Analisi dei post su Facebook dei politici italiani tramite LDA

Walter Genchi e Riccardo Menoli

1 Introduzione

L'analisi condotta prende in esame i post pubblicati su Facebook dal 1 gennaio 2017 al 30 aprile 2017 da 45 politici italiani, suddivisi per partito di appartenenza: si è scelto di considerare i 4 maggiori partiti politici italiani (Partito Democratico, Movimento 5 Stelle, Lega Nord e Forza Italia) e di ognuno sono stati selezionati i 10 politici con più "Mi Piace" sulla propria pagina pubblica; inoltre, sono stati presi in esame i leader di 5 partiti minori (Fratelli d'Italia, Movimento Democratico e Progressista, Possibile, Alternativa Popolare e Campo Progressista).

Il risultato preliminare è un corpus di 3.9 Megabyte composto da 39035 parole, che è stato ridotto a 1.3 Megabyte e 2814 parole dopo la fase di pre-processing.

L'obiettivo dell'analisi è l'identificazione degli argomenti (*topic*) discussi su Facebook dai maggiori esponenti della politica italiana. In quest'ottica lo strumento statistico-probabilistico scelto è il *topic model*. In particolare si è utilizzata la Latent Dirichlet Allocation (LDA).

2 Latent Dirichlet Allocation (LDA)

Nella teoria del *natural language processing* (NLP), la Latent Dirichlet Allocation è un modello statistico generativo di un corpus. L'idea di base è che ogni documento (l'insieme dei post di ciascun politico) può essere rappresentato come una mistura casuale di *topic* latenti, dove ciasun *topic* è caratterizzato da una distribuzione di parole, cioè da un vocabolario.

Nei modelli statistici generativi assumiamo che i dati provengano da un processo generativo che include *variabili latenti*. Tale processo definisce una distribuzione di probabilità congiunta sia delle variabili osservate che delle variabili latenti. Pertanto l'analisi statistica viene svolta usando la distribuzione congiunta per calcolare la distribuzione condizionale delle variabili latenti date le variabili osservate. Tale distribuzione condizionale viene detta *distribuzione a posteriori*.
LDA si colloca perfettamente in questo contesto: le variabili osservate sono le parole dei documenti, le variabili latenti rappresentano la struttura latente dei *topic*, mentre il processo generativo è quello descritto nella sezione 2.1.

2.1 Modello generativo

Un modello LDA con K argomenti assume il seguente processo generativo per un corpus D composto da M documenti, ciasuno di lunghezza N_i:

1. Generare θ_i da $\boldsymbol{\theta} \sim Dirichlet\,(\alpha)$ per $i = 1, \dots M$.
2. Generare φ_k da $\boldsymbol{\varphi} \sim Dirichlet\,(\beta)$ per $k = 1, \dots, K$.
3. Per ciascuna parola j nel documento i, dove $j = 1, \dots, N_i$ e $i = 1, \dots, M$,
   (a) Generare il *topic* z_{ij} da $Z_i \sim Multinomiale\,(\theta_i)$.
   (b) Generare la parola w_{ij} da $W_i \sim Multinomiale\left(\varphi_{z_{ij}}\right)$.

La lunghezza dell'i-esimo documento (N_i) è indipendente dalle altre variabili che hanno generato i dati.

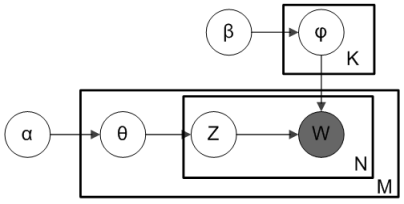

Figura 1: Modello grafico probabilistico per LDA

La proporzione dei *topic* per l'i-esimo documento è rappresentata dal vettore K-dimensionale θ_i, mentre la distribuzione delle parole (il vocabolario) per il k-esimo *topic* è descritta dal vettore V-dimensionale φ_k, dove V è il numero di parole (distinte) presenti nel corpus. Ne segue che i vettori α e β sono iperparametri del modello, rispettivamente di dimensione K e V. Infine, z_{ij} è l'assegnazione del *topic* per la parola j-esima dell'i-esimo documento del corpus e analogamente w_{ij} è la medesima parola, scelta all'interno del vocabolario del *topic* z_{ij} precedentemente selezionato.

