ANALYSIS DOCUMENT

Fatjona Bekollari[1], Luca Bassanese[1], Andrea Ongaro[1]

[1] *Università degli studi di Milano-Bicocca*

27 gennaio 2019

1 Introduzione, Motivazione e estrazione dati

È sotto gli occhi di tutti, i social network sono parte integrante della nostra quotidianità. Le relazioni di oggi spesso nascono sui social network, quali Twitter, Instagram, Facebook. Tali piattaforme sono fondamentali per instaurare relazioni ma anche per affermarsi nel mondo del lavoro. Lo dimostra il fatto che le aziende investono una grande parte del loro capitale in pubblicità che avviene soprattutto mediante l'utilizzo dei social. Sembrerebbe che il mondo dei social sia diventato un vero e proprio business. Anche la politica è stata travolta da questa ondata social: spesso i politici sono star del web. La logica è la stessa per tutti: riuscire a fare presa su più persone possibili per raggiungere i propri obiettivi. Questo porta a porci delle domande:

Come comunicano davvero i politici italiani su Twitter? Questa comunicazione varia secondo il consenso elettorale del proprio partito?

Per rispondere a tali domande si è deciso di utilizzare i dati Twitter poiché secondo molti giornali tra cui "Il Sole 24 Ore" twitter per l'anno 2018 si conferma il social più politico in Italia. Lo dicono gli hashtag più utilizzati e gli account twitter più menzionati. Tra questi vi sono gli account di Matteo Salvini, Matteo Renzi, Luigi Di Maio e gli hashtag del movimento 5 stelle, Matteo Salvini e Partito Democratico. Si è notato che nella maggior parte dei casi i post pubblicati su Twitter venivano poi ripostati su Facebook ma poiché i dati di Facebook, in seguito allo scandalo di Cambridge Analytica, non sono più disponibili si è deciso di tener conto solo dei dati di Twitter.

I social network, in particolare Twitter, rappresenta la piazza più grande che i politici hanno mai avuto per poter esprimere i pensieri, le idee, le reazioni in seguito a qualche accadimento, i valori, i principi sui quali si fonda il loro partito o più nello specifico il loro modo di vedere la politica e di voler migliorare il paese. Twitter rappresenta un'arma potente e proprio per la sua potenza si immagina che le parole usate per arrivare ai cittadini siano di fondamentale importanza.

Twitter permette di avere oltre ad un account twitter uno chiamato "Developer" tramite il quale, una volta specificato il nome dell'applicazione che si vuole creare, lo scopo che si vuole raggiungere con l'utilizzo dei dati Twitter e il motivo per il quale si intende fare ciò, è possibile avere per ogni account:

- *Created_at* : data e ora del tweet
- *Text* : testo del tweet
- *hashtags* : hashtags utilizzati nei tweet
- *followers_count* : seguaci
- *friends_count* : seguiti
- *account_created_at*: data di creazione dell'account twitter

L'account developer di base, utilizzato da noi tramite il pacchetto rtweet, consente l'estrazione degli ultimi 3200 tweets di un determinato account. Per questo motivo per gli account di Grillo e Salvini, che hanno nel periodo considerato più di 3200 tweet, abbiamo dovuto utilizzare l'account premium gratuito di Twitter. Esso consente di estrarre 100 tweet alla volta, indicando però la data desiderata. Quindi, sempre tramite R, abbiamo dovuto estrarre 100 tweet a volta per riuscire a ricostruire tutto l'arco temporale considerato. Per i dati relativi ai sondaggi elettorali abbiamo utilizzato il sito YouTrend che contiene lo storico relativo al periodo 2017-2018. In particolare abbiamo solamente analizzato i sondaggi dell'agenzia SWG perché erano quelli più numerosi.

In primo luogo si è deciso di analizzare l'andamento del consenso elettorale dei maggiori partiti italiani:

- Movimento 5 stelle
- Lega Nord
- Partito Democratico

- Forza italia
- Fratelli D'Italia
- Liberi e Uguali

Da inizio 2017 fino al periodo delle politiche del 4 marzo 2018, per quasi tutti i partiti si nota un trend costante. Dopo lo shock delle elezioni si evidenzia un cambiamento di direzione e intensità dei trend.

Successivamente sono stati considerati gli account Twitter delle figure politiche più importanti, concentrandoci poi su quelli più influenti dove per influenti si intende coloro che rappresentano i 4 partiti più importanti in Italia ovvero M5S, LN, PD, FI.

