Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Analisi congiunta delle opinioni espresse via Twitter su alcuni fatti di attualità politica

Relatore Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatore Prof. Alessio Farcomeni
Dipartimento di Sanità Pubblica e Malattie Infettive
Sapienza - Univeristà di Roma

Laureando Mattia Uttini
Matricola N 1104030

Anno Accademico 2015/2016

"The answer my friend
is blowin' in the wind"
Bob Dylan

Indice

Introduzione

Il fenomeno di massa più rilevante di questo inizio di XXI secolo sono sicuramente i social media. Tra i più noti ed utilizzati vi è Twitter, un servizio di microblogging che permette di condividere con la propria rete di seguaci (*follower*) pensieri, notizie, immagini e quant'altro. Il portale *statista.com* ha stimato che nel secondo quadrimestre del 2016 gli utenti attivi su Twitter siano stati 315 milioni. Oltre a permettere a questa fetta della popolazione mondiale di interagire e scambiare opinioni e punti di vista, in breve tempo questo social media è diventato anche un canale ufficiale di comunicazione per le istituzioni pubbliche oppure un ulteriore canale commerciale per aziende private. Per citare alcuni esempi, basti pensare all'abilità comunicativa di leader mondiali come Barack Obama (la foto dell'abbraccio con la moglie Michelle dopo la rielezione nel 2008 [1] conta più di 830.000 retweet), alle comunicazioni ufficiali e di semplici cittadini in occasione di attentati terroristici o calamità naturali, alle campagne virali di un'azienda ormai celeberrima sul web come Ceres [2].
Da un punto di vista statistico, l'ammontare di testi condivisi quotidianamente rappresenta una fonte di dati molto appetibile. Gli utilizzi, infatti, possono spaziare dal campo del marketing (Lassen et al. (2014)) a quello sociologico e politico (Taddy (2013b)), fino a quelli finanziario, di intelligence ed addirittura medico-sanitario (Paul and Dredze (2014)). In quest'ottica il lavoro presentato in questo elaborato mira ad indagare il modo in cui gli utenti italiani di Twitter si sono espressi riguardo a fatti di attualità avvenuti negli ultimi mesi. L'obiettivo delle analisi è quello, innanzitutto, di classificare gli utenti a seconda del loro pensiero riguardo gli argomenti in questione e, ove possibile, di legare e spiegare il *sentiment*

[1] https://twitter.com/BarackObama/status/266031293945503744

[2] https://compassunibo.wordpress.com/2016/03/14/strategia-ceres-marketing-geniale-sdegno-vero-o-impegno-sociale/

attraverso alcune covariate relative agli utenti stessi. Con questi scopi verranno implementati due modelli: in primo luogo attraverso un modello marginale si cercherà di indagare il legame presente tra le caratteristiche dell'utente ed il suo attivarsi o meno per alcuni temi presenti nel dibattito politico italiano. L'utilizzo di questa metdologia permetterà inoltre, a differenza dei tradizionali modelli, di stimare congiuntamente anche le interazioni tra gli *outcome*, così da potere determinare quali tematiche siano maggiormente connesse tra loro. In secondo luogo si indagherà circa i contenuti dei tweet per ogni tema, attraverso l'implementazione della regressione multinomiale inversa. Essa permette di individuare i termini chiave per ogni *sentiment* individuato e di effettuare una proiezione dei testi in un nuovo spazio di dimensione ridotta.

Nel **Capitolo 1** viene riportata la procedura di *crawling* che ha permesso di scaricare da Twitter i dati utilizzati per le analisi e di passare dal dato grezzo ad una tipologia di dato semi-strutturata. Twitter permette attraverso un accesso tramite API di scaricare tweet tramite *query*, cioè sottoponendo un insieme di parole chiave da ricercare nei testi.

Nel **Capitolo 2** viene fornita una visione d'insieme dei dataset costruiti. In primo luogo si concentra l'attenzione sugli utenti e sulle loro carattersitiche: da sesso e zona di provenienza ad aspetti più legati all'utilizzo del social network come ad esempio numero di follower e numero di tweet scritti. In un secondo momento si focalizza l'attenzione sui tweet e sul loro contenuto. Dopo aver fornito alcune indicazioni spaziali e temporali ricavate in fase di *crawling*, si passa all'analisi delle parole più utilizzati per ciascuno dei temi affrontati.

L'obiettivo del **Capitolo 3** è quello di cercare legami tra le covariate dell'utente e la propensione ad esprimersi su un tema piuttosto che su un altro e capire come i temi siano interconnessi tra di loro. Per fare ciò si utilizza una metodologia esposta da Glonek and McCullagh (1995), il cosiddetto modello marginale o logit multivariato. Oltre a modellare le probabilità marginali esso permette attraverso una particolare parametrizzazione di definire *odds ratio* locali o globali tra coppie di variabili.

Il secondo modello utilizzato per analizzare i dati raccolti è la regressione multinomiale inversa, esposta nel **Capitolo 4**. Essa permette di effettuare una riduzione dimensionale delle matrici di rappresentazione dei documenti, proiettando i testi in un nuovo spazio di dimensione inferiore. Questa operazione viene però compiuta

attraverso il *sentiment* espresso nei documenti. Diventa quindi possibile interpretare le proiezioni in ottica dei diversi *sentiment* per capire quali particolari termini siano più emblematici di un particolare atteggiamento. Grazie a questa operazione diventa agevole operare una previsione del *sentiment* attraverso un modello di regressione multinomiale.
Le operazioni descritte vengono applicate su un insieme di dati fortemente connotati da un punto di vista semantico, date le tematiche trattate. Tuttavia si tratta di approcci facilmente esportabili anche in altri ambiti: si pensi ad esempio alla stessa applicazione relativa al lancio di un prodotto per sondare le reazioni sul web, ad un'indagine di mercato per capire il posizionamento dei prodotti presso i consumatori oppure alla ricerca di particolari trend e tendenze sul web.

Capitolo 1

Da Twitter ai dataset di analisi

La creazione dei dataset di analisi si è sviluppata principalmente su due filoni: quello degli utenti e quello dei testi. Dopo avere effettuato il *crawling* dei tweet, infatti, sono state create apposite funzioni che permettessero di avere in forma strutturata un dataset che racchiudesse gli utenti e le loro caratteristiche "twitter-biografiche" ed un altro che contenesse le informazioni relative ai tweet che essi avevano pubblicato relativamente agli argomenti selezionati.

1.1 *Crawling* dei tweet

A partire dal mese di febbraio 2016 sono stati selezionati alcuni argomenti, a mano a mano che emergevano ed acquisivano importanza nel dibattito pubblico, ritenuti interessanti per le finalità di analisi prestabilite. Una volta selezionati gli ambiti su cui estrarre tweet, è stato messo in funzione un *crawler*, scritto nel linguaggio R, che producesse ogni 10 minuti un file *.json* contenente i tweet ed i relativi attributi (da caratteristiche dell'utente a variabili relativi a luogo ed ora del tweet). La costruzione del *crawler* si è poggiata in larga misura sul pacchetto streamR[1], scritto e manutenuto dal prof. Pablo Barberà (*University of Southern California*). All'interno del pacchetto è presente la funzione filterStream, la quale permette di aprire una connessione all'API Streaming per scaricare i tweet richiesti in tempo reale.

[1] https://github.com/pablobarbera/streamR

La selezione del sottoinsieme di tweet da scaricare è stata basata su *query* stabilite a priori, cioè sono stati selezionati tutti i tweet contenenti almeno uno dei termini presenti nella lista di interrogazione. In Tabella 1.1 sono riportati i temi selezionati, alcuni dei termini presenti nelle *query* costruite per individuare i tweet attinenti e le date in cui il *crawler* è stato attivo.

Il tema "Talk show", cioè il *crawling* dei tweet legati ai principali talk show televisivi, è stato introdotto con lo scopo di catturare alcune tematiche non decise a priori ma comunque presenti nel dibattito pubblico.
Una volta generati i file *.json* è stato possibile costruire una funzione che li leggesse e creasse un dataset contenente tutte le informazioni scaricate. Questi dataset, uno per argomento, sono stati ulteriormente processati per giungere ad avere i dati utilizzati come input nelle analisi presentate.

1.2 Dataset di analisi: utenti

Per costruire il dataset degli utenti a partire dai tweet scaricati è stata implementata una procedura in due passi. Inizialmente, per ognuno dei sette temi sono stati isolati tutti gli utenti unici per i quali fosse presente almeno un tweet. Per gli utenti presenti con più di un tweet per tema, si è scelto di tenere le covariate più recenti relative all'attività sul social media. Per quanto riguarda, invece, le covariate biografiche ci si è attenuti alla classificazione descritta in seguito, tenendo in considerazione, nel caso un utente cambiasse nome o località, la modalità presente il maggior numero di volte. In un secondo momento, trattando nello stesso modo gli utenti presenti su più temi a livello di covariate, è stato creato un dataset che contenesse tutti gli utenti unici che avessero twittato su almeno uno dei temi selezionati.
Agendo in questo modo, il dataset descritto in seguito è composto da **214603** utenti unici.

1.2.1 Covariate biografiche

Le informazioni biografiche relative all'utente rese disponibili scaricando i tweet attraverso la funzione filterStream del pacchetto *streamR* sono pressochè nulle. Tuttavia, è stato possibile ricostruire due caratteristiche degli utenti: il sesso e la

Dibattito legge Cirinnà e approvazione
17 febbraio 10:45 - 17:00 11 maggio 20:45 - 13 maggio 00:05
#unionicivili unioni civili #cirinnà #cirinnamoreremo
#stopcirinnà #lovewins #stepchildadoption
Dimissioni ministro Guidi
31 marzo 20:20 - 03 aprile 01:20
#Guidi dimissioni #Guididimettiti #MEB #Boschi Guidi
Referendum abrogativo trivelle
03 aprile 00:20 - 19 aprile 11:35
#referendum17aprile #iovotosì #iovotono #iononvoto #quorum
#noquorum #battiquorum #17aprile trivella trivelle #notriv
Talk show
18 aprile - 11 giugno (orari di trasmissione)
#ballarò #piazzapulita #portaaporta #virusrai2 #quintacolonna
#dimartedì #lagabbia #ottoemezzo #gazebo #fuorionda
Ritorno marò Salvatore Girone
26 maggio 09:35 - 29 maggio 20:00
#marò #girone marò Girone
Elezioni amministrative: ballottaggi
06 giugno 08:45 - 23 giugno 16:40
#elezionicomunali #amministrative2016 #Sala #Parisi #Milano
#Roma #Giachetti #Raggi #Torino #Appendino #Fassino
Referendum Brexit
23 giugno 21:50 - 26 giugno 23:45
#brexit brexit #maratonabrexit #britishreferendum #euref

Tabella 1.1: Caratteristiche delle operazioni di *crawling* effettuate

provenienza geografica a partire da nome utente e *location*. In quest'ultimo caso si è deciso di utilizzare l'indicazione geografica soggettivamente inserita dall'utente nel momento dell'iscrizione in quanto i tweet geolocalizzati rappresentavano un sottoinsieme molto limitato, circa il 2.8% del totale.
In particolare, ai fini di garantire una classificazione efficace, è stata riscritta per l'occasione la funzione *classificaUtenti* presente nel pacchetto R TextWiller[2], sviluppato dal prof. Livio Finos ed alcuni colleghi e manutenuto da Dario Solari. La funzione in questione compie una classificazione basata su dizionari, previa una pulizia del testo. Riscrivendola per l'occasione, si è cercato di migliorarne l'efficacia prevedendo tre scan differenti del testo in input. In un primo momento, si è deciso di cercare nel testo corrispondenze con i termini presenti nel vocabolario e collegati alle modalità prestabilite per la classificazione. In seguito, per gli utenti non classificati, vengono tolti dal testo i numeri e viene effettuato uno split sulle lettere maiuscole (es: "MarioRossi" diventa "Mario Rossi", in modo da permettere il riconoscimento di "Mario"). Infine, configurando nel modo opportuno i parametri della funzione, è possibile effettuare per gli utenti non ancora collocati in alcuna categoria un cosiddetto "scan interno". Attraverso la funzione grepl, infatti, non viene più ricercata una corrispondenza tra i termini del vocabolario e le intere parole contenute nel testo, ma viene effettuata la stessa ricerca anche all'interno dei termini stessi. Questa operazione è stata resa opzionale perchè comporta una grande quantità di falsi positivi: per limitare questo problema è possibile impostare la lunghezza minima dei termini del vocabolario o, in alternativa, una singola categoria per cui effettuare questa ricerca. Per la classificazione del sesso sono state previste tre categorie: "maschio" "femmina" e "ente" (per permettere di classificare mass media, istituzioni pubbliche, partiti politici ed altri). Per la classificazione geografica, invece, a partire dalla location si è cercato di riconoscere il comune per poi assegnare l'unità a "Nord-Ovest", "Nord-Est", "Centro", "Sud", "Isole" o "Estero". Di seguito, vengono forniti alcuni esempi di funzionamento della funzione.

```
classificaUtenti(c("Mario Rossi", "Maria Clara", "Corriere della Sera"))
mario rossi         maria clara    corriere della sera
"masc"              "femm"         "ente"
```

[2] https://github.com/livioivil/TextWiller

```
classificaUtenti(c("SanSepolcro", "Milano", "Palermo"), vocabolario=
    vocabolarioLuoghi)
sansepolcro     milano        palermo
"Centro"        "Nord–ovest"   "Isole"
classificaUtenti(c("Liviofinos", "alessiofarcomeni"), scan_interno = TRUE)
liviofinos    alessiofarcomeni
"masc"         "masc"
```

Come accennato in precedenza, sono stati riscontrati casi in cui, nel tempo intercorso tra due *crawling*, un utente avesse cambiato nome o località. In questi casi, tra tutte le classificazioni riconducibili a quell'id, si è tenuta quella modale. In totale, è stato classificato il sesso per il 70.17% degli utenti, mentre per la zona geografica il discorso è differente. Infatti, solo il 58.05% degli utenti ha inserito la sua provenienza al momento dell'iscrizione: di questi, il 64.10% è stato classificato (pari al 37.21% del totale degli utenti). Al netto di tutto, la classificazione congiunta di sesso e zona geografica è stata possibile per il 28.49% degli utenti, pari a 61150 unità. Occorre sottolineare come in diversi casi la mancata classificazione sia dovuta ad un'indicazione fallace fornita dall'utente (*location*="somewhere over the rainbow" oppure nome utente="Parlare d'altro") più che ad una errata definizione della funzione e del vocabolario.

1.2.2 Covariate relative agli account Twitter

Le informazioni relative all'attività su Twitter tenute in considerazione sono sei: numero totale di stati, numero di follower e following, data di creazione dell'account, data dell'ultimo stato osservato e se l'account in questione fosse o meno verificato. In particolare, per quanto riguarda le prime tre covariate, si è tenuto in considerazione il numero massimo osservato durante le diverse fasi di *crawling*. In Tabella 1.2 sono riportate le percentuali di dati mancanti presenti nelle variabili.

Inoltre, sono state calcolate due ulteriori covariate ritenute interessanti: il numero di giorni di iscrizione, calcolato come differenza tra data di creazione e data dell'ultimo stato osservato, ed il numero di tweet per giorno, calcolato come rapporto tra numero totale di tweet e giorni dall'iscrizione. Quest'ultima informazione è stata calcolata al fine di avere una misura dell'attività dell'utente che non risentisse della sua "anzianità" sul social media.

Numero stati	0.015%
Numero follower	0.010%
Numero following	0.011%
Data creazione	0%
Data stato	0%
Account verificato	0.008%

Tabella 1.2: Percentuale di dati mancanti per le variabili di account

1.3 Dataset di analisi: testi

Nella Sezione 1.1 è stato descritto il processo di *crawling* che ha permesso di scaricare i tweet per mezzo di *query* prestabilite. In seguito, è stato possibile creare due dataset per ogni tema (uno per i tweet ed uno per i retweet) contenenti i testi e le informazioni degli utenti. Infatti, all'interno dei file *.json* ottenuti tramite il *crawler*, nel caso il tweet scaricato fosse un retweet (cioè la condivisione tramite il proprio account di un contenuto pubblicato da un altro utente), venivano salvate anche tutte le caratteristiche relative al tweet originale. Attraverso una piccola modifica della funzione parseTweets del pacchetto streamR è stato possibile, perciò, aggiungere ai tweet scaricati anche i cosiddetti tweet originali, ovviamente eliminando i doppioni. Avendo scelto di considerare anche i tweet originali parte del corpus di documenti considerati per le analisi, sono stati isolati per ogni tema i "tweet unici", cioè eliminando i duplicati internamente ai tweet originali e tra tweet scaricati ed originali. In Tabella 1.3 sono riportate le numerosità totali dei tweet unici per tema.

1.3.1 Pulizia del testo e *stemming*

Una volta ottenuti i tweet unici per tema, è stata creato il vero corpus di documenti. Per ogni tema, sono stati concatenati i tweet per utente, in modo che la dimensione del corpus coincidesse con il numero di utenti unici. A questo punto, si è resa necessaria un'operazione di pulizia del testo. In primo luogo, sono stati isolati hashtag e tag, dato che non sembrava opportuno sottoporre all'operazione

Dibattito legge Cirinnà e approvazione	90991 tweet unici
Dimissioni ministro Guidi	89898 tweet unici
Referendum abrogativo trivelle	308919 tweet unici
Talk show	347138 tweet unici
Ritorno marò Salvatore Girone	27930 tweet unici
Elezioni amministrative: ballottaggi	186440 tweet unici
Referendum Brexit	412186 tweet unici

Tabella 1.3: Numerosità totali di tweet unici per tema

di pulizia stringhe che ricoprono un ruolo così importante nella struttura del social media analizzato. Mentre per gli hashtag l'approccio può sembrare più naturale, per i tag si è seguito quanto fatto per un'analisi simile, almeno riguardo la fonte dati, presentata in Taddy (2013b). Una volta tolte queste stringhe è stata applicata una prima volta al testo la funzione normalizzaTesti del pacchetto TextWiller. Questa funzione permette di compiere differenti operazioni sul testo: dalla rimozione della punteggiatura all'eliminazione delle *stop words*, cioè di quei termini non significativi come articoli, locuzioni, preposizioni, ausiliari, attraverso un vocabolario[3]. Inoltre, la funzione permette di standardizzare tutti i link presenti nei tweet, sostituendo all'indirizzo originario la dicitura *wwwurlwww*, e di trasformare in testo le emoticon, codificate come *emotebad*, *emotelove* ed altre ancora. Una volta ripuliti i testi, è stata applicata la cosiddetta operazione di *stemming*. Essa consiste in una standardizzazione del testo al fine di identificare i termini secondo la loro radice, eliminando differenti generi, numeri, tempi verbali o simili. Uno degli algoritmi di stemming più utilizzati in letteratura è sicuramente lo *snowball stemmer*, la cui versione inglese è nota come *Porter stemmer*. Negli ultimi anni è stato sviluppato anche uno stemmer per la lingua italiana[4], implementato nel pacchetto R SnowballC. In Tabella 1.4 sono riportati alcuni esempi di termini e del loro equivalente stemmato.

Lo scopo principale dell'applicazione dello *stemmer* è fare in modo che non vengano distinti termini che in realtà portano lo stesso contenuto. Come visto nella Tabella, un utente che scrive "approvata" ed uno che usa il termine "approvare"

[3]http://snowball.tartarus.org/algorithms/italian/stop.txt
[4]http://snowball.tartarus.org/algorithms/italian/stemmer.html

politica	*polit*
politici	*polit*
politicanti	*politic*
legge	*legg*
leggi	*legg*
approvata	*approv*
approvare	*approv*

Tabella 1.4: Esempi di funzionamento dell'algoritmo di stemming *snowball*

intendono lo stesso concetto, quindi è necessario uniformare le parole ai concetti. Una volta compiuta questa operazione, si è passati alla costruzione delle *term-by-document matrix*, ovviamente una per ogni tema distinto. Esse permettono la rappresentazione dei documenti come vettori e, di conseguenza, dell'intera collezione come una matrice. Come prima cosa, viene creato il vocabolario associato al corpus di documenti, cioè l'insieme di tutti gli stem presenti in almeno un testo. Per compiere questa operazione esistono differenti approcci: una soluzione è la creazione della cosiddetta *bag of words*, cioè l'insieme degli stem indipendentemente dalla loro posizione nella frase, un'altra è l'impiego dei cosiddetti *N-grammi*. Un N-gramma consiste nell'assegnazione di una misura di probabilità non alla singola parola, ma alla co-occorrenza di N termini in una determinata sequenza (Jurafsky and Martin (2008)). Nel caso in analisi, si è scelto di operare tenendo in considerazione, oltre alle singole parole, anche i bigrammi. Costruito il vocabolario come illustrato, sono stati concatenati ad ogni testo i rispettivi hastag e tag isolati in precedenza, ovviamente aggiungendoli anche al vocabolario. Perciò, in sintesi, il vocabolario è composto dalla *bag of words* degli stem presenti nel testo, dai bigrammi osservati e da hashtag e tag.
La rappresentazione vettoriale di ogni documento avviene creando un vettore di lunghezza pari alla dimensione del vocabolario, in cui viene assegnata ad ogni termine (o bigramma/hastag/tag) la frequenza assoluta con cui viene osservato nel testo. Si arriva perciò ad avere una matrice sparsa della forma:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

dove x_{ij} è la frequenza osservata per il j-esimo termine del vocabolario ($j = 1, \dots, p$) per l'i-esimo documento, in questo caso utente ($i = 1, \dots, n$). In Tabella 1.5 viene riassunto il *preprocessing* dei testi, dai tweet (ovviamente fitizi) alle *term-by-document matrix*, dalla quale si nota come anche con un numero di documenti ridotto e di scarse dimensioni è possibile avere vocabolari molto estesi e matrici molto sparse. Per limitare questo problema, ed in accordo con quanto esposto a più riprese in letteratura, si è deciso di tenere nelle matrici solamente i termini presenti in almeno il 5‰ dei documenti della collezione. Tenendo conto degli utenti che hanno twittato per ogni argomento e dei termini selezionati (*token*, utilizzando un termine proprio del *Natural Language Processing*), in Tabella 1.6 sono riassunte le dimensionalità delle *term-by-document matrix* dei 7 temi.

Testi originali

Id Utente	*Tweet*
123	Evviva! È stata approvata la legge! :) #bravi
123	#complimenti a tutti per la legge! Bravi! @parlamentare1
999	@parlamentare1 avete davvero approvato la legge? #vergogna

Concatenazione per utente e trattamento # e @

Id Utente	*Testo*	*Hashtag e tag*
123	Evviva! È stata approvata la legge! :) a tutti per la legge! Bravi!	#bravi #complimenti @parlamentare1
999	avete davvero approvato la legge?	@parlamentare1 #vergogna

Pulizia testo e *stemming*

evviva stata approvata	⇒	evviv stat approv
legge emotegood legge bravi		legg emotegood legg brav
davvero approvato legge	⇒	davver approv legg

Corpus definitivo

Id Utente	*Documento*
123	evviv stat approv legg emotegood legg brav #bravi #complimenti @parlamentare1
999	davver approv legg @parlamentare1 #vergogna

Term-by-document matrix

Vocabolario = ("@parlamentare1","#bravi", "#complimenti", "#vergogna", "approv", "approv legg", "brav", "davver", "davver approv", "emotegood", "emotegood legg", "evviv", "evviv stat", "legg", "legg brav", "legg emotegood", "stat", "stat approv")

Utenti = ("123", "999")

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Tabella 1.5: *Pre-processing* dei testi: esempio completo

	Documenti (utenti)	*Token*
Dibattito legge Cirinnà e approvazione	36694	920
Dimissioni ministro Guidi	24091	1208
Referendum abrogativo trivelle	67658	1271
Talk show	34886	2395
Ritorno marò Salvatore Girone	14000	848
Elezioni amministrative: ballottaggi	53229	836
Referendum Brexit	106284	1257

Tabella 1.6: Dimensioni delle *term-by-document matrix* per tema

Capitolo 2

Analisi preliminari

In questo capitolo si cercherà di indagare la composizione dei dataset costruiti. In particolare, nella Sezione 2.1 l'attenzione si concentrerà sugli utenti, analizzando la distribuzione delle principali variabili che compongono il dataset. Nella Sezione 2.2, invece, il focus si sposterà sui testi, sui tweet scaricati: si cercherà di capire quali temi sollecitino l'attenzione di particolari sottoinsiemi di utenti e come essi si siano espressi gli stessi su alcuni argomenti in particolare.
Per quanto riguarda i dati analizzati nella Sezione 2.1, salvo dove specificato, vengono considerati solo utenti per cui sia il sesso che la zona sia stato classificato. La scelta comporta una notevole riduzione delle numerosità (da 214603 a 61134 unità), però viene motivata tramite la volontà di osservare come l'attivazione degli utenti sui temi ed il *sentiment* espresso sia condizionato non solo da come si sia espresso l'autore dei post, ma anche dalle sue caratteristiche personali. Invece, per quanto riguarda i testi, in questa sezione si considera l'intero corpus dei tweet, sempre salvo eccezioni indicate esplicitamente.

2.1 Utenti

Eliminando gli utenti per cui fossero mancanti sesso o zona geografica, il dataset di analisi si compone di 61134 unità su 13 variabili, di cui 8 di reale interesse. Le altre dimensioni, infatti, sono chiavi primarie (id utente) oppure variabili che sono servite per ricostruire il sesso (il nome utente), la zona (la *location* inserita tra

le proprie informazioni personali) o il numero di giorni trascorsi dall'iscrizione al social media.

2.1.1 Variabili biografiche

La prima variabile esaminata è il sesso degli utenti. Ad un prima analisi sull'intero dataset si nota - Figura 2.1 - come la maggioranza degli *user* osservati sia maschio (60.62%) con una prevelanza quasi doppia rispetto alle donne (32.09%), mentre gli enti, con il 7.29% degli utenti osservati, risultano una esigua minoranza.

