Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea in

Statistica per le Tecnologie e le Scienze

Un'analisi delle opinioni degli italiani sui vaccini contro il SARS-Cov-2 attraverso l'applicazione di tecniche di *Text Mining* ai dati di Twitter

Relatrice: Prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Marius Viorel Parvu
Matricola n. 1218762

Anno Accademico 2021/2022

Indice

Introduzione

Oramai da oltre due anni, è in corso una delle più gravi emergenze sanitarie globali di cui l'uomo abbia memoria. Ci si sta riferendo alla pandemia di COVID-19, una malattia respiratoria acuta grave causata dal SARS-CoV-2 (*Severe Acute Respiratory Syndrome-Coronavirus-2*). Questo coronavirus, di cui oggi si contano almeno tredici varianti, è stato segnalato per la prima volta il 31 dicembre 2019, nella regione cinese di *Wuhan*. A partire da quella data, i casi sono aumentati con un andamento esponenziale, prima in Cina, poi nel resto del mondo. In un primo momento, si è creduto che questo coronavirus si potesse trasmettere agli umani solo da animali infetti, ma il 20 gennaio 2020 si sono avute le prime smentite con l'individuazione di alcuni contagi da umano a umano. Pochi giorni dopo, diverse regioni cinesi hanno attuato dei *lockdown* nel tentativo di arginare la diffusione del virus.

Se la Cina ha deciso di adottare sin da subito delle strategie di contenimento drastiche e pervasive, gli altri paesi hanno di molto sottovalutato il fenomeno. La situazione ha ricevuto da questi ultimi la giusta attenzione soltanto quando si sono registrati i primi casi al di fuori del territorio cinese. I vari governi, allora, hanno disposto alcune misure, tra le quali: la sospensione di molte attività in presenza e, quando possibile, la loro conseguente digitalizzazione; la restrizione della libertà di movimento dei cittadini; l'obbligo di indossare dispositivi di protezione individuale.

A causa dell'emergenza pandemica, sono stati fatti enormi investimenti per identificare delle strategie farmacologiche utili a prevenire e ad arginare l'ulteriore diffusione del virus, sicché già a dicembre 2020 sono stati autorizzati i primi vaccini. Attualmente, quelli approvati nel nostro paese dall'Agenzia Italiana del Farmaco (AIFA) sono: Comirnaty di Pfizer-BioNtech (22 dicembre 2020); Spikevax di Moderna (7 gennaio 2021); Vaxzevria di AstraZeneca (30 gennaio 2021); Janssen di Johnson & Johnson (12 marzo 2021); Nuvaxovid di Novavax (22 dicembre 2021).

In Europa, la campagna vaccinale ha avuto ufficialmente inizio il 27

dicembre 2020, data nota come "V-Day".

Quanto agli obiettivi perseguiti, limitandosi alla prospettiva europea, si è puntato a ridurre il più possibile il numero di decessi e ad alleggerire la pressione sui servizi ospedalieri, specialmente sui reparti di terapia intensiva. Per raggiungere questi scopi, la popolazione è stata suddivisa in gruppi, tenendo conto del rischio, da un lato, di contrarre la malattia e, dall'altro, di sviluppare sintomi gravi. In prima battuta, sono stati vaccinati gli operatori sanitari, gli anziani e le persone con gravi disabilità o con patologie critiche, mentre in un secondo momento le vaccinazioni sono state estese a tutti i cittadini.

Benché sia la letteratura che l'evidenza avessero, in un primo momento, sostenuto e dimostrato l'efficacia dei vaccini, in breve tempo è emerso che quest'ultima, contrariamente alle speranze, è principalmente legata al fatto di contrarre la malattia in forma grave. Inoltre, la durata della protezione vaccinale è limitata a pochi mesi, dopo i quali è necessario procedere a un richiamo (ISS, 2021).

Una novità importante che ha riguardato le vaccinazioni è stata l'introduzione, avvenuta il 1° luglio 2021, del *green pass*, ossia un certificato che consente gli spostamenti all'interno dell'Unione Europea. Questo documento è stato reso necessario per l'accesso a servizi come ristoranti, bar, palestre, *etc.* Non si può non citare, poi, l'introduzione nel nostro paese dell'obbligo di vaccinazione per gli *over* 50 (sanzionato a partire dal 1° febbraio tramite sanzione amministrativa) e dell'obbligo vaccinale per specifiche tipologie di lavoratori.

La campagna vaccinale, il suo esito e le azioni politiche intraprese in suo favore hanno innegabilmente scatenato l'opinione pubblica nell'arco degli ultimi due anni. Da una parte, essa è stata accolta con sollievo da molti cittadini. Dall'altra, sono sorti molti dubbi sulla sua efficacia, sui rischi connessi, nonchè sulle politiche attuate.

L'obiettivo di questo lavoro è esplorare mediante l'applicazione di appropriate tecniche statistiche, l'evoluzione dell'opinione pubblica in Italia in merito ai vaccini contro il SARS-CoV-2. A questo fine, è stato svolto uno studio statistico sui dati testuali opportunamente scaricati dal *social network* Twitter.

Data la moltitudine di informazioni a disposizione si è deciso di concentrare l'attenzione solo su tre periodi collegati ad avvenimenti che hanno sollevato pesanti polemiche:

- 27/12/2020 - 02/01/2021: inizio della campagna vaccinale in Italia;

- 15/03/2021 - 21/03/2021: sospensione temporanea dell'utilizzo del vaccino AstraZeneca nel nostro paese;

- 07/01/2022 - 13/01/2022: introduzione, con il DPCM del 7 gennaio 2022, dell'obbligo vaccinale per gli *over* 50.

Quanto alle tecniche utilizzate, si è fatto uso del *Text Mining* per estrapolare la distribuzione dei sentimenti all'interno di ogni periodo e comparativamente al variare dei periodi. Il termine *Text Mining* (Tan *et al.*, 1999) si riferisce al processo di trasformazione di un testo non strutturato, ossia che non presenta uno schema predefinito, in dato strutturato da cui sia possibile estrarre informazione. Infatti, alcuni dei possibili obiettivi raggiungibili con queste tecniche sono: l'individuazione degli argomenti trattatati nei testi; la loro classificazione in gruppi; l'estrazione di particolari informazioni.

Il lavoro è organizzato come segue: nel primo capitolo sono esposte le modalità attraverso le quali sono stati raccolti i dati, il modo con cui si è proceduto al *pre-processing* degli stessi e, infine, le analisi descrittive preliminari; nel secondo capitolo ci si concentra sulle analisi non supervisionate; nel terzo capitolo si descrive l'applicazione del modello supervisionato grazie a una previa classificazione manuale.

Capitolo 1

Raccolta e pretrattamento dei dati

1.1 Creazione del *dataset*

1.1.1 Twitter

Twitter[1] è una piattaforma *online*, resa disponibile a partire da luglio 2006, che permette di creare delle reti sociali virtuali volte all'interazione tra gli utenti. Si tratta di un *social network* che consente il cosiddetto *microblogging*, ossia una peculiare forma di pubblicazione di brevi contenuti, i *tweet*, i quali possono consistere in messaggi di testo (lunghi al massimo 280 caratteri), immagini, video, audio, *link*, *etc.* I *tweet* possono essere pubblici, cioè visibili da chiunque, oppure privati, ovvero accessibili solo da chi segue, previa accettazione, il profilo del loro autore.

Uno strumento largamente impiegato all'interno della piattaforma è l'*hashtag*: una etichetta costituita da un cancelletto seguito da una sequenza di caratteri alfanumerici non separati da spazi. L'*hashtag* dà la possibilità all'autore di un *tweet* di indicare e categorizzare parole chiave, così che gli altri utenti possano individuare facilmente gli argomenti a cui sono interessati.

La società Twitter Inc., proprietaria dell'omonimo servizio, mette a disposizione un'ulteriore piattaforma, Twitter Developer[2]. Da quest'ultima si possono ottenere le informazioni rese liberamente disponibili dagli utenti, al fine di utilizzarle per vari scopi, come la ricerca scientifica.

[1] www.twitter.com

[2] developer.twitter.com

Tale piattaforma consente l'acquisizione di suddette informazioni attraverso le *Application Programming Interfaces* (API). Generalizzando, le API rappresentano dei meccanismi mediante i quali una società fornisce l'accesso ai propri prodotti, servizi e/o dati da parte di un'entità terza. In altre parole, sono dei protocolli informatici che servono a risolvere i problemi di interazione tra dei *computer* oppure tra dei *software* o loro parti, consentendo la comunicazione tra questi (Reddy, 2011).

In conclusione, si è deciso di fare ricorso a Twitter principalmente per due motivi: il primo, è legato al fatto che questo *social network* sia quello oggi più usato per commentare i fatti di cronaca; il secondo, di natura più pragmatica, è legato alla facilità con la quale sia possibile scaricare i dati, grazie soprattutto alle parole chiave, ossia i citati *hashtag*. Inoltre, è bene precisare che la scelta di utilizzare quest'unica piattaforma potrebbe, inevitabilmente, aver introdotto una distorsione. Infatti, non tutta la popolazione italiana utilizza questo *social network* e non è detto che tutti gli utenti pubblichino i loro pensieri (per esempio, è più probabile che li esprimano le persone con una forte opinione). Tuttavia, nonostante la possibile introduzione di questo *bias*, si è deciso, per semplicità, di non reperire ulteriori dati da altre piattaforme.

1.1.2 R e RStudio

R (R Development Core Team, 2011) è un linguaggio di programmazione, disponibile sotto forma di *software open source*, che permette l'applicazione di una vasta gamma di modelli statistici e tecniche grafiche. Uno dei suoi punti di forza risiede nel fatto di poter essere esteso utilizzando i pacchetti contenuti nel *Comprehensive R Archive Network* (CRAN) e in altri siti *internet*.

RStudio (RStudio Team, 2022), invece, è un ambiente di sviluppo integrato (*Integrated Development Environment*, IDE) per il linguaggio R. Al suo interno presenta un *editor* di testo, che agevola la scrittura (in particolare, evidenzia l'esatta sintassi del codice), uno strumento per il *plotting*, un altro per la cronologia, un altro ancora per il *debug* e, per finire, un *tool* per la gestione dell'area di lavoro.

1.1.3 Aspetti operativi

Per l'estrazione dei dati ci si è avvalsi del pacchetto `rtweet` (Kearney, 2019).

Il procedimento può essere così riassunto:

- creazione di un *account* Twitter;
- creazione dell'*account* Twitter Developer;
- registrazione di un'applicazione in Twitter Developer;
- utilizzo delle funzioni di `rtweet` per il *download* dei dati.

Al fine di creare un *account* Twitter è sufficiente accedere alla suddetta piattaforma, dal sito o dall'*app mobile*, e compilare il *format* inserendo le informazioni richieste.

Per quanto concerne il secondo e il terzo punto, si deve accedere a Twitter Developer e seguire i vari passaggi per la registrazione. Si tenga presente che, durante la procedura, è necessario dichiarare il motivo per il quale si vogliono scaricare i *tweet*. Una volta effettuati questi passaggi, sarà possibile gestire l'accesso alle API necessarie l'acquisizione dei dati pubblici degli utenti.

Le funzioni della libreria `rtweet` si differenziano in base al lasso temporale di interesse:

- `search_tweets()` e `search_tweets2()` permettono l'estrazione di *tweet* pubblicati negli ultimi 6/9 giorni;
- `search_30day()` fornisce i *tweet* pubblicati negli ultimi 30 giorni;
- `search_fullarchive()` consente l'accesso all'intero storico dei *tweet* pubblicati.

Quanto ai parametri, le funzioni sono molto simili. Poiché ai fini dell'analisi è stata utilizzata `search_fullarchive()`, di seguito si riportano solo i parametri di questa:

- `q`, la *query* tramite cui filtrare i *tweet*;
- `n`, il numero di *tweet* che si desidera scaricare per chiamata;
- `fromDate` e `toDate`, le date e le ore di inizio e fine del periodo all'interno del quale ricercare i *tweet*;
- `env_name`, il nome del ambiente di sviluppo scelto all'interno dell'*account* Twitter Developer;
- `safedir`, la *directory* in cui salvare ogni oggetto di risposta;

- `parse`, il parametro logico per stabilire se convertire i dati in un *dataframe* (*TRUE* per convertire, *FALSE* altrimenti);
- `token`, il *token* associato all'applicazione creata dall'utente sempre nell'*account* Developer.

Bisogna, infine, tenere in considerazione le limitazioni, imposte da Twitter, relative al numero di *tweet* scaricabili in un determinato lasso di tempo. Ad esempio, per la funzione `search_fullarchive()` è possibile effettuare 50 chiamate/mese, ossia scaricare al massimo 5000 *tweet*/mese.

1.1.4 Specificazione dei criteri di selezione

Data la mole di dati disponibili, per comprendere l'evolversi dell'opinione pubblica nei confronti dei vaccini, si è deciso di concentrare l'attenzione sui tre archi temporali ritenuti più rappresentativi: il primo (27/12/2020 - 02/01/2021) s'incentra sull'inizio della campagna vaccinale; il secondo (15/03/2021 - 21/03/2021) riguarda l'episodio a seguito del quale si sono levati probabilmente i più forti dibattiti e le più accese discussioni circa la pericolosità dei vaccini, ossia la sospensione del vaccino AstraZeneca per dei casi sospetti di trombosi; il terzo (07/01/2022 - 13/01/2022) è relativo a una tra le più discusse decisioni del Governo, cioè l'introduzione dell'obbligo vaccinale per gli *over* 50.

Oltre alla selezione di soli tre periodi, si è scelto di considerare lassi temporali relativamente brevi: ciascun periodo, infatti, è composto da sette giorni. Questa decisione è stata dettata principalmente dalla velocità dei tempi di informazione e reazione nell'attuale società. Inoltre, come già descritto, Twitter impone delle importanti restrizioni al *download* dei contenuti. A ciò si aggiunga che la funzione `search_fullarchive()` effettua la ricerca dei *tweet* seguendo un ordine cronologico inverso: ad esempio, se si fornisce alla funzione un intervallo temporale che va dalla data "a" (meno recente) alla data "b" (più recente), essa inizia a scaricare i *tweet* pubblicati a partire dall'ultima data, proseguendo poi all'indietro fino alla meno recente. Considerando un intervallo più ampio si sarebbe, quindi, rischiato di non selezionare proprio i *tweet* pubblicati come reazione all'evento di interesse. Non solo, se alla funzione viene fornito un *set* di parole chiave da ricercare, questa comincerà l'estrazione dei *tweet* partendo dalla prima *key word* presente all'interno dell'insieme fornito in *input*. Anche in questo caso, pertanto, vi è il rischio di una selezione non casuale di *tweet*. Per sopperire a quest'ultimo eventuale problema, si è deciso di effettuare la ricerca una parola chiave per volta.

In più, si è optato per richiamare la funzione inizialmente sui primi due giorni di ciascun periodo (quelli più vicini all'evento significativo e quindi quelli nei quali, presumibilmente, si sarebbe concentrata la maggior parte dei *tweet*), poi sui restanti cinque.

Per far comprendere meglio la procedura seguita, si riportano i passaggi per il periodo 27/12/2020 - 02/01/2021:

1. attivazione della funzione `search_fullarchive()`, impostando i parametri di *input* sotto riportati:
   - `q`, pari alla prima parola chiave e alla lingua d'interesse, in questo caso l'italiano;
   - `n`, pari a 100;
   - `fromDate` e `toDate` , pari, rispettivamente, a 27/12/2020, ore 00:00, e 28/12/2020, ore 23:59;
   - `env_name` e `token`, al cui interno sono state inserite le informazioni personali relative al nome dell'ambiente di sviluppo e al *token* associato all'applicazione creata all'interno della piattaforma sviluppatore;
   - `safedir` e `parse`, pari, rispettivamente, al nome della *directory* scelta e a *TRUE*.
2. Iterazione di 1. tante volte quante sono i *pattern* d'interesse, cambiando di volta in volta solo il parametro `q`;
3. Ripetizione di 1. e 2. cambiando, esclusivamente, i parametri `fromDate` (28/12/2020, ore 00:00) e `toDate` (02/01/2021, ore 23:59).

Il procedimento per gli altri periodi è stato il medesimo.

Per ottenere un numero di *tweet* sufficientemente alto, si è dovuto procedere con ulteriori richiami alla funzione per quei giorni che contenevano un numero esiguo di dati.

Per quanto concerne le parole chiave, si è deciso di utilizzare gli *hashtag*, poiché permettono di individuare in maniera semplice e veloce i *tweet* che trattano di una determinata tematica. Ciò, però, ha portato a un ulteriore problema: visto che gli *hashtag* sono utilizzati per pochi giorni, se non per poche ore, è stato molto arduo individuarne di utilizzati in tutte le fasce temporali. Sempre in tema, si consideri che la scelta delle parole chiave è stata uno dei punti nodali della fase iniziale del lavoro, poiché da questa dipendeva l'omogeneità delle popolazioni nei tre gruppi. In questa ricerca sono stati utilizzati gli *hashtag* riportati nella Tabella 1.1.