Il modello generativo per LDA sopra descritto corrisponde alla distribuzione congiunta sia delle variabili osservate (w_{ij}) che di quelle latenti (z_{ij}, θ_i e φ_k):

$$p\left(\varphi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M}\right) = \prod_{k=1}^{K} p\left(\varphi_k\right) \cdot \left[\prod_{i=1}^{M} p\left(\theta_i\right) \cdot \left(\prod_{j=1}^{N} p\left(z_{ij} \mid \theta_i\right) p\left(w_{ij} \mid \varphi_{1:K}, z_{ij}\right)\right)\right] \tag{1}$$

La distribuzione della (1) specifica una struttura di dipendenza intrinseca nel modello generativo considerato: per esempio, la parola w_{ij} dipende non solo da z_{ij}, l'assegnazione del *topic* per la medesima parola, ma anche da $\varphi_{1:K}$, che stabilisce il vocabolario per i diversi *topic*. Un modo utile per rappresentare modelli generativi di questo tipo è il *modello grafico probabilistico*, così come si può vedere in Figura 1.

2.2 Calcolo della distribuzione a posteriori

Usando la notazione introdotta precedentemente, la distribuzione della struttra dei *topic* dato il corpus osservato è detta distribuzione a posteriori e può essere scritta per il teorema di Bayes come

$$p\left(\varphi_{1:K}, \theta_{1:M}, z_{1:M} \mid w_{1:M}\right) = \frac{p\left(\varphi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M}\right)}{p\left(w_{1:M}\right)} \tag{2}$$

Il numeratore è la distribuzione congiunta di tutte le variabili coinvolte e può essere calcolata facilmente per qualunque struttura delle variabili latenti. Il denominatore invece è la *distribuzione marginale* delle parole osservate, che corrisponde alla probabilità di osservare il corpus in questione sotto tutte le possibili combinazioni di variabili latenti.

Tale ampiezza di possibilità rende il calcolo ingestibile e impone un cambio di prospettiva: i metodi moderni bayesiani permettono infatti di ottenere una approssimazione della distribuzione a posteriori. Gli algoritmi usati a questo fine nel *topic modeling* si dividono generalmente in due categorie: algoritmi basati sul campionamento e metodi variazionali.

L'obiettivo dei primi è quello di ottenere dei campioni dalla distribuzione a posteriori per approssimarla con la distribuzione empirica. In questo contesto, l'algoritmo più usato per il *topic modeling* è il *campionamento di Gibbs*, dove si costruisce una *catena di Markov*, la cui distribuzione limite converge alla distribuzione a posteriori cercata. Si veda Steyvers and Griffiths [3] per una descrizione del campionamento di Gibbs per LDA.

I metodi variazionali rappresentano un'alternativa deterministica agli algoritmi basati sul campionamento. Invece di approssimare la distribuzione a posteriori con dei campioni, i metodi variazionali assumono una famiglia di distribuzioni parametriche per le variabili latenti e in seguito trovano il membro della famiglia che è più vicino alla distribuzione a posteriori, dove tale vicinanza è misurata con la divergenza di Kullback-Leibler. In questo modo il problema inferenziale viene ridotto ad un problema di ottimizzazione. Si veda Blei et al. [1] per un algoritmo *coordinate ascent* di inferenza variazionale per LDA.