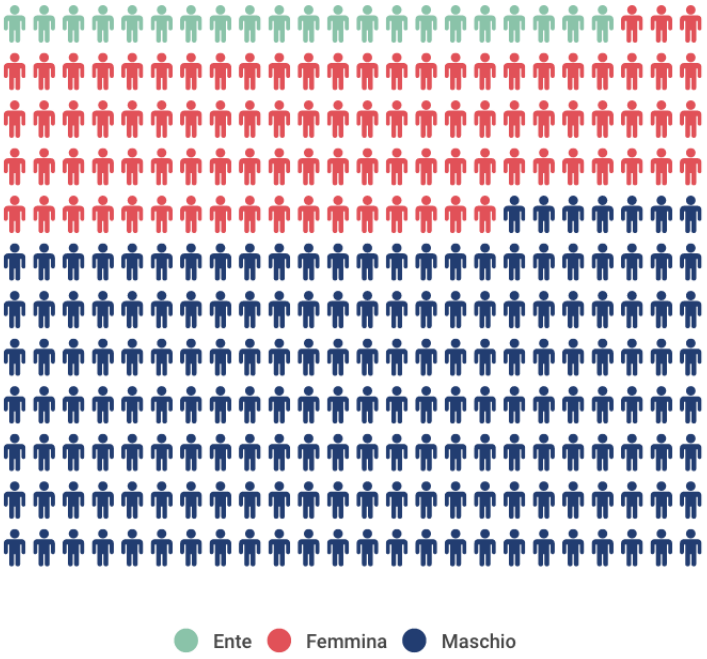

Figura 2.1: Utenti: distribuzione della variabile sesso

In Figura 2.2 è rappresentata la prevalenza degli utenti per zona, considerato che la quota di utenti all'estero è pari al 3.77%. Si nota subito come la zona più densa di *user* sia il Nord-Ovest (28.79%), mentre nelle isole vi sia la prevalenza minore (8.11%). Sorprende un po' il dato del Nord-est (16.25%), sopravanzato sia dal Centro (26.05%), che risente del traino di due grandi centri urbani come Firenze e Roma, che dal Sud (17.03%).

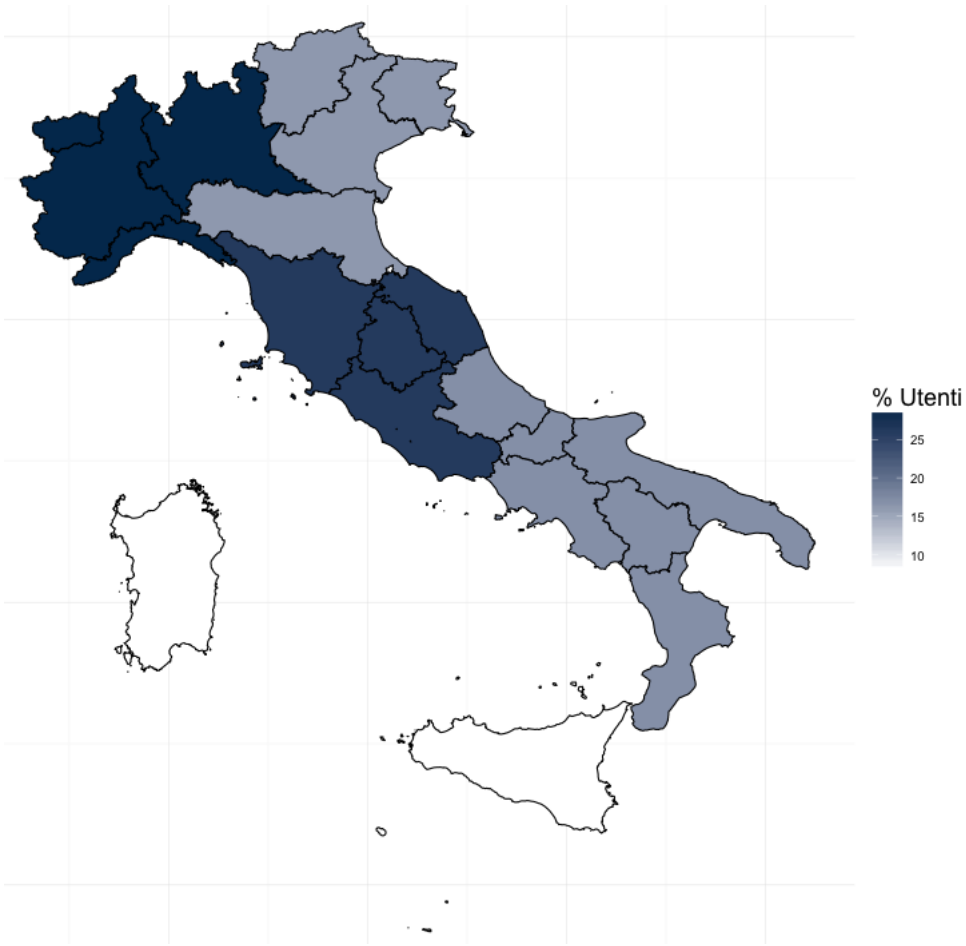

Figura 2.2: Prevalenza degli utenti per zona geografica

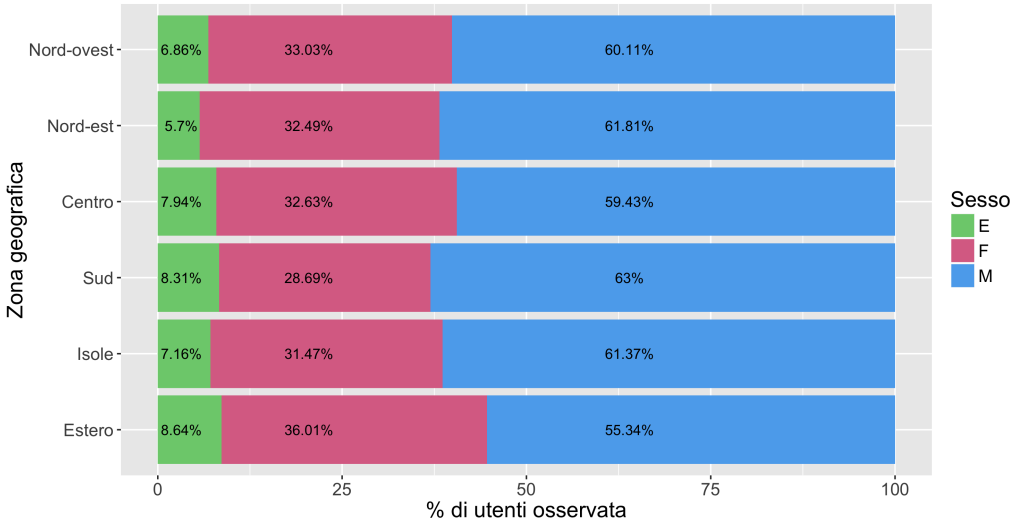

Figura 2.3: Utenti: distribuzione della variabile sesso per zona geografica

Passando ad analizzare come varia la distribuzione del sesso sulle differenti zone, in Figura 2.3 si può osservare come non sussistano enormi differenze. Tuttavia è possibile notare come Sud ed Estero siano le due zone in cui il divario tra maschi e femmine è rispettivamente massimo e minimo. Infatti nelle regioni meridionali si ha la quota più alta di presenze femminili e contemporaneamente più bassa di uomini, mentre accade l'esatto contrario per i tweet provenienti dal territorio extraitaliano. In questo caso, inoltre, spiccano le quote rosa, con gli utenti di sesso femminile che toccano il 36%. Per quanto riguarda gli enti si nota come ve ne siano pochi situati nel Nord-Est, raffrontati con la quota complessiva, mentre ve ne siano numerosi al Sud ed all'Estero. Mentre quest'ultimo fatto può sembrare più logico e legato ad associazioni di italiani all'estero, la prevalenza maggiore al Sud rispetto alle altre zone può spiegarsi con la presenza tra i temi del referendum sulle trivelle, tema molto sentito nelle regioni meridionali e che ha coinvolto numerosi enti non governativi.

2.1.2 Variabili relative al profilo Twitter

Le variabili relative al profilo Twitter prese in considerazione sono cinque: numero di follower e following, numero di tweet al giorno, verifica dell'account e giorni

dall'iscrizione. Osservando le densità delle prime 3 variabili appare evidente, come facilmente immaginabile, una fortissima asimmetria positiva, cioè una distribuzione troncata a 0 ma con una coda lunghissima a destra. Questo problema potrebbe avere importanti ripercussioni nel momento in cui si va ad inserire questa covariata tra le esplicative di un modello, dato che per colpa di questi valori estremi la stima del relativo coefficiente in un modello parametrico può risultare complicata. Per questo motivo si è ritenuto opportuno applicare una trasformazione logaritmica alle variabili in questione. In Figura 2.4 sono mostrate le tre densità empiriche prima e dopo la trasformazione. È immediato osservare il miglioramento in termini di simmetria. Permane qualche osservazione sulla coda sinistra per quanto riguarda follower e following, ma questo è dovuto alla presenza di valori pari a 0 a cui è stato aggiunto un valore molto piccolo (10^{-10}), come a tutti gli altri valori, per permettere la trasformazione.

Come spesso accade, in statistica ma non solo, occorre però guardare l'altro lato della medaglia. Infatti, se le variabili trasformate risultano molto più simmetriche e gestibili a livello di modellistica, l'analisi delle correlazioni fa sorridere meno. Mentre tra le variabili originali si raggiunge il valore massimo di correlazione tra follower e following, come lecito attendersi, per un valore di 0.14, dopo le trasformazioni la stessa correlazione risulta pari a 0.54. Inoltre anche gli altri due valori aumentano notevolmente, raggiungendo i valori di 0.50 per tweet/follower e 0.30 per tweet/following contro i rispettivi 0.033 e 0.098 sulle variabili originarie. Ma, citando una cara professoressa, "in statistica, come nella vita, nessuno regala niente".

In conclusione viene mostrata in Figura 2.5 la densità empirica del numero di giorni dall'iscrizione. In teoria non ci sarebbero forti motivazioni per aspettarsi una distribuzione differente da un'Uniforme ed il picco presente nella densità in corrispondenza dei 1600-1700 giorni non sembra giustificato. Riportando quel numero in date, si tratta di utenti iscrittisi a Twitter nella seconda metà del 2011. Considerando l'ambito in cui ci si muove, si può pensare che questo picco di iscrizioni sia in corrispondenza della crisi economica e politica che ha colpito il nostro paese in quei mesi. Ricollegandosi a quanto detto nell'Introduzione probabilmente anche in Italia in quel momento si è sviluppata la funzione di Twitter come social media con cui tenersi informati sulla realtà ed avere addirittura notizie in anteprima rispetto ai media più tradizionali. Inoltre proprio durante quell'anno alcune

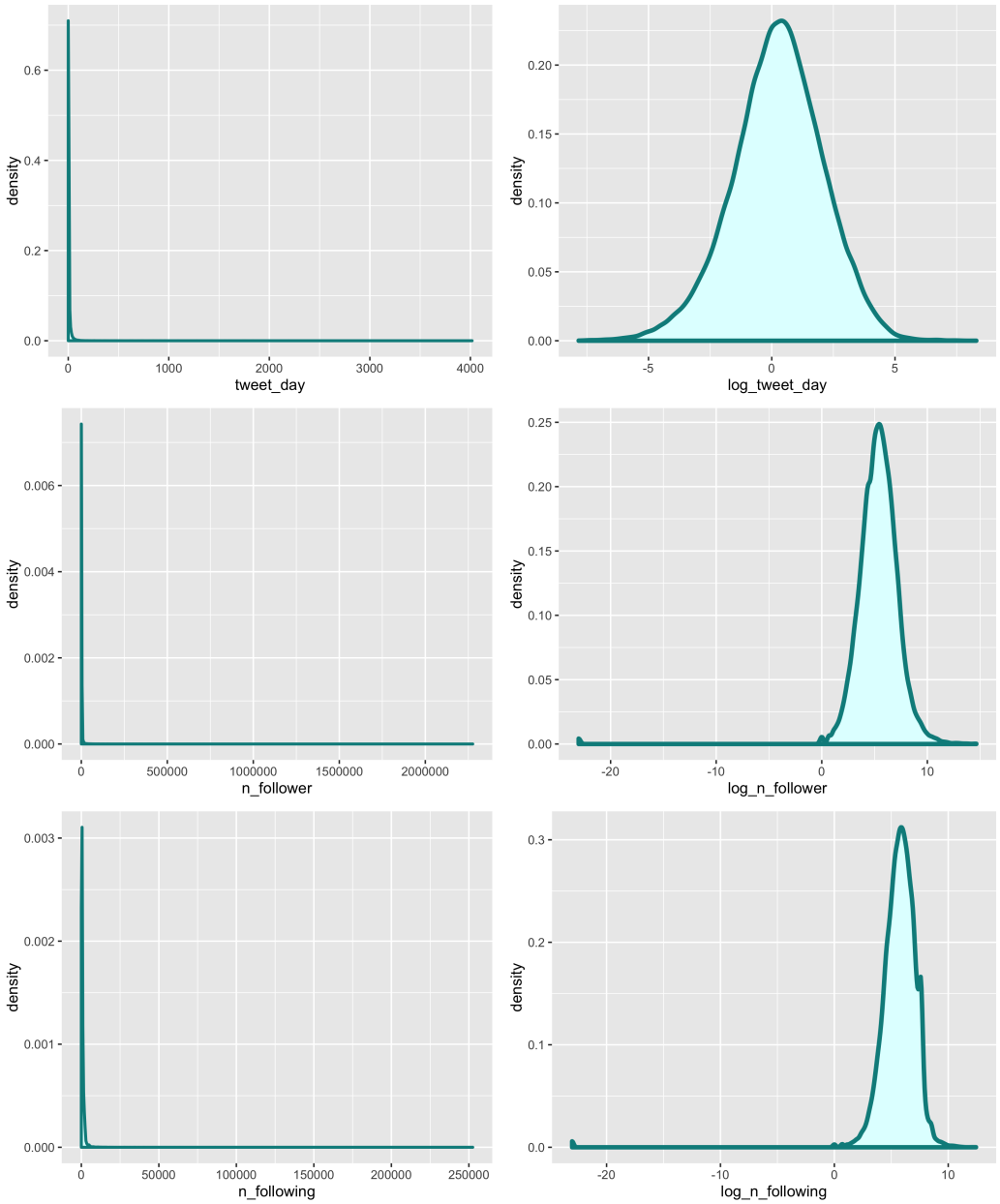

Figura 2.4: Utenti: densità empiriche delle variabili Twitter prima e dopo la trasformazione logaritmica

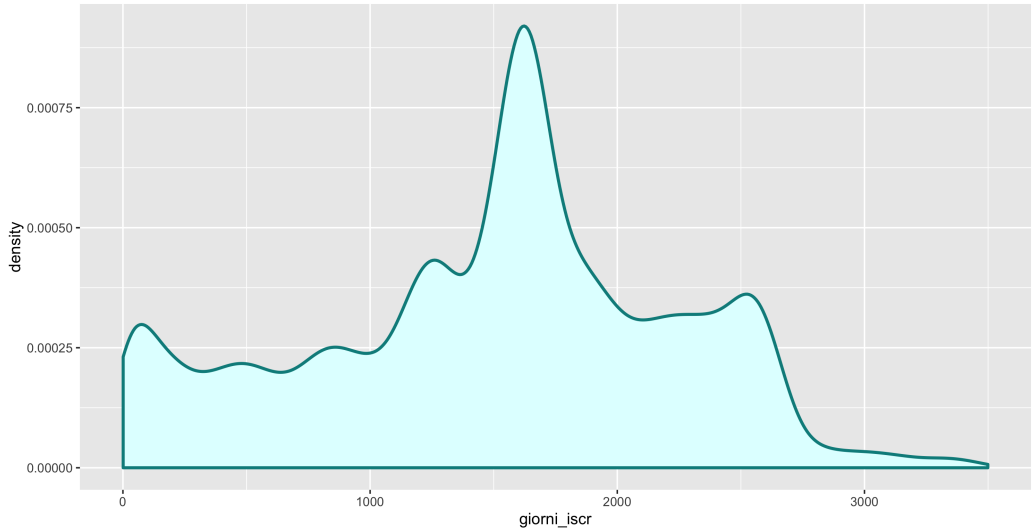

Figura 2.5: Utenti: densità empirica del numero di giorni trascorsi dall'iscrizione

star del mondo dello spettacolo decisero di promuovere le loro gesta utilizzando Twitter come mezzo. L'insieme di questi due fenomeni può giustificare la moda osservata nella distribuzione dei giorni trascorsi dall'iscrizione a Twitter.

2.2 Tweet

Dopo avere esaminato le caratteristiche degli utenti osservati, è il momento di passare ai tweet e, di conseguenza, alle tematiche che essi affrontano. Come illustrato nel Capitolo 1, il *crawling* dei tweet è avvenuto su sette temi. Ad un primo esame si è deciso di restringere a soli 4 temi il campo di analisi. Il tema Amministrative presentava problemi in fase di definizione delle query, con hashtag come #Milano o #Roma troppo generici; ciò ha comportato l'inserimento nel dataset di troppi tweet che nulla hanno a che vedere con il tema desiderato. Il tema Talk Show, invece, avrebbe richiesto l'adozione di un topic model per provare a distinguere la moltitudine di tematiche affrontate in 3 mesi di dibattiti televisivi, forse troppi. Perciò si è deciso, per gli scopi di questo lavoro, di ignorare anche questo tema. Infine le dimissioni del ministro Guidi non sono sembrate un argomento su cui misurare il posizionamento degli italiani. I quattro argomenti restanti sono perciò il **referendum Brexit**, la **legge Cirinnà**, il **ritorno del marò Girone**

ed il **referendum sulle trivelle**. Si è ritenuto che essi fossero congiuntamente in grado di fornire uno spaccato del panorama politico italiano, in quanto si vanno a toccare 4 tematiche tra loro distinte come politica estera, diritti civili, politica interna/difesa ed ambiente.

2.2.1 Flusso temporale e distribuzione geografica

In primo luogo l'attenzione si è concentrata sul flusso dei tweet, cioè sulla loro collocazione temporale negli archi di download. In Figura 2.6 è mostrato il numero di tweet scaricato ogni ora dal *crawler* relativamente al referendum sulle trivelle. Purtroppo, per un problema legato al server che ospitava il *crawler*, è presente un buco nei dati tra le giornate del 12 e 13 aprile.

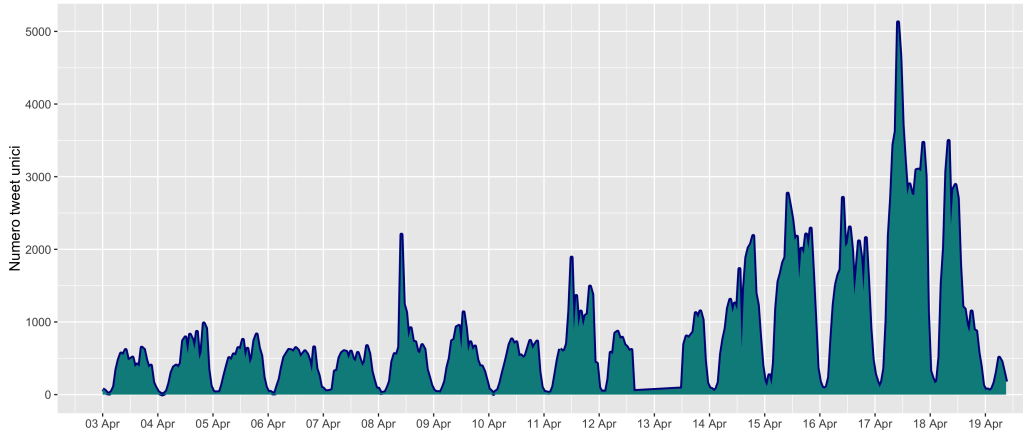

Figura 2.6: Referendum trivelle: numero di tweet scaricati ogni ora

I dati presentano ovviamente una stagionalità sulle 24 ore: durante la notte si hanno dei picchi negativi, mentre l'anadamento durante le ore diurne sembra ripresentare bene o male la stessa forma, con un picco nella mattinata ed uno nelle ore serali. Ovviamente il giorno con maggiore traffico è il 17 aprile, giorno della convocazione dei seggi, con il picco alle 10 di mattina (5125 tweet unici tra le 10:00 e le 11:00). Altre due note interessanti sono il trend crescente con l'avvicinarsi della data del referendum e l'assenza del cosiddetto "silenzio elettorale". È infatti tradizione che, nelle 24 ore che precedono la votazione, non si faccia campagna elettorale attiva. Evidentemente l'avvento dei social network ha rotto

anche questa tradizione. Infine non trova apparente spiegazione il picco riscontrato la mattina dell'8 aprile: esaminandone il contenuto, non sembra esserci un trend comune che lega questi tweet.
Accantonando la dimensione temporale e riprendendo un discorso già accennato nella Sezione 2.1, è interessante provare ad osservare quali aree siano state maggiormente attratte da un tema piuttosto che da un altro. Ciò è possibile risalendo alla zona attraverso l'utente che ha pubblicato il tweet. In Figura 2.7 viene riportato lo scarto tra la frequenza relativa rilevata nella zona per ogni tema e quella complessiva sull'insieme dei tweet. In questo modo è possibile osservare su quale tema si attivino maggiormente gli utenti di ogni area.

Centro ed Isole non sembrano segnalare rilevanti variazioni rispetto alla frequenza complessiva. Gli utenti del Nord-Ovest sembrano invece attivarsi maggiormente per Brexit e Cirinnà, mentre sono più freddi rispetto al ritorno del marò ed il referendum sulle trivelle. L'esatto opposto avviene invece per il Sud: spicca la partecipazione sul referendum sulle trivelle, visto l'impegno in prima linea di regioni come Puglia e Basilicata. Brexit e legge Cirinnà non sembrano attivare particolarmente invece questi *user*. Infine gli utenti del Nord-Est sono particolarmente sensibili a discussione ed approvazione della legge Cirinnà, mentre non sembrano molto attivi sul referendum sulle trivelle.

2.2.2 Termini più frequenti

Focalizzando l'attenzione sui contenuti, invece, si è scelto di rappresentare le *term-by-document matrix* servendosi di *wordcloud*. Questa tipologia di rappresentazione grafica permette di visualizzare stringhe testuali con dimensioni proporzionali ad una quantità, solitamente la loro presenza all'interno di un testo. Nel caso in esame le quantità su cui sono state costruite le *wordcloud* sono le somme per colonna delle matrici, cioè le frequenze assolute con cui i termini (o meglio gli stem) compaiono all'interno dei tweet. La prima *wordcloud* esaminata è quella costruita sui tweet relativi al referendum Brexit (Figura 2.8).

Ovviamente a farla da padrone sono i due termini "#brexit" e "brexit", che racchiudono la tematica trattata. Molto importante anche il ruolo dello stem "europ", che racchiude al suo interno molti termini riferiti all'Europa ed all'UE, e dell'hashtag "#ue", segno che, come lecito attendersi, buona parte del dibattito si

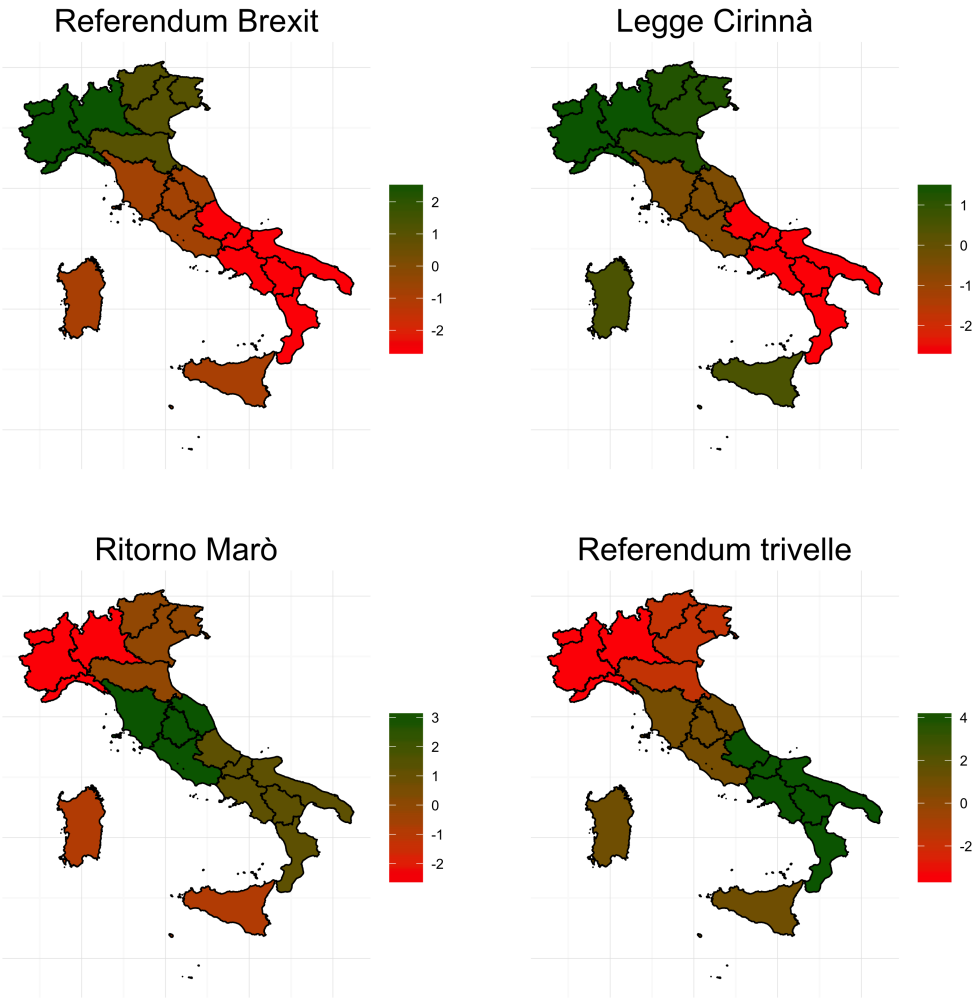

Figura 2.7: Numero di tweet per tema al variare della zona geografica

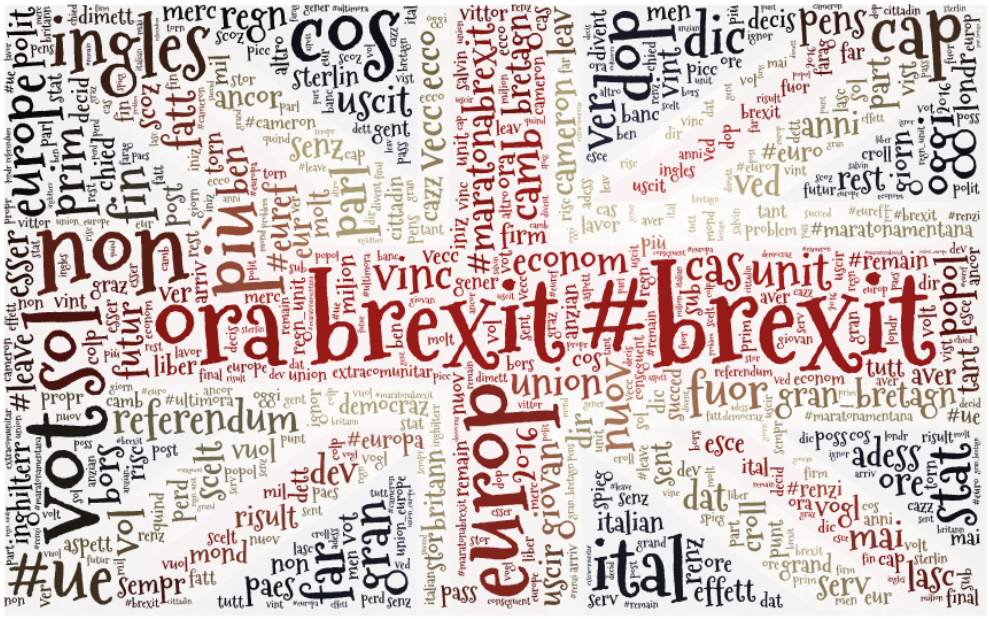

Figura 2.8: Termini maggiormente utilizzati nei tweet sul referendum Brexit

è concentrata anche sull'Europa, sul suo ruolo e sul suo futuro. Presenti anche stem come "ingles", "bretagn", "inghilterr", "gran bretagn" e "regn unit" (tra i pochi bigrammi visibili). Nella parte bassa compare l'Italia con "ital", somma di chi si preoccupa per il futuro del paese e chi invece evoca l'emulazione della Gran Bretagna per l'uscita dall'UE. Curioso vedere come assume un peso significativo viene assunto dallo stem "vecc" (in alto al centro). Questo fatto è dovuto alla polemica in seguito alle pubblicazioni di sondaggi secondo i quali la parte più anziana della popolazione sia stata quella maggiormente favorevole alla Brexit. Questo fatto ha fatto molto parlare Twitter e la wordcloud lo rivela. Un altro aspetto evidente è quello relativo alle conseguenze finanziarie del referendum, sottolineato dalla presenza di stem come "bors" o "sterlin". Infine spazio anche alla volontà popolare che ha avuto la meglio con termini come "popol", "referendum", "democraz" e "scelt" e "vot".

