rappresentare le parole in word embeddings, ed infine restituire l'output del modello. Il preprocessing si limita ad eliminare la punteggiatura ed estrarre gli unigrammi, motivo per cui questa fase non è stata effettuata in fase di preprocessing di questo progetto, ma delegata per intero al package stesso. Inoltre sono state mantenute le stopwords.

Tweet	Score positivo	Score negativo
polizza per danni collaterali da proposta in azienda. miocardite, infarto, paresi, ictus, neoplasia maligna... ma come? non era sicuro?? alla faccia dei complottisti!!	0.07	0.73

Tabella 5: Esempio di score prodotti da Sentita.

Per ciascun tweet è stato ottenuto quindi uno score positivo e uno negativo. Allo scopo di ottenere un'unica metrica che potesse riassumere entrambi e rappresentare quindi la polarità del tweet, è stato seguito il metodo proposto da Basile e Nissim [9] e Magnini et al. [11]. La differenza è che, mentre nei casi riportati la procedura si applica ai singoli synset, nel progetto è stata applicata ai tweet, in quanto gli score di sentiment positiva e negativa erano associati proprio ad essi nella loro interezza anziché alle singole parole. Siccome gli score ottenuti con Sentita rispettano la proprietà $score_{positivo} + score_{negativo} \leq 1$, che è la stessa valida per i synset (terms) di SentiWordNet, questa scelta risulta consistente sul piano teorico.
L'intuizione è la seguente: ciascun tweet può essere rappresentato come vettore in uno spazio bidimensionale che ha come coordinate $(score_{positivo}, score_{negativo})$.

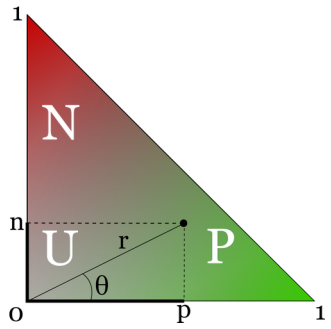

Figura 7: Rappresentazione dei tweet nello spazio vettoriale.

Passando al sistema di coordinate polari si ricavano angolo e modulo del vettore:

$$\theta = \arctan\left(\frac{score_{negativo}}{score_{positivo}}\right)$$

$$r = \sqrt{score^2_{negativo} + score^2_{positivo}}$$

Quindi polarity e intensity vengono definite come:

$$polarity = 1 - \frac{4\theta}{\pi} \quad \in [-1, +1] \begin{cases} -1, & \text{completamente negativo} \\ +1, & \text{completamente positivo} \end{cases}$$

$$intensity = r \quad \in [0, +1] \begin{cases} 0, & \text{completamente neutro} \\ +1, & \text{completamente polarizzato} \end{cases}$$

La statistica di sintesi utilizzata è stata ricavata come:

$$score_{sentiment} = polarity * intensity$$

Infine, come ultimo step, questi valori sono stati discretizzati nel seguente modo:

$$sentiment = \begin{cases} negative, & \text{se} \quad score_{sentiment} < -0.5 \\ positive, & \text{se} \quad score_{sentiment} > 0.5 \\ neutral, & \text{altrimenti} \end{cases}$$

A questo punto è stato possibile aggregare i risultati ottenuti contando il numero di tweet per sentiment all'interno di ciascun cluster. Tali conteggi sono stati trasformati in termini percentuali di modo da eliminare la dimensione dei cluster come fattore di distorsione. È ovvio infatti che i cluster più grandi riportassero valori maggiori in termini assoluti.

Figura 8: Distribuzione della sentiment all'interno delle community.