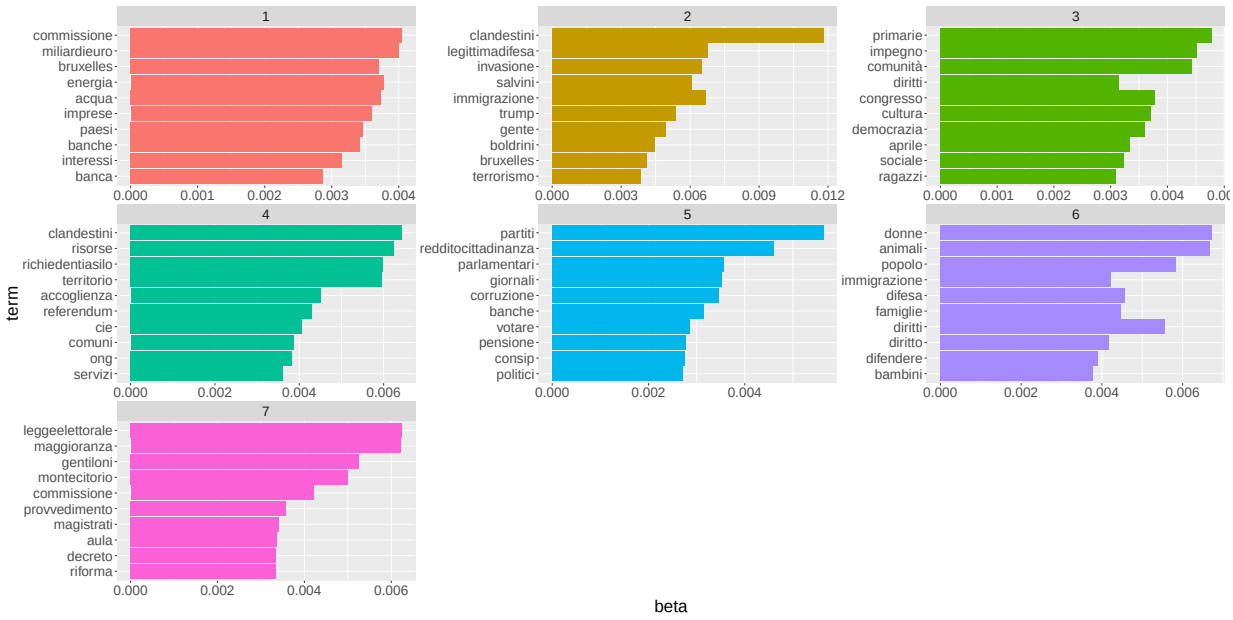

Figura 2: Distribuzione delle prime 10 parole per 7 *topic*

3 Applicazione di LDA ai dati

Lo scaricamento dei post è stato possibile grazie alla funzione `getpage` della liberia `Rfacebook`. Dopo aver ottenuto un unico elemento `character` per ciascun politico è iniziata la fase di *preprocessing*.

3.1 *Preprocessing*

Il lavoro di *preprocessing* può essere riassunto nelle seguenti 3 fasi:

1. Riduzione di lettere maiuscole in minuscole, eliminazione di numeri, segni di punteggiatura, spazi bianchi, nomi propri, nomi dei partiti, caratteri speciali del Web e stopwords. Alla fine di questa fase il corpus consiste in 33925 parole.

2. Individuazione e aggiunta al corpus dei bigrammi con frequenza superiore a 30. Tale soglia è stata scelta in base alla sensatezza dei bigrammi più frequenti nel corpus. Sono stati aggiunti un totale di 126 bigrammi.

3. Eliminazione delle parole che all'interno del corpus compaiono troppo frequentemente e troppo raramente. Nel primo caso è stata scelta come soglia il quantile 0.975 della distribuzione di frequenza delle parole, mentre nel secondo caso si è scelto di eliminare le parole presenti in meno di 10 documenti (45 documenti in totale). Alla fine di questa fase il corpus consiste in 2814 parole, dove 72 parole sono state eliminate da quelle troppo frequenti e 31162 parole sono state rimosse da quelle rare.

Si è scelto di non introdurre l'operazione di stemming in quanto questo portava ad una interpretazione del modello LDA poco chiara, dovuta principalmente alla bassa efficacia dell'implementazione di default dello stemming per la lingua italiana.

3.2 Stima del modello

L'implementazione dell'algoritmo per LDA è stata svolta dalla funzione `LDA` della libreria `topicmodels`. Gli argomenti in input utilizzati sono:

- `x`, una matrice `DocumentTermMatrix`.