In Figura 2.9 è mostrata la *wordcloud* con i termini più utilizzati nei tweet relativi a discussione ed approvazione della legge Cirinnà. Come lecito aspettarsi, i due termini più utilizzati sono "#unionicivili" ed il bigramma "union civil", presenti nella query di ricerca e che rappresentano l'argomento centrale. Dividendo le altre

Figura 2.9: Termini maggiormente utilizzati nei tweet sulla legge Cirinnà

parole in gruppi tematici, è possibile notare stem come "ital", "diritt", "oggi", "liber", con i quali l'accento viene posto sull'aspetto dei nuovi diritti riconosciuti. Un altro filone è costituito da termini come "vot", "legg", "approv", "#opensenato", "parl" (che rappresenta lo stem di "parlamento"), "ddl", usati per esaminare e discutere degli aspetti relativi all'approvazione parlamentare della legge. C'è poi chi mette al centro il tema dell'amore, con "#lovewins", "emotelove" (cioè le emoticon "<3" e ":-*"), "amor". Invece, dall'altro lato, ci sono termini che indicano meno entusiasmo nei confronti della legge, come "#stopcirinnà", "schifezz" o "vergogn".

Figura 2.10: Termini maggiormente utilizzati nei tweet sul ritorno del marò Girone

Passando alla *wordcloud* relativa al caso marò, illustrata in Figura 2.10, la situazione cambia notevolmente. Infatti non è possibile individuare uno o due termini chiave, che spiccano all'interno della mappa. Questo è probabilmente dovuto

Figura 2.11: Termini maggiormente utilizzati nei tweet sul referendum sulle trivelle

alla numerosità totale dei tweet, in questo caso molto inferiore rispetto agli altri temi esaminati. Si riesce comunque ad individuare alcuni filoni, a partire dai termini "#marò", "marò", "giron", "salvator", "mar giron" che vanno a caratterizzare il tema. Si passa poi dal giubilo espresso da stem come "final", "bentorn", "ora", "felic" a termini relativi alla trafila giudiziaria come "indi", "indian", "pescator", "cort", "suprem", "#india". Spunta anche il tema politico, con "#renzi", "renz", "ministr", "pinott", e fanno la loro comparsa hashtag che identificano i principali telegiornali ed emittenti *all news*: "#skytg", "#tgla7","#tgcom", "#rainews".

In conclusione, in Figura 2.11, è mostrata l'ultima *wordcloud*, quella che riassume i tweet relativi al referendum sulle trivelle. Trattandosi, appunto, della con-

sultazione, a farla da padrona è "#referendum", seguito da "vot" e da "17 april", data del referendum. Balzano all'occhio numerosi termini facenti riferimento alla campagna comunicativa del fronte del sì, come "#iovotosi", "#votasì", "#notriv", "#stoptrivelle" ed il bigramma "vot sì". In seguito alla forte polemica scoppiata nei giorni immediatamente precedenti la convocazione dei seggi sull'astensionismo, si segnalano forti, oltre a "vot", hashtag come "#noquorum", "#battiquorum" "#quorum", "#iononvoto", "astension". Un altro termine rilevante presente in molte discussioni è "mar", dato che il quesito rigurdava le concessioni marittime. Connesso al mare, con un bel gioco di parole, assume un peso importante "#unmaredisì". Passando agli aspetti più politici, presente in molti tweet il riferimento al presidente del Consiglio ("renz" o "#renzi"), mentre sputna nella parte bassa della mappa "#ciaone". Il peso di questo hashtag non è alto ma è legato ad una forte polemica scoppiata nella giornata delle votazioni, quindi solo nell'ultima parte del *crawling*. Detto ciò, in poche ore l'hashtag in questione riesce a ritagliarsi un posto importante nella mappa.

2.3 Creazione di un *subset* di utenti

Nella Sezione 2.2 si è spostata l'attenzione sulla frequenza assoluta nell'intero insieme dei tweet dei vari termini utilizzati. Nel Capitolo 4 si cercherà, attraverso l'applicazione di modelli come la regressione multinomiale inversa, di capire quali siano i termini che, più di altri, connotano una linea di pensiero specifica. Per fare ciò occorre attribuire ai documenti, cioè all'insieme di tweet per tema del singolo utente, un *sentiment*, individuando quale opinione viene espressa attraverso il documento. L'operazione deve essere compiuta manualmente leggendo tutti i documenti, perciò si è reso necessario determinare un *subset* di utenti sui quali compiere questa operazione. La numerosità scelta per il campione è risultata pari a 5001 utenti, ma la scelta degli utenti ha voluto preservare le proporzioni presenti nel dataset descritto in Sezione 2.1 per quanto riguarda l'espressione sui temi. In Tabella 2.1 vengono riportate le frequenza assolute osservate per combinazioni di temi toccati, cioè sui quali un utente ha twittato almeno una volta. Sono stati esclusi dalla Tabella gli utenti che non scrivono su alcun tema.

Per creare il *subset* sono state replicate le frequenze relative della distribuzione congiunta dei temi, in modo da avere un campione rappresentativo del comporta-

				Referendum Trivelle			
				✓		X	
				Cirinnà		Cirinnà	
				✓	X	✓	X
Referendum Brexit	✓	Marò	✓	1111	810	210	587
			X	2274	5538	2121	19826
	X	Marò	✓	107	278	108	958
			X	1126	4127	11403	-

Tabella 2.1: Distribuzione degli utenti per temi

mento dell'intero dataset degli utenti. Inoltre sono stati aggiunti 5001 utenti come controllo, cioè estratti casualmente tra tutti coloro che non hanno twittato su alcuno di questi argomenti. Questo è il dataset su cui verranno applicati i modelli dei capitoli 3 e 4.

Capitolo 3

Analisi delle associazioni tra temi: modello marginale

Il primo obiettivo dell'analisi presentata è cercare di capire quali siano le caratteristiche che portino gli utenti a twittare su un tema piuttosto che su un altro. Per rispondere a questo quesito si è pensato che fosse sì importante la modellazione marginale dei quattro temi ma, ancora di più, risultasse interessante capire quali siano le dinamiche che intercorrono tra i diversi temi. Nella Sezione 3.1 si presenta brevemente la teoria dei modelli marginali, o riprendendo la definizione dell'articolo che li introdusse (Glonek and McCullagh (1995)) logit multivariati. In seguito vengono presentati i risultati dell'applicazione di questa tipologia di modelli ad uno specifico sottoinsieme dei dati.

3.1 Modellazione congiunta di più *outcome* categoriali

Siano $Y_1, Y_2, \ldots, Y_D$ D variabili categoriali, ciascuna avente rispettivamente r_d livelli, con $d = 1, \ldots, D$. Per il momento si supponga $r_d = 2$ $\forall d = 1, \ldots, D$. Essendo disponibili anche una serie di covariate X, una comune applicazione è la ricerca del legame presente tra le probabilità $\mathbb{P}[Y_d = 1]$ e le covariate stesse. La prima e più semplice classe di modelli parametrici che indaga questa relazione è sicuramente il modello logistico, dove

$$\eta = X\beta = logit(p_d) = \log\left(\frac{p_d}{1 - p_d}\right) \tag{3.1}$$

dove $p_d = \mathbb{P}[Y_d = 1]$.
Nel caso in cui però l'obiettivo sia non tanto la singola probabilità marginale quanto $\boldsymbol{\pi}$, cioè il vettore delle probabilità congiunte (dimensione $\prod r_d$), esistono due differenti approcci: il modello log-lineare oppure il cosiddetto modello marginale o logit multivariato. Contrariamente a quanto presente nella maggior parte della letteratura, si farà qui riferimento al modello logit multivariato nell'accezione data da Glonek and McCullagh (1995). Il termine *multivariato*, infatti, non viene più utilizzato per indicare la numerosità delle covariate X, bensì delle variabili risposta. Per illustrare i due esempi si prende il caso in cui $D = 3$. Il modello log-lineare si basa su una trasformazione $\pi \rightarrowtail \lambda$ che porta a definire le equazioni in funzione di π nei termini

$$\log(\pi_{ijk}) = \alpha + \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \cdots + \lambda_{ijk} \tag{3.2}$$

dove ogni λ rappresenta il logaritmo della probabilità sul rispettivo margine univariato o bivariato (es: $\lambda_{ij} = \log(\pi_{ij.})$). Ovviamente, inserendo il predittore lineare $\boldsymbol{\lambda} = X\boldsymbol{\beta}$, si giunge alla formulazione completa del modello log-lineare.
Il logit multivariato, invece, si basa sulla trasformazione $\pi \rightarrowtail \eta$ descritta in seguito (sempre nel caso $D = 3$)

$$\begin{gathered} \eta_1 = logit(\pi_{1..}) \qquad \eta_2 = logit(\pi_{.1.}) \qquad \eta_3 = logit(\pi_{..1}) \\ \eta_{12} = \log\big(\frac{\pi_{11.}\pi_{00.}}{\pi_{10.}\pi_{01.}}\big) \quad \eta_{13} = \log\big(\frac{\pi_{1.1}\pi_{0.0}}{\pi_{1.0}\pi_{0.1}}\big) \quad \eta_{23} = \log\big(\frac{\pi_{.11}\pi_{.00}}{\pi_{.10}\pi_{.01}}\big) \\ \eta_{123} = \log\frac{\pi_{111}\pi_{001}\pi_{010}\pi_{100}}{\pi_{011}\pi_{101}\pi_{110}\pi_{000}} \end{gathered} \tag{3.3}$$

dove . nel d-esimo pedice indica la somma sui livelli della variabile Y_d. In forma compatta il modello si può scrivere come

$$\boldsymbol{\eta} = C\log(M\boldsymbol{\pi}) \tag{3.4}$$

dove C e M sono due matrici costruite attraverso opportuni prodotti di Kronecker per permettere di esplicitare le relazioni esemplificate in (3.3). L'algoritmo di costruzione viene illustrato nella Sezione 3.2.1. In particolare, l'aggiunta della matrice M a quella dei contrasti C serve a permettere l'operazione di marginalizzazione, quella cioè che differenzia questa tipologia di modelli dai log-lineari.
Osservando le due parametrizzazioni appare evidente come il modello marginale

riesca a coniugare la modellazione delle D distribuzioni marginali con i cosiddetti *log-odds ratio* che permettono di modellare le dipendenze. Il modello log-lineare, invece, che si propone come estensione del modello in (3.1), risulta incompatibile con il logit univariato in quanto la relazione tra esplicative e covariate non è lineare su scala logistica.
Un altro aspetto che porta a privilegiare il modello marginale rispetto al log-lineare è la possibilità per il primo di esprimere logit diversi rispetto a quelli locali, semplicemente modificando opportunamente il criterio di costruzione delle matrici C e M. Bartolucci and Farcomeni (2009), riprendendo la definizione di Agresti (2002), illustrano tre differenti tipologie di logit, pensati per permettere anche il trattamento di *outcome* ordinali:

- locale: $\eta_d(y;l) = \log \frac{\mathbb{P}[Y_d=y+1]}{\mathbb{P}[Y_d=y]}$
- globale: $\eta_d(y;g) = \log \frac{\mathbb{P}[Y_d\geq y]}{\mathbb{P}[Y_d<y]}$
- *continuation:* $\eta_d(y;c) = \log \frac{\mathbb{P}[Y_d\geq y]}{\mathbb{P}[Y_d=y-1]}$

Nel caso di variabili categoriali nominali, l'unico logit sensato risulta essere quello locale, in quanto non vi è modo di imporre un ordinamento tra le categorie. Questa specificazione rappresenta però un vantaggio notevole a livello di flessibilità per il modello marginale. Inoltre nel momento in cui si definiscono i *log-odds ratio* è possibile coniugare differenti tipologie di logit sulle singole variabili. Ad esempio, nel caso in cui Y_1 sia nomiale e Y_2 ordinale, è possibile definire $\eta_{12}(y_1,y_2;l,g) = \log \frac{\mathbb{P}[Y_1=y_1,Y_2\geq y_2]\mathbb{P}[Y_1=y_1-1,Y_2<y_2]}{\mathbb{P}[Y_1=y_1,Y_2<y_2]\mathbb{P}[Y_1=y_1-1,Y_2\geq y_2]}$, combinando un logit locale ed uno globale.
Sintetizzando quanto detto finora, il modello marginale permette di modellare congiuntamente più variabili risposta mantenendo l'interpretabilità del logit univariato, anche in presenza di variabili ordinali. D'altro canto, però, la stima del modello risulta molto più complicata rispetto al modello log-lineare, per cui il passaggio attraverso un modello di regressione di Poisson rende facilmente trattabile la verosimiglianza. Inoltre, potendo rientrare nella classica teoria dei GLM, per quest'ultimo sono disponibili metodi asintotici per test e confronti tra modelli basati sulla devianza. Nella Sezione 3.2 viene definita la funzione di verosimiglianza del modello marginale, esaminando i problemi relativi alle stime dei parametri coinvolti.

3.2 Stima del modello marginale

Si ritorni nel caso generale in cui $Y_1, \ldots, Y_D$ siano variabili categoriali (ordinali o nominali) con rispettivamente r_d livelli. Sia $\boldsymbol{\pi}$ il vettore delle probabilità congiunte di tutte le possibili configurazioni delle variabili risposta in ordine lessicografico. Ad esempio, nel caso di tre variabili binarie, l'ordine lessicografico risulta essere (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0) e (1,1,1). Riprendendo l'equazione (3.4), il modello marginale viene espresso in forma generale come $\boldsymbol{\eta} = C \log(M\boldsymbol{\pi})$. La matrice C va ad identificare i contrasti, cioè determina i sottoinsiemi di variabili considerati in ogni equazione. La matrice M, invece, permette la marginalizzazione. È necessario sottolineare come il vettore $\boldsymbol{\eta}$ sia composto dalle possibili interazioni tra le variabili più un elemento nullo che garantisce il vincolo di somma unitaria, vincolando $\log(\sum \boldsymbol{\pi}) = 1$, e che la trasformazione $\boldsymbol{\pi} \rightarrowtail \boldsymbol{\eta}$ sia di rango pieno.
Considerando le osservazioni $Y_i \sim M(n_i, \boldsymbol{\pi}_i)$, cioè come frequenze di una Multinomiale, si ha che la funzione di log-verosimiglianza risulta pari a $\ell(\boldsymbol{\beta}, Y) = \sum_{i=1}^{m} y_i^T \log \boldsymbol{\pi}_i$. L'ordine m della somma corrisponde alle differenti configurazioni possibili di (y_i, x_i), quindi nel caso in cui si abbia almeno una covariata continua si ha $m = n$ e $n_i = 1 \ \forall i = 1, \ldots, n$. Il vero problema sorge nell'esplicitare la trasformazione $\boldsymbol{\eta} \rightarrowtail \boldsymbol{\pi}$, cioè nel calcolare il vettore di probabilità dato il vettore dei predittori lineari. Risulta necessario invertire la relazione presentata in 3.4, ma non esiste alcuna soluzione esplicita. Inoltre, dovendo assicurarsi che $\boldsymbol{\pi} > 0$, si decide di lavorare con $\boldsymbol{\nu} = \log \boldsymbol{\pi}$. Glonek and McCullagh (1995) applicano una procedura iterativa di Newton-Raphson per risolvere in $\boldsymbol{\nu}$ l'equazione $\boldsymbol{\eta} = C \log(M \exp \boldsymbol{\nu})$. Innanzitutto occorre definire la derivata della trasformazione. Da 3.4 si ha immediatamente che

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}} = CD^{-1}M \tag{3.5}$$

dove $D = diag(M\boldsymbol{\pi})$. A questo punto è possibile definire l'algoritmo NR per ottenere stime di $\boldsymbol{\pi}$ dato $\boldsymbol{\eta}$. Partendo da un punto ν_0, l'equazione di aggiornamento risulta pari a

$$\boldsymbol{\nu}_k = \boldsymbol{\nu}_{k-1} - \{CD_{k-1}^{-1} M diag(\boldsymbol{\nu}_{k-1})\}^{-1} \{C \log(M \exp \boldsymbol{\nu}_{k-1}) - \boldsymbol{\eta}\} \tag{3.6}$$

dove $D_{n-1} = diag(M \exp \boldsymbol{\nu}_{k-1})$. Data la natura iterativa dell'algoritmo, occorre fissare dei criteri di convergenza per determinare l'arresto della procedura. Una volta

applicato questo algoritmo per ogni configurazione di (y_i, x_i) è possibile calcolare la verosimiglianza e, di conseguenza, massimizzare la funzione.

3.2.1 Costruzione delle matrici di contrasto e marginalizzazione

Colombi and Forcina (2001) illustrano nel dettaglio la costruzione delle matrici C e M utilizzate per la costruzione e specificazione del modello marginale.
Sia B una matrice dal numero di righe pari alla dimensione di $\boldsymbol{\eta}$ (escludendo il termine nullo) e numero di colonne pari a D. Il valore B_{ij} è pari a 1 se la j-esima variabile è inclusa nel set di variabili che definiscono l'i-esimo *odds ratio*. Si definisce allora la matrice C_i come

$$C_i = \bigotimes_j C_{ij} \qquad dove \qquad C_{ij} = \begin{cases} [-I_{r_j-1} I_{r_j-1}] & \text{se } B_{ij} = 1 \\ 1 & \text{se } B_{ij} = 0 \end{cases} \tag{3.7}$$

dove I_{r_j-1} è la matrice identità di dimensione $r_j - 1$. C viene costruita come diagonale a blocchi, dove ogni blocco corrisponde ad una matrice C_i.
Analogamente M è una matrice a blocchi, dove ogni blocco M_i è definito, nel caso di logit locale, come

$$M_i = \bigotimes_j M_{ij} \qquad \text{dove} \qquad M_{ij} = \begin{cases} I_{2(r_j-1)} & \text{se } B_{ij} = 1 \\ 1'_{b_j} & \text{se } B_{ij} = 0 \end{cases} \tag{3.8}$$

3.2.2 Implementazione del modello marginale

L'implementazione del modello si è resa possibile grazie ad alcune funzioni contenute nel pacchetto `ggm`, sviluppato dal prof. Giovanni Marchetti dell'università degli Studi di Firenze (per dettagli Marchetti, Drton, and Sadeghi (2015)). Non essendo presente una funzione che restituisse le stime in situazioni diverse da quella in cui $D = 2$ e $r_1 = r_2 = 2$, sono state utilizate delle funzioni interne per costruire un ottimizzatore. In particolare, `binve` permette di calcolare la trasformazione $\boldsymbol{\eta} \rightarrowtail \boldsymbol{\pi}$ fornendo in input il vettore $\boldsymbol{\eta}$ e le matrici di contrasto e marginalizzazione. È stato necessario, dati alcuni problemi riscontrati, modificare la funzione aggiungendo un controllo per evitare una situazione di *overflow*: nel caso in cui tutta la probabilità si concentrasse su una cella, raggiunta una certa iterazione,

la funzione internamente restituiva valori non finiti che impedivano la convergenza. L'altra funzione utilizzata è stata marg.param per il calcolo delle matrici C e M. Una volta ottenuti questi valori è stato possibile massimizzare numericamente la verosimiglianza multinomiale. Data la necessità di applicare l'algoritmo di Newton-Raphson un numero di volte pari al numero di unità in presenza di covariate continue, si è reso necessario parallelizzare numerose funzioni al fine di contenere i tempi computazionali.

3.3 Applicazione

Il modello marginale è stato applicato al *subset* di utenti creato nella sezione 2.3, contenente 5001 utenti attivi su almeno uno dei quattro temi analizzati e 5001 utenti di controllo estratti casualmente dagli utenti attivi sui tre temi scartati. Sono state modellate le distribuzioni marginali con le covariate twitter-biografiche e le interazioni a coppie solo tramite la specificazione dell'intercetta. Le interazioni di ordine superiore al secondo, invece, sono state vincolate a 0. Lo scopo di questa analisi era, per il momento, cercare di capire quali temi fossero legati tra loro, dando più importanza all'aspetto interpretativo che a quello previsivo.

3.3.1 Stime di massima verosimiglianza

Dato che l'ottimizzazione della funzione di verosimiglianza associata al modello è stata effettuata per via numerica, è stato necessario tutelarsi circa la presenza di massimi locali. Per questo motivo è stata lanciata una prima volta la funzione dall'origine dello spazio $\mathbb{R}^{58}$ (numero totale di parametri coinvolti). Una volta ottenuto il punto di massimo, sono state effettuate 100 perturbazioni casuale dello stesso tramite una Normale a componenti indipendenti a media 0 e deviazione standard 0.1 e sono stati utilizzati i 10 punti che hanno resituito un valore di verosimiglianza maggiore come punti di partenza per l'algoritmo di ottimizzazione. Purtroppo, data la dimensione elevata dello spazio da indagare e l'onerosità computazionale dell'ottimizzazione, non è stato possibile operare una soluzione più elegante per la ricerca del massimo globale. La scelta di lanciare la prima ottimizzazione con punto di partenza nell'origine è stata presa in un'ottica conservativa, dato appunto l'elevato numero di parametri in gioco. In Tabella 3.1 sono riportate

		Brexit	**Cirinnà**	**Marò**	**Trivelle**
Intercetta		-1.11498***	-0.82932***	-2.19532***	-3.67523***
Sesso	*F*	0.55252***	0.20099***	0.37401***	-0.38538***
(rif=*Ente*)	*M*	0.61145***	-0.04781	0.17877***	-0.03211
Zona	*Centro*	-0.13673**	0.10522*	-0.0225	0.3579***
(rif=*Nord-ovest*)	*Estero*	-0.19098***	-0.63587***	-0.53086***	-0.61581***
	Isole	-0.0242	0.28443***	0.24966***	0.25447***
	Nord-Est	0.18152***	0.27488***	0.15797*	-0.15744***
	Sud	-0.1264*	0.20914***	0.02551	0.23981***
log(# follower)		0.06988***	-0.04003*	*-0.02494*	*0.02368*
log(# following)		-0.06076***	-0.08295***	-0.06839***	*0.02634*
Verificato		-0.19199***	0.22054***	0.10609***	0.12949***
Giorni Iscr.		*0.000001*	0.00038***	0.00028***	*0.00008*
log(tweet/giorno)		0.18459***	0.2004***	0.12451***	0.23157***
Cirinnà		0.00702***	-	-	-
Marò		0.22571***	-0.05791***	-	-
Trivelle		0.10958*	-0.03340***	0.21624***	-
$\ell(\boldsymbol{\beta})$	-16544.07				

Tabella 3.1: Stima di massima verosimiglianza per il modello marginale completo

le stime di massima verosimiglianza dei parametri del modello. La prima osservazione che occorre fare riguarda la significatività della maggior parte dei coefficienti. Le quattro intercette non sono direttamente interpretabili a livello di valore assoluto perchè per le variabili in scala logaritmica non è realistico assumere il valore 0 come riferimento. Andando poi con ordine, la Brexit sembra essere il tema su cui il sesso incide maggiormente. Gli enti non si sono molto espressi su questo tema, prova ne sono i valori elevati dei due coefficienti per M e F e la loro forte significatività. Per la legge Cirinnà, invece, non emergono significative differenze tra maschi ed enti, mentre le donne sono più portate ad esprimeri su questo tema. L'opposto di quanto avviene invece per il referendum sulle trivelle: qui a farla da padrone sono gli enti, senza significative differenze con gli uomini. Il coefficiente relativo alle donne è invece significativo ma di segno negativo. La prevalenza degli enti in questo tema si può spiegare con la presenza nel dataset di numerose ONG quali Greenpeace, Legambiente e simili, ovviamente molto attive nei giorni pre-

cedenti il referendum. Per quanto riguarda il ritorno del Marò Girone si osserva un comportamento simile alla Brexit ma con valori assoluti più contenuti ed una prevalenza, a dire il vero inaspettata, del genere femminile. Passando alla zona, dove la modalità di riferimento è il Nord-Ovest, si notano comportamenti molto diversi a seconda del tema. Una costante è sicuramente la scarsa propensione degli utenti residenti all'estero nel trattare i temi in questione, con coefficienti negativo e significativi per tutti e 4 i temi. Per quanto riguarda la Brexit la zona più "attiva" sembra essere il Nord-Est, seguita da Nord-Ovest ed Isole senza significative distinzioni e, nell'ordine (almeno stando ai valori dei coefficienti), Sud, Centro ed Estero. In realtà la differenza tra Sud e Centro è talmente esigua che qualche dubbio circa la sua significatività sorge. Isole e Nord-Est la fanno invece da padroni sulla legge Cirinnà, seguiti dal Sud e dal Centro, anche se la significatività di quest'ultimo coefficiente è scarsa. Poco attivo invece il Nord-Est, prova ne è il fatto che l'unico coefficiente negativo sia proprio quello relativo agli utenti provenienti dall'estero. Più incerta la situazione relativa al ritorno del Marò, con le Isole in prima linea seguite dal Nord-Est, mentre Sud, Nord-Ovest e Centro non sembrano apportare differenze significative alla probabilità di trattare l'argomento. Tutti i significativi i contrasti con il Nord-Ovest per il referendum sulle trivelle: l'unica zona, oltre al solito Estero, ad essere meno coinvolta sul tema è il Nord-Est. Come lecito attendersi, data la mobilitazione dei presidenti di molte regioni dell'Italia centro-meridionale e la connessione del tema trivelle alla salvaguardia dei mari, i coefficienti di Sud, Isole e Centro sono positivi e significativi. Passando alle caratteristiche relative all'account Twitter, sembra che a trattare il tema Brexit siano utenti mediamente molto seguiti ma con pochi *following*, mentre il possedere un account verificato (caratteristica degli account di personaggi famosi e popolari) è associato negativamente al tema. Passando alla legge Cirinnà l'essere connesso ad altri utenti, in entrambe le direzioni, ha un effetto negativo sulla probabilità di trattare il tema, mentre i personaggi pubblici e gli utenti più "anziani" sembrano essere particolarmente sensibili alla questione. Anzianità dell'account ed account verificato portano, seppure in misura minore a parità di altre covariate, ad aumentare la probabilità di trattare anche il tema Marò, mentre c'è associazione negativa con il numero di profili seguiti. Le connessioni di rete non vanno invece ad influire sulla probabilità di trattare il tema Trivelle. L'unica covariata che influisce, tra quelle trattate finora, è la verifica dell'account, anche in questo caso con segno

positivo. Discorso a parte merita invece la variabile che racchiude, in scala logaritmica, il numero di tweet medi giornalieri scritti dagli utenti dalla loro iscrizione al social network. Tutti e quattro i valori sono significativi e positivi: questo fatto è abbastanza intuitivo in quanto ci si aspetta che se un utente più attivo sia in generale più portato a trattare un maggior numero di temi. A parità di attività e delle altre covariate in gioco, però, sembra essere il referendum sulle trivelle il tema su cui più facilmente si attiva un utente, seguito da Cirinnà e Brexit.
La novità apportata dal modello utilizzato sta però nelle interazioni tra le variabili risposta. Risulta dalla seconda parte di Tabella 3.1 che vi siano interazioni significative tra quasi tutti gli *outcome*. I due valori maggiori si hanno per le coppie Brexit-Marò e Trivelle-Marò, in entrambi i casi positivi. La positività del predittore lineare sull'interazione sta a significare che è maggiore la probabilità che il comportamento (scrivere o no del tema) dell'utente verso le due tematiche sia la medesima rispetto a due comportamenti differenti, qualsiasi essi siano. Date anche le numerosità in gioco verrebbe da dire che chi è attivo su Brexit o Trivelle ha alta probabilità di esserlo anche sui Marò. Seppure con valore assoluto minore e significatività al 5%, è significativa e positiva anche l'associazione tra Brexit e Trivelle. Particolare è il comportamento della variabile relativa alla legge Cirinnà: l'associazione con la Brexit è positiva e significativa ma quantitativamente esigua, mentre con Marò e Trivelle l'associazione è negativa.