Tabella 1.1: *Hashtag* utilizzati per il *download* dei dati

Hashtag	
#vaccino	#vaccini
#astrazeneca	#pfizer
#moderna	#novax
#iomivaccino	#provax

1.2 Descrizione dei dati e operazioni preliminari

A seguito del processo di estrazione dei *tweet*, si dispone di tre insiemi di dati, ciascuno con una dimensione dell'ordine di alcune migliaia di elementi, come riportato in dettaglio nella Tabella 1.2. Ogni dato contiene varie informazioni, nello specifico novanta, riferite a un determinato *tweet*.

Le variabili più rilevanti sono:

- `text`, il testo;
- `lang`, la lingua;
- `screen_name`, il *nickname* dell'autore (inizia con il simbolo "@" ed è univoco);
- `name`, il nome dell'autore (non è univoco e può corrispondere al nome vero o a uno pseudonimo);

Tabella 1.2: Variazione nelle numerosità dei campioni a seguito della rimozione dei dati duplicati e di quelli non provenienti da utenti privati

Periodo	***Numerosità originali***	***Numerosità a seguito delle modifiche***
1	3842	3488 (di cui 2691 utenti unici)
2	3473	2983 (di cui 2230 utenti unici)
3	4410	3870 (di cui 2499 utenti unici)
Numero di utenti unici comuni a tutti i periodi (a seguito delle modifiche)		83

- `location`, la località di provenienza indicata nel profilo;
- `description`, la descrizione del profilo;
- `protected`, la variabile dicotomica pari a *TRUE* se l'*account* è privato, *FALSE* se l'*account* è pubblico;
- `followers_count`, il numero di seguaci;
- `friends_count`, il numero di utenti seguiti;
- `verified`, la variabile dicotomica pari a *TRUE* se l'*account* è verificato, *FALSE* se l'*account* non è verificato;
- `place_full_name`, la posizione del *tweet* (disponibile solo se, quando è stato scritto, era attiva la geolocalizzazione);
- `country`, sigla dello stato dal quale è stato pubblicato il *tweet* (sempre se era stata attivata la geolocalizzazione);
- `created_at`, la data e l'ora di creazione;
- `favorite_count`, il numero di *like* ricevuti;
- `retweet_count`, il numero di *retweet* (condivisioni);
- `reply_count`, il numero di commenti;
- `quote_count`, il numero di *tweet* di citazione (*retweet* con commento);
- `is_retweet`, la variabile dicotomica pari a *TRUE* se il testo è un *retweet*, *FALSE* in caso contrario;
- `is_quote`, la variabile dicotomica pari a *TRUE* se il testo è un *retweet* con commento, *FALSE* se non lo è.

Inoltre, nel caso in cui il *tweet* in questione sia un *tweet* con citazione si hanno a disposizione altre informazioni, relative al *tweet* originale, tra cui:

- `quoted_text`, il testo;
- `quoted_created_at`, la data e l'ora di creazione;

- `quoted_favorite_count`, il numero di *like* ricevuti;
- `quoted_retweet_count`, il numero di *retweet*;
- `quoted_screen_name`, il *nickname* dell'autore;
- `quoted_name`, il nome dell'autore;
- `quoted_followers_count`, il numero di seguaci dell'utente;
- `quoted_friends_count`, il numero di profili seguiti dall'utente;
- `quoted_location`, la località di provenienza indicata nel profilo;
- `quoted_description`, la descrizione del profilo dell'utente;
- `quoted_verified`, la variabile dicotomica pari a *TRUE* se l'*account* è verificato, *FALSE* se l'*account* non è verificato.

Poi, nel caso di un *retweet*, si hanno le stesse informazioni detenute nel caso di *tweet* con citazione, con la sola modifica dei nomi delle variabili.

Poiché l'oggetto di studio è l'opinione degli individui i tre *dataset* presentano due criticità: la presenza di dati duplicati e di dati non provenienti da *account* di persone private (es. *blog*, testate giornalistiche, istituzioni, *etc.*). Dunque, si sono rese necessarie, prima del loro utilizzo, alcune operazioni preliminari di pulizia.

Per quanto concerne il primo problema, questo è dovuto alla scelta, effettuata in fase di *download*, di procedere con la chiamata alla funzione `search_fullarchive()` utilizzando come parametro una parola chiave alla volta. Per comprendere meglio la dinamica si riporta il seguente esempio: si immagini un *tweet* con due *hashtag*, #vaccino e #pfizer, entrambe parole chiave utilizzate per il *download*; in questo caso, è possibile che questo *tweet* venga scaricato due volte.

In ordine al secondo problema, è stata fatta una verifica preliminare di tutti gli utenti che presentavano, all'interno dei *dataset*, un numero di *tweet* superiore a due. Inoltre, si è fatto uso della funzione `classificaUtenti()`, implementata nella libreria `TextWiller` (Solari *et al.*, 2016), la quale, basandosi prima sugli `screen_name`, dopo sui `name` dei vari profili Twitter, cataloga gli *user* in 'maschio', 'femmina' ed 'ente' (dove per 'ente' si intende un profilo appartenente a un giornale, *blog*, *etc.*). Nella pratica, la funzione confronta i nomi degli autori dei *tweet* con dei nomi già classificati nelle tre categorie e, una volta che si verifica il *matching*, si riesce a individuare il sesso del proprietario

del profilo. Tuttavia, a causa del fatto che molti profili hanno dei nomi di fantasia, la funzione, in diverse occasioni, non è riuscita a categorizzarli adeguatamente. Analizzando i tre *dataset*, sono stati individuati 109 profili non appartenenti a persone private. Questo ha portato a un'ulteriore diminuzione delle numerosità dei campioni.

Nella Tabella 1.2 sono riportate le numerosità dei tre insiemi di dati a seguito della rimozione dei duplicati e dei *tweet* non provenienti da utenti privati.

1.3 Trattamento preliminare dei testi

1.3.1 Panoramica

Un testo è una composizione di parole, ossia di segni che possiedono dei significati e che rappresentano dei concetti. A partire dalle parole, e in base agli obiettivi, l'unità di analisi può essere:

- una forma grafica, ossia si decide di utilizzare una parola così come compare nel testo;
- un lemma, ovvero si decide di trasformare una parola nell'esponente che compare nei vocabolari (l'infinito per i verbi, il maschile singolare per gli aggettivi, *etc.*);
- un poliforme, cioè una sequenza di due o più parole che costituiscono un'unità di senso indipendente;
- un tema, ossia si decide di ridurre la parola alla radice comune (es. si riconducono "correre", "corri", "affrettarsi", al tema "corr");
- una forma testuale, come in questo scritto, ovvero un misto delle precedenti.

Tornando al testo, questo racchiude la quasi totalità delle informazioni che possono essere ritenute di interesse, ma prima di riuscire a ottenerle è necessario che i testi subiscano alcune modifiche. Quest'ultime vengono effettuate durante la fase preliminare, detta di pre-trattamento, il cui scopo è trasformare l'informazione testuale in un insieme di dati quantitativi. A tal fine, il testo può essere sottoposto a diverse procedure, tra le quali le più importanti sono: la normalizzazione e lo *stemming*. Per ulteriori approfondimenti si vedano Bolasco *et al.* (2004) e Misuraca (2018).

1.3.2 Normalizzazione

Attraverso la normalizzazione, che consiste in un insieme di azioni volte a uniformare il testo, si opera su tutte le parti di uno scritto così da:

- riconoscere ed eliminare le forme grafiche prive di contenuto informativo (le cosiddette *stop word* o parole vuote);
- trasformare tutti i caratteri in minuscolo ed eliminare la punteggiatura;
- codificare adeguatamente le *emoticon*;
- sostituire gli *slang* con le parole originarie o con stringhe alternative.

Il primo punto è forse quello più importante poiché i testi presentano un numero elevato di "forme strumentali" che non portano con sé un grande contenuto informativo (si pensi, ad esempio, agli articoli, alle congiunzioni, agli avverbi, *etc.*). Per riuscire a individuare le *stop word* i *software* sfruttano delle liste, chiamate *stop list*, contenenti le parole vuote di una specifica lingua; una volta verificato il *matching* tra la forma testuale presente nella lista e quella contenuta nel testo, si procede alla sua eliminazione. Si hanno a disposizione delle *stop list* già pronte all'uso, ma vi è anche la possibilità di crearne di più specifiche, in base all'argomento trattato nei testi.

Quanto al secondo punto, questo presenta delle criticità perché la trasformazione di tutti i caratteri in minuscolo può introdurre delle ambiguità (ad esempio, non si sarebbe più in grado di distinguere tra "Perla", nome proprio, e "perla", gemma). In ogni caso, però, nonostante si vada a perdere dell'informazione, si preferisce comunque attuare la modifica perché permette di ridurre la complessità del problema.

Infine, il terzo e il quarto punto sono particolarmente utili perché consentono di cogliere lo spirito del testo (scherzoso, ironico, serio, *etc.*). Per ulteriori approfondimenti si vedano Bolasco *et al.* (2005) e Canale e Scarpa (2022).

All'interno della Tabella 1.3 viene riportato un esempio di normalizzazione. Quest'ultimo mette in luce anche un problema della funzione: in certi casi modifica senza un'apparente logica le parole (ad esempio, "#Vaxday" viene trasformata in "#va emoteamaze ay"). Di conseguenza, per rimediare a questa stranezza, è stata compiuta una verifica a campione dei *tweet* normalizzati e sono state apportate le correzioni necessarie ad eliminare questi errori.

Tabella 1.3: Esempio di applicazione della normalizzazione

Tweet	***Testo***
Originale	Nuovi metodi per prendere sonno: contare quante persone ci sono in pista prima di te per prendere il vaccino#sonno #vaccino #vaccinoCovid #VDAY #Vaxday https://t.co/EHuv8ivpof
Normalizzato	nuovi metodi prendere sonno contare persone pista prima te prendere vaccino #sonno #vaccino #vaccinocovid #vday #va emoteamaze ay wwwurlwww

Inoltre, prima di procedere con le altre procedure di *pre-processing*, si è deciso di togliere il carattere "#" da tutti i testi così da trasformare gli *hashtag* in forme testuali. La trasformazione è stata attuata in quanto, al fine di questa indagine, non vi è alcuna utilità nel tenere divisi quest'ultimi dalle altre parole che compongono il *tweet* (per esempio non c'è alcun beneficio nel considerare separatamente la forma "#vaccino" dalla forma "vaccino").

1.3.3 *Stemming*

L'operazione di *stemming* consiste nel trasformare le parole in temi (o stilemi o *stem*), ovvero attribuire una medesima etichetta alle parole che possiedono lo stesso significato (per esempio le parole "cibarsi", "sfamarsi", "saziarsi", "rifocillarsi", possono essere ricondotte al tema "mangiare"). In tal modo, si riesce a ridurre il testo a un ristretto insieme di termini, rendendo le analisi successive computazionalmente meno onerose. Bisogna, tuttavia, valutare molto attentamente la scelta di attuare questa procedura in quanto si andrà a perdere gran parte dell'informazione.

Uno stilema può essere costituito da una singola forma, un unigramma, oppure da un insieme di forme: dei bigrammi, una coppia di parole, o dei trigrammi, una terna di termini; usualmente non si va oltre.

Per rendere più chiara la comprensione delle espressioni unigramma, bigramma e trigramma, si riportino alcuni esempi:

- le parole "vaccino", "vaccini" e "vaccinazione" possono essere considerate degli unigrammi, riconducibili a uno stilema comune, come ad esempio "vaccin";

- la coppia di parole "carta" e "bianca" può essere vista come un bigramma, riconducibile allo stilema "carta_bianca";

- le tripla "vaccino", "anti" e "covid" può essere unita in un trigramma, riconducibile allo stilema "vaccino_covid" oppure "vaccin".

Spesso, la fase di *stemming* viene realizzata per mezzo di un algoritmo automatizzato. Tuttavia, gli algoritmi di *stemming* mostrano i loro limiti nel caso di multilinguismo oppure nel caso di parole che possono essere ricondotte allo stesso stilema nonostante abbiano significati diversi (si pensi al caso delle parole "pésca", l'attività sportiva, e "pèsca", il frutto, che hanno lo stesso stilema). Inoltre, tale procedura rende tipicamente la comprensione del testo molto difficoltosa. Ancora, in diverse situazioni, gli algoritmi non riescono a ricondurre le parole con stesso significato al medesimo tema.

Per tutti i suddetti motivi, si è scelto di procedere con una operazione concettuale di *stemming*: per ottenere questo risultato le parole sono state ordinate per frequenza decrescente, per poi essere ridotte a stilemi. Si tenga presente che all'interno di questo elaborato, da questo momento in poi, per *stemming* si intenderà lo *stemming* concettuale applicato.

Nella Tabella 1.4 vengono riportate le modifiche apportate agli unigrammi, mentre nelle Tabelle 1.5 e 1.6 quelle relative, rispettivamente, ai bigrammi e ai trigrammi.

Tabella 1.4: *Stemming* concettuale

Stilema	***Parola***
vaccino	vaccinoanticovid, vaccinocovid, covid19vaccino, vaccine, vaccini
vaccinato	vaccinata
vaccinazione	vaccinazioni
sivax	provax, vaccinationdone
novax	ilvaxuccide, iosononovax

Continua nella prossima pagina

Tabella 1.4 – *Continua dalla pagina precedente*

Stilema	***Parola***
vaxday	vaccinoday, vday, 27dicembre, vaccineday
covid19	coronavirus, covid, covid19italia, covid1919
pfizer	pfizervaccino, pfizeritalia, pfizerbiontech, pfizerbaby
astrazeneca	astrazenaca, vaccinoastrazeneca, astrazenica, astrazeneka, astrazenenca, astrazenecavaccino, astrazenika, astrazeca
johnsonandjohnson	johnsonandjohnsonvaccino
ue	europa, europeista, europeo, europei, europe
lockdown	lockdownitalia
infermiere	infermieri, infermiera
persona	persone
popolazione	popolo
giorno	giorni
odio_draghi	draghivattene, draghingalera, draghistan, draghibugiardo

Continua nella prossima pagina

Tabella 1.4 – *Continua dalla pagina precedente*

Stilema	***Parola***
moderna	modernavaccine
libertà	libeà
goldman_sachs	goldmansachs
emoticon_vomito	0001f92e
emoticon_ditoversoilbass	0001f447
emoticon_ ditoversodestra	0001f449
emoticon_partyface	0001f973
emoticon_sorrisoconocchiali	0001f60e
emoticon_siringa	0001f489
emoticon_pensierosa	0001f914
emoticon_ditoversoalto	0001f446
emoticon_nerdface	0001f913
emoticon_risata	emote_TearsOfJoy, emote_RollingOnFloorLaughing

Tabella 1.5: Bigrammi e loro trasformazione in unigrammi

Unigramma	***Bigramma***
covid19	covid 19 covid19 19
ue	unione europea unione ue
reazioni_avverse	reazioni avverse
anti_covid19	anti covid19
astrazeneca	astra zeneca
novax	no vax
economia_italiana	economia italiana
effetti_collaterali	effetti collaterali
campagna_vaccinale	campagna vaccinale campagna vaccinazione
operatori_sanitari	operatori sanitari
regno_unito	regno unito
paesi_ue	paesi ue paesi uei

Continua nella prossima pagina

Tabella 1.5 – *Continua dalla pagina precedente*

Unigramma	***Bigramma***
johnsonandjohnson	johnson johnson
figliuolo	generale figliuolo
ministro_speranza	ministro speranza ministro robersperanza
accademia_crusca	accademia crusca
scienziati_criminali	scienziati criminali
reazione_allergica	reazione allergica
casa_riposo	casa riposo
no_party	no party
leader_opposizione	leader opposizione
media_mobile	media mobile
uni_oxford	università oxford
sistema_immunitario	sistema immunitario
goldman_sachs	goldman sachs

Tabella 1.6: Trigrammi e loro trasformazione in unigrammi

Unigramma	***Trigramma***
vaccino	vaccino anti covid vaccino anti covid19
bere_bicchier_acqua	bere bicchier acqua

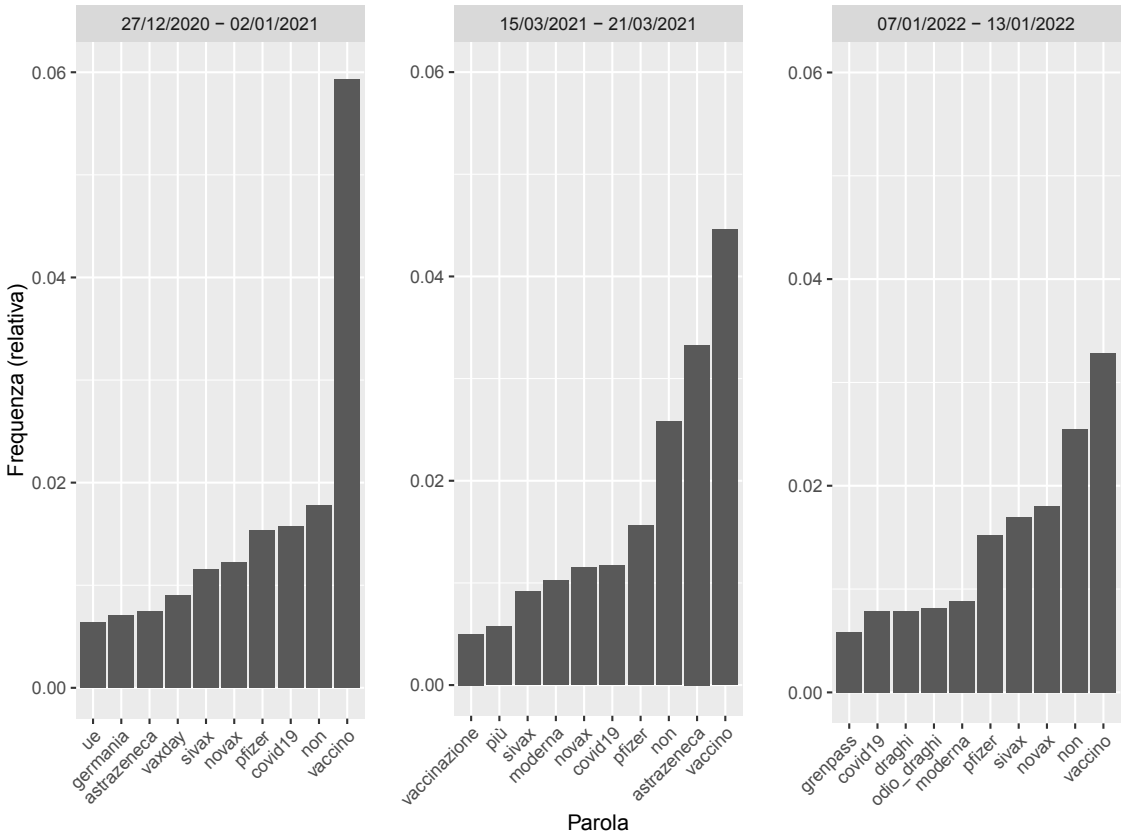
1.4 Analisi esplorative

Si riportano ora le analisi descrittive svolte sui *tweet*, una volta normalizzati e sottoposti allo *stemming* concettuale, e sulle altre informazioni accessorie.