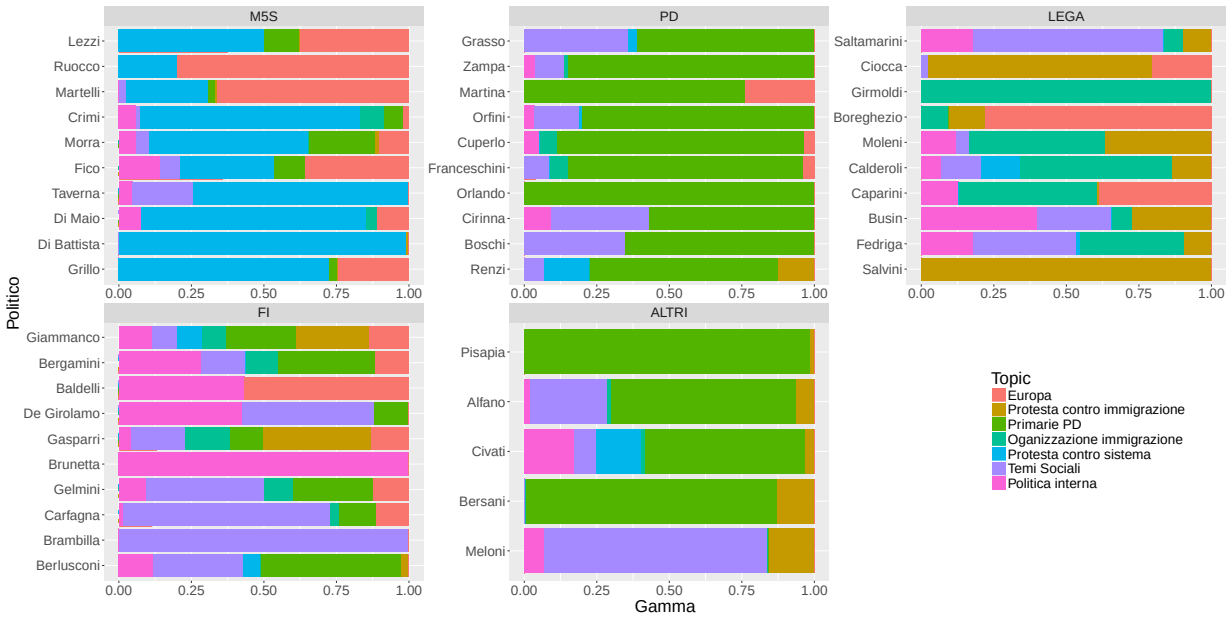

Figura 3: Distribuzione dei *topic* nei politici

- `k`, il numero di *topic* scelto.
- `method`, dove si è scelto di usare il default `"VEM"`, ossia il metodo variazionale ben descritto da Blei et al. [1] L'alternativa è `"Gibbs"`, ossia l'algoritmo basato sul campionamento esposto nella sezione (2.2).

L'output della funzione `LDA` è un oggetto di tipo `S4` che contiene diversi elementi. I due elementi di nostri interesse sono:

- `beta`, una matrice $K \times V$, nel nostro caso 7×2693[1]. La k-esima riga di `beta` corrisponde esattamente alla stima di massima verosimiglianza del vettore φ_k introdotto nella sezione (2.1), ovvero la distribuzione delle parole per il k-esimo *topic*.
- `gamma`, una matrice $M \times K$, nel nostro caso 45×7, dove l'i-esima riga di `gamma` è la stima del vettore θ_i introdotto nella sezione (2.1), ossia la proporzione dei *topic* per l'i-esimo documento.

Si veda Hornik and Grün [2] per un approfondimento sulla funzione `LDA`.

3.3 Risultati

La Figura 2 rappresenta la distribuzione delle prime 10 parole (quelle con i 10 maggiori valori in ciasuna riga della matrice `beta`) per un numero scelto di 7 *topic*. Se consideriamo per esempio la distribuzione delle parole per il Topic 4, la parola "clandestini" ha una probabilità stimata di essere presente nel Topic 4 pari a 0.00643 e si tratta della parola con più alta probabilità all'interno del Topic 4. La stessa parola ha probabilità di essere presente nel Topic 2 pari a 0.0118 ed è anche qui la parola con più alta probabilità all'interno del *topic*.

Se osserviamo le altre parole presenti all'interno dei *topic* in Figura 2 possiamo "naturalmente" assegnare alcune etichette ai diversi *topic*: Topic 1: Temi legati all'Europa; Topic 2: Protesta contro l'immigrazione; Topic 3: Primarie del PD*;* Topic 4: Organizzazione del fenomeno dell'immigrazione; Topic 5: Protesta contro il sistema; Topic 6: Temi sociali; Topic 7: Politica interna.

[1]Si sono eliminati ulteriori 121 parole rispetto alle 2814 parole presenti al punto 4 delle sezione (3.1), al fine di permettere una maggiore interpretabilità dei risultati.