Nella figura 8 vengono riportati i dati relativi alle prime 18 community, ovvero quelle contenenti almeno 20 tweet ciascuno.
Dai risultati ottenuti si può osservare come in tutte le community la componente di tweet neutrali sia predominante. Si può quindi affermare che, come è lecito aspettarsi, non si verifichi mai che i tweet negativi si isolino in maniera assoluta dal resto della rete. Al tempo stesso è innegabile che in quasi tutte le community analizzate la componente di tweet negativi sia di gran lunga superiore rispetto alla componente dei tweet positivi. In particolare spiccano:

La community 0, in cui la maggior parte dei tweet è contraria al Green pass in quanto ritenuto uno strumento discriminatorio. L'aspetto più interessante è che la maggior parte dei tweet è esplicitamente contraria a tale misura ma non al vaccino, dimostrando come i due argomenti, pur essendo estremamente correlati, possano essere divisivi ciascuno a modo proprio.

```
Io da oggi detentrice di #supergreenpass #COVID19 senza averlo
↪ richiesto, provo un disagio tremendo.Le distinzioni tra
↪ cittadini/e di serie A e B, provocano fratture pericolose.Non
↪ voglio far parte di un'élite che non ho scelto

@corpodelledonne La capisco. Sono vaccinato e il #greenpass lo uso
↪ solo per il lavoro (e già lo trovo umiliante). Ristoranti e bar mi
↪ hanno perso come cliente. Non farò #terzadose. Voglio la
↪ promozione a #novax. Non voglio avere nulla in comune con chi
↪ discrimina una parte della popolazione.
```

```
Io da free vax rispetto tutti, sinceramente non mi frega nulla se ti
↪ sei vaccinato o meno. Ma chi è favorevole al #greenpass ha tutto
↪ il mio disprezzo possibile, in quel caso non ti rispetto, ma sei
↪ solamente un miserabile discriminatore. #ObbligoVaccinale
↪ #nogreenpass
```

La community 7 in cui la maggior parte dei tweet più che parlare di Super green pass (pur utilizzandone l'hashtag associato), sembra focalizzarsi prevalentemente contro il vaccino, sostenendo la tesi che sia inefficace.

```
Sono un po' troppi 'sti #novax in #veneto. Sto scherzando, ormai da
↪ tempo anche i muri sanno che il vaccinato contagia. Solo al
↪ governo fingono ancora di non saperlo e una bugia ripetuta, si
↪ sa... #supergreenpass https://t.co/GdQpAdyhuD

Ma se dopo 3 dosi di " vaccino" una persona si ammala ugualmente..Nn
↪ vi sentite un filo presi x i fondelli dal Gov?\nAltra conferma ke
↪ il #greenpass è solo dannoso e nn serve ad evitare il contagio
↪ @carlosibilia chiaro?

E' la prima volta nella storia che un #vaccino ti garantisce l'entrata
↪ al ristorante, al cinema, ai concerti ma non ti garantisce
↪ l'immunità contro la malattia per il quale lo fai ... Pensateci
↪ #COVID19 #GreenPassrafforzato #vaccinoCovid
```

Nella community stessa, compaiono anche tweet che sostengono, con tono polemico, che le istituzioni dovrebbero occuparsi di altri problemi più urgenti. Anche in questo caso, sebbene in forma implicita, traspare una certa avversione nei confronti del Green pass.

```
Con un simile dispiegamento di forze dell'ordine impegnato a
↪ controllare #greenpass, sarà un felicissimo #Natale per i
↪ delinquenti.

È bello sapere che si possono effettuare controlli a tappeto
↪ dispiegando così tanti agenti in tutta Italia. Sono certo che
↪ oltre per il #supergreenpass il Ministro Lamorgese vorrà
↪ utilizzarli anche per combattere degrado e delinquenza. Se si può
↪ fare, si può fare sempre.

Su 1.250 alberghi a Roma ben 380 sono ancora chiusi e la metà di
↪ questi probabilmente non riaprirà. Poi #greenpass e
↪ #supergreenpass danno il colpo di grazia al settore in favore
↪ degli abusivi. E i prezzi degli alberghi crollano del 30% ROSCIOLI
↪ Federalberghi su @LaVeritaWeb
```

La community 10, in cui oltre che una sentiment fortemente negativa, non sembra distinguersi per qualche caratteristica peculiare.