Nella Figura 1.1 vengono presentate le parole più frequenti nei vari periodi. Da questa emerge come:

- in tutti e tre i periodi, la forma testuale più utilizzata sia "vaccino";
- all'interno del secondo periodo, la parola "astrazeneca" abbia acquistato molto peso;

Figura 1.1: Frequenze delle parole maggiormente utilizzate nei tre periodi all'interno dei testi sottoposti a normalizzazione e *stemming*

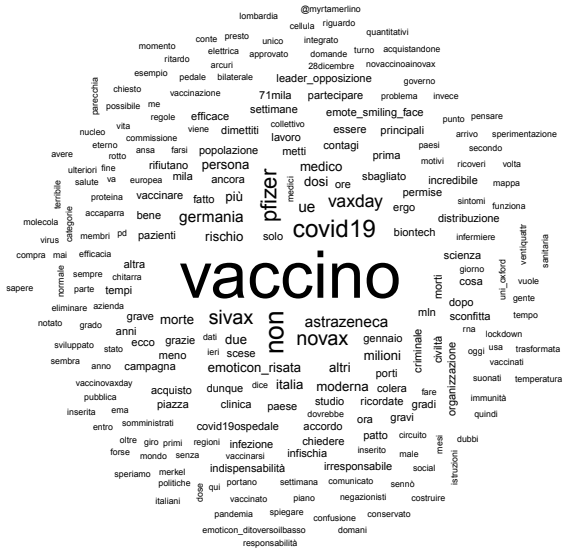


- nel terzo periodo, le parole riferite al Presidente del Consiglio dei Ministri, Mario Draghi, ricorrano con grande frequenza (si ritrova, infatti, sia la parola neutra "draghi", che lo stilema negativo "odio_draghi"). Inoltre le parole "novax", "sivax" e "greenpass", assumono maggiore valenza.

É possibile estrarre analoghe informazioni a partire dalle *word cloud* presenti nella Figura 1.2.

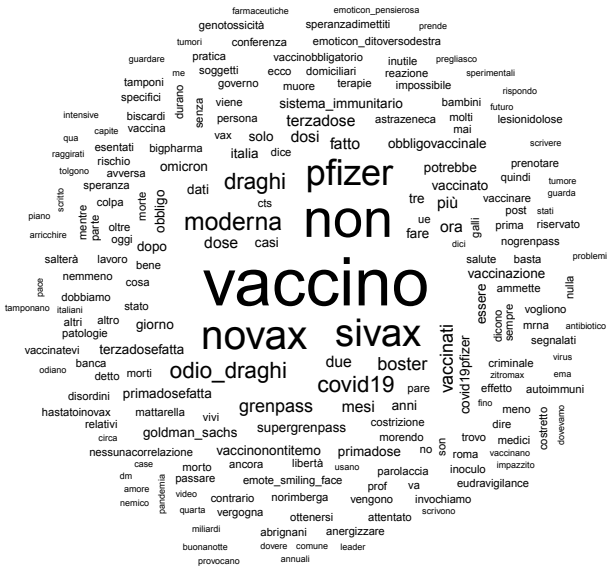
Dall'analisi dei bigrammi, presenti all'interno della Tabella 1.7, è possibile notare che: nel secondo intervallo la discussione verteva prevalentemente sui diversi tipi di vaccini (il che è un risultato atteso vista la preoccupazione scatenata dal vaccino AstraZeneca); nel terzo, invece, si parlava maggiormente delle varie tipologie di *green pass* e del numero di dosi inoculate (molto probabilmente in risposta agli obblighi imposti dal governo).

Figura 1.2: *Word cloud* delle parole maggiormente utilizzate nei tre periodi all'interno dei testi sottoposti a normalizzazione e *stemming*



(a) *Word cloud* primo periodo

(b) *Word cloud* secondo periodo



(c) *Word cloud* terzo periodo

Tabella 1.7: Bigrammi più frequenti all'interno dei tre periodi e loro corrispondenti frequenze relative

Periodo	***Bigramma***	***freq. relativa***
1	vaccino pfizer	(0.00523)
	metti rischio	(0.00361)
	sbagliato lavoro	(0.00361)
	clinica studio	(0.00359)
	contagi porti	(0.00359)
	covid19ospedale clinica	(0.00359)
	criminale dunque	(0.00359)
	dimettiti vaccino	(0.00359)
	dunque dimettiti	(0.00359)
2	vaccino astrazeneca	(0.00581)
	astrazeneca pfizer	(0.00552)
	pfizer moderna	(0.00497)
	moderna johnsonandjohnson	(0.00299)
	astrazeneca vaccino	(0.00232)
	covid19 vaccino	(0.00192)
	vaccino covid19	(0.00171)
3	pfizer moderna	(0.00416)
	grenpass supergrenpass	(0.00298)
	vaccino sivax	(0.00271)
	boster vaccino	(0.00268)
	sivax covid19pfizer	(0.00259)
	covid19pfizer pfizer	(0.00254)
	terzadose terzadosefatta	(0.00252)
	moderna sivax	(0.00248)
	primadose vaccinonontitemo	(0.00248)
	primadosefatta primadose	(0.00248)

Successivamente, si sono effettuati dei confronti tra le parole appartenenti ai vari periodi, al fine di capire quanto spesso venivano utilizzate. Scendendo nei dettagli, sono state calcolate le frequenze relative delle parole all'interno di ciascun *set* di dati e poi sono stati realizzati tre grafici di dispersione:

- dal raffronto tra i primi due periodi (Figura 1.3) si può notare come: dal primo periodo, non emergano sentimenti chiari, ma si vedano, principalmente, riferimenti alla distribuzione e gestione delle dosi nei vari stati europei ("ue", "germania", "campagna", "paesi"); dal secondo periodo, ci siano riferimenti ai "bambini", probabilmente per via dell'inizio delle sperimentazioni dei vaccini su di loro, alla possibilità di rendere obbligatorio il vaccino ("obbligovaccinale"), all'inizio dei movimenti "nomask", alle reazioni avverse ("reazione") causate dai vaccini e alla commercializzazione del vaccino Johnson & Johnson. In comune, invece, si ritrovano i vocaboli che fanno riferimento ai vaccini ("astrazeneca", "moderna", "pfizer"), ai movimenti *novax* e *sivax* ("novax", "sivax"), alla libertà ("libertà"), *etc.*;

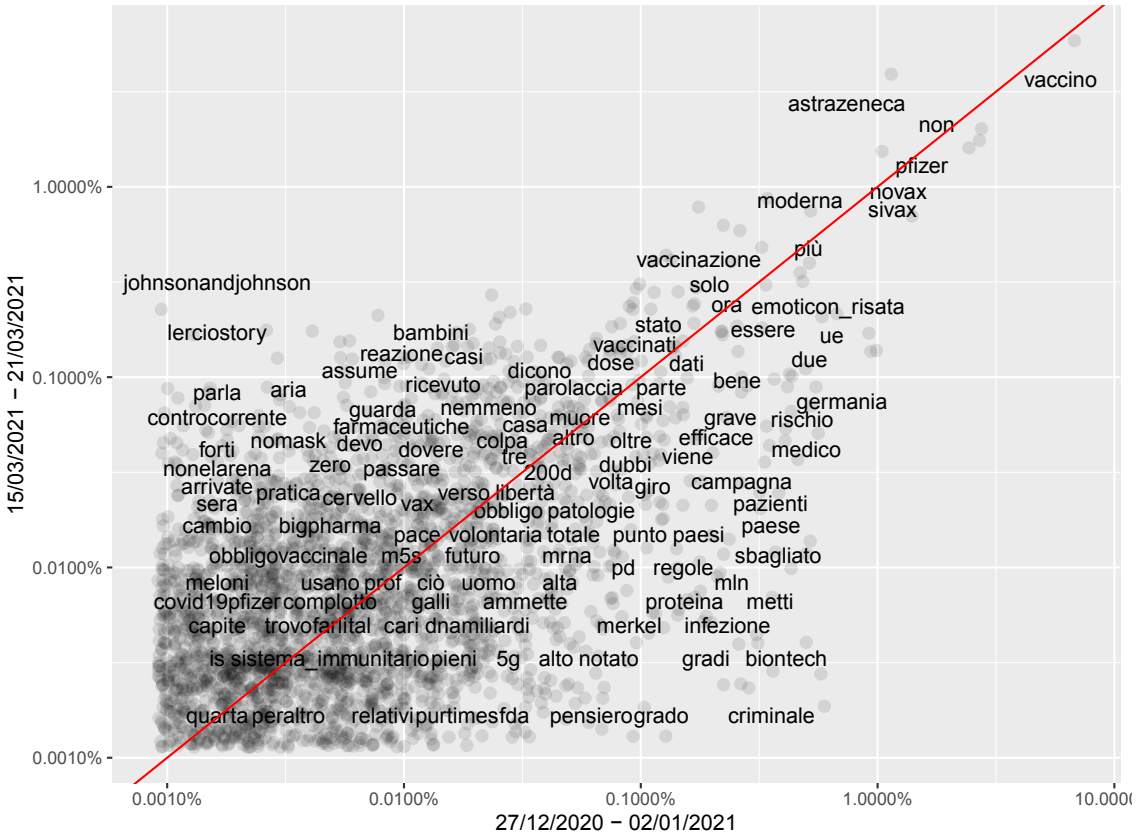
- dal raffronto tra il primo e il terzo periodo (Figura 1.4) si osserva come: nel primo prevalga il sentimento delle preoccupazione ("vittime", "ospedale", "malattie", "efficacia", "vaccinarsi"), mentre nel terzo emerga la rabbia per gli obblighi imposti dal governo e, conseguentemente, verso il governo stesso ("obbligovaccinale", "squallido", "imbecilli", "mattarella");

- dal raffronto tra il secondo e il terzo periodo (Figura 1.5) si deduce come: dal secondo sembra emergere un sentimento di speranza ("grazie", "scienza", "speriamo"), mentre nel terzo continui a prevalere il sentimento di rabbia ("criminale", "norimberga", "regime").

Al fine di determinare una misura di probabilità di ciascun vocabolo in un periodo rispetto a un altro si è fatto ricorso al log-rapporto delle quote:

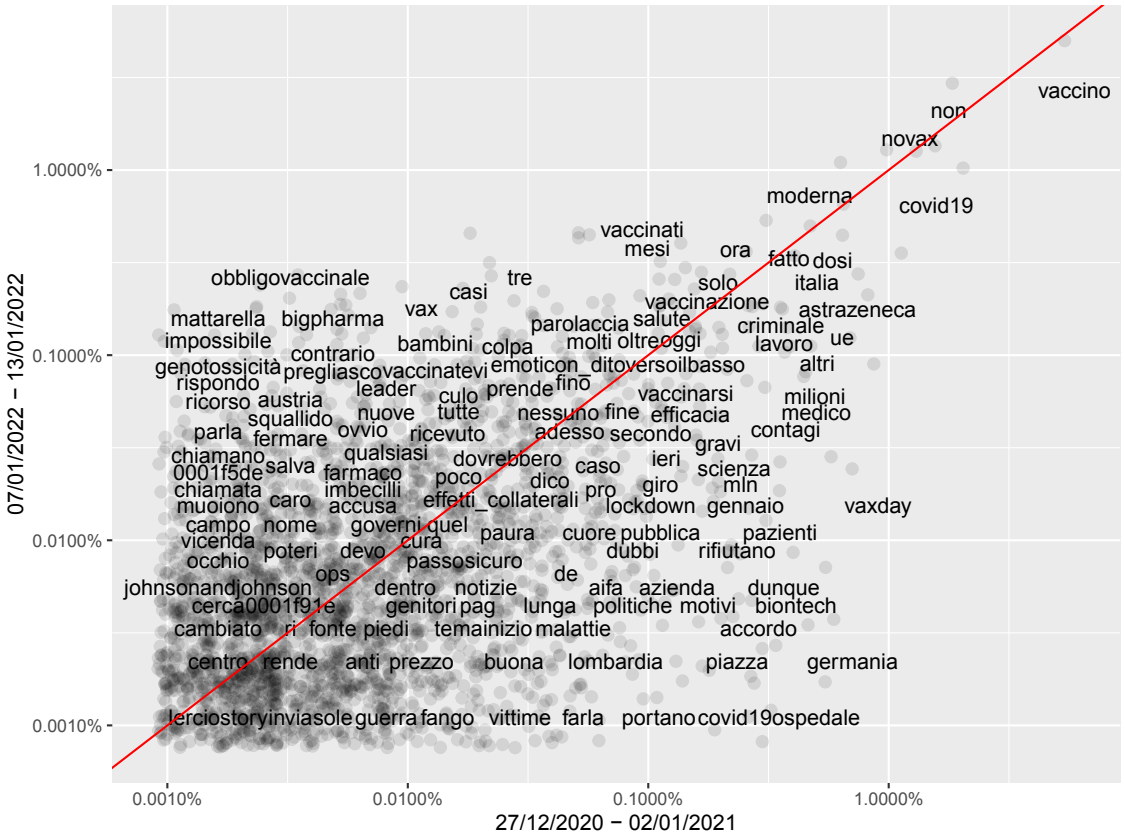
$$log \frac{\frac{n_j+1}{totale_j+1}}{\frac{n_i+1}{totale_i+1}} \tag{1.1}$$

Figura 1.3: Grafico di dispersione delle frequenze delle forme grafiche nel primo e secondo periodo



con n_i e n_j, la frequenza con cui il vocabolo di interesse compare nel periodo i e nel periodo j, e $totale_i$ e $totale_j$, la frequenza complessiva delle parole usate nel periodo i e j. I risultati ottenuti sono stati riportati nelle Figure 1.6, 1.7 e 1.8. Dalla prima si vede come le parole più positive nei confronti dei vaccini ("vaxday", "civiltà", "indispensabilità") abbiano maggiore probabilità di provenire dal primo *set* di dati che dal secondo, mentre le forme che fanno riferimento al farmaco di AstraZeneca e ai suoi effetti collaterali ("trombosi", "astrazeneca") dal secondo insieme più che dal primo. Dalla seconda figura si evince come le parole legate al Presidente del Consiglio Draghi ("draghi", "odio_draghi"), al numero di dosi fatte ("primadosefatta", "primadose", "terzadosefatta",) e alle tipologie di *green pass* abbiano maggiore probabilità di provenire dal terzo periodo che dal primo. Infine, dall'ultima figura si vede come gli stilemi "odio_draghi", "terzadose", "boster", abbiano più probabilità di provenire dal terzo periodo che dal secondo, mentre le parole "trom-

Figura 1.4: Grafico di dispersione delle frequenze delle forme grafiche nel primo e terzo periodo