3.3.2 Selezione delle variabili

Per selezionare le variabili del modello, nonostante le significatività siano numerose, è stata proposta una procedura *one-block backward*. Essa consiste nella rimozione contemporanea di tutte le variabili non significative (soglia posta per p-value pari a 0.10). La soluzione è stata proposta perchè computazionalmente sarebbe risultato eccessivamente oneroso procedere con una usuale procedura *backward*. Ripetendo l'esplorazione compiuta in precedenza, perturbando in questo caso il vettore $\hat{\boldsymbol{\beta}}$ dei parametri illustrati in Tabella 3.1 privato dei parametri non significativi (in corsivo nella tabella), si è raggiunto un valore di log-verosimiglianza massimo pari a $\ell(\beta) = -16457.40$. Applicando un test rapporto di verosimiglianza si ottiene il valore $\chi_{6,obs} = 173.34$, il cui p-value è pari a 0. Ciò significa che la perdita di informazione dovuta alla riduzione del modello è significativa, perciò si

rifiuta l'uguaglianza tra il modello ristretto ed il modello completo.

3.3.3 Accuratezza del modello: curve ROC

Nonostante lo scopo dell'analisi non sia prettamente previsivo, può essere un'indicazione di bontà del modello indagare la sua accuratezza. In casi di classificazione binaria come quelli trattati una buona indicazione può essere fornita dalla curva ROC e dall'indice AUC. La curva ROC, a fronte di una previsione compiuta sulla probabilità che la variabile risposta sia pari a 1, si costruisce ordinando il vettore delle previsioni e muovendo la soglia oltra a cui si considera classificata come successo l'unità. Per ogni valore della soglia si registrano due valori: la *sensitivity*, pari al tasso di veri positivi, e la *specificity*, pari al tasso di veri negativi. La curva si costruisce unendo i punti corrispondenti alle coordinate (1-*specificity*, *sensitivity*) per ogni valore assunto dalla soglia. In Figura 3.1 sono illustrate le quattro curve ROC relative al modello marginale implementato per i 4 *outcome*. La migliore accuratezza sembra essere raggiunta per la previsione del comportamento circa il ritorno del Marò Girone, mentre per il referendum non si hanno ottime performance. Quest'ultimo fatto sembra essere abbastanza in accordo con il numero elevato di variabili non significative nel predittore marginale del logit di questa probabilità. Una misura di sintesi delle curve mostrate è l'indice AUC, che calcola l'area sottostante la curva ROC. In caso di assegnazione randomica l'indice assume indice pari a 0.5 (la curva è la bisettrice del quadrante). In Tabella 3.2 sono riportati i valori dell'indice per le quattro curve illustrate, che confermano le impressioni avute dai grafici delle curve.

Brexit	0.618
Legge Cirinnà	0.612
Ritorno Marò	0.702
Referendum Trivelle	0.596

Tabella 3.2: Modello marginale: valori dell'indice AUC per le quattro variabili risposta

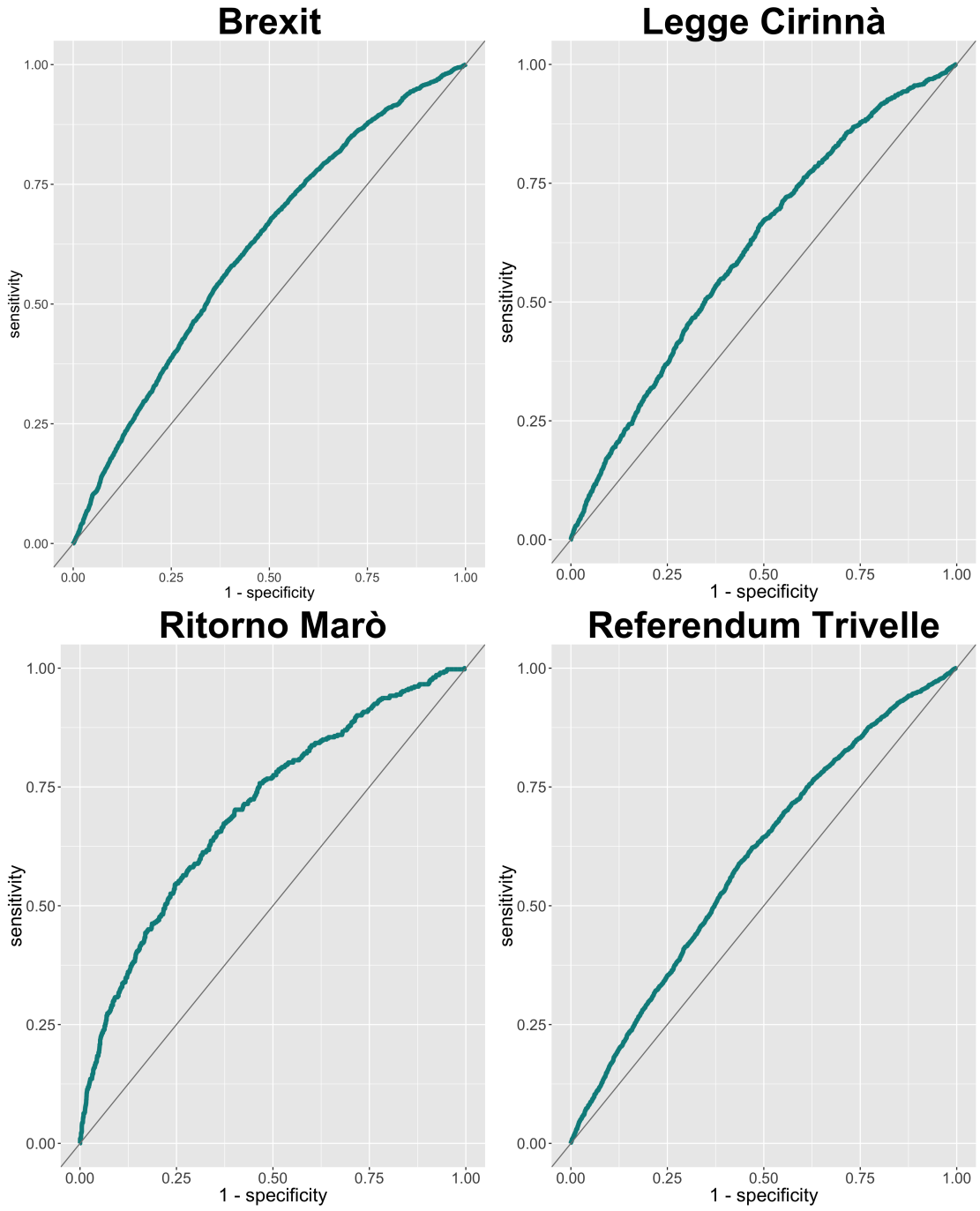

Figura 3.1: Curva ROC relativa alle quattro variabili previste tramite modello marginale

Capitolo 4

Sentiment analysis: regressione multinomiale inversa

Dopo avere esaminato quali sono le caratteristiche che portano gli utenti a scrivere su un tema piuttosto che su un altro e quali temi sono tra loro più strettamente connessi, è giunto il momento di esaminare il contenuto dei tweet. Per fare ciò si è deciso di utilizzare un modello sviluppato da Matt Taddy (University of Chicago) intorno al 2013 e noto come regressione multinomiale inversa. Grazie a questa metodologia è possibile ridurre dimensionalmente le *term-by-document matrix* caratteristiche della collezione di documenti. Si passa dalle frequenze di tutti i termini ad un numero limitato di fattori basati sui *sentiment* espressi e che possiedono particolari proprietà. In questo modo diventa intuitivo lavorare con quello che viene chiamato modello *forward* (Taddy (2013a)), in cui si cerca di spiegare e/o prevedere il sentiment degli utenti.

4.1 Regressione multinomiale inversa

La maggior parte delle procedure di *text analysis* presenti in letteratura si basa sulle *term-by-document matrix* ed utilizza le frequenze dei *token* come predittori. Agendo in questo modo non ci si preoccupa particolarmente della natura dei dati analizzati e si rischia di utilizzare modelli non adatti. Si ricorda, ad esempio, che nel caso delle *tbd matrix*, cioè le rappresentazioni vettoriali dei documenti nello

spazio dei *token* (Capitolo 1), ci si trova dinanzi a matrici con elevati livelli di sparsità. Una strada particolarmente indicata sviluppata di recente porta alla definizione di modelli dove sia prevista una riduzione dimensionale *text-specific* basata sulla distribuzione multinomiale. Un tipico esempio di questo tipo di modello è la *Latent Dirichelet Allocation* (LDA). Si tratta di un *topic model* che tratta i documenti come estratti da una distribuzione multinomiale il cui vettore di probabilità è individuato come combinazione lineare di quantità specifiche per ogni topic. Il vantaggio si ha quindi nel considerare un solo vettore di probabilità $\boldsymbol{\theta}$ per topic e non per documento, con una notevole riduzione delle dimensionalità in gioco. Passando ad analisi supervisionate, come nel caso della *sentiment analysis*, il concetto chiave è che non è possibile modellare in maniera efficiente la risposta condizionata $Y|X$ date le dimensioni di X. Per questo motivo si cercano soluzioni alternative, di cui la regressione multinomiale inversa rappresenta sicuramente un importante esempio.

4.1.1 Specificazione del modello

Sia $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]$ la rappresentazione vettoriale di un documento nello spazio dei p *token* del vocabolario, con $i = 1, \ldots, n$ indice del documento. La *term-by-document matrix* viene allora definita come $X = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_n]'$. Inoltre sia $m_i = \sum_j x_{ij}$ il numero totale di *token* utilizzati in un documento. Sia $\boldsymbol{y} = [y_1, \ldots, y_n]$ il vettore dei *sentiment* associati ai documenti. Volendo dare una definizione di *sentiment*, non fermandosi alla semplice polarità, Taddy (2013a) parla di "qualità sensibile" correlata con il testo, in modo da potere includere più sfaccettature espresse nel documento.

Regressione Inversa e riduzione di sufficienza

Il concetto chiave della MNIR (*MultiNomial Inverse Regression*) risiede nel compiere inferenza sulla distribuzione di X/Y al fine di individuare una riduzione dimensionale della matrice del documento. Sulla scorta di quanto teorizzato in Cook (2007), la regressione inversa viene formulata come

$$\boldsymbol{x}_i = \Phi \boldsymbol{v}_i + \boldsymbol{\epsilon}_i \tag{4.1}$$

dove $\boldsymbol{v}_i$ è un vettore di "fattori risposta" di dimensione K, costruito applicando opportune funzioni ad y_i. Φ è la matrice di dimensione $p \times K$ che contiene i coefficienti della regressione inversa, mentre $\boldsymbol{\epsilon}$ è un vettore di termini di errore. Attraverso Φ è dunque possibile calcolare la proiezione dei vettori $\boldsymbol{x}_i$ (documenti) in un nuovo spazio: $\boldsymbol{z}_i = \Phi' \boldsymbol{x}_i$. La proprietà forte derivante dalla regressione inversa (e dimostrata in Taddy (2013a)) risiede nella riduzione di sufficienza fornita dai vettori $\boldsymbol{z}_i$, cioè

$$y_i \perp\!\!\!\perp \boldsymbol{x}_i | \boldsymbol{z}_i = \Phi' \boldsymbol{x}_i \tag{4.2}$$

In pratica, nel momento in cui si conosce il valore dei vettori $\boldsymbol{z}$, il *sentiment* ed il testo diventano indipendenti, quindi si può utilizzare la matrice $Z = [\boldsymbol{z}_1 \dots \boldsymbol{z}_K]'$ come predittore. Operando in questo modo si ha una riduzione dimensionale da p a K covariate, perdendo i problemi derivati dalla sparsità delle *tbd matrix*.

Contestualizzando al caso di interesse, cioè di analisi testuale, Taddy (2013a) applica i concetti visti finora nel caso in cui i predittori seguano una distribuzione multinomiale. Data $\boldsymbol{x}_i$ la rappresentazione vettoriale dell'i-esimo documento con numero totale di token utilizzati pari a m_i, si ha

$$\boldsymbol{x}_i \sim MN(\boldsymbol{q}_i, m_i) \quad con \quad q_{ij} = \frac{e^{\eta_{ij}}}{\sum_{j=1}^{p} e^{\eta_{ij}}} \quad e \quad \eta_{ij} = \alpha_j + u_{ij} + \boldsymbol{v}_i' \varphi_j \tag{4.3}$$

dove $j = 1, \dots, p$ indica i token e $i = 1, \dots, n$ i documenti. Viene applicato il classico legame multilogit ai predittori lineari composti da tre parti:

- α_j è l'intercetta specifica del termine
- u_{ij} è l'effetto del soggetto sull'uso del termine
- $\boldsymbol{v}_i' \varphi_j$ è l'effetto del sentiment sull'uso del termine

In questo modo, stimando la matrice Φ è possibile ricavare le proiezioni della *tbd matrix* X nel nuovo spazio. In realtà è possibile un'ulteriore semplificazione del modello: al posto delle singole $\boldsymbol{x}_i$ è possibile modellare i conti collassati sulle possibili configurazioni di $\boldsymbol{v}_i$ (se discreto), quindi $\boldsymbol{x}_v$. In questo modo si suppongono intercette comuni ai soggetti ($u_{ij} = 0 \quad \forall (i,j)$) e non è possibile introdurre, di conseguenza, covariate specifiche del soggetto nell'equazione. Nelle analisi presentate,

al fine di controllare meglio il comportamento del singolo utente, si è deciso di adottare l'approccio illustrato in 4.3. Inoltre, seguendo come traccia quanto illustrato in Taddy (2013a) e Taddy (2013b), nel caso di *sentiment* espressi come variabili categoriali si segue l'approccio per cui $v_i = y_i$. Di conseguenza il predittore lineare in (4.3) diventa $\eta_{ij} = \alpha_j + u_{ij} + y_i'\varphi_j$ e si ha che K è uguale al numero di livelli di y_i.

4.1.2 Stima della MNIR: regressione *gamma-lasso*

La procedura di stima della regressione multinomiale inversa illustrata in Taddy (2013a) richiede innanzitutto la specificazione di distribuzioni a priori per i parametri coinvolti. L'algoritmo *gamma-lasso* prevede infatti la massimizzazione della distribuzione a posteriori per Φ, fornendo le cosiddette stime *MAP* (*Maximum A Posteriori probability*). La scelta di percorrere questa strada viene giustificata attraverso l'eccessiva onerosità computazionale dell'applicazione di un *Gibbs sampler*, date le dimensioni in gioco. La struttura delle distribuzioni a priori risulta essere:

- $\alpha_i \sim N(0, \sigma_\alpha^2)$ con $\alpha_k \perp\!\!\!\perp \alpha_w \quad \forall k \neq w$
- $e^{u_{ij}} \sim Gamma(1, 1)$
  Modellando e^u e non direttamente u si applica un moltiplicatore ai rapporti trovati; la distribuzione ha moda in 0, quindi permette di individuare termini non utilizzati dagli utenti, media in 1, il che centra il modello su una intercetta comune, e ha una coda pesante in modo da permettere che alcuni utenti utilizzino termini più rari
- $\varphi_{jk} \sim Laplace(\lambda_{jk})$ per $j = 1, \ldots, p \quad k = 1, \ldots, K$
  La scelta della Laplace sui coefficienti Φ serve, seguendo la letteratura sull'interpretazione bayesiana del lasso con penalizzazione L_1, per fare in modo che molti coefficienti siano posti uguali a 0 compiendo un'operazione di *shrinkage* ma, contemporaneamente, consentire valori alti nelle code pesanti
- $\lambda_{jk} \sim Gamma(s, r)$
  L'iperpriori sul parametro di precisione della Laplace viene posta, dato l'alto nuermo di predittori ($j = 1, \ldots, p$), per evitare la sovrapenalizzazione che si

potrebbe creare con un singolo valore fissato e quindi consentire maggiore flessibilità

Riferendosi alla struttura di priori ed iperpriori posta su Φ si parla di priori Gamma-Laplace, dove

$$GL(\varphi_{jk}, \lambda_{jk}) = \frac{\lambda_{jk}}{2} e^{-\lambda_{jk}|\varphi_{jk}|} \frac{r^s}{\Gamma(s)} \lambda_{jk}^{s-1} e^{-r\lambda_{jk}} \tag{4.4}$$

In alcune prove eseguite e riportate in Taddy (2013a) e Taddy (2013b) viene sottolineato come i risultati risultino robusti a variazioni dei valori dei parametri s e r. Spesso i valori dei parametri vengono scelti anche al fine di migliorare la convergenza dell'algoritmo di stima e fare in modo che venga effettuata l'operazione di *shrinkage* auspicata. A questo punto diventa possibile esplicitare la distribuzione a posteriori

$$p(\Phi, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{U}) \propto \prod_{i=1}^{n} \prod_{j=1}^{p} q_{ij}^{x_{ij}} \pi(u_{ij}) N(\alpha_j; 0, \sigma_\alpha^2) \prod_{k=1}^{K} GL(\varphi_{jk}, \lambda_{jk}) \tag{4.5}$$

Per la massimizzazione della distribuzione a posteriori viene costruito un algoritmo *ad hoc*, dopo alcune osservazioni. Taddy (2013a) sottolinea come sia possibile riscrivere la struttura di a priori individuata per Φ e λ come penalizzazione della log-verosimiglianza in Φ. Si ottiene che la stima MAP per Φ e λ è equivalente alla stima di massima verosimiglianza per Φ penalizzata per

$$c(\Phi) = \sum_{j=1}^{p} \sum_{k=1}^{K} c(\varphi_{jk}) \qquad dove \qquad c(\varphi_{jk}) = s \log\left(1 + \frac{varphi_{jk}}{r}\right) \tag{4.6}$$

In Figura 4.1 è possibile osservare la forma di questa funzione penalizzazione nel caso univariato per due differenti coppie di valori per s e r. La cosa evidente è sicuramente il punto angoloso in 0, causa della nullità di alcuni parametri di Φ. L'intensità dell'operazione di shrinkage dipende evidentemente dal valore dei parametri s e r. La massimizzazione di questa verosimiglianza penalizzata non avviene attraverso l'usuale procedura di risoluzione data un griglia di valori di *lambda*, data che il calcolo della verosimiglianza sarebbe troppo oneroso. Si procede perciò con un algortimo cosiddetto *coordinate descent*: si procede aggiornando ogni parametro condizionandosi al valore degli altri. Data l'impossibilità di ottenere stime in forma chiusa per i minimi condizionati, inoltre, si utilizzano approssimazioni

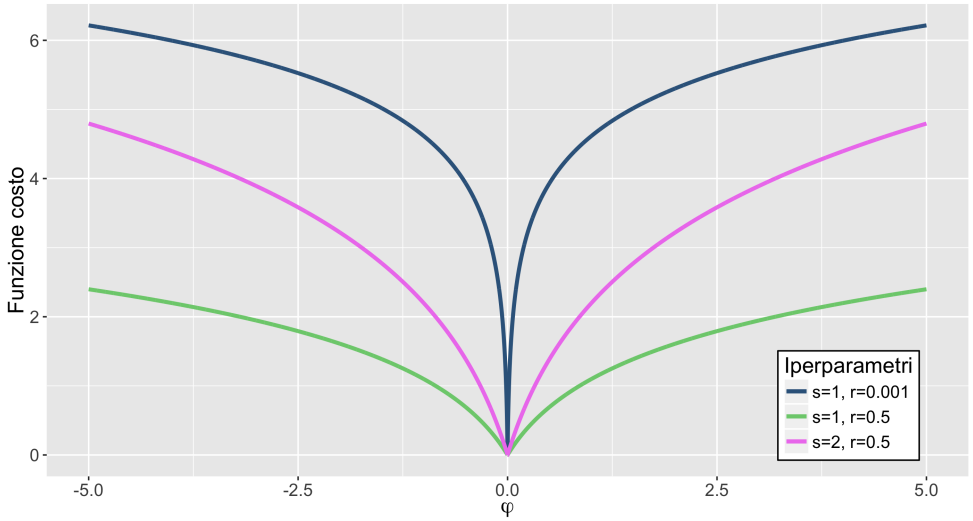

Figura 4.1: Penalizzazione della log-verosimiglianza per φ: alcuni esempi

della funzione da ottimizzare più facili da trattare, tipicamente uno sviluppo di Taylor al secondo ordine.
L'algoritmo descritto viene implementato nel pacchetto textir, sviluppato dallo stesso Matt Taddy (per dettagli Taddy (2015)). Il pacchetto consente di stimare un modello di regressione inversa in pochi secondi anche su collezioni estese di documenti, grazie anche alla parallelizzazione consentita dalle funzioni.

4.2 Regressione *forward*

Una volta applicata la regressione multinomiale inversa è possibile utilizzare le proiezioni ottenute ($\boldsymbol{z}_k$) come regressori per quella che Taddy (2013a) chiama regressione *forward*. Nel caso in esame il modello prescelto, al fine di privilegiare l'aspetto interpretativo, è una semplice regressione multilogit. Sia S la variabile contente il *sentiment* espresso nei documenti, quindi una variabile categoriale a K livelli, Z la matrice delle proiezioni fornite dalla MNIR e X_{bio} la matrice delle covariate twitter-biografiche descritta nella sezione 1.2. Il modello viene allora

specificato come

$$\begin{aligned}\mathbb{P}[S_i = k|Z = z_i, X_{bio} = x_i] &= \frac{e^{\eta_{ik}}}{1 + \sum_{t=1}^{K-1} e^{\eta_{it}}} \\ \mathbb{P}[S_i = K|Z = z_i, X_{bio} = x_i] &= \frac{1}{1 + \sum_{t=1}^{K-1} e^{\eta_{it}}}\end{aligned} \tag{4.7}$$

dove $\eta_{ik} = \alpha + \beta_k z_i + \gamma_k x_i$ è il predittore lineare che consente di tenere conto sia del testo con le proiezioni Z che delle carateristiche proprie dell'utente.

4.3 Applicazione

La prima applicazione presentata in questo capitolo riguarda l'implementazione del modello di regressione multinomiale inversa alle *term-by-document matrix* introdotte nella Sezione 1.3 e presentate nella Sezione 2.2. Innanzitutto, ancor prima di concentrarsi sui risultati del modello, vengono presentati i livelli di classificazione del *sentiment* e le distribuzioni osservate nel *subset* di analisi, frutto della classificazione a mano da parte dell'autore. Occorre specificare come si è scelto, in casi dubbi, un atteggiamento più conservativo (modalità Neutrale). Nella presentazione dei risultati l'attenzione viene posta in un primo tempo sui valori dei coefficienti, al fine di individuare quali siano i termini più probabili nel loro utilizzo per *sentiment* espesso. In secondo luogo l'attenzione si sposta sull'associazione tra il *sentiment* ed altre quantità coinvolte nelle regressioni, come il numero di termini utilizzati o le proiezioni stesse. Viene poi applicato il modello *forward* descritto in Sezione 4.2 e ne vengono valutate le prestazioni. Infine si cerca di capire come si comporta l'intero campione di utenti attivi sul tema prevedendo il *sentiment* grazie al modello *forward*. Questa procedura viene applicata ai quattro temi selezionati.