Figura 4: Confronto delle parole con maggiore differenza nella matrice `beta` per due coppie di *topic*

In Figura 3 è stata rappresentata la distribuzione dei *topic* nei 45 politici, contenuta nella matrice `gamma`. Si noti una forte prevalenza del Topic 4 tra gli esponenti politici del M5S, mentre una eterogenerità di *topic* all'interno dei politici della Lega, dove alcuni si concentrano più sulla protesta contro l'immigrazione (Salvini e Ciocca), mentre altri discutono in misura maggiore sull'organizzazione del fenomeno dell'immigrazione (Grimoldi e Calderoli). Inoltre il tema sociale è trattato in prevalenza da Brambilla (animalista), Meloni (nazionalista) e Saltamarini (attenta alle tutele della maternità).

È possibile confrontare la diversità tra i processi generatori dei *topic* di due politici, ossia $Z_i \sim Multinomiale(\theta_i)$ e $Z_j \sim Multinomiale(\theta_j)$ per $i, j = 1, \dots 45$ e $i \neq j$, utilizzando la distanza di Bhattacharyya[2]. Nel caso di Brambilla e Meloni tale misura di distanza fornisce un valore pari a 0.123, tra Brambilla e Saltamarini 0.198 e tra Meloni e Saltamarini è pari a 0.395. La distanza massima è 5.211 e viene osservata tra Grimoldi e Brunetta, mentre la distanza minima pari a 0.007 si ha tra Orlando e Pisapia. La distanze media è pari a 1.347.
Un ulteriore confronto è stato effettuato analizzando la compattezza del partito in base alle differenze di *topic* discussi dal leader e dagli appartenenti al medesimo partito, ottenendo quindi nove confronti per ciascun partito. La distanza media di Bhattacharyya tra Grillo e gli altri politici del M5S è pari a 0.151; la stessa misura è pari a 0.221 per il PD, 1.459 per la Lega e 0.499 per Forza Italia.

La Figura 4 mostra in forma di wordcloud le parole che presentano la maggiore differenza nella matrice `beta` tra i Topic 4 e 2 e tra i Topic 1 e 6.
Concentrandosi d'ora in avanti sul confronto tra Topic 4 e Topic 2, tale differenza può essere calcolata basandosi sulla quantità $\log(\varphi_4/\varphi_2)$, che permette una migliore visualizzazione delle parole per cui il rapporto φ_4/φ_2 è molto grande. Al fine di restringere l'analisi ad un insieme di parole rilevanti, è bene fissare una soglia per entrambi i *topic* e considerare solo quelle parole che hanno un valore $\varphi_4 > s_4$ e $\varphi_2 > s_2$. Le parole in blu sono quelle per le quali $\log(\varphi_4/\varphi_2) > 0$, cioè quelle caratterizzanti il Topic 4, mentre le parole in rosso hanno $\log(\varphi_4/\varphi_2) < 0$ e descrivono il Topic 2. La grandezza delle parole è proporzionale a $\log(\varphi_4/\varphi_2)$: più grande è la dimensione della dimensione della parola, maggiore è la sua capacità di discriminare tra i due *topic*.

3.4 Conclusioni

Una delle maggiori potenzialità di LDA consiste nella flessibilità del metodo di ricerca dei *topic*: è prevista infatti la possibilità di assegnare la stessa parola a più *topic*, caratteristica molto rilevante quando il corpus contiene testi che trattano diverse tematiche. Inoltre, nel caso specifico è stato possibile ottenere una buona fotografia del panorama politico italiano, incluse le discrepanze degli argomenti affrontati da politici dello stesso partito.

Il limite maggiore di LDA risiede nella scelta del numero dei *topic*. Trattandosi di un metodo *non supervisionato*, la scelta di K è arbitraria e necessita di verifiche a posteriori. Nel caso specifico il criterio seguito per la scelta di K è stato l'interpretabilità dei *topic*, eliminando le parole poco discriminanti. Esistono inoltre alcune assunzioni di LDA che possono risultare troppo stringenti: CTM (*Correlated Topic Model*) assume per esempio la possibilità che vi sia una correlazione tra i *topic*; nel caso dei post dei politici, appare ragionevole includere la possibilità che vi sia una certa correlazione tra i due *topic* sull'immigrazione.

[2]La distanza di Bhattacharyya gode della proprietà di simmetria ed è stata per questa ragione preferita alla divergenza di Kullback-Leibler.

Riferimenti bibliografici

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Kurt Hornik and Bettina Grün. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

[3] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.