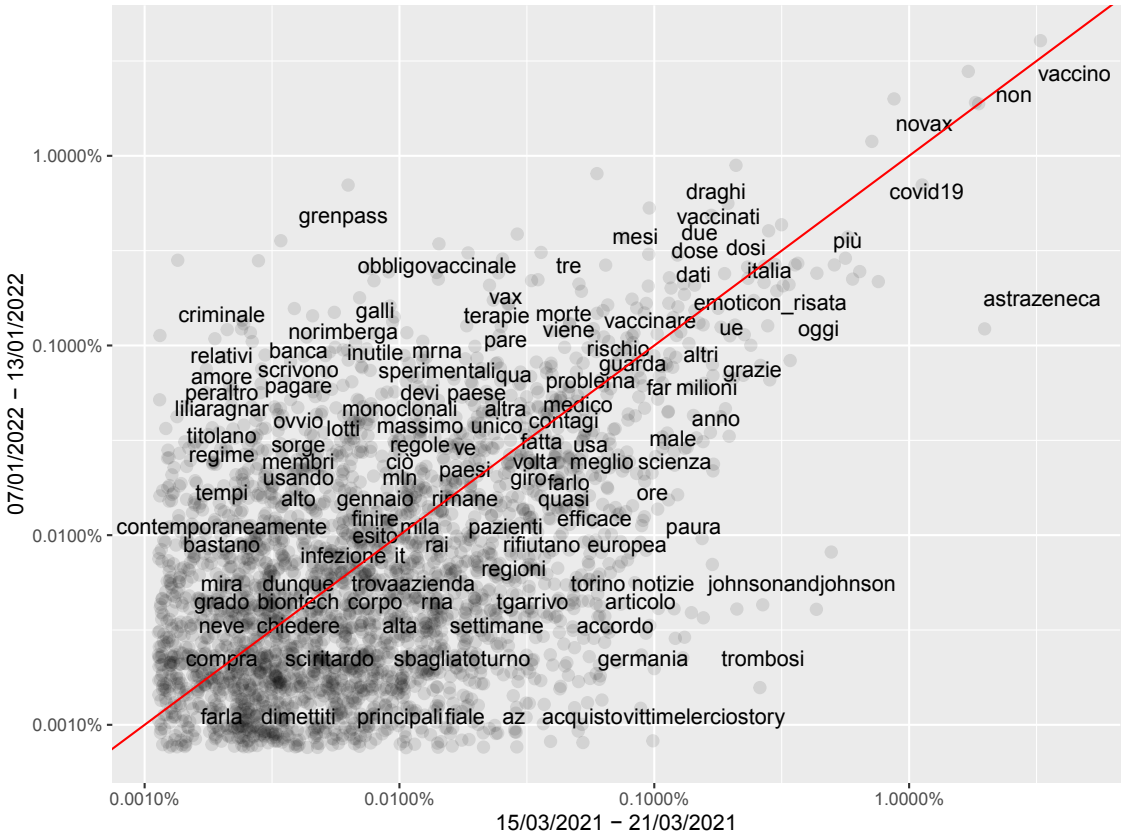


bosi" e "aifa" provengano con maggior probabilità dal secondo periodo che dal terzo.

Dalle osservazioni appena svolte si può trarre una primissima e preliminare considerazione, che poi, ovviamente, potrà essere confermata o smentita dalle analisi successive: parrebbe che, in punto di sentimenti, vi sia stata, con il passare del tempo, una progressiva deriva verso forme di rabbia e odio nei confronti dei vaccini e del governo. L'iniziale euforia e la gioia - derivanti prima dalla scoperta dei vaccini, poi dell'inizio della campagna vaccinale e dei primi risultati positivi con riguardo al numero di contagi - sono state gradualmente rimpiazzate dalla preoccupazione per i possibili effetti collaterali di detti farmaci e dal risentimento verso i sempre più pervasivi obblighi e le pesanti restrizioni adottate dalle istituzioni.

In conclusione, i testi sono stati trasformati in una matrice che prende il nome di *Document Term Matrix* (DTM), la quale contiene le frequenze

Figura 1.5: Grafico di dispersione delle frequenze delle forme grafiche nel secondo e terzo periodo



delle parole presenti nei testi. In detta matrice le righe corrispondono ai documenti, le colonne corrispondono agli stilemi. Nel caso in esame ne sono state realizzare tre, una per ciascuno periodo. Quanto alla loro dispersione si ha che la prima è densa al 0.233%, la seconda al 0.216% e la terza al 0.219%.

Passando allo studio delle altre variabili, si è controllato quanti *tweet* sono presenti per ciascun utente e si è visto che l'utente più produttivo all'interno del primo periodo presenta 27 testi; quello all'interno del secondo presenta 15 testi; infine, quello all'interno del terzo presenta 30 testi.

In aggiunta, sono stati rilevati i valori medi e mediani di altre variabili, quali: il numero di *like*, il numero di *retweet*, il numero di commenti e il numero di volte in cui è stato citato un *tweet* (Tabella 1.8). Risulta subito evidente come, in tutti i casi, si abbiano dei valori mediani nulli.

Figura 1.6: Parole che hanno maggiore probabilità di provenire dal primo periodo (nero) o dal secondo periodo (grigio) secondo il log-rapporto di quote

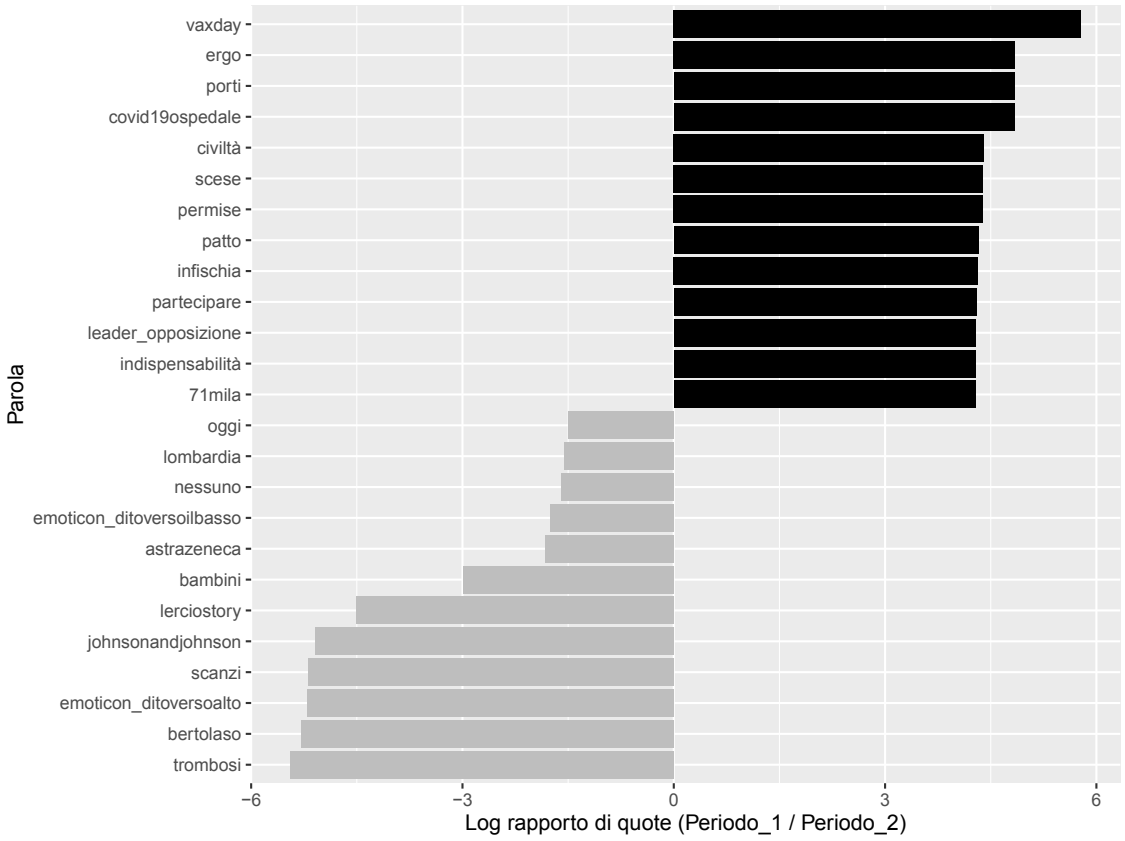


Tabella 1.8: Informazioni aggiuntive legate ai *tweet* dei tre periodi

Periodo	***Variabile***	***Max***	***Min***	***Media***	***Mediana***
1	n° *like*	996	0	2.26	0
	n° *retweet*	147	0	0.44	0
	n° commenti	89	0	0.25	0
	n° citazioni	20	0	0.05	0
2	n° *like*	374	0	1.71	0
	n° *retweet*	125	0	0.37	0
	n° commenti	61	0	0.18	0
	n° citazioni	7	0	0.03	0
3	n° *like*	1380	0	1.57	0
	n° *retweet*	405	0	0.42	0
	n° commenti	64	0	0.16	0
	n° citazioni	11	0	0.03	0

Figura 1.7: Parole che hanno maggiore probabilità di provenire dal primo periodo (nero) o dal terzo periodo (grigio) secondo il log-rapporto di quote

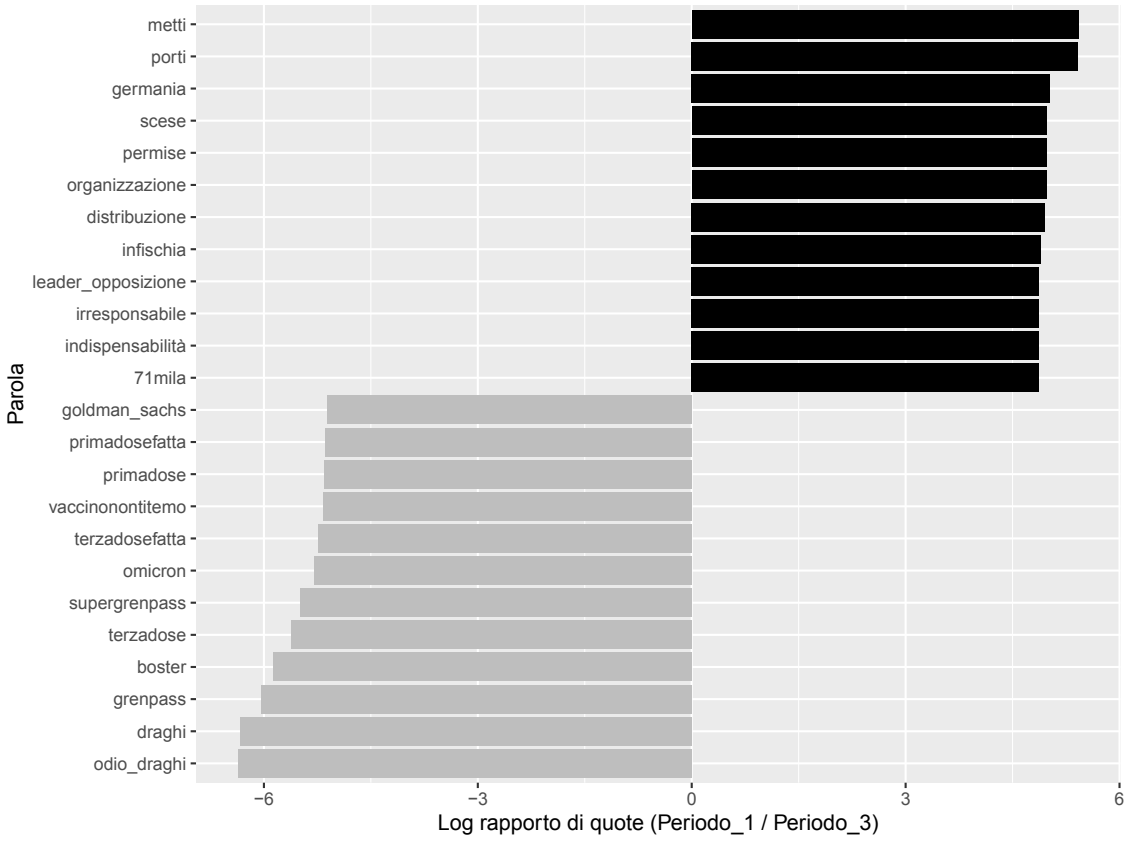
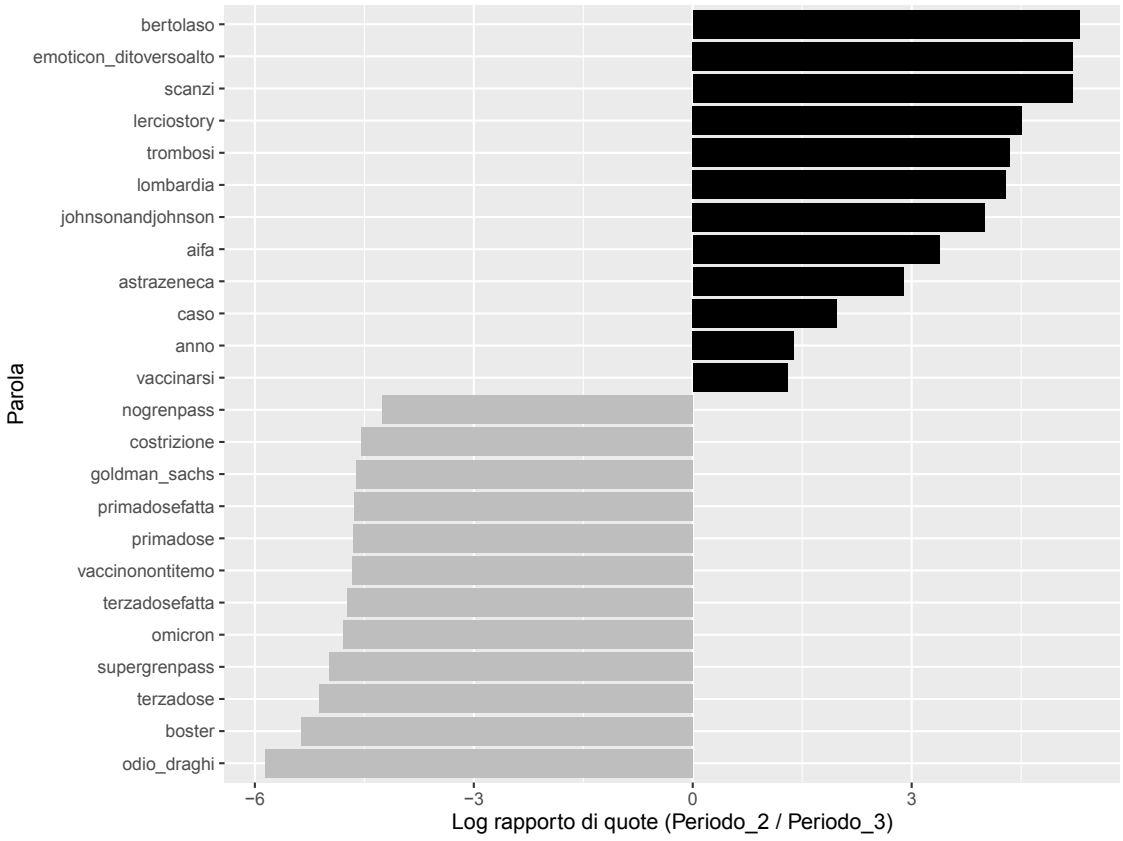


Figura 1.8: Parole che hanno maggiore probabilità di provenire dal secondo periodo (nero) o dal terzo periodo (grigio) secondo il log-rapporto di quote



Capitolo 2

Analisi non supervisionate

2.1 Obiettivi

Al fine di riconoscere gli argomenti di discussione e le opinioni espresse nei *tweet* dei tre periodi, si è fatto ricorso, in prima analisi, all'applicazione di metodi non supervisionati. Sono così definiti in statistica degli approcci volti a individuare, in un insieme tipicamente multivariato di dati, una o più dimensioni latenti sottostanti. Tra questi, si considerano in questo lavoro i metodi di *clustering*, il cui obiettivo è quello di identificare gruppi omogenei di osservazioni.

Poiché i dati di interesse, in forma di *Document Term Matrix*, derivano dalla prevalente osservazione di variabili dicotomiche che descrivono la presenza o assenza di un termine dentro un *tweet*, ai metodi di *clustering* è stata affiancata anche l'applicazione di tecniche non supervisionate di estrazione di variabili, volte a produrre una rappresentazione numerica di dati anche qualitativi.

2.2 Metodi per l'individuazione di variabili latenti

2.2.1 Panoramica generale

Tra i metodi statistici per l'individuazione di variabili latenti, l'attenzione è ora rivolta alle tecniche di estrazione (o proiezione) delle variabili, volte a creare nuove dimensioni, attraverso trasformazioni algebriche di quelle osservate, in modo tale da riuscire a lavorare, potenzialmente in uno spazio di dimensione ridotta. Queste risultano particolarmente utili anche nel caso in esame, in quanto consentono di ottenere delle rappre-

sentazioni numeriche di variabili che per loro natura sono categoriali. Alcune tra le più famose tecniche di proiezione sono:

- l'Analisi delle Componenti Principali (*Principal Component Analysis*, PCA), basata sulla scomposizione in valori singolari oppure sulla decomposizione spettrale, che si occupa di costruire, attraverso delle combinazioni lineari delle p variabili quantitative originarie, un insieme di nuove p variabili, chiamate componenti principali, che hanno varianza massima e sono incorrelate tra loro;

- il *Multi-Dimensional Scaling* (MDS), basato sulla minimizzazione di una funzione chiamata "funzione di *stress*", che ha l'obiettivo di determinare, a partire dalla matrice delle distanze o dissimilarità tra i dati in *input*, una nuova rappresentazione dei dati utilizzando un numero inferiore di coordinate, le quali preservino il più possibile le distanze (Ayesha *et al.*, 2020);

- l'Analisi delle Corrispondenze (CA), basata sulla scomposizione in valori singolari, che permette di individuare delle nuove variabili numeriche a partire da un insieme di variabili categoriali. Tale metodo verrà spiegato più nel dettaglio nella prossima sezione, perchè largamente impiegato nelle analisi che seguono.