Infine si osserva che per la community 11, quella associata all'hub 'CislNazionale', la Sentiment Analysis, in un certo senso, trova un riscontro concreto rispetto a quanto osservato nel capitolo precedente: tale cluster, infatti, è composto interamente da tweet neutrali, il che rappresenta un risultato che ci si può attendere rispetto al profilo ufficiale di una confederazione sindacale. Semplicemente i tweet contenuti in tale community si limitano a riportare fatti oggettivi.

4.2 Emotion Recognition

L'attività di Emotion Recognition, detta anche Emotion Detection, consiste nel riconoscere la presenza di una o più emozioni all'interno di un dato non strutturato. Trattasi di un sotto-task a granularità più fine rispetto alla Sentiment analysis: in letteratura lo spettro di emozioni considerato in quasi tutte le ricerche è quello proposto da Plutchik, che prevede 8 emozioni di base: aspettativa, gioia, fiducia, paura, sorpresa, tristezza, disgusto e rabbia, e dalla cui unione se ne possono inferire di più complesse, riportate in figura 9. Ne consegue che questo tipo di analisi permette di cogliere informazioni più dettagliate (c'è una differenza significativa tra il provare paura rispetto al provare rabbia, pur essendo entrambe negative) e, a propria volta, di sviluppare un processo di decision-making più raffinato.

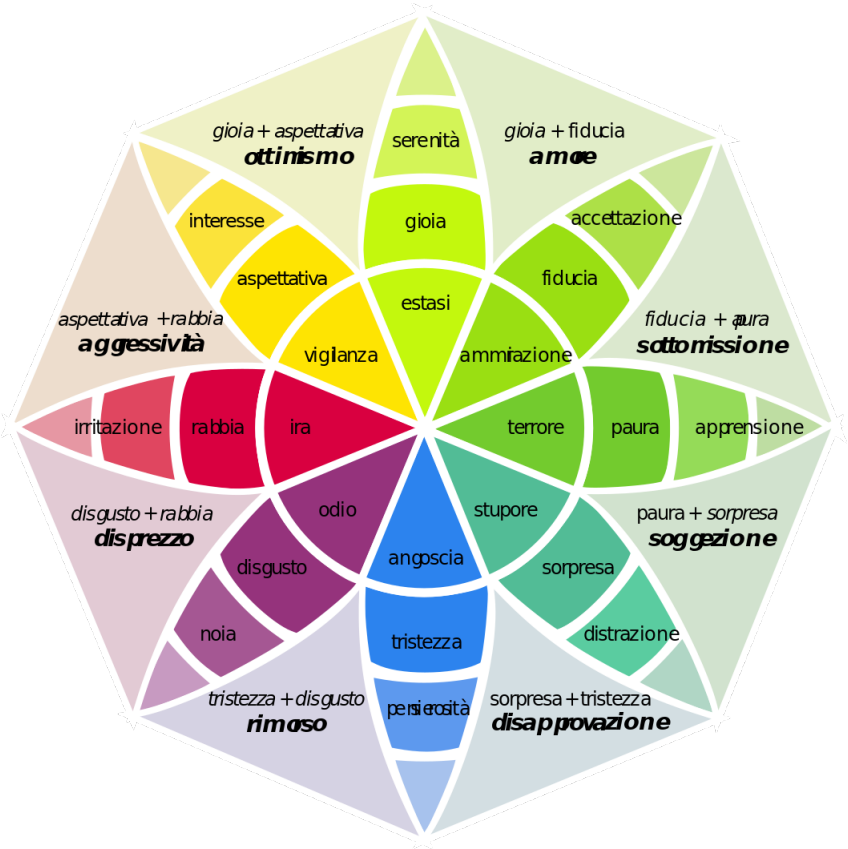

Figura 9: Ruota di Plutchik.