4.3.1 Brexit

Il primo tema analizzato è il referendum svoltosi in Gran Bretagna il 23 giugno 2016 circa la possibilità di abbandonare l'Unione Europea. I livelli scelti per esprimere le diverse opinioni emerse sono stati 4:

- **Neutrale:** l'utente non prende direttamente posizione riguardo il referendum ed il suo risultato;

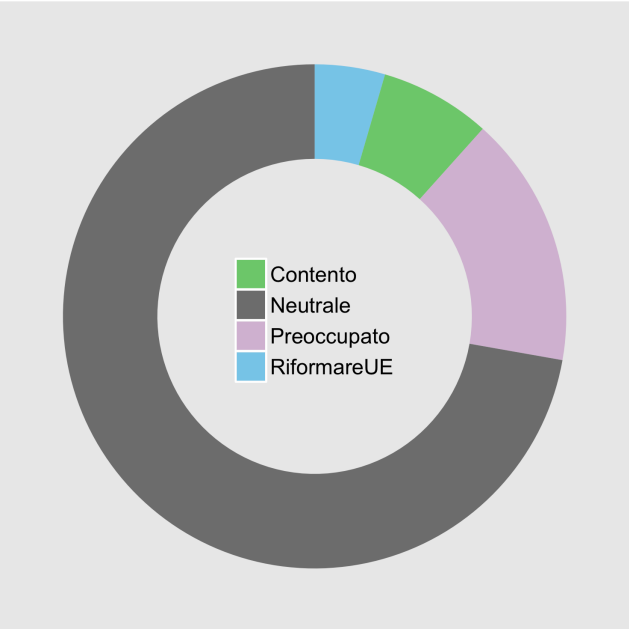

Figura 4.2: Distribuzione del *sentiment* circa la Brexit nel *subset* di analisi

- **Contento:** l'utente esprime soddisfazione circa il risultato del referendum oppure evoca l'emulazione dell'uscita dall'UE anche da parte dell'Italia;

- **Preoccupato:** l'utente esprime la sua preoccupazione o tristezza circa l'esito del referendum;

- **RiformareUE:** l'utente individua la necessità di sfruttare questo avvenimento per avviare una riforma dell'istituzione europea, preservando comunque fiducia nell'UE.

La distribuzione nel campione di 3211 utenti dei *sentiment* è esplicitata in Figura 4.2. I neutrali rappresentano una netta maggioranza (72.22%), dimostrando come la maggior parte degli utenti in questo caso ha utilizzato Twitter per condividere informazioni senza esprimersi. Tra coloro che invece prendono posizione circa la Brexit i preoccupati rappresentano il 16.13%, seguiti da contenti (7.13%) e, infine, da coloro che esprimono la necessità di riformare l'UE (4.52%).

Regressione Multinomiale Inversa

Nella Figura 4.3 sono mostrati i termini delle *tbd matrix* che hanno restituito i valori più elevati dei coefficienti φ_{jk}, dove j indica il termine e k il *sentiment*. L'interpretazione da attribuire a questi valori, sulla scorta di quanto esposto nella Sezione 4.1, è una variazione in scala logistica della probabilità di utilizzare il termine nel documento dato il *sentiment*. In verde sono rappresentati i coefficienti positivi ed in rosso quelli negativi, mentre il valore assoluto del coefficiente è proporzionale alla dimensione delle parole nelle *wordcloud*. Si è scelto, per consentire una facile lettura dei grafici, di rappresentare i 20 termini con valori più alti ed i 20 con valori più bassi. Partendo dal sentiment neutrale, i valori più elevati vengono assunti da termini e bi-grammi relativi ad un tweet ironico pubblicato dall'utente @guglielmoscilla in cui si faceva ironia sul fatto che la Brexit riducesse le possibilità di ricevere una lettera da Hogwarts, sede della scuola di magia della saga di Harry Potter. Dati i numerosi retweet ricevuti il coefficiente associato a questi termini assume valori molto elevati, superando il "filtro" degli effetti casuali specifici del soggetto. Altre tematiche affrontate nei tweet neutrali sono il futuro della lingua inglese dopo l'uscita della Gran Bretagna dall'UE (bi-gramma "impar ingles") ed i risultati dei primi sondaggi ("risult sondagg"). Come lecito attendersi compare anche il tag relativo ad un canale all-news, in questo caso Rai News, tra i termini associati positivamente al *sentiment* neutrale. Passando ai coefficienti negativi, invece, compaiono token connotati sia negativamente - "schif", "foll", "dispiac" - che positivamente - "valor", "più fort". Il valore assoluto dei coefficienti negativi risulta però molto più basso rispetto ai positivi, segno che i neutrali si esprimono con un vocabolario proprio e non in negazione ad altri *sentiment*. Coloro che si dichiarano contenti della Brexit, invece, tendono a non utilizzare principalmente termini propri di altre opinioni. Una polemica non toccata da questo gruppo di utenti è certamente quella relativa alla distribuzione per età del voto. Prova ne è la colorazione rossa di bigrammi come "fasc 18", "sopr 65" ed "elettor fasc". Altri due temi non affrontati in questi documenti sono la reazione delle borse nei giorni successivi ("brexit bors" e "caos") e la comparsa sul web di una petizione per ripetere il referenfum ("petizion nuov"). Tra i coefficienti positivi la fanno da padrone token a forte connotazione positiva e che vanno ad elogiare la scelta compiuta dal popolo britannico, come "coragg liber", "graz uk", "liber cittadin". Anche la preoccupa-

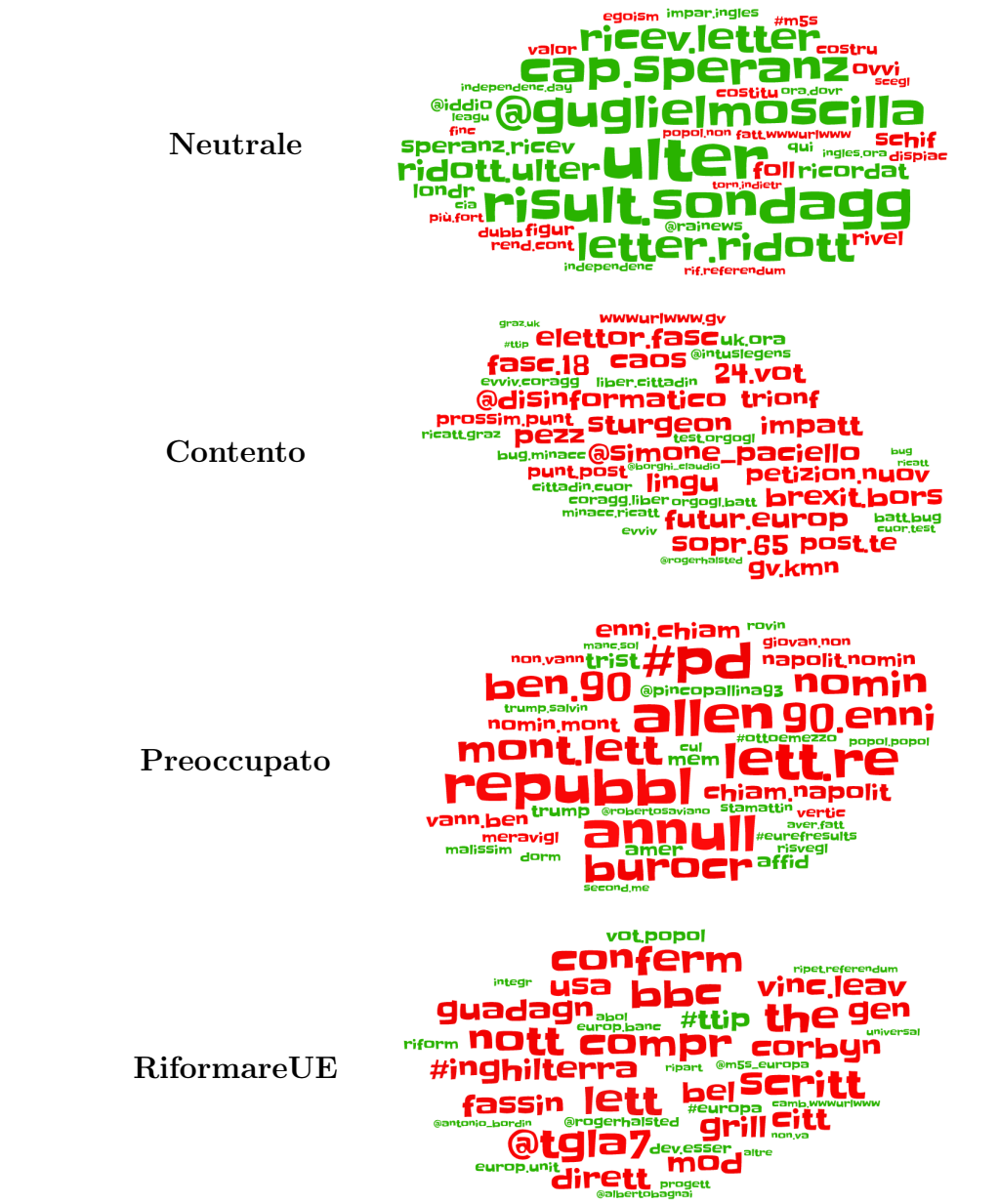

Figura 4.3: Brexit - valori più elevati (in valore assoluto) dei coefficienti φ_{jk}

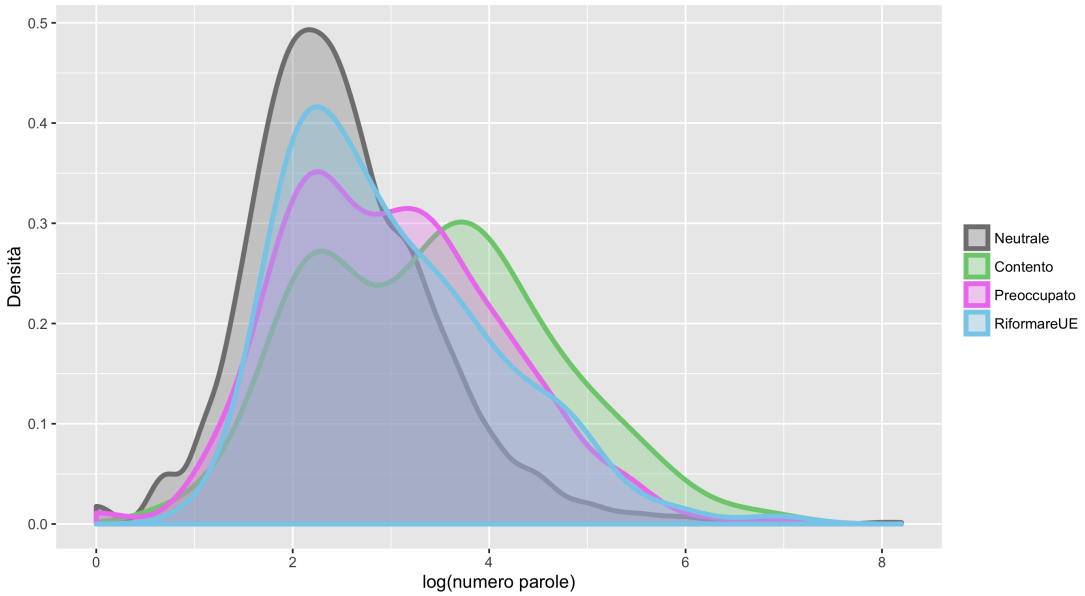

Figura 4.4: Brexit - densità del numero di token nei documenti per *sentiment*

zione per le conseguenze della Brexit viene espressa principalmente in negazione rispetto agli altri *sentiment*. In questo caso i termini più influenti sono riferiti agli schieramenti ed ai personaggi politici che vengono visti più legati all'Unione Europea, come "mont lett", "napolit nomin" e "#pd", ed alla burocrazia ("burocr") vista come una malattia dell'UE. I termini associati positivamente, invece, sono soprattutto indice di stati d'animo negativi ("trist", "malissim", "rovin") e tag a personaggi che hanno espresso opinioni simili, come @robertosaviano. Infine anche per i "riformisti" i coefficienti a valori assoluti più elevati sono negativi. In questo caso non sembrano emergere trend particolari nei termini, a parte il riferimento ad alcune emittenti ("@tgla7" e "bbc") e politici ("fassin", "grill" e "corbyn"). Nei termini positivi, invece, emerge la volontà di riformare l'Unione Europea in token come "europ unit", "riform", "progett". In Figura 4.4 viene mostrata la distribuzione del logaritmo del numero di parole utilizzate in ogni documento, distinte per *sentiment* espresso. La prima impressione che emerge è che ci sia una distinzione abbastanza marcata tra le distribuzioni. I neutrali sembrano essere coloro che scrivono meno, dato che la distribuzione sembra essere stocasticamente dominata dalle altre tre. Riformisti e preoccupati si collocano a metà strada: l'unimodalità

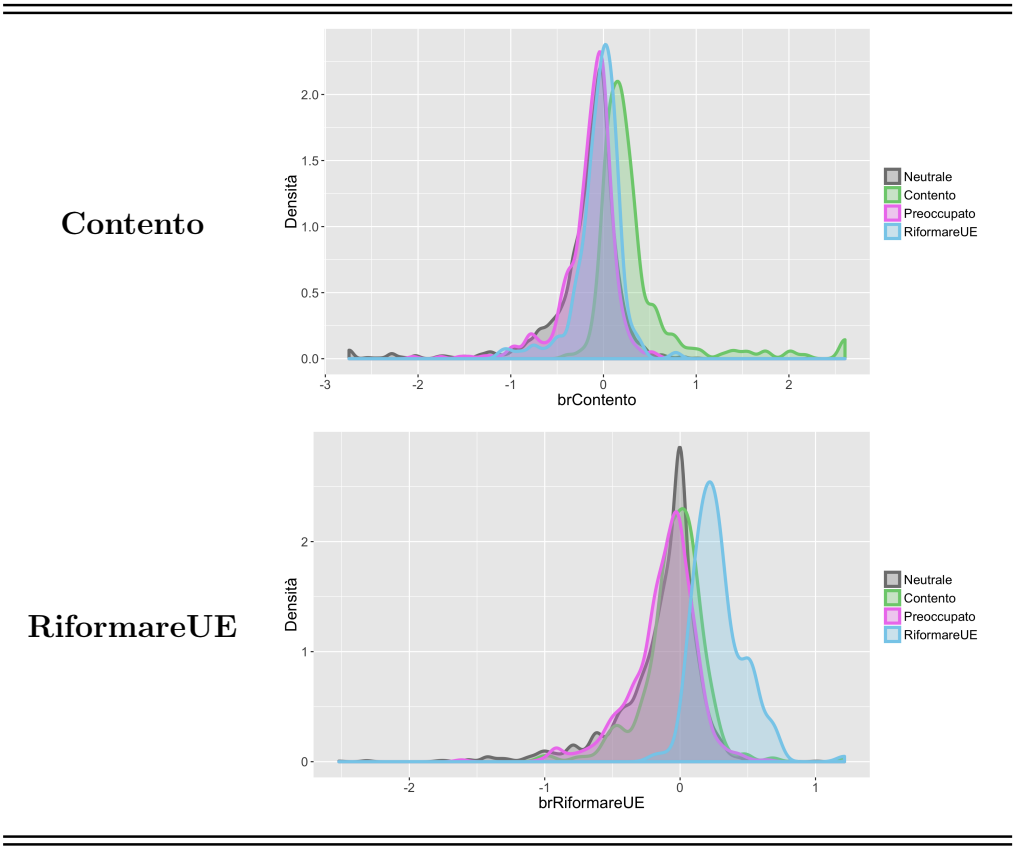

Figura 4.5: Brexit - confronto delle densità delle proiezioni z_k fornite dalla MNIR per i *sentiment* $k = Contento$ e $k = RiformareUE$

dei primi su un valore abbastanza basso li equipara ai neutrali, mentre per i preoccupati si osserva una bimodalità a parità o quasi di coda destra. Infine, i contenti mostrano una spiccata bimodalità con una coda destra molto pesante ed una moda intorno alle 40-45 parole usate. Passando alle proiezioni delle *tbd matrix* nello spazio individuato dai vettori di coefficienti $\boldsymbol{\varphi_k}$, nella Figura 4.5 sono rappresentate le densità per argomento delle proiezioni nelle dimensioni relative ai *sentiment* Contento e RiformareUE. In pratica si tratta di valori che potremmo riassumere come risposta alla domanda "Quanto un utente parla come un utente "tipo" di quel *sentiment*?". Si nota dal confronto tra i due grafici come i riformisti usino un vocabolario (sia in positivo che in negativo come visto) più proprio rispetto ai contenti, i cui valori nella propria dimensione sono sì più alti di quelli degli utenti degli altri sentiment ma non così tanto. In entrambi i casi si evidenzia però come

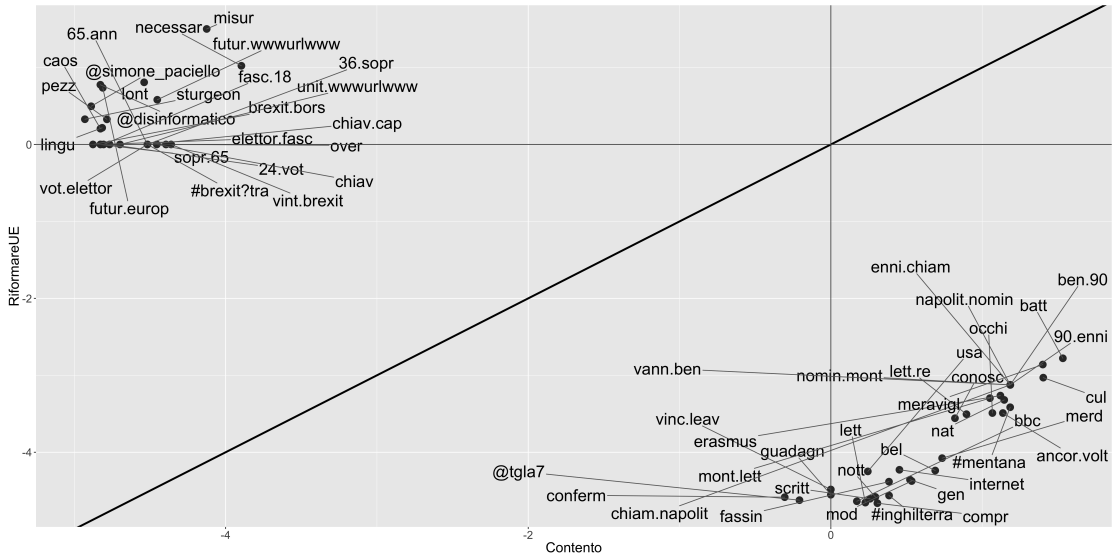

Figura 4.6: Termini maggiormente discriminanti tra Contento e RiformareUE

riformisti e contenti sembrino essere due *sentiment* che condividono una parte di vocabolario utilizzato. Questo fatto può essere giustificato guardando i due profili come due differenti critiche mosse all'assetto attuale dell'UE: una più costruttiva ed una più distruttiva. Un aspetto interessante può essere cercare di capire quali siano i termini che maggiormente discriminano i due *sentiment*. Da un punto di vista applicativo può risultare utile sapere quali siano gli argomenti che distinguono due diverse opinioni o atteggiamenti per capire, nel caso si campagne pubblicitarie o elettorali, quali argomenti possono spostare fette di consumatori o elettori, a seconda del campo di applicazione. Figura 4.6 mostra i 60 termini per cui la differenza assoluta tra i due coefficienti $|\varphi_{j,Contento} - \varphi_{j,RiformareUE}|$ risulta maggiore. Come lecito attendersi i termini si collocano nel secondo e quarto quadrante, cioè in porzioni del piano in cui i coefficienti associati ai due *sentiment* hanno segni opposti. Tra i termini che maggiormente discriminano a favore di RiformareUE troviamo di nuovo "futur europ", visto nelle wordcloud di Figura 4.3, così come il riferimento al premier scozzese Sturgeon, che ha espresso le sue perplessità circa il risultato del referendum e la volontà della Scozia di restare nell'Unione Europea. In questa porzione di piano sono numerosi anche i termini riferiti alla polemica generazionale ("65 ann", "sopr 65", "sopr 18"). Tra i termini positivamente correlati

con il *sentiment* contento si trovano riferimenti chiari a figure politiche. Inoltre emergono termini che esprimono chiaramente un giudizio positivo sull'esito della consultazione come "meravigl".

Modello *forward*

A questo punto è possibile utilizzare le proiezioni nel nuovo spazio determinato dalla matrice Φ per la regressione *forward*. Viene applicato il modello multilogit esposto nella Sezione 4.2 sui 3211 utenti attivi e viene effettuata una selezione delle variabili basata sull'indice AIC che porta al modello illustrato in Figura 4.7. La modalità di riferimento per il *sentiment* è Neutrale, quindi i coefficienti sono

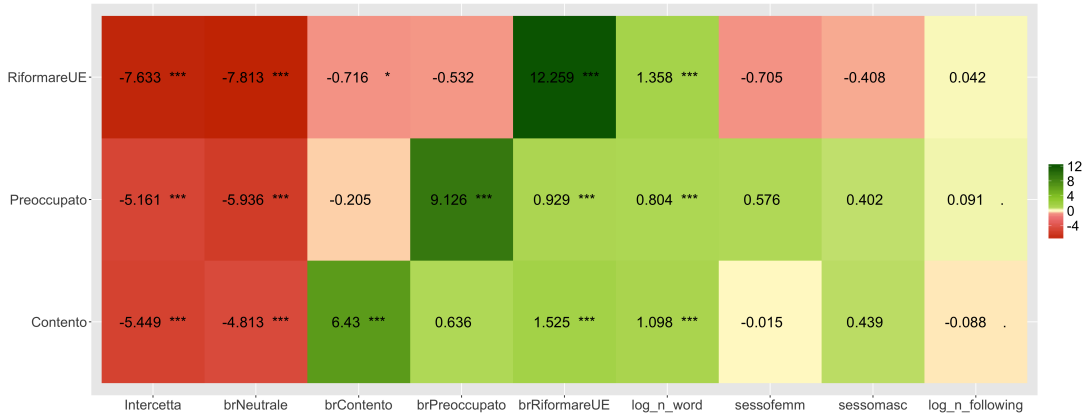

Figura 4.7: Brexit - regressione *forward* (categoria di riferimento: *Neutrale*)

da interpretarsi come variazione rispetto alla probabilità di essere Neutrale. Oltre al testo, come logico attendersi superstite alla selezione delle variabili, resistono il sesso ed il numero di following, cioè di persone seguite, tra le covariate twitter-biografiche. Partendo proprio da queste ultime si nota come, sebbene conservata nel modello, la variabile sesso (con la categoria *ente* come baseline) non presenta coefficienti significativi. In realtà il p-value relativo al sesso femminile per il livello preoccupato è pressochè pari a 0.10, quindi l'impatto di questa variabile può essere visto su questo livello. L'unica significatività si può allora individuare nel coefficiente che indica una maggiore propensione per le donne ad essere preoccupate per l'esito del referendum. Riguardo il numero di following, invece, tra neutrali e riformisti non si evidenziano differenze significative, mentre i preoccupati sono

mediamente più connessi ed i contenti seguono un numero di utenti minore rispetto agli altri. Per quanto riguarda il testo si nota come i coefficienti sulla proiezione $z_{Neutrale}$ siano fortemente negativi e significativi, sintomo che non vi è sovrapposizione di vocabolario in questo caso. Questo comportamento si ripete nei tre valori in diagonale colorati di verde scuro: sono i coefficienti di ogni sentiment sulla propria proiezione. Ovviamente più alto è il valore nella propria dimensione e più si tende ad esprimere quel sentiment. Tuttavia non sempre è così o perlomeno non con la stessa intensità e non come unica indicazione. Si noti come, ad esempio, i riformisti si individuino anche in negazione ai contenti, segnale che rafforza quanto osservato in precedenza. Dato che i due *sentiment* sembrano sovrapposti è importante scrivere come un riformista ma anche non utilizzare il vocabolario proprio dei contenti. Opposto, ed anche abbastanza curioso, il comportamento dei contenti. I tre coefficienti sono tutti positivi ed addirittura significativo quello sulla proiezione $z_{RiformareUE}$. Questo significa che per i contenti usare anche un po' del vocabolario dei riformisti è normale. Si evidenzia quella che potremmo definire una sorta di sovrapposizione asimmetrica. Infine si nota come i neutrali siano coloro che, al netto dei contenuti, utilizzano un minor numero di parole nei loro documenti. In questo ambito i più prolissi sono i riformisti, seguiti da contenti e preoccupati. Questa indicazione ribalta quanto osservato nelle densità marginali di Figura 4.4 ma si tratta di due indicazioni differenti: in un caso si osservano le densità marginali per gruppo, nell'altro si esamina, al netto delle altre covariate, l'incidenza del numero di parole utilizzate sui differenti *sentiment*.

Osservando le previsioni sul campione utilizzato per testare il modello, si riscontra una accuratezza (complemento a 1 del tasso di errata classificazione) pari all'80.69%. In Tabella 4.1 viene riportata la matrice di confusione su cui è stata calcolata questa quantità. In linea con l'atteggiamento conservativo della classificazione a mano si osserva come vengano classificati come Neutrale molti documenti che non lo sono, in quantità in alcuni casi superiore a quanti classificati correttamente. Questa è stata una scelta in fase di classificazione che può essere rivista per evitare questo fenomeno: molto dipende dallo scopo dell'analisi e dall'importanza che assumono falsi positivi e falsi negativi.

		Osservato			
		N	C	P	R
Previsto	N	2190	116	280	66
	C	33	102	3	4
	P	82	4	230	6
	R	14	7	5	69

Tabella 4.1: Brexit - matrice di confusione osservati vs previsti

Previsione sull'intero campione

Lo scopo principale dell'analisi presentata è interpretativo, però pensando ad una seconda fase previsiva può essere utile fornire un primo *benchmark*. Estendendo allora la previsione del modello all'intero campione di 32477 utenti attivi sulla Brexit tra coloro che hanno sesso e zona classificati (Tabella 2.1) si osserva la distribuzione riassunta in Figura 4.8.