In ultimo, risulta importante rimarcare il fatto che l'analisi delle componenti principali è, per costruzione, adatta per dati di natura quantitativa, mentre l'MDS può essere utilizzato per qualsiasi tipo di variabile, sia quantitativa, sia qualitativa. La CA, invece, è adatta esclusivamente a dati qualitativi. Per ulteriori approfondimenti, si veda Zebari *et al.* (2020).

2.2.2 Analisi delle corrispondenze

L'analisi delle corrispondenze (*Correspondence Analysis*, CA) rappresenta un metodo, basato sulla scomposizione in valori singolari, per la creazione di nuove variabili numeriche a partire da un *set* di variabili categoriali. Esistono due tipologie di analisi delle corrispondenze: la CA (semplice), che permette di analizzare due sole variabili; la *Multiple correspondence analysis* (MCA), che consente di studiare più di due variabili. Quest'ultima, però, esula dai fini di questo lavoro.

Concentrandosi, dunque, sulla prima, l'oggetto analizzato è la matrice di contingenza, la quale contiene il numero di volte in cui le modalità

di due variabili sono state usate insieme. L'obiettivo è quello di individuare i legami che intercorrono tra le modalità delle due variabili e per fare ciò si studiano le relazioni tra le righe, tra le colonne e tra le righe e le colonne. Tali legami vengono presentati graficamente: le righe e le colonne, a seguito di trasformazioni algebriche, vengono considerate come punti geometrici all'interno di due, distinti, spazi multidimensionali. In ciascuno spazio questi punti creano delle nuvole, ma per riuscire a coglierne le strutture è necessario proiettarle in un sottospazio ridotto, come, ad esempio, un piano. Tuttavia, grazie alle trasformazioni algebriche applicate sia alle righe che alle colonne, si riescono a far coincidere i due sottospazi e, dunque, si ottiene un'unica mappa nella quale è possibile studiare la vicinanza tra le proiezioni di questi punti.

Nel seguito si analizzerà il metodo da un punto di vista più formale, facendo riferimento alle sole righe della matrice di contingenza $\underset{n\times p}{\mathrm{X}} = [x_{ij}]$. Per rappresentare anche le colonne, invece, basterà ripercorrere i medesimi passaggi sulla matrice X^T.

Dapprima, bisogna dividere ogni riga per la sua marginale, ottenendo i cosiddetti profili riga; la matrice che li contiene è definita come $R = diag\{X \underset{p\times 1}{1}\}^{-1}X$. In questo modo i testi potranno essere descritti mediante le frequenze relative delle parole utilizzate. Poi, per individuare le differenze o le somiglianze tra le righe è necessario definire un profilo medio, detto baricentro (o centroide o centro di gravità) di riga $c^T = (\underset{1\times n}{1} X \underset{p\times 1}{1})^{-1} \underset{1\times n}{1} X$ e, poi, studiare la matrice $Y = R - (\underset{n\times 1}{1} c^T)$ delle deviazioni dal profilo medio. Proseguendo, si assegna a ciascuna riga una massa $m = (\underset{1\times n}{1} X \underset{p\times 1}{1})^{-1} X \underset{p\times 1}{1}$, che esprime la sua importanza nel campione, e a ciascuna colonna un peso $w = [w_j] = [c_j^{-1}]$, che esprime il suo potere discriminante tra le righe.

Grazie alla notazione introdotta poc'anzi, è possibile inquadrare la questione come un problema di decomposizione a valori singolari generalizzata. Y, infatti, viene scomposta nel seguente modo:

$$Y = P\Delta Q^T \text{ soggetto a } P^T D_m P = Q^T D_c^{-1} Q = I, \qquad (2.1)$$

con P la matrice dei vettori singolari destri, Q la matrice dei vettori singolari sinistri, Δ la matrice diagonale degli autovalori, D_m una matrice diagonale che contiene le masse e D_c^{-1} una matrice diagonale dei pesi delle colonne. Le righe della matrice X sono rappresentate dai loro punteggi fattoriali, ossia le proiezioni delle unità statistiche sui vettori singolari di X, i quali sono contenuti all'interno della matrice $F = P\Delta$,

di dimensioni n x L (con L il numero di valori singolari non nulli). Ora è possibile rappresentare graficamente i risultati facendo un grafico dei punteggi fattoriali, nel quale ogni punto rappresenta una riga della matrice X. Per definire la distanza tra due profili riga, i e $i^{'}$, si utilizza la distanza del χ^2, definita come

$$d^2_{i,i'} = \sum_{l}^{L} (f_{i,l} - f_{i',l})^2, \tag{2.2}$$

con $f_{i,l}$ i punteggi fattoriali e L il numero dei fattori derivanti dalla CA.

Per determinare la variabilità dei profili riga rispetto al baricentro c si utilizza l'inerzia I, definita come,

$$I = \sum_{i=1}^{n} m_i d^2_{c,i} = \sum_{l}^{L} \lambda_l, \tag{2.3}$$

con λ_l un autovalore.

Per ulteriori approfondimenti si vedano Cilione (2011) e Abdi e Williams (2022).

2.3 *Clustering*

I metodi di analisi dei gruppi o *clustering* si occupano di allocare un determinato numero di elementi, non precedentemente classificati, all'interno di *cluster*, genericamente definibili come insiemi di elementi simili tra loro e, contemporaneamente, dissimili da quelli appartenenti ad altri.

In virtù della generalità dell'obiettivo del *clustering*, la letteratura riporta una pluralità variegata di approcci. In questa relazione l'attenzione è stata limitata ai metodi classici di stampo prevalentemente euristico e basati sui concetti di distanza e dissimilarità tra due unità statistiche. Si supponga di avere a disposizione un *dataset* con n righe e p colonne e di essere interessati a determinare quanto siano simili due unità statistiche, $x_i = (x_{i1}, ..., x_{ip})$ e $x_r = (x_{r1}, ..., x_{rp})$, descritte per mezzo delle p variabili. Per raggiungere l'obiettivo si dovrà scegliere, sulla base delle variabili a disposizione e delle caratteristiche del fenomeno, se utilizzare una misura di distanza o un indice di dissimilarità. Dati due vettori, x_i e x_r, si definisce distanza una funzione che soddisfa le seguenti proprietà:

- $d(x_i, x_r) = d(x_r, x_i)$, simmetria;
- $d(x_i, x_r) \geq 0$, non negatività;
- $d(x_i, x_r) = 0$ se e solo se $x_i = x_r$, identità;
- $d(x_i, x_r) \leq d(x_i, x_s) + d(x_s, x_r)$, disuguaglianza triangolare.

Nel caso in cui siano verificate le prime tre proprietà si ha a disposizione un indice di dissimilarità, mentre nel caso in cui siano tutte valide si parla di distanza metrica.

La distanza metrica più nota è quella euclidea, definita come

$$d_2(x_i, x_r) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{rk})^2}. \tag{2.4}$$

Un esempio di indice di dissimilarità è quello del coseno:

$$d_{ir} = 1 - cos\alpha = 1 - \frac{x_i^t x_r}{||x_i|| ||x_r||}. \tag{2.5}$$

Un ulteriore indice di dissimilarità è quello di Gower, utile nel caso di variabili di tipo misto,

$$d_G(x_i, x_r) = 1 - \frac{\sum_{j=1}^{p} \delta_{ir}(j) s_{ir}(j)}{\sum_{j=1}^{p} \delta_{ir}(j)} \tag{2.6}$$

con

$$s_{ir}(j) = \begin{cases} 1 - \frac{|x_{ij} - x_{rj}|}{\text{range j-esima variabile}} & \text{se } x_j \text{ è quantitativa} \\ I(x_{ij} = x_{rj}) & \text{se } x_j \text{ è binaia/nominale} \\ 1 - |x_{ij} - x_{rj}| & \text{se } x_j \text{ è ordinale} \end{cases} \tag{2.7}$$

e

$$\delta_{ir}(j) = \begin{cases} 1 & \text{se } i, r \text{ confrontabili rispetto } x_j \\ 0 & \text{altrimenti.} \end{cases} \tag{2.8}$$

Due unità statistiche sono confrontabili rispetto alla variabile x_j se è presente almeno un valore nullo in una delle due oppure se si ha coassenza 0-0 e la variabile j è binaria.

A partire da un campione di n unità statistiche x_i, $i = 1, \ldots, n$, si definisce matrice di dissimilarità (distanza) una matrice nxn che riporta nella cella (i, j) la dissimilarità (distanza) tra x_i e x_j. Per altri approfondimenti su distanze e dissimilarità si vedano Solari (2022) e Azzalini e Scarpa (2012).

Tornando all'analisi dei gruppi, le tecniche di *clustering* vengono principalmente suddivise in gerarchiche, nelle quali i dati, sulla base della matrice di dissimilarità, vengono organizzati in una struttura gerarchica, e partizionali, nelle quali i dati vengono assegnati a K *cluster* senza alcuna struttura gerarchica. I metodi gerarchici si suddividono a loro volta in agglomerativi e divisivi.

Nei metodi agglomerativi, si parte con n singolette (ossia n *cluster* dove ciascuno di essi contiene una sola unità) e si procede con l'unirli due per volta, fino ad ottenere un unico *cluster* che racchiude l'intero campione. L'aggregazione tra *cluster* è determinata sulla base di una opportuna distanza tra *cluster*. Ad esempio, nel metodo del legame singolo, la distanza tra due *cluster* è determinata tramite il minimo delle distanze tra le unità statistiche appartenenti a due *cluster*; il metodo del legame completo considera il massimo tra tali distanze; e il metodo del legame medio ne determina la media.

Nei metodi divisivi, invece, si parte dall'intero campione e si procede per suddivisioni successive, fino ad ottenere un *cluster* per ciascuna unità statistica. Nel contesto del *Text Mining*, è spesso usato il metodo divisivo, noto come Classificazione Gerarchica Discendente (CGD), introdotto da Reinert (1983) per il raggruppamento di unità statistiche su cui sono osservate variabili binarie (quindi $x_{ij} = 1$ se il vocabolo j è presente all'interno del testo i, 0 altrimenti). Benché usato nell'analisi dei testi, ha senso anche per altri dati, purché binari. A partire dal singolo gruppo, composto da tutte le unità statistiche, si determinano tutte le possibili partizioni delle unità in due gruppi, $g = 1, 2$, e, tra queste, si seleziona quella che massimizza la distanza

$$\chi^2 = \sum_{g=1}^{2} \sum_{j=1}^{p} \frac{(f_j(g) - \frac{\sum_{j=1}^{p} f_j(g)}{N} s_j)^2}{f_j(g)}, \tag{2.9}$$

con $f_j(g)$, la frequenza del vocabolo j nella classe g, N, la frequenza complessiva delle parole, e s_j, la frequenza con cui il vocabolo j compare nella matrice. Successivamente si suddivide il gruppo più numeroso usando lo stesso procedimento. Si procede fintanto che non si raggiunge il numero di classi prestabilito. La suddivisione sarà tanto migliore

quanto meno si avrà la condivisione di termini tra le varie classi. Per ulteriori approfondimenti sul metodo di Reinert si vedano anche Lapalut (1995) e Corridoni (2019).

Quanto alle criticità dei metodi gerarchici classici si citano: la poca robustezza; il fatto che una volta che una unità viene assegnata a un *cluster* non può più essere riallocata; l'elevato costo computazionale.

Tra le tecniche di *clustering* partizionali il metodo più famoso è il k-*means*, il quale, nella sua versione più semplice, può essere così descritto:

- si suddividono casualmente le unità statistiche in K *cluster* e si calcolano le medie dei gruppi (centroidi);
- si assegna l'unità statistica al *cluster* con la media più vicina. La distanza viene solitamente calcolata tramite la distanza euclidea;
- si ricalcolano le medie dei *cluster* che hanno perso o che hanno guadagnato una unità;
- si ripetono il secondo e terzo passaggio fino a quando non si contano più cambiamenti.

Il metodo k-*means* soffre di alcune criticità, come: la necessità di stabilire a priori il numero di gruppi; la non garanzia di trovare un ottimo globale; la poca robustezza nel caso di valori anomali. Nel corso del tempo, però, sono state proposti diversi miglioramenti: per esempio, per rendere la procedura più robusta è stato ideato il k-*medoids* che utilizza la mediana come centroide dei *cluster*.

Per ulteriori approfondimenti si vedano Xu e Wunsch (2005) e Johnson e Wichern (2007).

2.4 Applicazione

Al fine di produrre una prima descrizione esplorativa dei temi trattati nei *tweet* nell'arco dei tre periodi, si è svolta un'analisi delle corrispondenze a partire dalla tabella di contingenza, che mette in relazione le frequenze d'uso di ciascuno dei termini del vocabolario usato nei *tweet* con il periodo di riferimento. Un estratto di tale frequenza è riportato nella Tabella 2.1.

La Figura 2.1 riporta una rappresentazione grafica delle prime due coordinate identificate dalla CA per caratterizzare i diversi lemmi del

Tabella 2.1: Frequenze dei vocaboli all'interno di ciascuno dei tre periodi di riferimento (per motivi di spazio vengono riportate solo le prime colonne)

Periodi-vocaboli	***nuovi***	***persona***	***prendere***	***prima***	***...***
Periodo 1	16	278	9	212	...
Periodo 2	12	142	32	145	...
Periodo 3	31	146	21	122	...

vocabolario. Sebbene sia evidente una generale omogeneità di rappresentazione, in cui non spiccano differenze sostanziali tra gli argomenti discussi nei tre periodi, è possibile trarre le considerazione di massima che seguono:

- la prima dimensione individuata, ossia l'asse orizzontale, sembra essere legata al tempo, in quanto i tre periodi sono presentati in ordine cronologico, da sinistra verso destra. Tale dimensione separa nettamente il primo e il terzo periodo. La seconda dimensione, cioè l'asse verticale, invece, isola il secondo periodo;

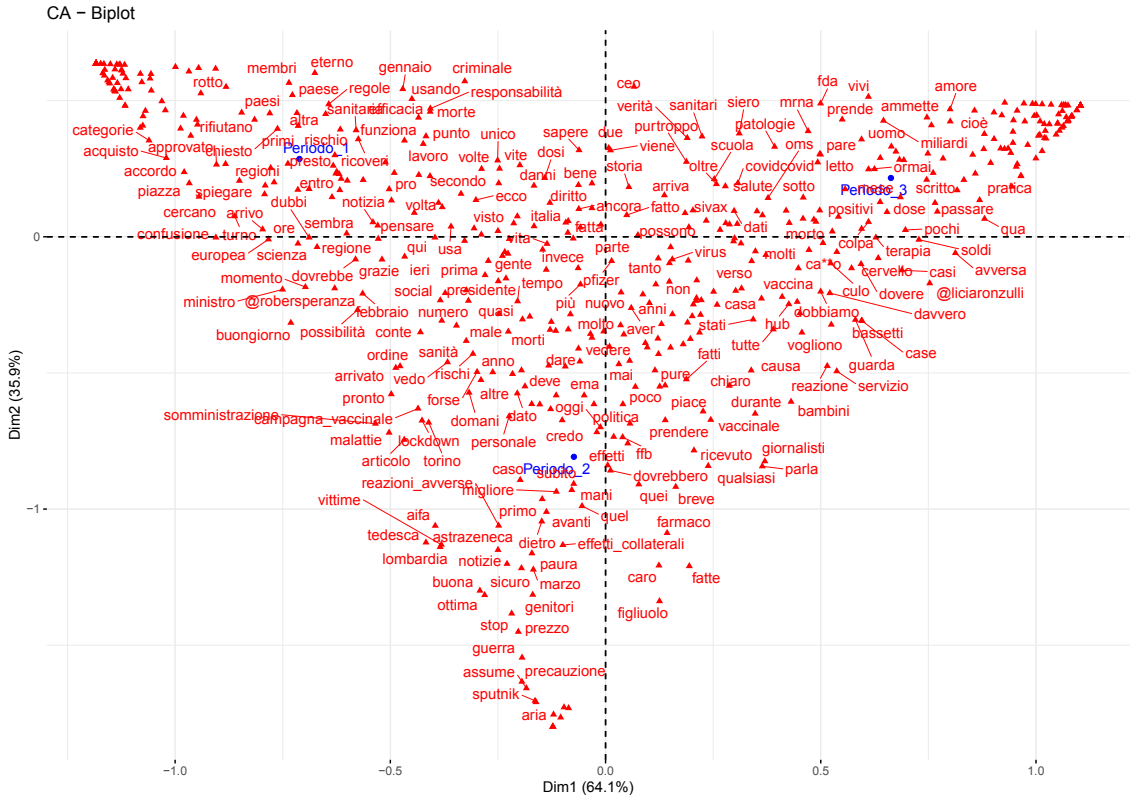
- il primo periodo è caratterizzato da termini che sembrano essere legati: a una incertezza generale sul tema dei vaccini ("confusione", "dubbi", "spiegare"); al senso del dovere ("responsabilità"); alla positività circa i vaccini ("pro", "funziona"); ad aspetti pratici legati alle vaccinazioni, come gli accordi per l'acquisto dei vaccini ("accordo", "acquisto") e alla suddivisione della popolazione in vari gruppi sulla base del rischio ("categorie");

- il secondo periodo è incentrato principalmente sui fatti di cronaca legati al vaccino di AstraZeneca e si trovano per la maggior parte riferimenti ai suoi effetti collaterali ("reazioni avverse", "effetti collaterali", "malattie") e alla conseguente paura collettiva ("paura", "precauzione");

- nel terzo periodo compaiono termini troppo generici per individuare degli argomenti di discussione. L'unico tema che parrebbe emergere con più forza tra i molti è legato ai "bambini", ma è possibile che il riferimento derivi dal rientro a "scuola".