La complessità di questo task, tuttavia, fa sì che la gamma di soluzioni disponibili sia più limitata rispetto a quella per la Sentiment analysis, a maggior ragione se si considera l'applicazione a testi in lingua italiana, per di più malformati come i tweet. In questo caso l'approccio adottato è di tipo lexicon-based, e ha fatto uso del lessico NRC Emotion Lexicon, una lista di parole tradotte in tutte le lingue e annotata manualmente, che a ciascuna parola associa una o più emozioni. Ogni parola quindi può essere rappresentata come un vettore binario di lunghezza 8. L'associazione delle emozioni a un tweet è stata effettuata utilizzando la libreria NRCLex. Semplicemente ogni tweet viene rappresentato anch'esso come un vettore di dimensione 8, in cui l'i-esimo elemento indica se è presente almeno una parola nel tweet a cui viene associata l'i-esima emozione.

Tweet	paura	rabbia	aspettativa	fiducia	sorpresa	tristezza	disgusto	gioia
un successone vero? turismo distrutto per aver voluto consentire il blando green pass facoltativo europeo e voluto fortemente dal ministro del turismo nonostante si intuisse la deriva autoritaria. ora ovunque disdette a iosa montagna aria pura vietata	1	1	1	0	0	1	0	0

Tabella 6: Esempio di vettore binario delle emozioni associato ad un tweet.

A questo punto, come per la Sentiment analysis, i risultati sono stati aggregati a livello di cluster effettuando la somma *element-wise* dei vettori associati ai tweet. Anche in questo caso tali valori in termini assoluti sono stati aggiustati per il numero di tweet del cluster allo scopo di eliminare il fattore di distorsione dovuto alla diversa numerosità dei cluster stessi. I risultati ottenuti sono riportati in figura 10.

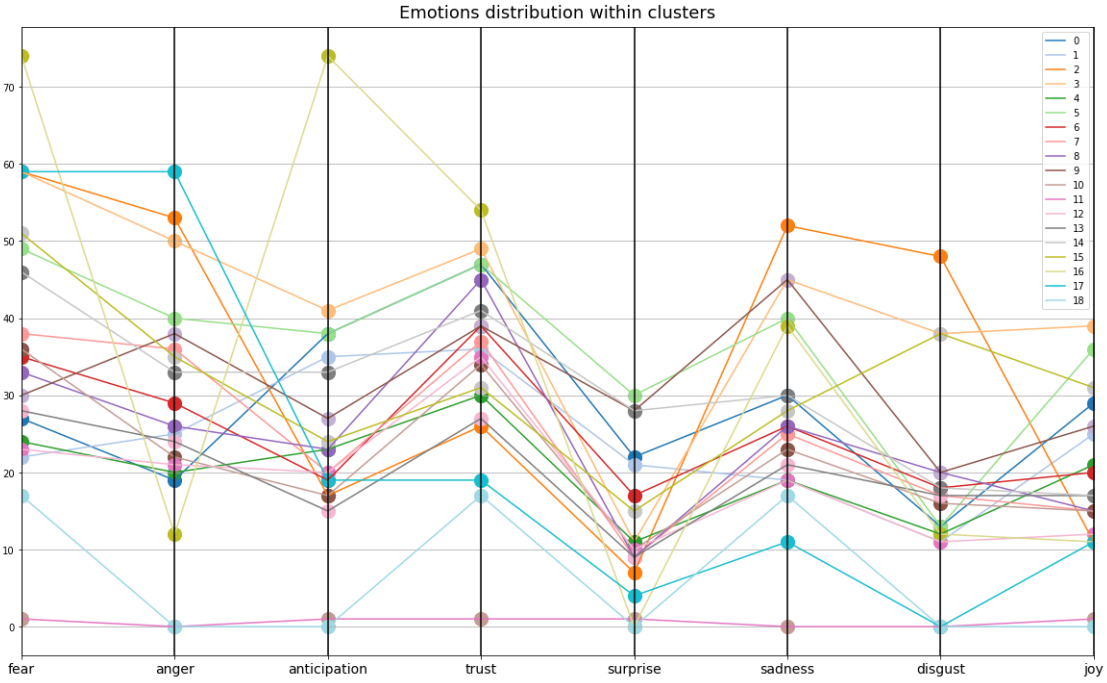

Figura 10: Distribuzione delle emozioni all'interno dei cluster.