4.3.2 Legge Cirinnà

Il secondo tema analizzato è relativo alla discussione e successiva approvazione della cosiddetta legge Cirinnà, cioè del decreto legge che ha inserito nell'ordinamento giuridico italiano le unioni civili, valide anche per persone dello stesso sesso. In questo caso i livelli della variabile che esprime il *sentiment* sono solamente tre:

- **Neutrale:** l'utente non esprime direttamente un giudizio di merito riguardo la legge;
- **Contrario:** l'utente esprime la sua contrarietà al contenuto del decreto oppure manifesta il suo malcontento circa l'approvazione dello stesso;
- **Favorevole:** l'utente si dichiara favorevole alla legge o si compiace per la sua approvazione.

La distribuzione dei tre atteggiamenti verso la legge nel *subset* di analisi è rappresentata in Figura 4.9. Anche in questo caso, come accade spesso in casi simili, la

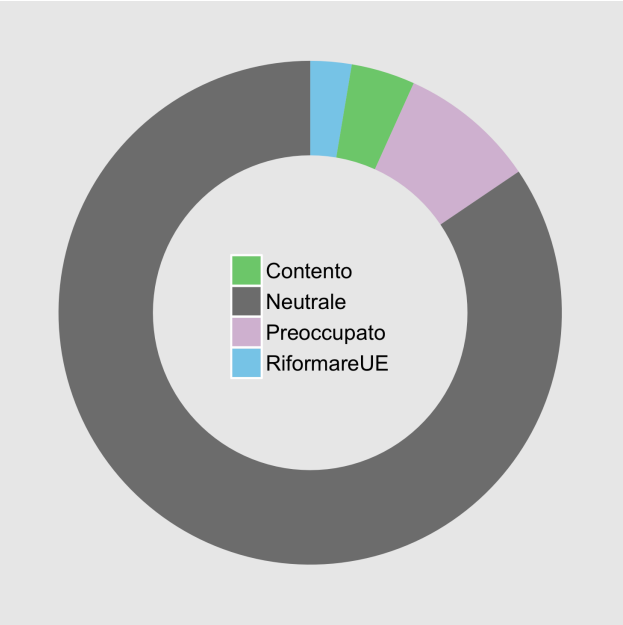

Figura 4.8: Previsione aggregata del *sentiment* circa la Brexit

maggioranza degli utenti è stata classificata come Neutrale (51.40%). La categoria Favorevole rappresenta comunque una fetta importante dei 1107 utenti osservati (39.93%) mentre il restante 8.67% si è dichiarato contrario alla legge.

Regressione Multinomiale Inversa

Andando ad indagare quali termini risultino più legati ad un *sentiment* piuttosto che ad un altro, in Figura 4.10 è possibile osservare le wordcloud relative ai coefficienti del modello MNIR applicato a questi documenti. Diversamente da quanto osservato per la Brexit, i Neutrali non sembrano possedere un loro particolare vocabolario. I termini in verde, infatti, non hanno un coefficiente particolarmente più elevato di quelli in rosso. Tra i token utilizzati non emergono particolari trend, mentre tra i termini non utilizzati dai neutrali vi sono connotazioni positive - "giorn storic", "libert", "progress" - come negative - "ridicol", "referendum abrog". Tra i contrari, invece, emergono due termini che hanno forte impatto sul *sentiment*: "giudic" e "lim" (stem riconducibile al termine "limite"). Il primo è un riferimento a ricorsi contro la legge approvata, mentre il secondo viene in realtà utilizzato poche volte in documenti molto lunghi ma esclusivamente da contrari al-

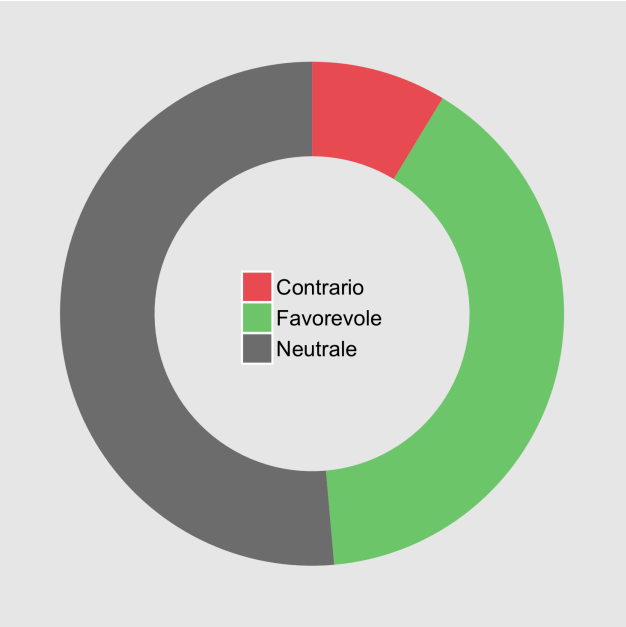

Figura 4.9: Distribuzione del *sentiment* circa la legge Cirinnà nel *subset* di analisi

la legge. Peso importante assumono anche l'emblematico hashtag "#stopcirinnà" ed "adozion gay", in riferimento alla polemica circa la *stepchild adoption*. Inoltre tra i tag compaiono il segretario della Lega Nord @matteosalvinimi e l'associazione @manifpourtousit, sezione italiana della famosa associazione nata in Francia in opposizione alla legge sui matrimoni per coppie omosessuali approvata nell'aprile 2013. Tra i termini associati negativamente si trova un riferimento al nome della relatrice della legge Monica Cirinnà, il tag @matteograndi, giornalista molto attivo su Twitter e schierato a favore della legge, e termini come "lov" e "civilt". La wordcloud dei favorevoli, invece, si distingue per la forte colorazione verde. Questo fatto è indice dell'utilizzo principalmente di un proprio vocabolario da parte di questo gruppo di utenti. Si evidenziano riferimenti politici, dall'hashtag utilizzato dal premier Renzi "#lavoltabuona" al tag @zanalessandro, deputato Pd che in aula è intervenuto con una commossa dichiarazione di voto, così come termini che racchiudono la soddisfazione per l'approvazione. Si trovano allora "amor vinc", "cuor", "bell giorn", "più giust", oltre a riferimenti più generali a "#diritti" e "#diritticivili". Passando ai termini rossi si nota l'hashtag "#canguro" relativo alla polemica politica che ha avuto luogo nel mese di febbraio in seguito alla deci-

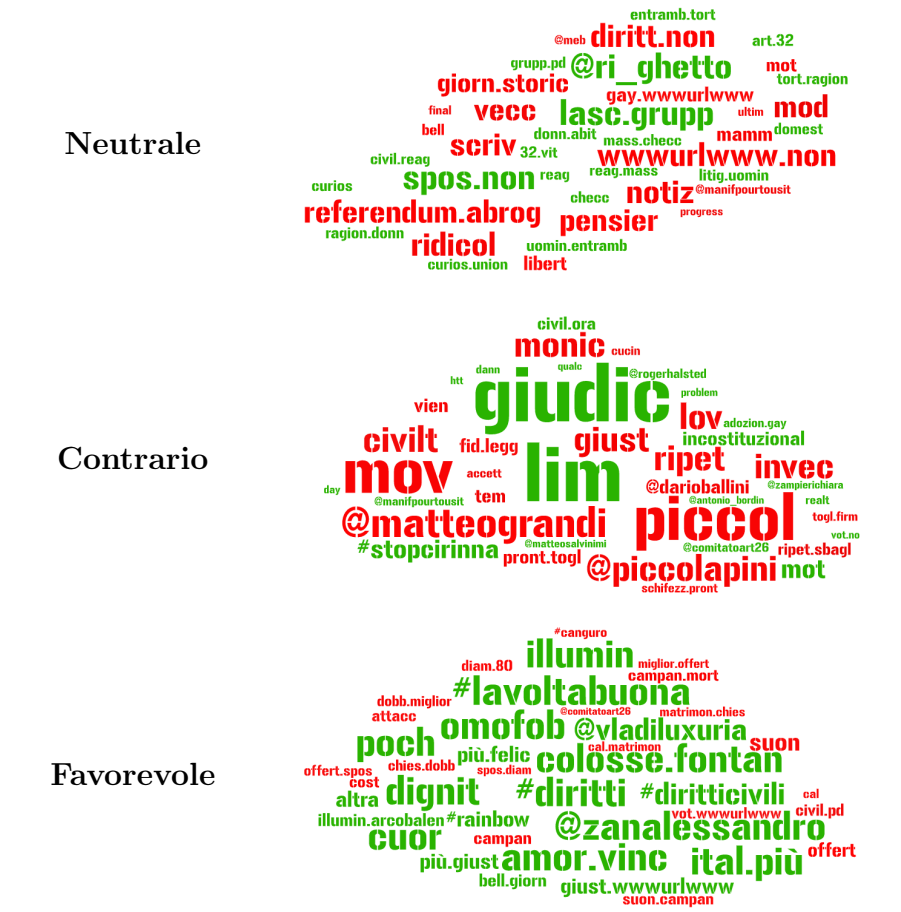

Figura 4.10: Cirinnà - valori più elevati (in valore assoluto) dei coefficienti φ_{jk}

sione del governo di applicare il cosiddetto canguro, cioè di saltare la discussione di alcuni emendamenti. Per il resto sembrano esserci bigrammi tra loro connessi come "suon campan" e "campan mort", segno che ci potrebbe essere un particolare tweet retwittato molte volte da coloro che non sono favorevoli alla legge.
Figura 4.11 mostra la distribuzione per *sentiment* del numero di token contenuto nel documento. Non sembrano emergere particolari differenze tra le tre distribuzioni. La più regolare e simmetrica sembra essere la densità dei Neutrali, mentre per Contrari e Favorevoli si hanno code più pesanti, soprattutto per questi ultimi. La conclusione sembra essere che non sussistono notevoli differenze tra i tre gruppi per quanto riguarda il numero totale di termini utilizzati.

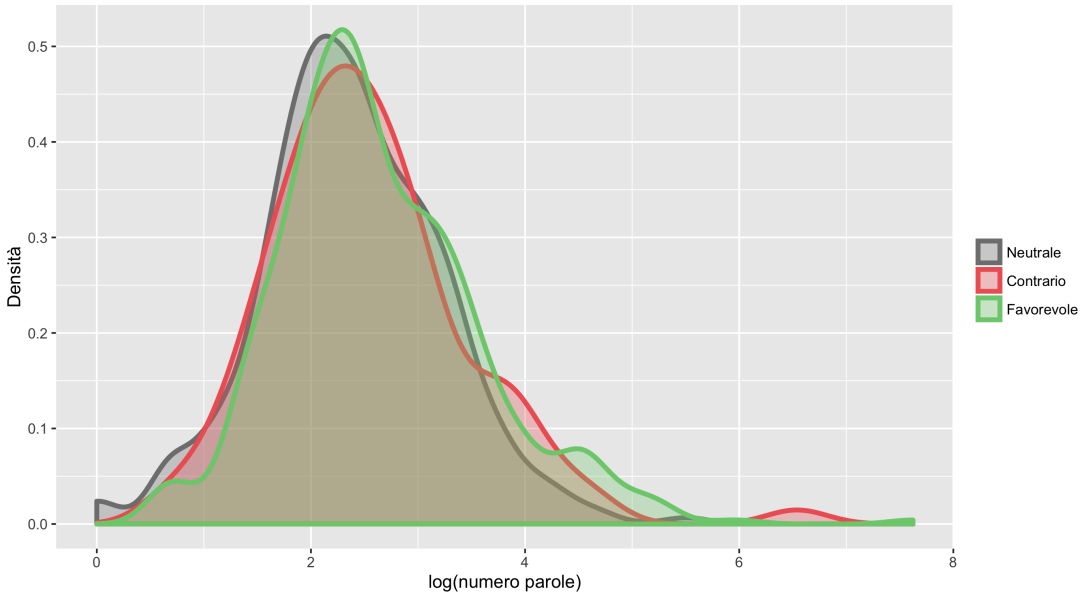

Figura 4.11: Cirinnà - densità del numero di token nei documenti per *sentiment*

Spostandosi nello spazio definito dalla matrice Φ è possibile notare come i tre *sentiment* siano caratterizzati abbastanza nettamente da un vocabolario proprio. Infatti la densità del gruppo a cui è riferita la dimensione è distinguibile chiaramente nella parte destra del grafico. Questo vale soprattutto per Contrario e, soprattutto, Favorevole, che evidenzia una moda molto più bassa come valore di densità ed una coda destra pesantissima, a conferma di quanto visto in Figura 4.10. Per i contrari, invece, si evidenzia una certa importanza attribuita ai pesi negativi: prova ne sono le code molto pesanti a sinistra delle altre due densità, sintomo che questi due gruppi utilizzano molti termini con coefficienti negativi per il sentiment *Contrario*. Rispondendo alla domanda "Come parlano gli utenti in riferimento al proprio *sentiment*?" si potrebbe dire che, in questo caso, i gruppi si esprimono con tre vocabolari quasi distinti, forse eccezion fatta per i neutrali. Tuttavia non è usuale osservare una coda destra così pesante per un *sentiment* che si caratterizza generalmente come negazione degli altri. Osservando i documenti nel nuovo spazio, allora, sarebbe lecito aspettarsi una separazione abbastanza netta tra i documenti, soprattutto per *Contrario* e *Favorevole*: gli utenti di questi sentiment dovrebbero posizionarsi su valori positivi nella dimensione relativa al proprio

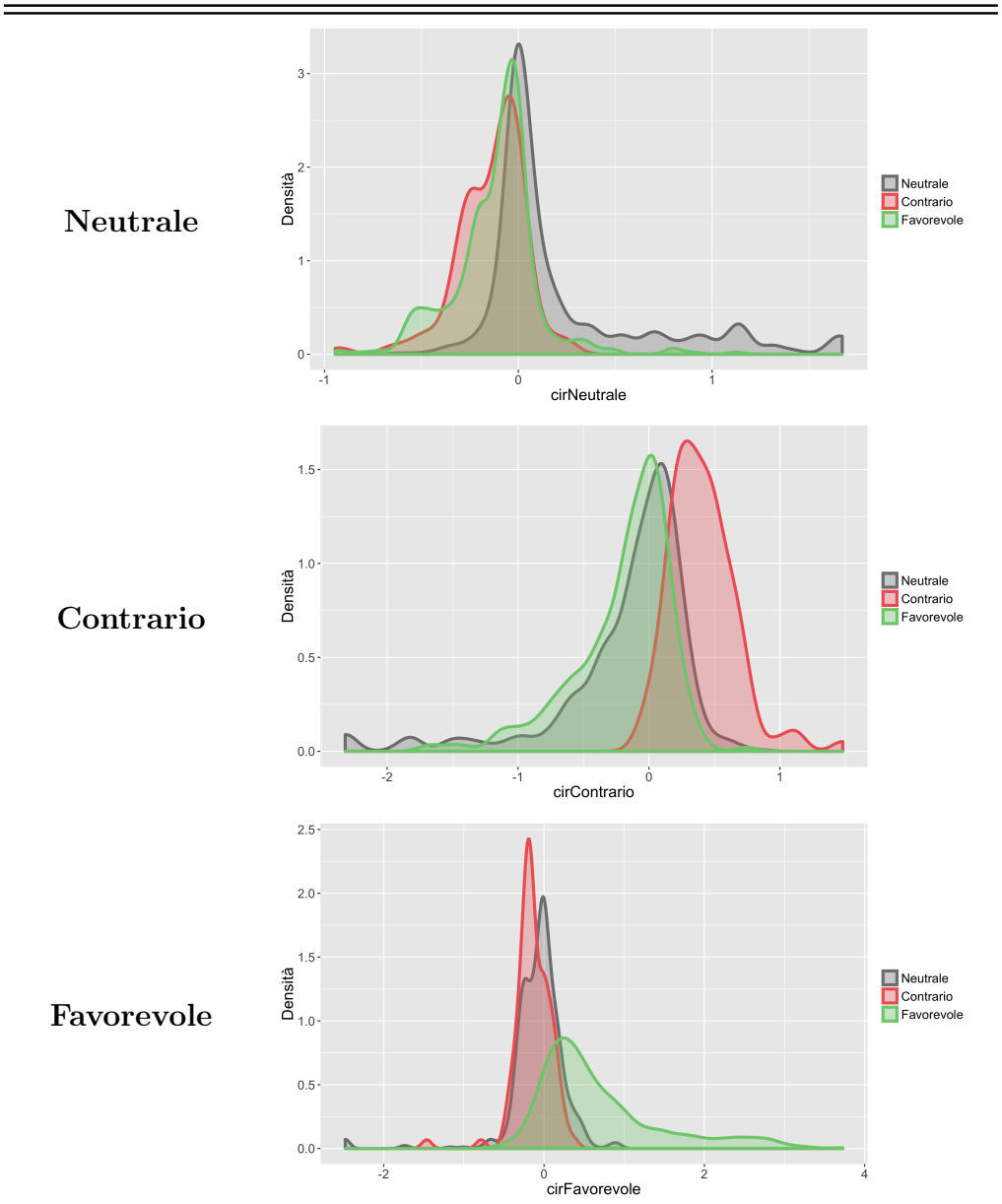

Figura 4.12: Cirinnà - confronto delle densità delle proiezioni z_k fornite dalla MNIR

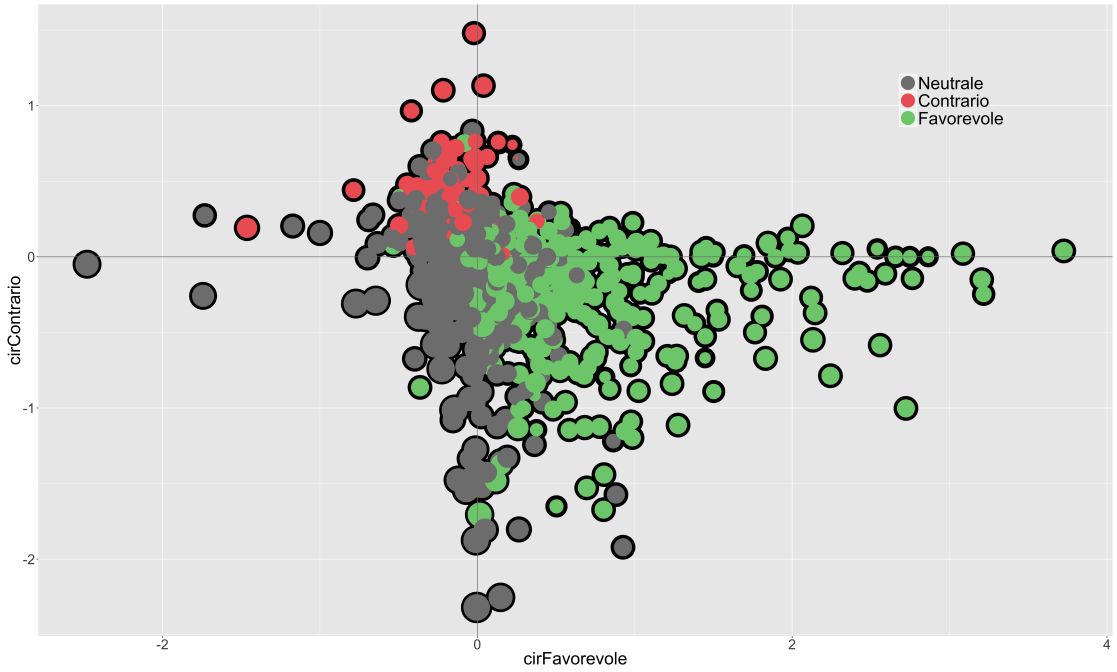

Figura 4.13: Cirinnà - rappresentazione dei documenti nello spazio generato da Φ

sentiment e negativa per gli altri. In Figura 4.13 sono rappresentati i documenti nello spazio generato dalla matrice Φ. Data la scarsa chiarezza grafica fornita dallo scatter plot tridimensionale, si è scelto di rappresentare i punti nel piano creato da $z_{Contrario}$ e $z_{Favorevole}$ e di rappresentare i punti con dimensione proporzionale al loro valore di $z_{Neutrale}$. Il comportamento dei documenti nello scatter plot va in direzione di quanto visto in Figura 4.12. Nel secondo e nel quarto quadrante si trovano, come atteso, rispettivamente Favorevoli e Contrari, con valori di $z_{Neutrale}$ che vanno diminuendo più ci si allontana dall'origine. Nel terzo quadrante, invece, si trovano per la maggior parte i documenti neutrali, con la dimensione del punto che va crescendo più ci sposta dall'origine. Tuttavia questo comportamento è meno marcato rispetto ai due evidenziati in precedenza, in accordo ancora una volta con quanto visto in Figura 4.12. Il primo quadrante, come lecito attendersi, risulta invece popolato da pochissimi punti e perlopiù schiacciati verso i due assi.

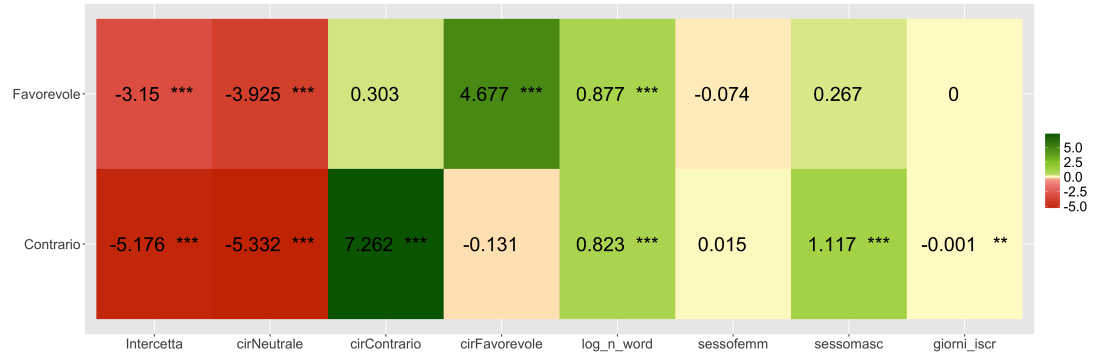

Figura 4.14: Legge Cirinnà - regressione *forward* (categoria di riferimento: *Neutrale*)

Modello *forward*

Ottenuta la matrice Z e recuperate le informazioni twitter-biografiche degli utenti attivi sulla legge Cirinnà e presenti nel *subset* di analisi, è possibile stimare il modello multilogit. È stata poi effettuata una selezione delle variabili secondo la procedura *backward* ed utilizzando come criterio di selezione l'indice AIC. I coefficienti sono rappresentati nella heatmap in Figura 4.14, tenendo presente che la categoria di riferimento è Neutrale. In questo caso le covariate twitter-biografiche utili per individuare il *sentiment* sono il sesso e l'anzianità del profilo Twitter (variabile *giorni_ iscr*). Nel primo caso si osserva un'associazione positiva al netto del testo tra l'essere maschio e l'esprimersi in maniera contraria alla legge, mentre per il Favorevole non si osserva grande differenza tra le categorie. Allo stesso modo sembra che gli utenti iscritti da più tempo non siano contrari alla legge ma si dividano tra Favorevoli e Neutrali senza grossa distinzione. Relativamente al testo la situazione in questo caso sembra molto chiara ed in linea con quanto visto in precedenza. Ogni *sentiment* si caratterizza per l'uso del proprio vocabolario, non vi sono sovrapposizioni o casi in cui è importante la negazione del vocabolario usato da un altro gruppo di utenti. Ovviamente, avendo come *baseline* Neutrale, risultano fortemente negativi i coefficienti per $z_{Neutrale}$. Il numero di parole va infine a differenziare le due categorie "polarizzate" dai neutrali: entrambi i coefficienti sono infatti positivi e significativi, sintomo che chi esprime un'opinione tende a scrivere di più, però i valori simili portano a pensare non ci siano differenze significative tra Contrario e Favorevole. La funzione utilizzata per stimare il modello (**multinom**

del pacchetto **nnet**) permette di restituire in output la matrice hessiana, grazie alla quale è stato calcolato lo standard error relativo alla differenza di cui sopra. Il valore della differenza è pari a 0.054 con uno standard error pari a 0.1398: il valore della statistica test risulta essere pari a 0.390, lontano da qualsiasi soglia di significatività. Provando ad osservare l'accuratezza delle previsioni del modello,

		Osservato		
		N	C	F
Previsto	N	510	32	95
	C	8	55	7
	F	51	9	340

Tabella 4.2: Cirinnà - matrice di confusione osservati vs previsti

riportate nella matrice di confusione in Tabella 4.2, si osserva una maggiore precisione rispetto a quanto riscontrato in precedenza per la Brexit. La quota di "falsi Neutrali" sembra meno persistente, prova ne è anche la migliore accuratezza del modello (81.75%).

Previsione sull'intero campione

Nel datset di utenti presentato nella Sezione 1.2 sono presenti 11184 utenti attivi sulla legge Cirinnà. Utilizzando il modello appena descritto per effettuare una previsione del *sentiment* anche per gli utenti non classificati a mano si ottiene la distribuzione aggregata rappresentata in Figura 4.15.

4.3.3 Ritorno Marò

Il terzo tema analizzato è il ritorno, avvenuto il 28 maggio 2016, di Salvatore Girone, uno dei due marò detenuti in India per omicidio. Gli utenti del *subset* attivi su questo tema sono 413. I livelli della variabile di *sentiment* legata a questo evento sono quattro:

- **Neutrale:** l'utente non prende una posizione relativamente al ritorno del marò oppure condivide tweet ironici;

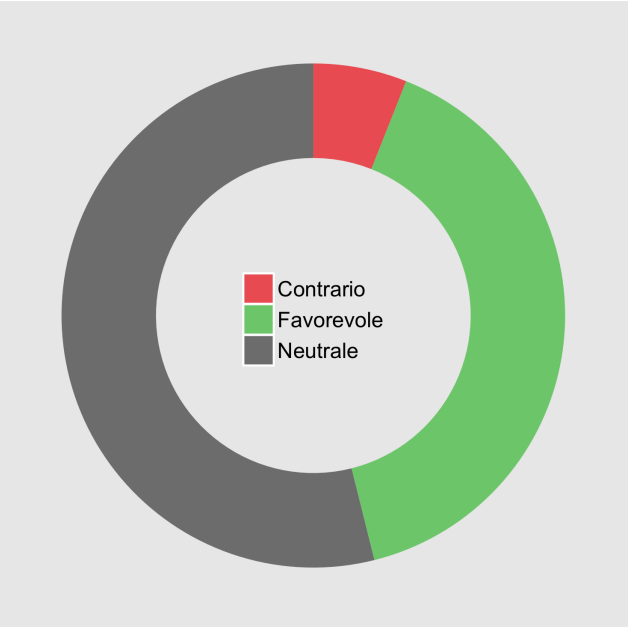

Figura 4.15: Previsione aggregata del *sentiment* circa la legge Cirinnà

- **Critico:** l'utente esprime principalmente la sua contrarietà riguardo la detenzione e l'operato degli ultimi governi sulla vicenda;
- **Felice:** l'utente esprime la sua felicità per il ritorno in Italia di Girone;
- **Scettico:** l'utente esprime il suo scetticismo per il trattamento riservato a Girone al rientro o più in generale sulla vicenda.