Successivamente, si è deciso di procedere con l'analisi dei gruppi. Le analisi sono state svolte sia sui dati originali, ovvero le DTM dei

Figura 2.1: Esito dell'applicazione dell'analisi delle corrispondenze alla tabella di contingenza che mette in relazione le frequenze d'uso di ciascuno dei termini del vocabolario usato nei *tweet* con il periodo di riferimento. In blu vengono rappresentate le righe, in rosso le colonne



tre periodi, sia su delle loro trasformazioni ottenute tramite il *multidimensional scaling* e l'analisi delle corrispondenze. Un riassunto delle metodologie utilizzate è presente all'interno della Tabella 2.2. Per ogni tipologia di dati - tanto quelli originali, quanto quelli frutto di una mappatura in uno spazio di dimensione ridotta - sono stati utilizzati diversi algoritmi di *clustering* e per ciascuno di essi si sono fatte varie prove, cambiando il numero dei *cluster*. Inoltre, al fine di creare la matrice di prossimità necessaria all'inizializzazione degli algoritmi gerarchici, si sono utilizzate diverse metriche.

Le procedure che hanno fornito i risultati più soddisfacenti sono quelle gerarchiche, in particolare il metodo CGD di Reinert, che è stato applicato direttamente ai dati originali, e il metodo di Ward, il quale ha ricevuto in input una matrice di dissimilarità costruita a partire dai tweet sottoposti all'analisi delle corrispondenze. Per la costruzione della

Tabella 2.2: Riassunto delle operazioni svolte nell'ambito dell'analisi dei gruppi

Dati	***Clustering***	***N° cluster***	***Metriche***
Originali	*K-means* *Spherical k-means* *K-medoids* Metodo Ward Metodo Reinert	da 2 a 15	Distanza euclidea Indice Gower Similarità coseno
Ridotti MDS	*K-means* *Spherical k-means* *K-medoids* Metodo Ward	da 2 a 15	Distanza euclidea Indice Gower Similarità coseno
Ridotti CA	*K-means* *Spherical k-means* *K-medoids* Metodo Ward	da 2 a 15	Distanza euclidea Indice Gower Similarità coseno

matrice di dissimilarità si è fatto uso dell'indice di Gower.

Queste tecniche hanno portato a dei risultati sostanzialmente sovrapponibili. Per tale ragione verranno discusse esclusivamente le risultanze ottenute con il metodo di Reinert.

Il *clustering* divisivo di Reinert, implementato all'interno del pacchetto `rainette` (Barnier, 2022), è stato utilizzato distintamente su tutti e tre i periodi: i risultati ottenuti sono presenti all'interno delle Figure 2.2, 2.3 e 2.4 che riportano il dendrogramma del partizionamento. Al fine di interpretare la composizione dei gruppi, sono riportati anche le numerosità dei *cluster*, i termini più frequenti di ciascuno e, per alcuni, la rete delle co-occorrenze. Dalle figure emerge immediatamente la presenza di un gruppo estremamente numeroso e di tanti altri nettamente meno popolati. Questa situazione si è presentata con tutte le tecniche di *clustering* provate e si è mantenuta anche all'aumentare del numero dei gruppi. Tale distribuzione potrebbe essere dovuta ai periodi e agli intervalli temporali dai quali si è deciso di estrarre i *tweet*: questi ultimi,

in tali lassi di tempo, erano, molto probabilmente, incentrati su di un unico fatto e l'opinione di coloro che scrivevano su Twitter era omogenea e verteva, in gran parte, sullo specifico episodio di cronaca.

Tornando al *clustering* basato sul metodo di Reinert, per quanto concerne il primo periodo (Figura 2.2), si nota che:

- il gruppo più numeroso (ossia il terzo *cluster*) è inevitabilmente il più eterogeneo. Spiccano, tuttavia, discussioni sull'episodio dell'acquisto di dosi di vaccino da parte della Germania, la quale ha fatto degli accordi con le case farmaceutiche senza considerare le direttive dell'Unione Europea. Emerge, oltre alla rabbia nei confronti della Germania e a considerazioni sulla debolezza dell'Unione Europea, quanto gli italiani fossero impazienti di ricevere i lotti di vaccino, in modo tale da potersi vaccinare;

- i gruppi che vanno dal quattro al dieci individuano dei micro argomenti di discussione, tutti, comunque, a favore dei vaccini;

- il primo e il secondo *cluster* sono quelli che contengono *tweet* scritti probabilmente da persone *novax*.

Quanto al secondo periodo (Figura 2.3), si nota che:

- il gruppo più numeroso, cioè il quinto gruppo, tratta principalmente del vaccino AstraZeneca, con particolare riguardo ai suoi effetti collaterali. Tuttavia, leggendo i *tweet* appartenenti a questo gruppo e analizzando la rete e le parole più rappresentative, si osserva che anche in questo caso il gruppo è eterogeneo e non emerge una netta prevalenza di *tweet* a favore o contro i vaccini;

- gruppi come il terzo, il settimo e il nono contengono *tweet* prettamente di stampo *novax*;

- altri gruppi, come il primo, il secondo, il quarto e il sesto contengono prevalentemente testi *provax*.

Infine, dal *clustering* sul terzo periodo (Figura 2.4), si nota che:

- all'interno del gruppo più numeroso, ossia il quinto, vi è una forte componente contraria ai vaccini. Nello specifico, il malcontento è rivolto verso l'obbligo vaccinale e il governo Draghi;

- all'interno di altri gruppi, come il quarto, il settimo, l'ottavo e il decimo si osserva sempre un forte risentimento, in particolare nel numero di richiami del vaccino e del Primo Ministro;

- si riesce comunque a individuare una componente, anche se piccola, di *provax* (si vedano il terzo e sesto *cluster*).

Provando a tirare le fila di queste analisi, sebbene emerga una notevole eterogeneità di opinione in tutti e tre i periodi, c'è una qualche evidenza empirica che riferisce una prevalenza di opinioni favorevoli ai vaccini nei primi periodi e una più netta opinione contraria nel terzo.

Figura 2.2: Applicazione del metodo di Reinert al primo periodo: dendrogramma e parole più rappresentative per ciascun *cluster* (sopra) e rete pesata del terzo *cluster* (sotto)

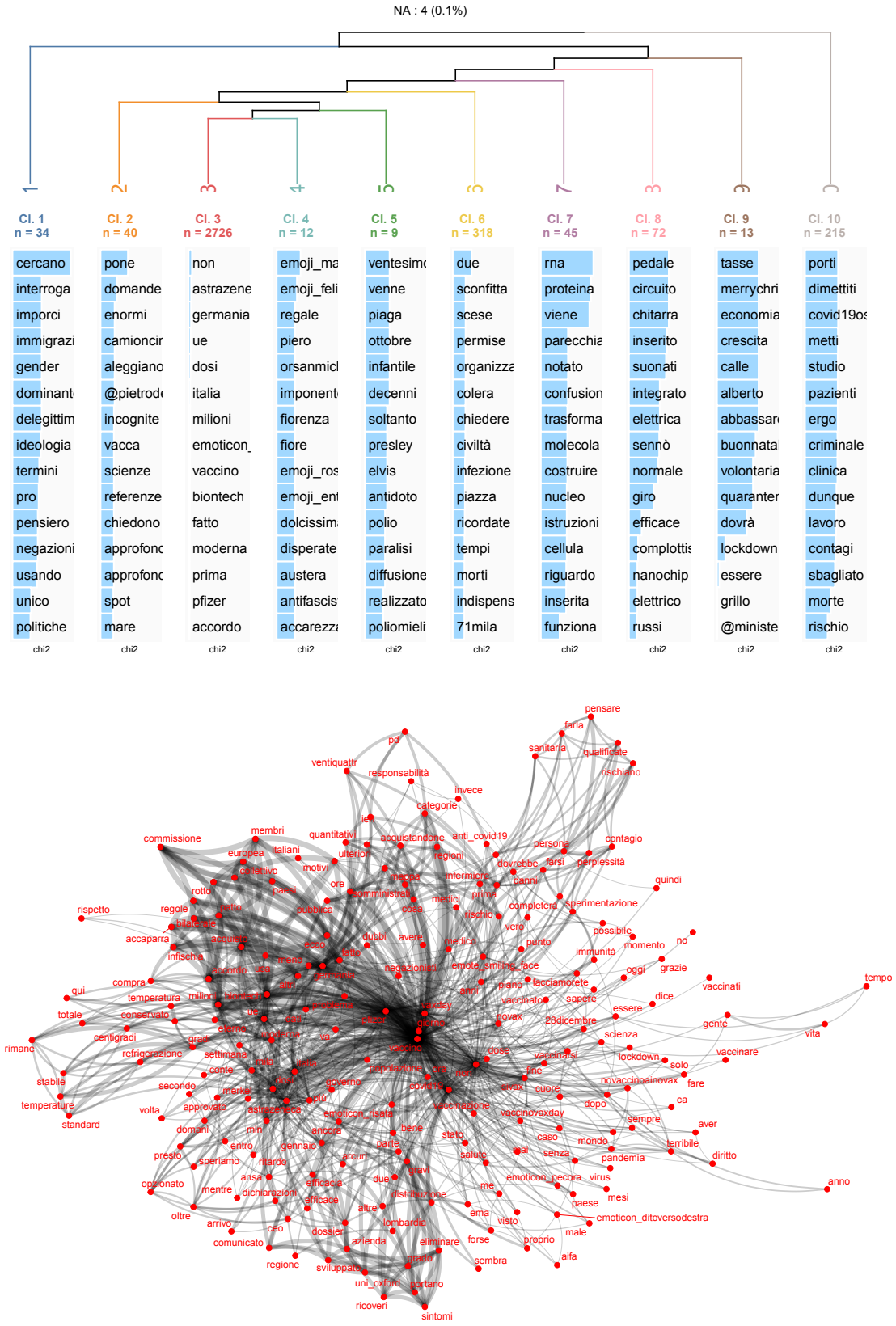
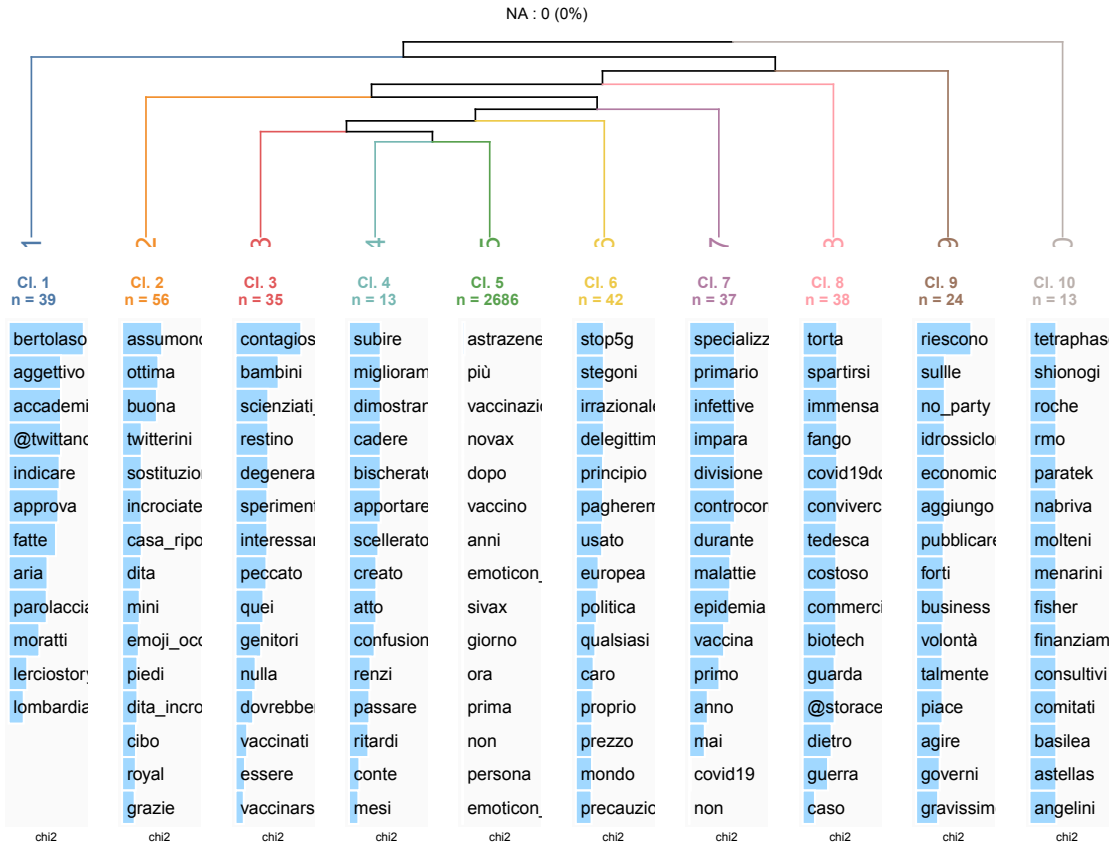


Figura 2.3: Applicazione del metodo di Reinert al secondo periodo: dendrogramma e parole più rappresentative per ciascun *cluster* (sopra) e rete pesata del quinto *cluster* (sotto)