Innanzitutto, come ulteriore dimostrazione del fatto che per la community 11 (CISL) le informazioni riportate siano oggettive, si osserva come tutte le emozioni siano praticamente assenti.
Inoltre si osserva che la community 16 emerge per le emozioni di paura, aspettativa e fiducia. Riprendendo la ruota di Plutchik è possibile inferire che, data la concomitanza di fiducia e paura, in questa community sia presente l'emozione più complessa della sottomissione.
La community 17 spicca per rabbia e paura, così come anche la community 2, la quale è caratterizzata inoltre da una forte presenza di tristezza e disgusto. Quest'ultime, portano all'affermazione dell'emozione più complessa del rimorso.

Le community 3 e 5 hanno distribuzioni simili per quanto concerne paura, rabbia, aspettativa e fiducia. Come affermato per la community 16, fiducia e paura comportano la sottomissione, mentre rabbia e aspettativa insieme implicano aggressività.
Per le community rilevanti per la sentiment come osservato nel paragrafo precedente non ci sono osservazioni particolari.
Infine si nota come, tendenzialmente, l'emozione meno presente in tutte le community sia la sorpresa, fatto che non stupisce essendo il Super green pass una misura preannunciata, e in parte anche la gioia.

5 Conclusioni

In merito a quanto emerso nelle diverse fasi di progetto è possibile concludere che è maggiore la percentuale di utenti contro le nuove misure. Sebbene nella realtà essi siano soltanto una minoranza, la forte componente emotiva presente nei loro post permette un efficace condivisione aumentando la loro visibilità sul social. Questo effetto si ripercuote sulla realtà in quanto fomenta una visione distorta del problema se ci si concentra soltanto su quanto avviene nei social, insinuando sempre più dubbi nelle persone vaccinate. Tutto ciò diminuisce l'efficacia delle misure contenitive e di prevenzione attuate dal governo. In tal senso, sarebbe interessante estendere l'analisi considerando più giorni, o anche un mese intero, per cercare di cogliere se il malcontento sia costante o se si evolve a seconda degli avvenimenti politici e sanitari.
Con l'intento di marginare il fenomeno, visto che nella rete non sono state trovate figure di riferimento pro Green pass, si potrebbero sviluppare strategie di comunicazione coinvolgendo influencer famosi che si rivolgano proprio alle persone contrarie al fine di influenzare positivamente le loro idee. Vista la forte disconnessione della rete questo approccio potrebbe essere complesso e/o vano, quindi si potrebbe allargare lo studio con lo scopo di individuare le persone che potrebbero cambiare idea e concentrare questo sforzo comunicativo solo su di loro.

Riferimenti bibliografici

[1] Python Core Team, *Python: A dynamic, open source programming language*, Python Software Foundation, 2019, python version 3.6.9. [Online]. Available: https://www.python.org/

[2] J. Roesslein, "Tweepy: Twitter for python!" *URL: https://github.com/tweepy/tweepy*, 2020.

[3] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[4] W. H. Institute, "Pyvis - a python library for visualizing networks," *URL: https://github.com/WestHealth/pyvis*, 2018.

[5] M. M. Bailey, "Nrclex," *URL: https://github.com/metalcorebear/NRCLex*, 2019.

[6] G. Nicola, "Sentita, a sentiment analysis tool for italian," *URL: https://nicgian.github.io/Sentita/*, 2018.

[7] R. PLUTCHIK, "Chapter 1 - a general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980, pp. 3–33. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780125587013500077

[8] F. Bianchi, D. Nozza, and D. Hovy, ""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, 2021.

[9] V. Basile and M. Nissim, "Sentiment analysis on italian tweets," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, 2013, pp. 100–107.

[10] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text." in *ICWSM*, E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, Eds. The AAAI Press, 2014. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14

[11] B. Magnini, M. Negri, E. Pianta, L. Romano, M. Speranza, L. Serafini, C. Girardi, V. Bartalesi, and R. Sprugnoli, "Ontotext," in *From Text to Knowledge for the Semantic Web: the ONTOTEXT project*, vol. 166, 01 2005. [Online]. Available: https://ontotext.fbk.eu/sentiwn.html