Il *donut plot* rappresentato in Figura 4.16 mostra la distribuzione di questa variabile. Il 57.14% degli utenti non mostra una polarità particolare, venendo classificato come Neutrale, mentre il 25.91% si dichiara Felice per il rientro in Italia di Girone. La restante fetta degli utenti si dichiara Scettica circa questo avvenimento per il 9.20% mentre il 7.75% critica l'atteggiamento della classe politica nella vicenda.

Regressione Multinomiale Inversa

Nelle *wordcloud* di Figura 4.17 sono rappresentati i valori dei coefficienti Φ per *sentiment*. In tutti e 4 i casi il colore dominante nelle *wordcloud* sembra essere il verde, sintomo che i 4 *sentiment* possiedono un vocabolario proprio piuttosto

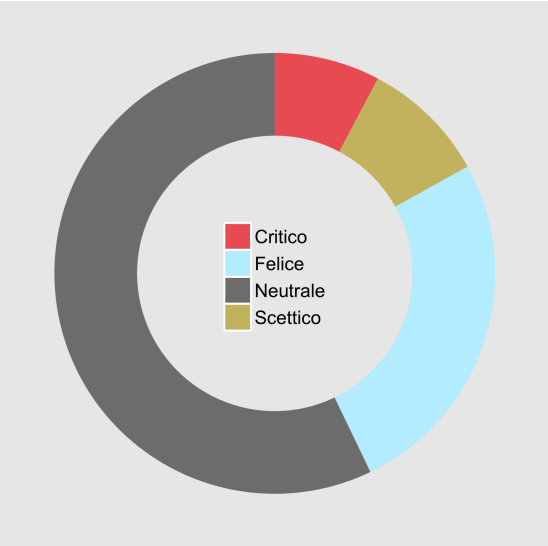

Figura 4.16: Distribuzione del *sentiment* circa il ritorno del marò Girone nel *subset* di analisi

che essere definiti come negazione di altre polarità. L'unica eccezione sembra essere rappresentata dagli utenti Critici, per i quali non sembra esserci discrimine in valore assoluto tra coefficienti positivi e negativi. La criticità verso i governi che si sono succeduti negli anni è esplicita in termini come "mont" o "govern incapac". Inoltre molto condiviso è un tweet in cui viene espresso il desiderio che nessun politico si vanti di questo risultato: "politic vant" e "vant sput" sono evidenti riferimenti a questo tweet. Tra i termini con coefficiente negativo si notano riferimenti all'accusa di omicidio ("indi accus", "accus aver" e "pescator wwwurlwww") e ad un tweet ironico sul cane Argo, rimasto in un primo tempo in India. I neutrali, invece, si concentrano sull'avvenimento con termini come "aeroport", "giorn arriv" e "aeroport ciampin". Spunta un riferimento alla parata del 2 giugno ("giugn rom"), al centro delle polemiche per la proposta di alcune forze politiche di fare sfilare i due marò. Parata del 2 giugno presente anche tra i coefficienti di segno negativo ("sfilat giugn"), insieme giudizi di merito sulla vicenda come "innocent" o reazioni come "felic" o "final cas", dove "final" è stem per "finalmente". La *wordcloud* più verde è sicuramente quella dei felici, che esprimono la loro soddisfazione con termini come "diam bentorn". Inoltre è molto retwittato il tweet del premier Renzi che esprime la sua gioia per l'avvenimento e che contiene "popol govern" e "govern diam". Tor-

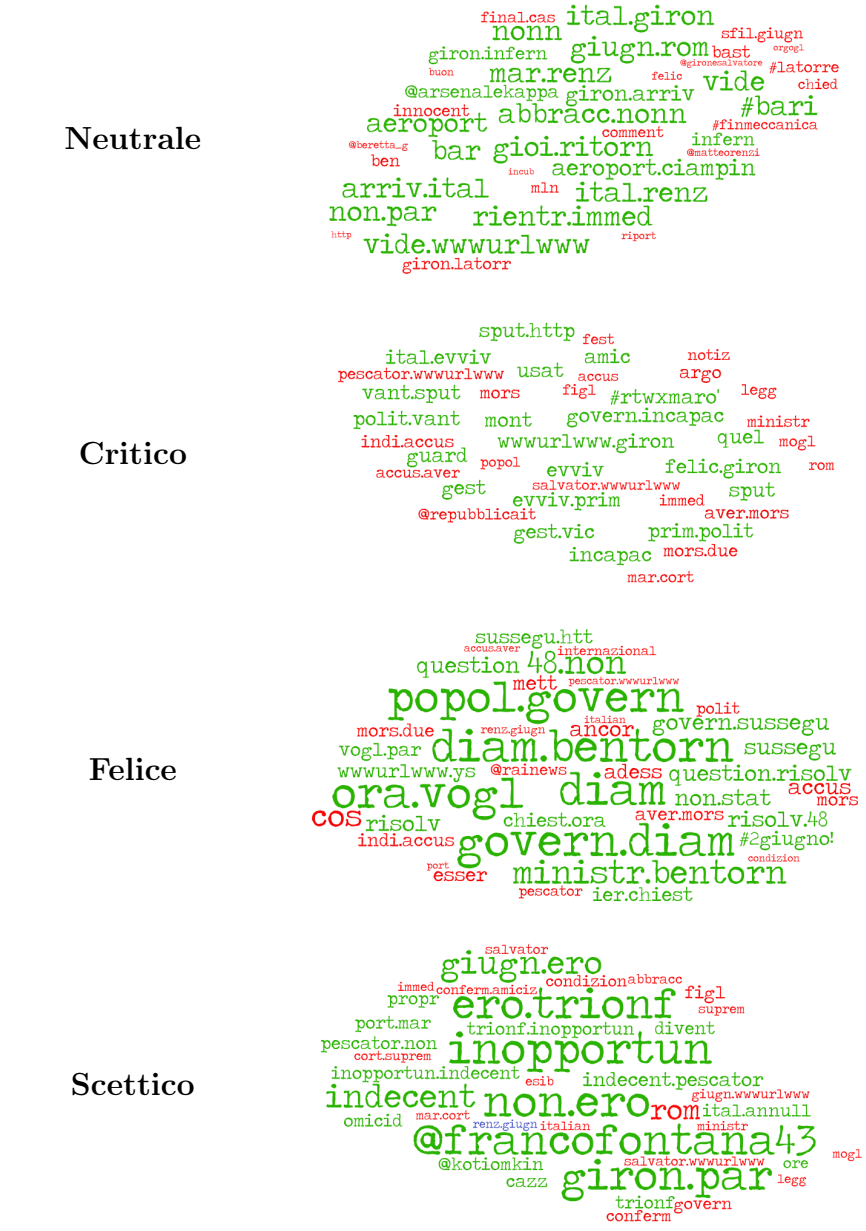

Figura 4.17: Marò - valori più elevati (in valore assoluto) dei coefficienti φ_{jk}

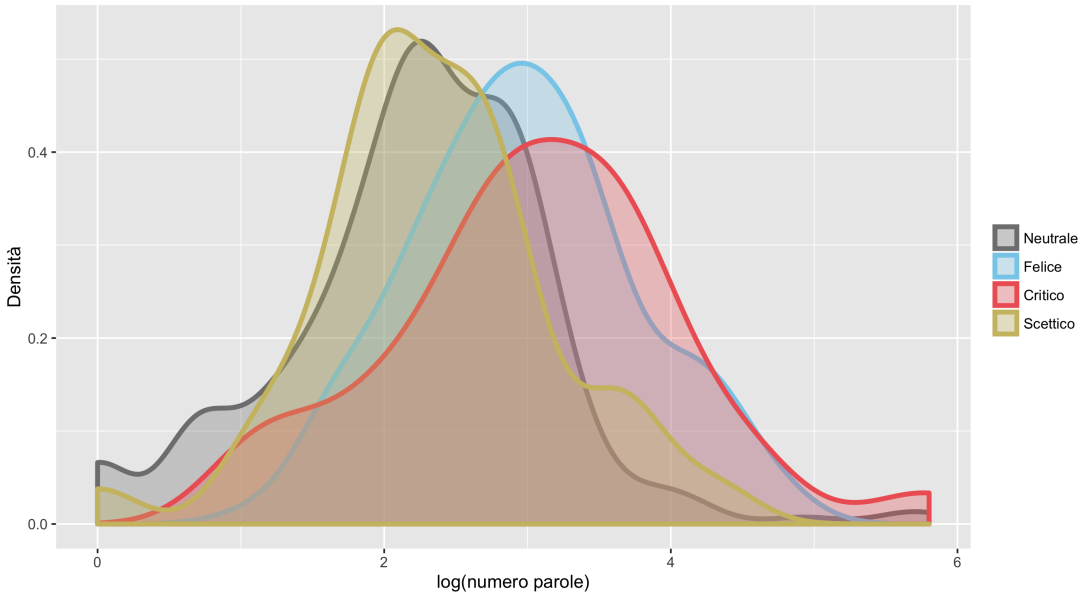

Figura 4.18: Marò - densità del numero di token nei documenti per *sentiment*

na, ovviamente tra i termini positivi, un riferimento alla parata con "#2giugno". "accus" e "pescator" sono invece tra i termini associati negativamente con questo *sentiment*, insieme ad altre parole riguardanti l'atto di accusa rivolto dall'India ai marò. Infine, lo scetticismo dell'ultimo gruppo di utenti riguarda principalmente il trattamento da eroi dei due militari ("ero trionf", "inopportun", "inoppotun indecent", "trionf inopportun"), mentre assumono rilievo termini come "omicid" e riferimenti ai pescatori ("pescator non" e "indecent pescator"). Non si evidenziano, invece, particolari tendenze nei termini associati negativamente a questo *sentiment*.

Guardando il numero di parole per *sentiment* attraverso le densità empiriche illustrate in Figura 4.18, appaiono comportamenti differenti. I critici hanno una distribuzione molto piatta, con valori a densità non nulla su entrambe le code. I felici sembrano essere coloro che scrivono di più, con una moda a densità più elevata dei critici ed una coda sinistra più leggera. Neutrali e scettici hanno invece comportamenti molto simili nella parte centrale, con i primi che tendono a scrivere meno in media dei secondi visto il confronto tra le due code.

La situazione mostrata in Figura 4.19 è invece molto chiara ed emblematica. Ap-

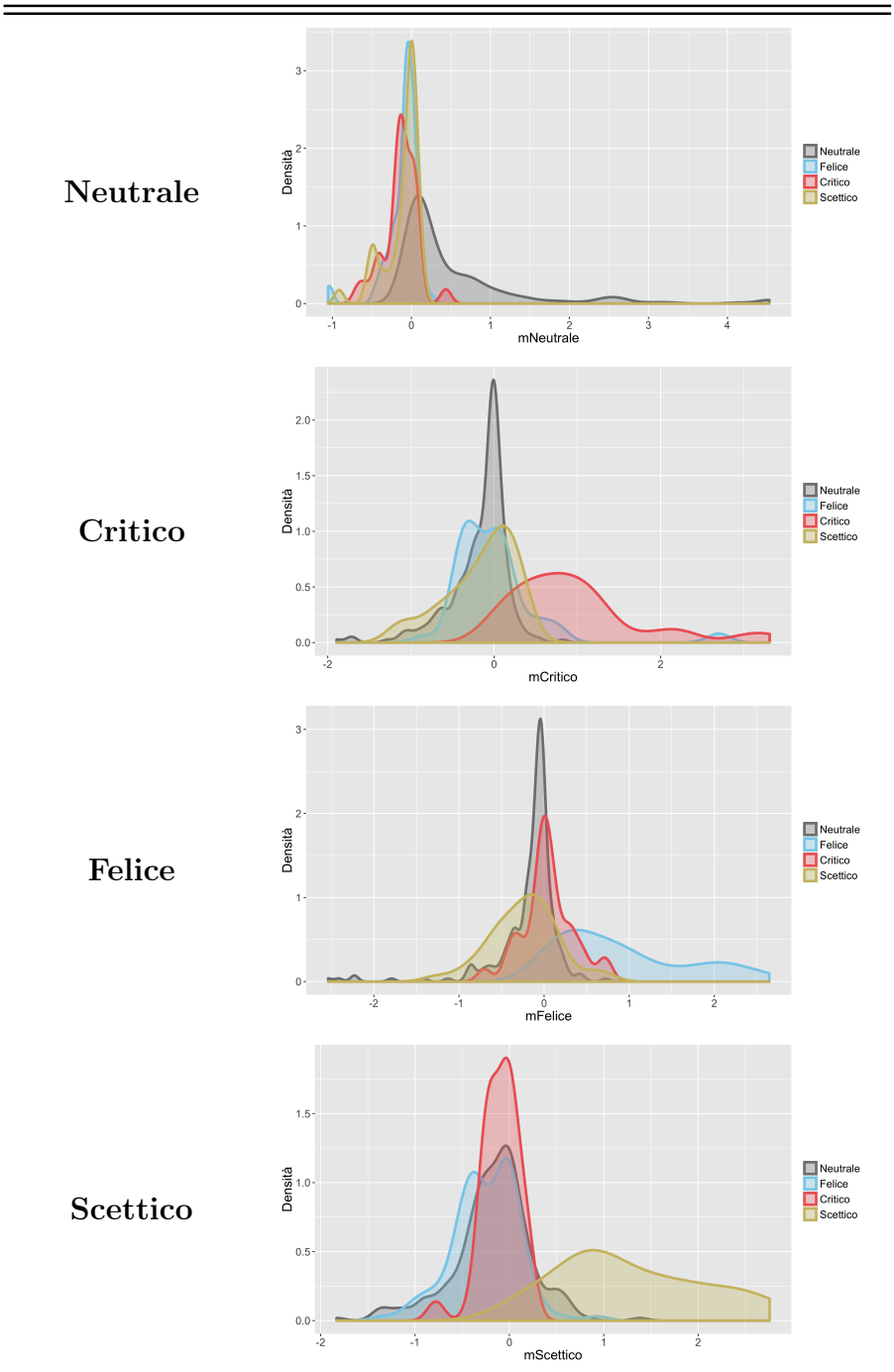

Figura 4.19: Marò - confronto delle densità delle proiezioni z_k fornite dalla MNIR

pare ancora più evidente quanto osservato in Figura 4.17, cioè la tendenza degli utenti dei gruppi di utilizzare un dizionario proprio. Tutte le densità del gruppo a cui si riferisce la proiezione risultano spostate verso destra e con una coda molto pesante in quella direzione. Ciò sta a rappresentare che nei documenti vengono utilizzati quasi esclusivamente termini con coefficiente positivo per quel *sentiment* o, in generale, i termini negativi vengono sopraffatti dai positivi. Inoltre tutte le code sinistre mostrano peso praticamente nullo a sinistra dello 0. L'unico gruppo che si distingue, per sua natura, è quello dei neutrali. La coda destra risulta molto lunga e pesante ma gli utenti degli altri gruppi assumono comunque valori positivi lungo questa dimensione. L'unica sovrapposizione che sembra esserci è tra Critico e Felice: nel secondo grafico, quello relativo alla proiezione $z_{Critico}$ si nota una "gobba" nel gruppo dei Felici intorno a 3. Spesso i critici, infatti, sono sì felici per il rientro ma si concentrano più sulla polemica politica.

Modello *forward*

Applicata la regressione multinomiale inversa per ottenere le proiezioni della *term-by-document matrix* nel nuovo spazio di dimensione $K = 4$, è stato possibile stimare la probabilità che ogni documento esprima un determinato *sentiment* attraverso il modello multilogit. I coefficienti del modello sono riportati nella *heatmap* in Figura 4.20.

Data la numerosità esigua del campione di utenti attivi su questo tema nel *sub-*

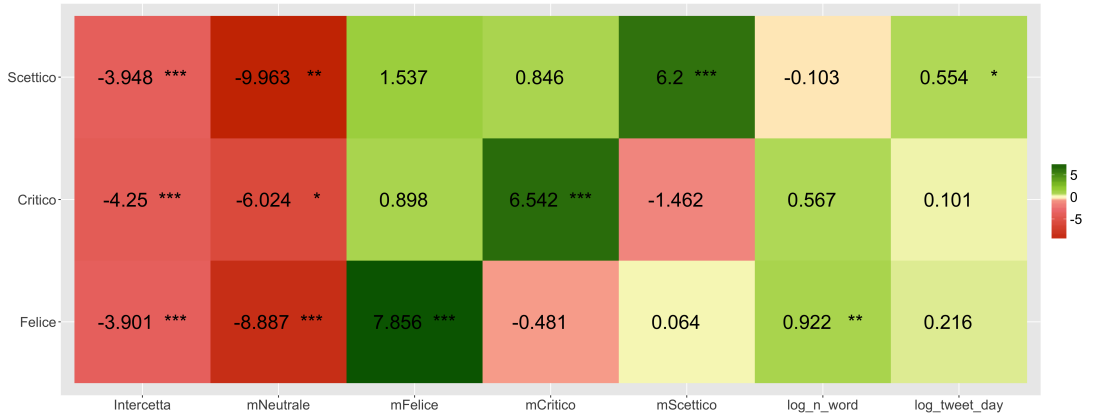

Figura 4.20: Ritorno Marò - regressione *forward* (categoria di riferimento: *Neutrale*)

set di analisi le stime risultano più incerte, quindi, di conseguenza, è più difficile riscontrare significatività. A differenza dei temi precedenti si osserva come la proiezione $z_{Neutrale}$ abbia pesi sostanzialmente differenti sui tre gruppi. L'impatto su Scettici e Felici è elevato, mentre la probabilità di essere Critico piuttosto che Neutrale sembra essere sì influenzata negativamente dal valore di $z_{Neutrale}$ ma non con la stessa intensità. Il coefficiente risulta inoltre significativo solo ad un livello $\alpha = 0.05$ (p-*value*$= 0.0299$). Per le altre tre proiezioni il comportamento è invece chiaro: coefficienti significativi e positivi sulla diagonale e non significativi altrove. Questo significa che ogni gruppo di utenti si esprime come proprio per il sentiment che esprime e non vi sono altre sovrapposizioni se non quella debole tra Neutrale e Critico. Passando al numero di parole sembra che i Felici siano maggiormente propensi a scrivere documenti più lunghi e di conseguenza più tweet rispetto a Neutrali e Scettici (coefficiente non significiativo ma negativo), mentre rispetto ai Critici la differenza tra i due coefficienti è pari a 0.3557 con uno standard error associato pari a 0.414, indicazione che fa propendere per la non significatività della stessa. Infine l'unica variabile superstite tra le covariate specifiche dell'utente è il numero di tweet scritti al giorno. Gli utenti più attivi hanno probabilità di essere Scettici significativamente maggiore rispetto alla neutralità, mentre nessuna delle due differenze tra Scettici e le altre due categorie è significativa.
Il modello in questione raggiunge ottime performance a livello di accuratezza, con un errore di errata classificazione pari all'8.74%.

Previsione sull'intero campione

Espandendo le previsioni operate dal modello *forward* presentato ai 4169 utenti attivi sul tema presenti nel dataset completo, si ottiene la distribuzione del *sentiment* rappresentata in Figura 4.21.

4.3.4 Referendum Trivelle

L'ultimo tema analizzato è relativo al referendum sulle concessioni per l'estrazione del petrolio in mare tenutosi il 17 aprile 2016. Il *sentiment* in questo caso è stato individuato in 5 livelli:

- **Neutrale:** l'utente non prende posizione riguardo la consultazione elettorale;

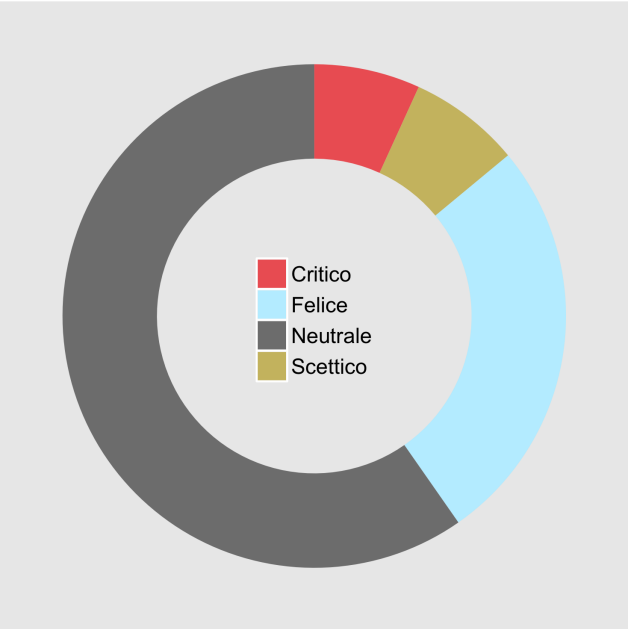

Figura 4.21: Previsione aggregata del *sentiment* circa il ritorno del Marò

- **Astensione:** l'utente esprime chiaramente la volontà di astenersi;
- **ContrarioNo:** l'utente esprime la sua perplessità circa il quesito referendario o manifesta l'intenzione di votare No;
- **Sì:** l'utente esprime la volontà di votare Sì;
- **Voto:** l'utente non entra nel merito della disputa ma esprime la volontà di andare a votare oppure condivide un appello al voto.

La distribuzione dei *sentiment* tra i 2238 utenti attivi su questo tema è mostrata in Figura 4.22. Ancora una volta il *sentiment* prevalente è Neutrale con 44.28% ma in questo caso il Sì segue a ruota con il 40.08% degli utenti. Il 9.96% si è invece speso per la partecipazione alle urne contro un 3.49% che ha esplicitato la propria volontà di astenersi dal voto. Infine, chiude il 2.19% che si è espresso per il No o, più in generale, ha manifestato perplessità circa la consultazione.

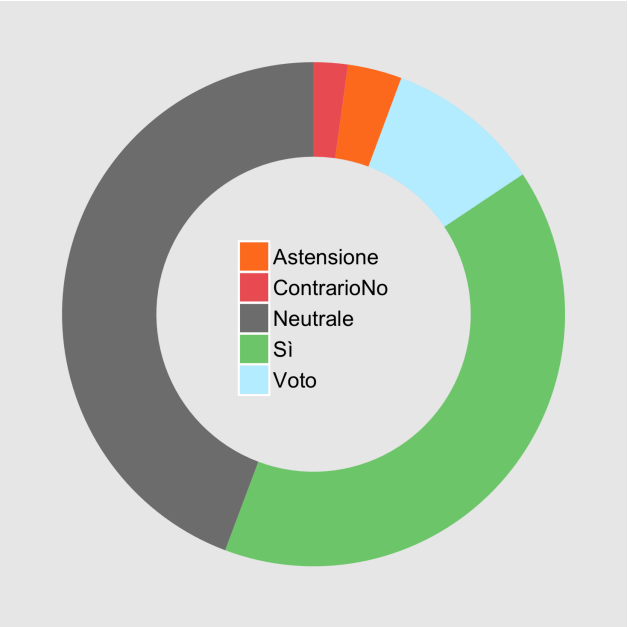

Figura 4.22: Distribuzione del *sentiment* circa il referendum sulle trivelle nel *subset* di analisi

Regressione Multinomiale Inversa

I termini più influenti per sentiment sono rappresentati nelle *wordcloud* di Figura 4.23. In questo caso non sembra esserci una tendenza per tutti i *sentiment* ad utilizzare un vocabolario proprio ma vi sono distinzioni. I neutrali, ad esempio, mostrano un coefficiente negativo su hashtag come "#iovotosì" o tag di politici di diversi schieramenti, come @luigidimaio e @giampaologalli, esponenti rispettivamente del Movimento 5 Stelle e del Partito Democratico. Sempre associati negativamente ai documenti neutrali sono bigrammi legati al voto come "andand vot" e "intenzion vot". Tra i termini verdi, invece, spiccano riferimenti al silenzio elettorale ed al suo rispetto ed il profilo di un quotidiano online come @linkiesta. Tra gli astenuti si ha equilibrio tra coefficienti positivi e negativi: tra le parole colorate di verde spicca sicuramente "#iononvoto" mentre stupisce la presenza di "#sivotasi". Tra gli altri token, sia positivi che negativi, invece, non si apprezza alcun particolare trend. Tra i contrari assumono maggiore rilevanza i termini non utilizzati, come "trivell vot" o "vot val", mentre tra i termini utilizzati spicca-

Neutrale

Astensione

ContrarioNo

Sì

Voto

Figura 4.23: Referendum Trivelle - valori più elevati (in valore assoluto) dei coefficienti φ_{jk}

no "mar pens" e "different". Anche in questo caso, comunque, non si evidenzia qualche riferimento particolare a tematiche del dibattito o personaggi politici. La *wordcloud* più informativa è sicuramente quella del Sì, la più verde tra le 5. Molto utilizzati, ovviamente, gli hashtag "#iovotosì" e "#unmaredisì", così come il più politico "#tuttiacasa". Tra i token utilizzati anche riferimenti alla tematica della salvaguardia dei mari con "amo mar", mentre l'hashtag "#iononmiastengo" anticipa una sovrapposizione tra i due *sentiment* Sì e Voto. I tag presenti fanno riferimento ad utenti molto attivi e fortemente schierati per il sì come @beatricedimadi e @dukana2. Tra i termini in rosso importante sottolineare la presenza di "#unmarediballe", hashtag utilizzato dagli oppositori al sì per confutarne le tesi, ed i riferimenti al silenzio elettorale. Infine a farla da padrone nella mappa del Voto è il colore rosso. Nessun trend chiaro tra questi termini, se non i riferimenti al ministro Guidi ed allo scandalo che la ha coinvolta. Tuttavia "#total", compagnia coinvolta nello scandalo, risulta tra i token positivamente associati al *sentiment*. Inoltre compaiono gli hashtag "#iovoto" e "#democrazia", oltre a riferimenti ad un voto consapevole come "vot inform" e "inform clicc". Lo scopo degli utenti di questo gruppo è, ovviamente, quello di "batt quorum" e di esercitare un proprio "diritt dov", citando due token utilizzati. In Figura 4.24 sono rappresentate le densità del numero di parole utilizzate per *sentiment*. I neutrali sembrano essere i più stringati nell'esprimersi, con una distribuzione pressochè simmetrica. D'altro canto gli astenuti sono coloro che utilizzano il maggior numero di parole, con una prevalenza importante verso valori elevati (più di 250 parole). Anche i fautori del Sì sembrano propensi a scrivere tanto, ma la coda è più *smooth* e la moda assume un valore minore rispetto all'Astensione. Per Voto e ContrarioNo le distribuzioni si concentrano su valori abbastanza alti, con una forma più piatta per i secondi ed in entrambi i casi un valore di densità che si azzera prima di 5 (circa 150 parole in scala originale).