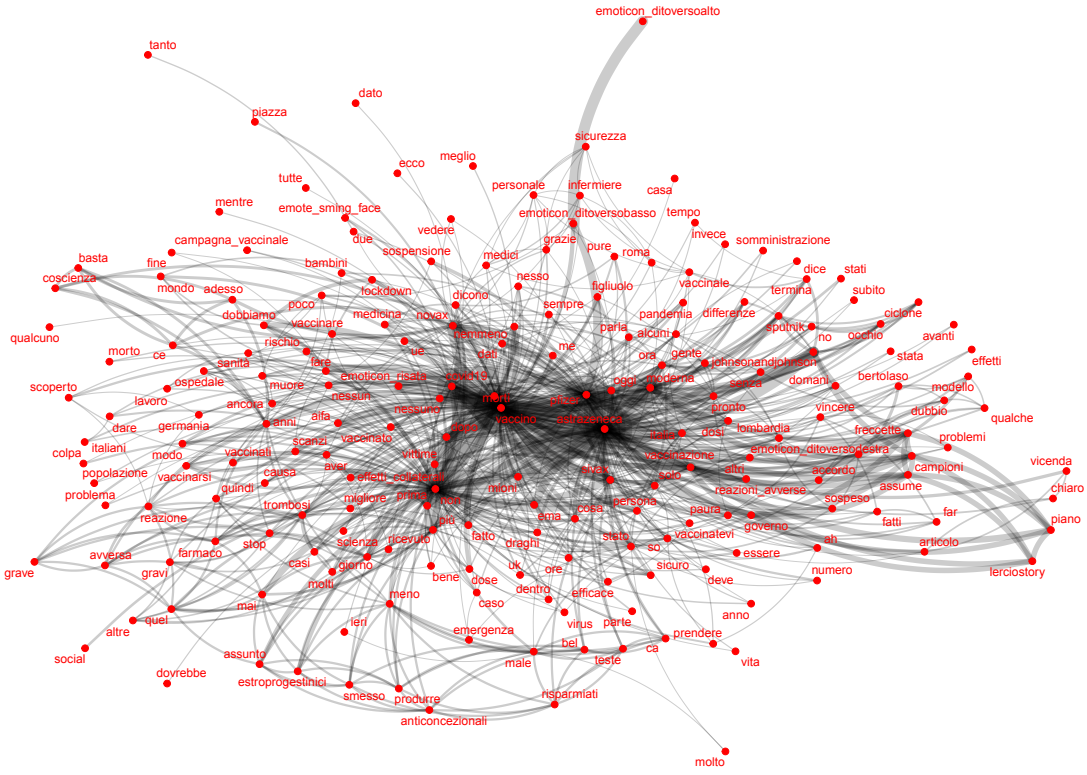
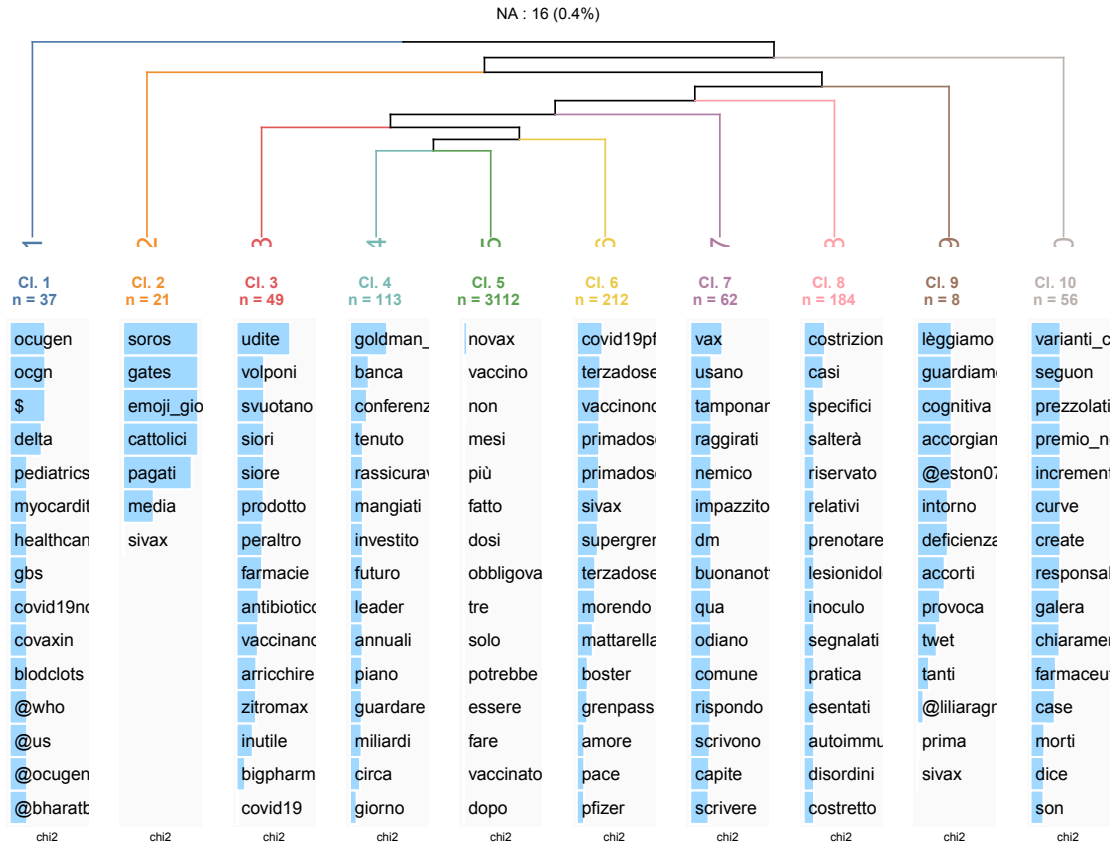
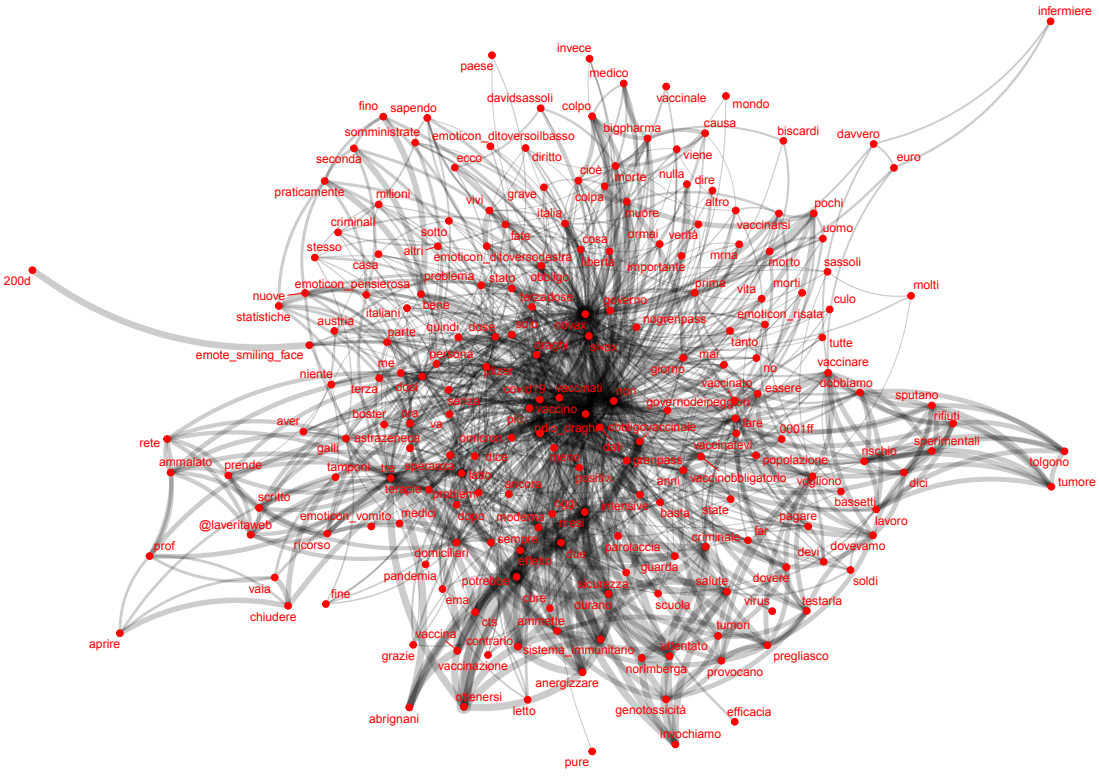


Figura 2.4: Applicazione del metodo di Reinert al terzo periodo: dendrogramma e parole più rappresentative per ciascun *cluster* (sopra) e rete pesata del quinto *cluster* (sotto)





Capitolo 3

Analisi supervisionate

3.1 Obiettivi

Un approccio alternativo a quello seguito finora è quello di cercare di classificare i *tweet* mediante l'applicazione di un metodo supervisionato.

Tali metodi si basano sull'osservazione, almeno per un campione di dati (chiamato *training set*), di una variabile risposta utilizzata in fase di stima per guidare la classificazione. Nell'applicazione in esame, non avendo a disposizione tale variabile risposta, si è proceduto alla classificazione manuale di 200 *tweet* per ciascuno dei periodi di osservazione.

3.2 Modello *logit* multinomiale

Il modello *logit* multinomiale ha lo scopo di assegnare ogni unità statistica a una specifica categoria, scelta all'interno di c possibili, utilizzando l'informazione proveniente da p variabili esplicative. Innanzitutto, si consideri la realizzazione della variabile risposta per l'i-esimo soggetto,

$$y_i = (y_{i1}, \ldots, y_{ij} \ldots, y_{ic}), \quad i = 1, \ldots, n, \tag{3.1}$$

con

$$y_{ij} = \begin{cases} 1 & \text{se il soggetto } i \text{ presenta la modalità } j \\ 0 & \text{altrimenti} \end{cases} \tag{3.2}$$

per $j = 1, \ldots, c$. Inoltre, $\sum_{j=1}^{c} y_{ij} = 1 \ \forall \ i$ in quanto ogni unità statistica è descritta da un'unica modalità della variabile dipendente. Risulta, quindi, possibile assumere che y_i sia una realizzazione di una variabile

casuale $Y_i = (Y_{i1}, ..., Y_{ic})$ con distribuzione multinomiale, la cui funzione di probabilità presenta la seguente forma

$$Pr(Y_i = y_i; \pi_i) = \pi_{i1}^{y_i 1} \dots \pi_{ic}^{y_i c}, \tag{3.3}$$

con $\pi_i = (\pi_{i1}, \dots, \pi_{ic})$ e $\pi_{ij} \in (0, 1)$, probabilità che il soggetto i-esimo presenti la modalità j. Il supporto $S = \{y_i \in \{0, 1\}^c : \sum_{j=1}^{c} y_{ij} = 1\}$ e $\sum_{j=1}^{c} \pi_{ij} = 1$.

In termini più formali, lo scopo è quello di analizzare la relazione che intercorre tra la variabile risposta Y_i, o meglio le probabilità π_{ij}, e p variabili esplicative x_{ir}, $r = 1, \dots, p$. Per raggiungere l'obiettivo si fa uso di un metodo che consiste: nell'utilizzare una delle modalità della variabile risposta come categoria di riferimento c; nel determinare per ogni j, il j-esimo *logit* rispetto alla modalità di riferimento, $log\frac{\pi_{ij}}{\pi_{ic}}$; nell'assumere che questi ultimi siano una funzione lineare dei predittori. Da queste premesse, il modello *logit* multinomiale è, allora, espresso dalla relazione

$$log\frac{\pi_{ij}}{\pi_{ic}} = x_i\beta_j = \sum_{r=0}^{p} \beta_{jr} x_{ir}, \quad j = 1, \dots, c-1, \tag{3.4}$$

con $x_i = (x_{i0}, \dots, x_{ip})$ che rappresenta il vettore delle p variabili esplicative per il soggetto i-esimo, con la convenzione che $x_{i0} = 1 \ \forall i$ in modo tale da includere un'intercetta, e $\beta_j = (\beta_{j0}, \dots, \beta_{jp})^T$ che indica il vettore dei coefficienti di regressione.

Il modello (3.4) può essere riferito direttamente alle probabilità originali, piuttosto che ai log rapporti di quote:

$$\pi_{ij} = \frac{e^{x_i\beta_j}}{1 + \sum_{h=1}^{c-1} e^{x_i\beta_h}} \quad j = 1, \dots, c-1. \tag{3.5}$$

I parametri vengono tipicamente stimati massimizzando la funzione di log-verosimiglianza

$$
\begin{aligned}
l(\beta) &= \sum_{i=1}^{n} y_{i1} x_i \beta_1 + \cdots + \sum_{i=1}^{n} y_{ic-1} x_i \beta_{c-1} \\
&\quad - \sum_{i=1}^{n} log(1 + e^{x_i \beta_1} + \cdots + e^{x_i \beta_{c-1}}) \\
&= (\sum_{i=1}^{n} y_{i1} x_i) \beta_1 + \cdots + (\sum_{i=1}^{n} y_{ic-1} x_i) \beta_{c-1} \\
&\quad - \sum_{i=1}^{n} log(1 + e^{x_i \beta_1} + \cdots + e^{x_i \beta_{c-1}}).
\end{aligned}
\tag{3.6}
$$

Per scopi predittivi si deve assegnare l'unità i alla modalità j, per la quale la probabilità stimata π_{ij} di appartenenza risulta massima. Per maggiori approfondimenti si veda, ad esempio, Salvan *et al.* (2020).

3.3 Penalizzazione dei metodi di stima: LASSO

Quando il modello di riferimento è caratterizzato da un elevato numero di parametri, una procedura diffusa in statistica è quella di introdurre una penalizzazione all'interno delle funzione obiettivo utilizzata per la stima. L'idea alla base di questo approccio è proprio quella di penalizzare i modelli che, anche a dispetto di un ottimo adattamento, abbiano determinate caratteristiche, ad esempio in termini di numero di parametri, in modo da trovare un ragionevole compromesso tra capacità di adattamento e parsimonia.

Le penalizzazioni dei metodi di stima sono utilizzate per:

- ridurre la complessità dei modelli;
- diminuire la variabilità delle stime (a discapito di un aumento della distorsione);
- consentire di ottenere delle stime anche quando vengo violate alcune assunzioni (ad esempio ci sono più variabili che unità statistiche).

Nell'ambito dei metodi supervisionati di penalizzazione, in questo lavoro si è deciso di limitare l'attenzione alla tecnica del LASSO (*Least Absolute*

Shrinkage and Selection Operator, Tibshirani (1996)), essendo questa molto usata nel caso di dati sparsi.

Mantenendo la notazione introdotta nel paragrafo precedente, un modello *logit* multinomiale con penalizzazione di tipo LASSO produce una stima dei parametri ottimizzando la log-verosimiglianza penalizzata

$$l_p(\beta) = l(\beta) - \lambda \sum_{j=0}^{p} |\beta_j| \tag{3.7}$$

dove $l(\beta)$ è la funzione di log-verosimiglianza (3.6), λ è un parametro di regolazione che definirà l'intesità della penalizzazione e $\sum_{j=0}^{p} |\beta_j|$ è la penalità assegnata alle stime.

Appare dunque evidente che la procedura tenderà a fornire modelli in cui i parametri meno rilevanti sono nulli, producendo in questo modo una selezione delle variabili.

La stima si ottiene applicando metodi numerici per un *range* di possibili valori del parametro di regolazione. Quest'ultimo viene selezionato mediante convalida incrociata. Per maggiori approfondimenti si vedano Hastie *et al.* (2009) e Azzalini e Scarpa (2012).

3.4 Applicazione

Al fine di applicare i metodi ora descritti, si è proceduto alla classificazione manuale dei *tweet* nelle seguenti categorie:

- *provax*, racchiude i *tweet* a favore dei vaccini;
- *novax*, contiene i *tweet* contrari ai vaccini;
- neutro, accoglie i *tweet* il cui contenuto non è esplicitamente schierato in nessuno dei due gruppi precedenti;
- *off-topic*, presenta i *tweet* che parlano di temi diversi da quello del vaccino.

É stato deciso di classificare 200 *tweet*, estratti casualmente, per ciascuno dei tre periodi. Successivamente, questi 600 *tweet* sono stati utilizzati per stimare e allenare il modello.

Proseguendo, è stata costruita la DTM, con pesi TF-IDF (*Term Frequency - Inverse Document Frequency*), a partire dai testi etichettati.

I pesi TF-IDF danno maggiore importanza alle parole meno frequenti a discapito di quelle più usate. Tali pesi derivano dal prodotto dei seguenti due punteggi:

- *Term Frequency*, definito come il rapporto tra la frequenza del vocabolo nel testo e il numero totale di vocaboli unici nel testo;
- *Inverse Document Frequency*, definito come il logaritmo del rapporto tra il numero totale di documenti e il numero di documenti che presentano la parola di interesse.

La funzione utilizzata per la stima del modello è stata `cv.glmnet`, contenuta nel pacchetto `glmnet` (Friedman *et al.*, 2010), scegliendo "neutro" come modalità di riferimento. La Figura 3.1 riporta una sintesi dei risultati al variare del parametro di regolazione: essa suggerisce che i valori attivi per λ siano tali che $-4.47 \leq log(\lambda) \leq -3.63$. Nelle Tabelle 3.1 e 3.2 si trovano i termini associati a coefficienti non nulli per le modalità *novax* e *provax*. Interessante notare come alcuni *tag* vengano utilizzati per catalogare i *tweet* (esempio di rilievo è il *tag* "@matteosalvini" dell'omonimo esponente politico, utilizzato per individuare i *novax*). Parole particolarmente significative per la categoria *novax* sono: "bugiardo", "combattimento", "contedimettiti", "controcorrente", "covid19buffoni", "governodeipeggiori", "incapaci", "lesionidolose", "obbligovaccinale", "odio_draghi". Per la categoria *provax*, invece: "allarmisti", "bufale", "campagna", "complottisti", "egoisti", "fiera", "follia", "intelligentoni", "sivax", "vincere".

La Tabella 3.3 riporta la matrice di confusione, al cui interno si trova il confronto tra la classificazione manuale e quella presunta dal modello stimato. Il tasso di accuratezza che si raggiunge è del 91%. Inoltre, dalla tabella, si nota che gli errori principalmente commessi si verificano per la previsione della modalità "neutro".

A partire dal modello stimato sono state effettuate le previsioni sui restanti *tweet* privi di etichette. Il primo *step* è stato quello di costruire la DTM con pesi TF-IDF per i *tweet* non utilizzati per stimare il modello. Le accortezze avute sono state quelle di: eliminare i vocaboli presenti nei nuovi testi , ma non in quelli usati per la stima; aggiungere delle colonne nulle per i termini dei *tweet* usati per la stima che non comparivano nei nuovi testi; ordinare i vocaboli della nuova DTM secondo lo stesso ordine con cui si trovavano all'interno della DTM usata per la stima del modello. Le classificazioni dei *tweet* hanno portato alla suddivisione riportata nella Tabella 3.4.