Riprendendo quanto detto in commento alla Figura 4.23, i due *sentiment* Voto e Sì, per loro definizione, sembrano mostrare lati in comune. Può essere allora informativo confrontare, come fatto per gli altri temi, le densità delle proiezioni nel nuovo spazio per vedere quanto marcata sia questa sovrapposizione. In Figura 4.25 viene riportato il confronto delle densità per i due temi in questione. Confronto che conferma l'impressione manifestata dalle *wordcloud*: i due gruppi sembrano esprimersi con vocabolari simili e, in particolare, chi vota sì utilizza termini propri anche

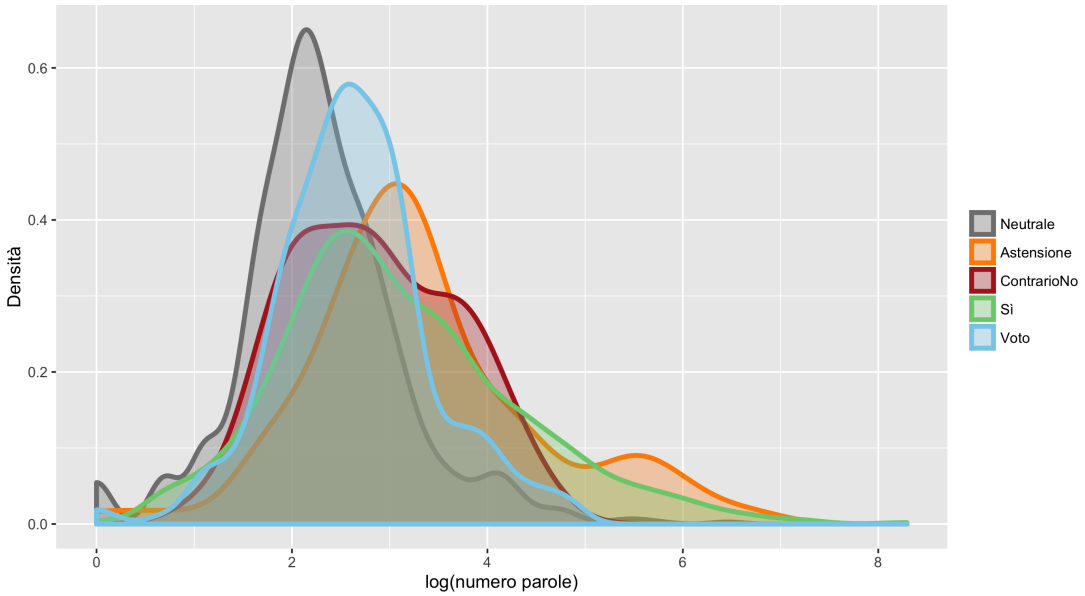

Figura 4.24: Referendum Trivelle - densità del numero di token nei documenti per *sentiment*

di chi invita al voto. Gli altri sentiment, invece, hanno distribuzioni simmetriche centrate in 0 senza particolari comportamenti degni di nota. Come accaduto per la Brexit diventa allora interessante capire quali termini discriminano maggiormente i due *sentiment*. In Figura 4.26 viene riportato lo scatter plot dei termini per cui la differenza assoluta tra i coefficienti $|\varphi_{j,Voto} - \varphi_{j,S}|$ risulta maggiore. Sono stati esclusi per chiarezza grafica i termini che avessero almeno uno dei due coefficienti pari a 0. I token più interessanti sono quelli che portano pesi di segno opposto ai termini. Ad esempio chi vota sì tende a fare più riferimenti a schieramenti e personaggi politici rispetto a chi fa appello alla partecipazione elettorale. Prova ne è la posizione di tag come "@pdnetwork", account ufficiale del PD, "@micheleemiliano", presidente della regione Puglia fortemente schierato per il sì, e l'hashtag "#boschi", riferimento al ministro Maria Elena Boschi. Entrando nel merito del quesito, fortemente discrimianti per il sì anche "ferm trivell" e "energ rinnov". Curiosa la differenza tra "#iovoto" e "#iovotosì": mentre il primo esprime valore positivo per entrambi i *sentiment*, il secondo è prerogativa di chi vota sì. Nel quarto quadrante, tra i termini discriminanti per Voto, si trovano riferimenti al quorum

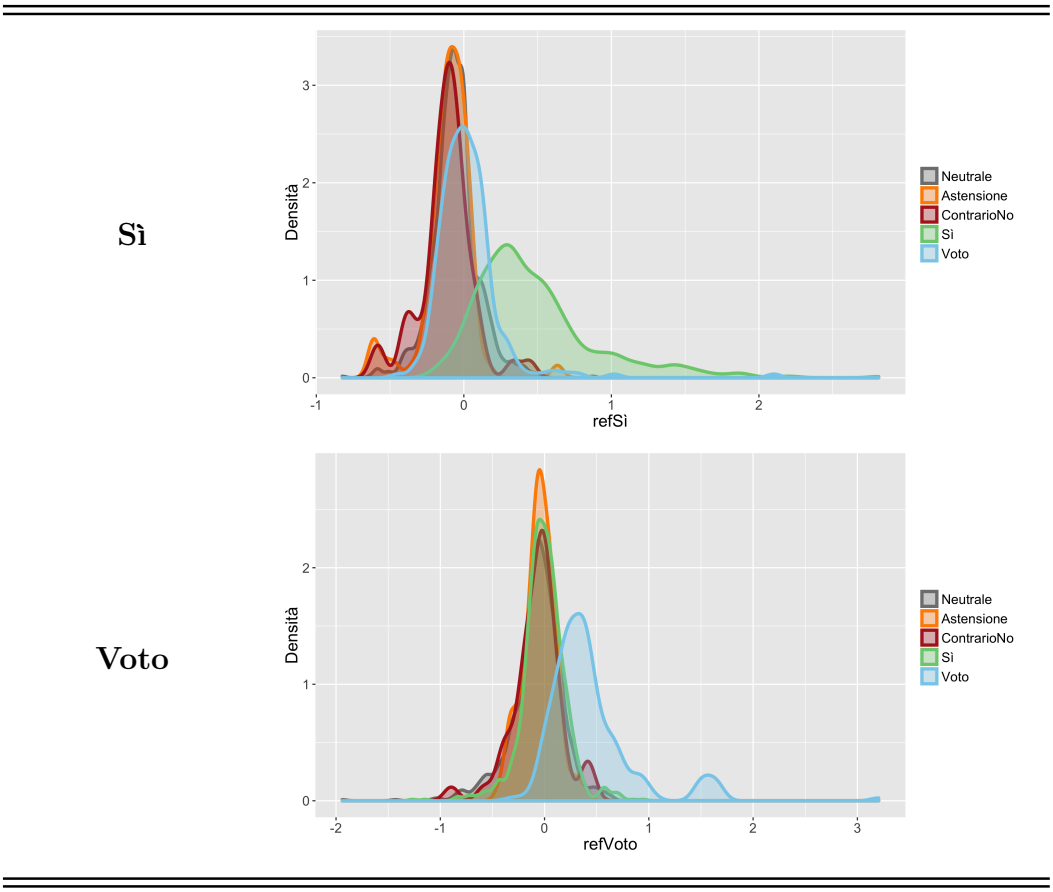

Figura 4.25: Referendum Trivelle - confronto delle densità delle proiezioni z_k fornite dalla MNIR per i *sentiment* $k = S$ e $k = Voto$

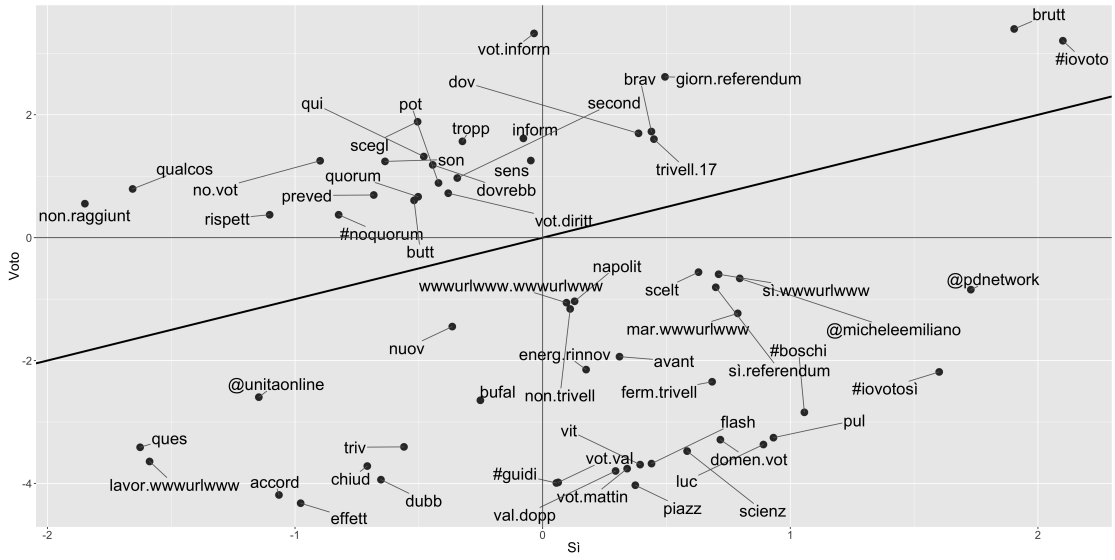

Figura 4.26: Termini maggiormente discriminanti tra Sì e Voto

("#noquorum", "quorum", "non raggiunt") e al voto informato, come già visto nelle *wordcloud* ("inform" e "voto inform").

Modello *forward*

Per concludere l'analisi delle applicazioni della regressione multinomiale inversa ai tweet raccolti è stato applicato il modello *forward* anche alle proiezioni relative al referendum sulle trivelle. In Figura 4.27 sono mostrati i coefficienti del modello multilogit con variabili selezionate con procedura *backward* e criterio di selezione AIC. Per quanto riguarda le proiezioni Z, andando con ordine, si notano comportamenti diversi su $z_{Neutrale}$. Sembra infatti che il Sì, pur avendo coefficiente fortemente significativo e negativo, si distingua dagli altri tre gruppi con un valore inferiore. Resta, come visto in tutti i temi e come lecito attendersi, la diagonale verde dell'incidenza di ogni proiezione sul proprio tema. Vi sono però coefficienti con significatività anche al di fuori di questa diagonale: ad esempio, chi manifesta la volontà di astenersi utilizza un vocabolario simile a quello dei Contrari, mentre non è vero il viceversa. Inoltre gli stessi astenuti si distinguono per il mancato utilizzo dei termini propri del Sì, nello stesso modo in cui i contrari non utilizzano termini propri di chi si appella al voto in un curioso incrocio di posizioni. In

Figura 4.27: Referendum Trivelle - regressione *forward* (categoria di riferimento: *Neutrale*)

modo quasi inspiegabile, poi, c'è un'associazione significativa tra valori positivi di $z_{ContrarioNo}$ ed il Sì. L'unica spiegazione potrebbe essere nel comune interessamento al merito della questione, con riferimenti a trivelle, mari e temi propri della consultazione. Passando oltre, si osserva con evidenza come i neutrali scrivano molto meno di tutti gli altri gruppi, a conferma dell'indicazione fornita da Figura 4.24. Sembra inoltre che chi prende posizioni contrarie o ostili al referendum scriva di più di chi invece sostiene la causa referendaria. In particolare, tra le quattro differenze (Astensione e ContrarioNo vs Sì e Voto) risulta significativa ad un livello del 5% la differenza tra ContrarioNo e Sì, mentre le altre mostrano significatività più marcate. Le covariate twitter-biografiche superstiti sono in questo caso sesso e numero di follower (su scala logaritmica). Per quanto riguarda la prima, il discorso è molto interessante. Lasciando un attimo da parte le significatività, è evidente la colorazione verde per *sentiment* contrari alla buona riuscita della consultazione e rosso per il Sì. Ciò si può spiegare con la presenza nel dataset di utenti di numerose associazioni ambientaliste come Greenpeace e Legambiente, molto attive sul fronte del Sì. Perciò per le altre due categorie della variabile la probabilità aumenta per posizioni lontane dal Sì mentre diminuisce per quelle più vicine. Chiudendo con il numero di follower, sembra che i neutrali siano i più conessi e seguiti. Soprattutto chi propende per il Sì tende ad avere meno seguaci, in apparente contraddizione con quanto osservato per il sesso. È opportuno ricordare che comunque gli enti

rappresentano un sottinsieme abbastanza ristretto degli utenti considerati (Figura 2.1).
Guardando le previsioni effettuate dal modello in confronto con quanto osservato emerge (Tabella 4.3) come la categoria che viene maggiormente riconosciuta dal modello è il Sì: il tasso di corretta classificazione condizionandoci a chi viene osservato come Sì è dell'85.17%. Quindi chi si esprime per il Sì è più facilmente prevedibile e "riconoscibile" rispetto alle altre categorie. Il tasso di accuratezza compessivo risulta comunque pari all'83.25%.

		Osservato				
		N	A	C	S	V
Previsto	N	889	20	19	115	67
	A	7	53	4	2	0
	C	8	1	21	2	0
	S	61	3	5	764	20
	V	26	1	0	14	136

Tabella 4.3: Referendum Trivelle - matrice di confusione osservati vs previsti

Previsione sull'intero campione

Il modello *forward* è stato inoltre utilizzato per prevedere il *sentiment* circa il referendum nell'intero campione dei 22647 utenti osservati ed attivi sul tema. In Figura 4.28 è riportata la distribuzione della previsione effettuata.

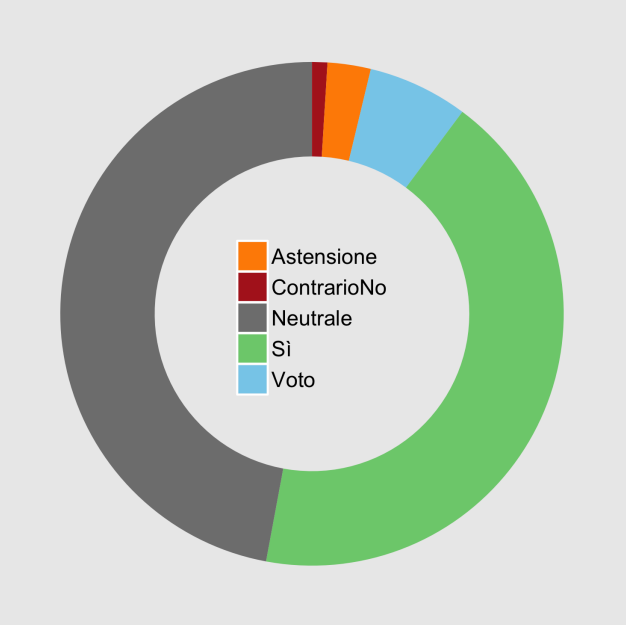

Figura 4.28: Previsione aggregata del *sentiment* circa il referendum sulle trivelle

Conclusioni

In un mondo sempre più interconnesso i social network rappresentano per chi analizza dati una fonte molto importante. La mole di informazioni che è possibile estrarre dal web è in continua crescita e non può essere ignorata in tutti quegli ambiti in cui è importante capire opinioni e comportamenti di una larga fetta della popolazione di riferimento. Nel caso di studio presentato l'attenzione si è focalizzata sul dibattito politico nel paese, analizzando le posizioni degli utenti Twitter su alcuni temi di attualità.
Se da un lato queste nuove fonti dati sono una potenziale miniera d'oro, d'altro canto, però, dati non strutturati come possono essere documenti testuali richiedono metodologie e strumentazioni apposite. In quest'ottica rivestono un ruolo importantissimo le fasi di *crawling* e *pre-processing* dei dati, in cui rispettivamente si acquisiscono i dati dal web e si opera il passaggio da testo grezzo a dati strutturati o semi-strutturati. L'importanza del *crawler* è evidente: una volta chiaro l'obiettivo di ricerca e lo scopo delle analisi è indispensabile procurarsi i dati corretti. Questa operazione può essere svolta in diversi modi: nel caso in esame si è deciso di procedere per *query*, cioè impostando per ogni tema analizzato un insieme di parole sulla base delle quali selezionare i tweet. Un'altra strada da percorrere potrebbe essere quella di scaricare un campione casuale di tutti i tweet che vengono pubblicati in una determinata finestra di tempo ed indagare i temi presenti attraverso un *topic model*.
Una seconda fase molto importante è quella di *pre-processing* e strutturazione (anche parziale) dei dati. Questo è forse il passaggio più delicato perchè definisce l'insieme dei dati da analizzare: si tratta di leggere i file scaricati tramite il *crawler* e creare dataset contenenti informazioni relative a testi, utenti e quant'altro. Inoltre, trattandosi di problemi di analisi testuale, diventa di fondamentale importanza

la costruzione delle *term-by-document matrix*, matrici contenenti rappresentazioni vettoriali dei documenti. Un'ulteriore problematica viene fornita dal linguaggio particolare proprio del social network (tag e hashtag): un insieme di problemi che devono essere gestiti per potere arrivare alle analisi vere e proprie con dati puliti in input. Il famoso rischio del *garbage in, garbage out* è incombente e, se possibile, in casi come quello esaminato si moltiplica.
La fase delle analisi è stata condotta seguendo uno scopo preciso: è stato infatti privilegiato sia nella scelta che nella costruzione dei modelli l'aspetto interpretativo rispetto a quello previsivo. Si è ritenuto fosse maggiormente di interesse, per una prima fase di indagine, valutare quali caratteristiche influissero sulla partecipazione dell'utente al dibattito su un tema e sulla sua opinione, o meglio sul suo *sentiment*, riguardo lo stesso. Riassumendo la questione, le domande di ricerca sono state "Chi scrive su ogni tema?" e "Coloro che ne scrivono, come ne parlano?".
Per rispondere al primo quesito si è fatto ricorso ad un modello marginale, il quale ha permesso di modellare congiuntamente logit marginali e *log-odds ratio* che esprimessero le interazioni a coppie tra i temi. Il risultato mostra come quasi tutte le variabili in gioco, sia quelle biografiche dell'utente che quelle relative all'account Twitter, influiscano sulla probabilità di esprimersi sui temi e, soprattutto, come le interazioni a coppie tra i temi siano significative. Ne consegue che un utente attivo sulla Brexit ha più probabilità di esserlo anche su tutti gli altri temi, mentre chi si esprime sulla legge Cirinnà tende a non interessarsi di altri argomenti se non, per l'appunto, la Brexit. Purtroppo, data l'onerosità computazionale del modello, non è stato possibile affinare l'indagine, ad esempio effettuando una selezione *stepwise* delle variabili oppure inserendo covariate nella modellazione delle interazioni. Uno sviluppo del lavoro presentato potrebbe sicuramente essere un miglioramento dell'efficienza computazionale del modello presentato.
La seconda domanda di ricerca è stata affrontata applicando alle matrici dei documenti una metodologia relativamente recente (Taddy (2013a)): la regressione multinomiale inversa. Essa ha permesso di proiettare i documenti in spazi di dimensione ridotta e di fornire una comoda interpretazione di questa trasformazione. La matrice di proiezione Φ, infatti, fornisce informazioni importanti circa le parole associate positivamente e negativamente ad ogni *sentiment*. Perciò, una volta classificato a mano un *subset* di documenti, è stato possibile fornire precise indicazioni circa il modo di esprimersi di ogni gruppo di utenti: chi si caratterizza

per l'utilizzo di un vocabolario proprio, chi per il mancato utilizzo di certi termini, chi per un mix dei due atteggiamenti. Le proiezioni permettono inoltre di valutare quanto le modalità di espressione dei gruppi siano, nel loro complesso, differenti tra loro e quali temi siano maggiormente sovrapposti. Ad esempio, nel caso del referendum sulle trivelle, si è evidenziata una parziale sovrapposizione tra le dimensioni relative a Voto e Sì. Grazie ai coefficienti di proiezione è stato però possibile mostrare quali siano i termini caratteristici e, di conseguenza, le tematiche maggiormente utilizzate dai diversi schieramenti. In secondo luogo è stato possibile unire alle proiezioni delle *term-by-document matrix* le covariate cosiddette twitter-biografiche per capire se ci fossero anche caratteristiche dell'utente che influissero sul *sentiment*. Il riscontro è stato positivo con forte incidenza del sesso e di alcune caratteristiche di rete (tweet per giorno e following). Un possibile sviluppo relativo a questa sezione è sicuramente legato alla parte previsiva. Nella Sezione 4.3 sono state mostrate le previsioni del *sentiment* per tema ottenute applicando i modelli *forward* illustrati per gli utenti esterni al *subset* classificato a mano. Questo approccio presenta alcuni limiti che possono essere superati applicando una metodologia semi-supervisionata. Applicando questo approccio è possibile inserire gli utenti non etichettati nel modello e, contemporaneamente, aggiornarne le componenti. Così facendo si coniugherebbero due importanti obiettivi: la previsione del comportamento di un ampio campione di utenti e l'interpretabilità dei modelli.
Concludendo, le metodologie illustrate ed applicate ad un ambito specifico possono sicuramente trovare spazio anche in altri settori. Basti pensare ad una campagna pubblicitaria o al lancio di un nuovo prodotto in ambito di marketing. È di sicuro interesse capire quali fasce di utenti vengano raggiunte ed attivate e confrontare l'impatto del proprio prodotto con altri di competitor nel mercato. È altresì importante capire quali siano i fattori che portano un utente a giudicare positivamente o negativamente un prodotto e potere agire di conseguenza. La semplicità divulgativa dei modelli utilizzati rappresenta sicuramente un vantaggio in situazioni come queste, dove l'applicazione delle metodologie può coinvolgere professionalità con competenze variegate. Altri ambiti di applicazione di una procedura simile a quella illustrata possono essere l'ambito finanziario, tecnologico-industriale o, tristemente di attualità, di intelligence, ambiti in cui può essere utile estraree conoscenza dal web per cogliere tendenze ed opionioni oppure per individuare particolari comportamenti di gruppi di utenti (si pensi alle indagini anti-terrorismo).

Bibliografia

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Wiley-interscience.

Azzalini, A., & Scarpa, B. (2012). *Data analysis and data mining.* Oxford University Press.

Bartolucci, F. (n.d.). *Analisi di dati categorici con modelli marginali epressi tramite vincoli di uguaglianza e disuguaglianza.* `http://http://www.stat.unipg.it/~bart/einaudi251002.pdf`.

Bartolucci, F., & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent markov heterogeneity structure. *Journal of the American Statistical Association*, *104*(486), 816-831.

Colombi, R., & Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, *88*(4), 1007-1019.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical science*, *22*(1), 1-26.

Glonek, G. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, *83*(1), 15-28.

Glonek, G., & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B, Methodological*, *57*(3), 533-546.

Hastie, T. J., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (Vol. 2. edition). Prentice Hall.

Lassen, N., Madsen, R., Vatrapu, R., Reichert, M., RinderleMa, S., & Grossmann, G. (2014). *Predicting iphone sales from iphone tweets.* Proceedings of

the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014).

Marchetti, G. M., Drton, M., & Sadeghi, K. (2015). *Package ggm.* https://cran.r-project.org/web/packages/ggm/ggm.pdf.

Paul, M., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS One*, *9*.

Schervish, M. J. (1995). *Theory of statistics.* Springer.

Taddy, M. (2013a). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, *108*(503), 755-770.

Taddy, M. (2013b). Measuring political sentiment on twitter: Factor optimal design for multinomial inverse regression. *Technometrics*, *55*(4), 415-425.

Taddy, M. (2015). *Package textir.* https://cran.r-project.org/web/packages/textir/textir.pdf.

Ringraziamenti

Un primo sentito ringraziamento va ai professori Livio Finos ed Alessio Farcomeni per la passione e l'interesse con cui hanno seguito questo lavoro di tesi, accogliendo idee e proposte e rendendosi sempre disponibili al confronto.

Grazie alla mia famiglia, a mamma Giovanna e papà Daniele per il sostegno in questi anni di studio, per non avermi mai fatto mancare il loro affetto e per essermi sempre stati vicini anche lontani chilometri, a mia sorella Martina, a cui forse dovrei più chiedere scusa per le mie assenze, e a mio fratello/cugino Stefano per esserci sempre.

Grazie alla mia famiglia padovana, ad Andrea ed Edoardo, compagni di viaggio, veri amici, per le risate, per i consigli e le chiacchierate, per le serate a condividere le nostre vite. Grazie di cuore ragazzi! Finisce un'avventura ma ciò che abbiamo costruito in questi due anni durerà per sempre.

Grazie ai compagni di viaggio padovani, ad Alice e Lucia, conosciute grazie al tutorato ma diventate subito vere amiche, ad Alvise, Arianna, Ciro, Chiara. Non finisce qua...

Grazi(e) ad Erica e Mara, grazi(e) di tutto, per esserci in ogni momento a strapparmi un sorriso, per ascoltare i miei sfoghi, per dimostrarmi cosa significhi la parola amicizia.

E infine grazie a te. Grazie perchè due anni fa questa mia avventura ci sembrava dura da affrontare ma ci ha solo dimostrato che insieme siamo invincibili. Grazie per sapermi leggere dentro sempre, per avermi aiutato a superare le mie debolezze e difficoltà, per essermi rimasta a fianco ogni giorno. Grazie per quello che siamo diventati in questi due anni. Semplicemente grazie, anche se in una parola non può essere contenuto tutto quello che vorrei dirti. Ti amo immensamente *Laura*...