Figura 3.1: *Cross-validation plot*: in ascissa, il logaritmo del parametro di regolazione; in ordinata, i valori dell'errore di classificazione. Partendo da sinistra, la prima linea verticale tratteggiata evidenzia il valore di λ che minimizza l'errore; la seconda linea tratteggiata indica il parametro che fornisce un modello più regolarizzato il cui errore non si discosta per più di un errore standard dal minimo

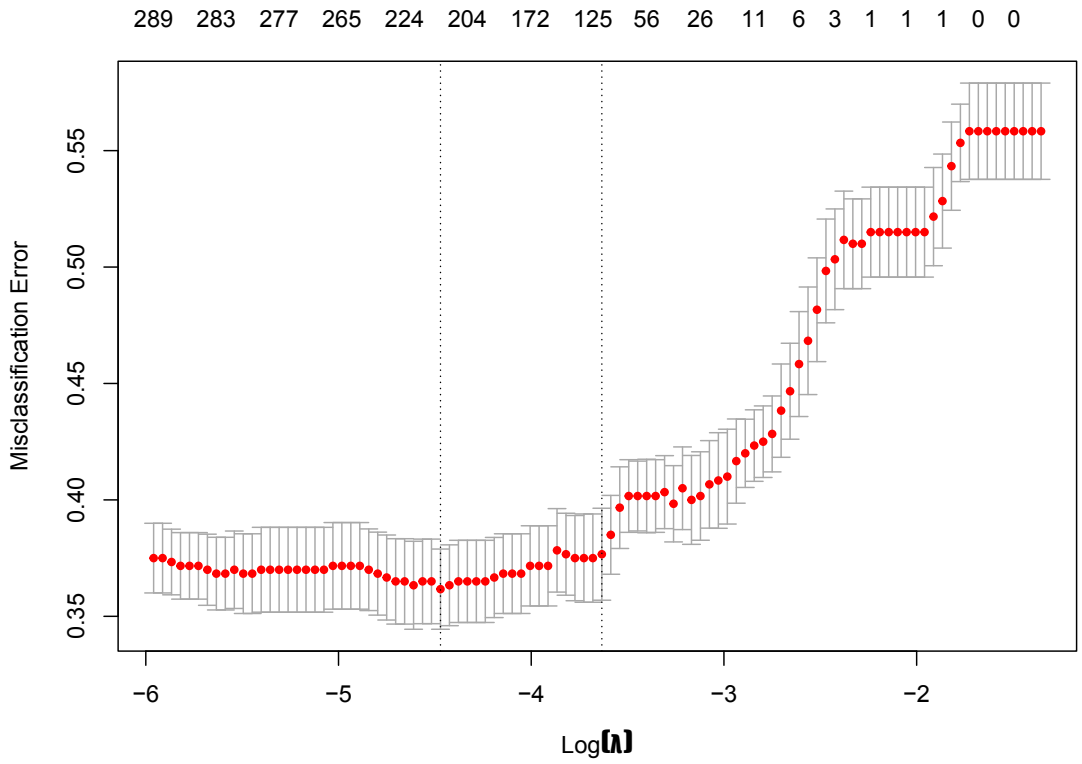


Tabella 3.1: Termini associati ai coefficienti non nulli per la modalità "novax" della risposta

@angy	@annaisa75640851	@controcorrente	@davide38296157
@eston07390779	@etherea	@eurallergico	@forza
@gasparripdl	@laveritaweb	@liliaragnar	@matteosalvini
@mediasettgcom24	@pietrodeleo	@silvia92745700	emoji_bacino
23ec	2a	abbandona	accorgiamo
affanno	afferma	alchimia	aleggiano
allenamento	alto	ammalato	ammaleranno
ammette	anziano	aspettando	aumentate
autoimmuni	autorizzati	barillari	barzelletta
bassetti	bofrost	bucati	bugiardo
bulli	cabinadiregia	cani	capendo
capisco	capro	case	centinaio
cinegiornali	claudio	collega	combattimento
compulsivo	conoscenza	considerato	consiglio
contedimettiti	conti	controcorrente	corrispondere
cos	costui	covid19buffoni	c**o
dannoso	dateci	decessi	dedonno
dichiarasse	dici	dirvi	disgrazia
disordini	divisione	dobbiamo	dogma
dopo	dovrei	economicamente	emoteohoh
escludono	esentati	esistente	essersi
evitare	fatevi	fidarmi	figli
freddo	galli	gestione	giornata
godbye	gonfiore	governodeipeggiori	guarito
guido	immunitario	importante	incapaci
informare	insegnante	janssen	ladri
lesionidolose	letti	linkedin	malati
neppure	obbligovaccinale	odio_draghi	opporsi
ostacola	palle	passando	passano
pensare	pericolosi	piacciono	pieghiamoci
plasma	poche	pochi	pro
propaganda	pulsossimetro	ricoverata	riepilogo
risolto	roba	salotto	sanzionatorie
sapendo	scadenza	segnalati	sempre
sgp	silvestri	sistema	smettetela
specializzazione	sperimentazione	termini	vaccina
vaers	vax	virologi	vivi
zero			

Tabella 3.2: Termini associati ai coefficienti non nulli per la modalità "provax" della risposta

@gmail	@hanan1868	@lory31511460	@marcocattaneo
@martaecca	emoji_rosa	emoji_occhiolino	emoji_stupore
emoji_microbo	12gennaio	12gennaio2022	800
abusando	accesso	aggiornamento	allarmisti
altro	andata	antibiotico	appena
assumono	astronauta	bar	beve
bloccano	brutta	bufale	caccc
campagna	capolavoro	carbonara	causare
cercando	chiarissima	chitarra	circuito
clinica	coloro	complimentoni	complottisti
conferma	confido	conosce	conquista
conta	contagiosa	corale	cov
covid19covid19	covid19dovrebbe	covid19ospedale	cruciale
davvero	dimettiti	disdicevole	diventare
dunque	egoisti	elettrica	ennesimo
enorme	erba	ergo	farmi
febbre	fermento	fiera	follia
fortuna	garantito	giro	grave
inserito	integrato	intelligentoni	lavoro
licenziato	male	mondo	nuovo
operatori_sanitari	pace	paese	passato
permesso	persona	possibilitã	poteva
qualsiasi	questione	raggiungere	reni
sars	sbattete	scherziamo	scienza
sivax	soluzione	storiella	succederà
succedere	tot	vaccinare	vincere
vite	voglia	volte	yomevacuno
zitromax			

Tabella 3.3: Matrice di confusione: sulle righe sono presenti i valori previsti dal modello, sulle colonne i valori osservati

	off-topic	***neutro***	***novax***	***provax***
off-topic	27	0	0	0
neutro	11	249	10	11
novax	3	2	111	0
provax	0	14	3	159

La Tabella 3.5 riporta la suddivisione di tutti i *tweet* classificati tramite il modello in base ai tre periodi. La stessa rappresenta una tabella di contingenza. A questa è stata applicata la tecnica dell'analisi delle corrispondenze. Il risultato ottenuto è riportato nella Figura 3.2: la disposizione dei periodi nella prima dimensione mostra la netta distanza tra i primi due periodi (a sinistra) e il terzo (a destra); se si guarda

Tabella 3.4: Esito dell'applicazione del modello *logit* multinomiale ai rimanenti *tweet* privi di etichetta

Categoria	***Numerosità***
off-topic	229
neutro	5009
novax	1789
provax	2714

Tabella 3.5: Suddivisione, in base al periodo, dei *tweet* classificati tramite il modello *logit* multinomiale con penalizzazione lasso

	off-topic	***neutro***	***novax***	***provax***
1° periodo	42	1880	338	1228
2° periodo	118	1801	410	654
3° periodo	96	1609	1157	1008

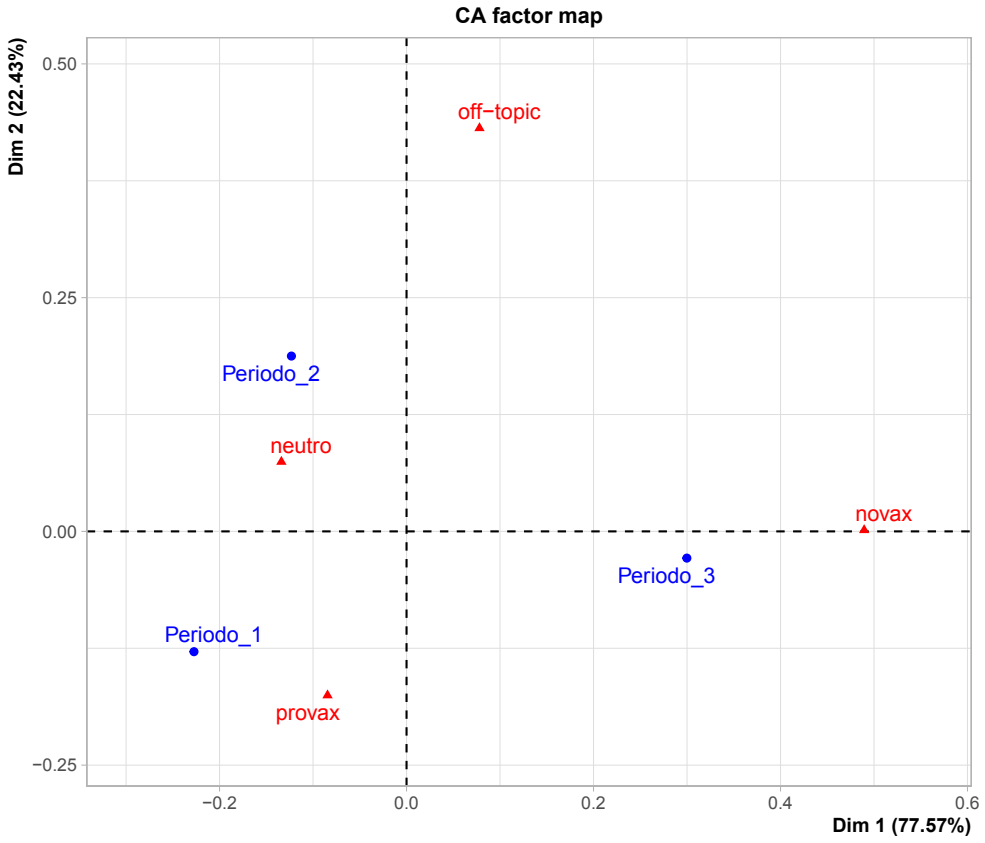
alla distribuzione nella seconda dimensione si nota, invece, la distanza che intercorre tra il primo e il secondo periodo. Il primo periodo è rappresentato principalmente dal gruppo "provax"; il secondo periodo è caratterizzato dal gruppo "neutro"; infine, il terzo periodo viene descritto dal gruppo "novax". Si potrebbe, con tutte le cautele del caso, darne la seguente interpretazione:

- nel primo periodo gli utenti di *Twitter* avevano un parere positivo circa i vaccini, il che era presumibilmente dovuto alla novità rappresentata dal farmaco, giudicato come una possibile via d'uscita dalla pandemia;

- il secondo periodo è come se fosse un periodo di transizione, nel quale la credibilità dei vaccini comincia ad essere messa in discussione a seguito della diffusione di false notizie e il verificarsi di preoccupanti fatti di cronaca (si pensi a quelli legati agli effetti collaterali di AstraZeneca). In ogni caso, comunque, sembra prevalere la neutralità;

- il terzo periodo, invece, è chiaramente caratterizzato da una evidente presenza di pareri negativi, molto probabilmente derivanti dai malcontenti crescenti dovuti ai vari obblighi vaccinali imposti dal governo e dalle restrizioni già da molti mesi in vigore e ancora vigenti.

Figura 3.2: Applicazione dell'analisi delle corrispondenze a una tabella di contingenza che contiene le frequenze dei *tweet* classificati per i vari periodi



Conclusioni

L'idea alla base di questo studio era l'individuazione di quali siano state le tendenze all'interno dell'opinione pubblica italiana riguardo alla tematica vaccinale. Nello specifico, si voleva capire l'evoluzione degli umori degli italiani circa i vaccini contro il SARS-CoV-2 dall'inizio della campagna vaccinale fino ai primi mesi di quest'anno. Per condurre quest'analisi, si è scelto, come fonte dei dati testuali, Twitter, dal quale grazie all'impiego di opportune tecniche di *Text Mining* si sono estrapolati all'incirca 10000 *tweet*, scritti in tre periodi di tempo particolarmente significativi. Quest'ultimi, infatti, sono stati identificati sulla base della rilevanza dei fatti di cronaca accaduti, ossia dell'intensità delle discussioni e delle polemiche che si sono levate nell'arco della settimana seguente all'evento (il via delle vaccinazioni, la scoperta di alcune gravi reazioni avverse provocate dal vaccino AstraZeneca, l'introduzione dell'obbligo vaccinale per gli *over* 50).

Già al termine della fase preliminare, durante la quale si sono svolte le indagini esplorative sui dati raccolti, si è delineata, abbastanza chiaramente, una possibile ipotesi ricostruttiva. In linea generale, si è tratteggiato un *trend*: un costante e progressivo spostamento da sentimenti perlopiù neutro/positivi verso sentimenti negativi. Più nel dettaglio, le analisi esplorative e l'applicazione di tecniche non supervisionate hanno suggerito che nel primo periodo a prevalere è: da un lato, la preoccupazione per il dilagare del virus, per l'alto tasso di malati ospedalizzati e per il numero delle vittime; dall'altro, l'euforia e il sollievo per l'inizio della campagna vaccinale (in più, sullo sfondo, c'è una critica per la gestione dei rifornimenti delle dosi). Nel secondo periodo si assiste alla separazione dell'opinione pubblica in due schieramenti contrapposti: c'è chi esprime la propria speranza nei vaccini e la propria gratitudine nei confronti degli scienziati e della scienza, e chi, all'opposto, solleva dubbi circa la sicurezza dei vaccini ed elogia la nascita di movimenti di contrasto (es. *novax* e *nomask*). Infine, nel terzo periodo c'è una più netta separazione tra *novax* e *provax* e un mutamento dell'oggetto del

dibattito, che qui è incentrato sul numero delle dosi da inoculare, sulle categorie di *greenpass* e sugli obblighi e restrizioni adottate dal Governo. È in questo lasso temporale che si riconoscono nitidamente dei sentimenti negativi, quali rabbia e odio, rivolti, per la maggior parte, però, verso le decisioni che hanno impattato sulla libertà individuale dei cittadini e non direttamente verso i vaccini.

L'ipotesi di cui si è dato conto è stata definitivamente suffragata dai risultati ottenuti a seguito dell'applicazione delle tecniche supervisionate.

Vale la pena sottolineare la necessità di mantenere una certa cautela nell'interpretare i risultati raggiunti e trarre conclusioni. Da una parte, la comunità di Twitter, plausibilmente, non rappresenta un campione casuale della popolazione italiana. Dall'altra, in ogni caso, si ritiene che l'opinione espressa con maggior forza non sia necessariamente quella prevalente. La speranza, al di là di tutto, è che ora che, finalmente, l'emergenza pandemica è almeno in parte rientrata, l'evidenza empirica convinca anche i più scettici ad affidarsi alla scienza.

Bibliografia

Abdi H.; Williams L. (2022). Correspondence analysis.

Ayesha S.; Hanif M. K.; Talib R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, **59**, 44–58.

Azzalini A.; Scarpa B. (2012). *Data analysis and data mining: An introduction*. OUP USA.

Barnier J. (2022). *rainette: The Reinert Method for Textual Data Clustering*. R package version 0.3.0.9000.

Bolasco S. *et al.* (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di statistica*, **7**, 17–53.

Bolasco S.; Bisceglia B.; Baiocchi F. (2004). Estrazione automatica d'informazione dai testi. *Mondo digitale*, **3**(1), 27–43.

Canale A.; Scarpa B. (2022). *Dispensa di Metodi statistici per i big data, Analisi dei testi (text mining), Università degli Studi di Padova.*

Cilione C. M. P. (2011). *Analisi delle corrispondenze. Il metodo e le applicazioni.*

Corridoni T. (2019). Analisi del linguaggio per lo studio del pensiero scientifico di bambini ed insegnanti. *Progress in Science Education (PriSE)*, **2**(1).

Friedman J.; Hastie T.; Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Hastie T.; Tibshirani R.; Friedman J. H.; Friedman J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

ISS (2021). Epidemia covid-19 aggiornamento nazionale 10 novembre 2021 – ore 12:00. *Istituto Superiore di Sanità.*

Johnson R.; Wichern D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Kearney M. W. (2019). rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software*, **4**(42), 1829. R package version 0.7.0.

Lapalut S. (1995). Text Clustering to Support Knowledge Acquisition from Documents. Relazione Tecnica RR-2639, INRIA.

Misuraca M. (2018). *Le basi della statistica testuale, Università degli Studi di Napoli Federico II*.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reddy M. (2011). Chapter 1 - introduction In *API Design for C++*. A cura di Reddy M., pp. 1–19. Morgan Kaufmann, Boston.

Reinert A. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données*, **8**(2), 187–198.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA.

Salvan A.; Sartori N.; Pace L. (2020). Modelli lineari generalizzati. In *Modelli Lineari Generalizzati*, pp. 67–119. Springer.

Solari A. (2022). *Dispensa di Analisi Statistica Multivariata, Distanze, Università degli Studi di Milano-Bicocca*.

Solari D.; Sciandra A.; Rinaldo M.; Redaelli M.; Finos. L. (2016). *TextWiller: Collection of functions for text mining, specially devoted to the italian language*. R package version 2.0.

Tan A.-H. *et al.* (1999). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases*, volume 8, pp. 65–70.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Xu R.; Wunsch D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, **16**(3), 645–678.

Zebari R.; Mohsin Abdulazeez A.; Zeebaree D.; Zebari D.; Saeed J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, **1**, 56–70.