La comparsa dei vari personaggi politici è dilazionata nel tempo, addirittura Berlusconi ha creato il suo account in vista delle elezioni, probabilmente per ampliare i suoi canali di comunicazione allineandosi così ai suoi rivali. Inoltre emerge che il numero dei follower non rispecchia esattamente le preferenze politiche degli italiani poiché l'utenza di twitter non è un campione rappresentativo della popolazione italiana.

In un'analisi preliminare dove sono stati considerati la somma dei tweet per mese emerge subito il fatto che: Berlusconi, dopo l'exploit del periodo elettorale, riduce drasticamente l'utilizzo di twitter infatti il numero di tweet mensili si è decisamente ridotto. Per Renzi, si nota il comportamento opposto, da quando è all'opposizione il suo utilizzo di Twitter è aumentato. Spiccano Grillo e Salvini.

Da questo momento l'attenzione è stata rivolta a Salvini, Renzi, Di Maio, Berlusconi. Nei due anni considerati questi quattro politici hanno avuto una differente evoluzione sia in termini di consenso elettorale che di ruolo istituzionale. Per Salvini e Di Maio, dopo le elezioni a loro favorevoli e la formazione del governo, è seguita un'evoluzione opposta dei loro voti: ad una crescita sostenuta del primo si è contrapposta una riduzione del secondo. Dall'altra parte Berlusconi e Renzi, che rappresentano oggi l'opposizione, hanno sofferto dell'erosione del loro consenso dei partiti al potere.

2 STATISTICAL METHOD

2.1 TEXT MINING

La parte di text mining si è concentrata principalmente sull'analisi delle parole più frequenti. A questo fine abbiamo utilizzato il pacchetto di r *tidytext* che fornisce strumenti molto utili per l'analisi dei dati testuali.

Questo pacchetto si basa principalmente sulla trasformazione del dataset in tipo *tidy*, che consente di gestire i dati in modo semplice ed efficace(soprattuto quando si tratta di dati testuali). Questi dati hanno una particolare struttura:

- Ogni variabile è una colonna
- Ogni osservazione è una riga
- Ogni tipo di unità è una tabella

I dataset in formato tidy consento la manipolazione tramite i vari pacchetti: *dplyr, tidyr, ggplot2 e broom*.

Fondamentale è il processo di tokenizzazione che consiste nel suddividere il testo in tokens, che rappresentano una parte d'interesse nel testo. Con l'utilizzo di questo tipo di dati, il testo risulta essere una tabella con un token per riga. Molto spesso il token è rappresentato da una parola, ma si possono utlizzare anche n-gram(composizioni di n parole), frasi o paragrafi.

Nel nostro caso l'analisi si è concentrata sulla singola parola. Abbiamo comunque provato a sviluppare la parte relativa agli n-gram(individuando anche la correlazione tra parole) che abbiamo però deciso di non inserire nel blog post.

La funzione che ci consente di eseguire questa operazione di tokenizzazione è *unnest_tokens*. Abbiamo in particolare utilizzato l'opzione *token = tweets* che consente di mantere i simboli di twitter quali # e @ . Dopo aver tokenizzato, abbiamo rimosso le stopwords(lista presa dal web) che rappresentano le parole non significative e che vengono utilizzate molto spesso(congiunzioni, verbi essere e avere, proposizioni, articoli, etc ...).

Dato che il nostro obiettivo era quello di catturare il cambiamento della comunicazione nel corso dei due anni d'analisi, abbiamo deciso di dividere il conteggio delle parole più frequenti in due macroperiodi: prima delle elezioni e dopo le elezioni. Abbiamo deciso di fare ciò perchè le elezioni sono l'evento che ha causato lo shock maggiore in termini di consenso elettorale, le percentuali di voto sono cambiate di molto, prima e dopo questa giornata. Per rappresentare le parole più frequenti in questi due periodi abbiamo sfruttato lo strumento dei wordcloud. Esso crea una *nuvola di parole* dove la grandezza della rappresentazione delle parole è proporzionale alla freqenza.

L'utilizzo di questa analisi ha evidenziato un cambiamento nell'uso di vocaboli da parte dei politici, chi più chi meno.

L'altra parte di text mining inserita consiste in un app shiny integrata da un grafico interattivo costruito tramite il pacchetto *plotly*. Shiny consente di costruire applicazioni che consentano all'utente di interagire con i dati. In particolare noi abbiamo costruito un app in cui è possibile inserire una parola del quale viene poi visualizzata la serie storica relativa all'utilizzo di essa da parte dei vari politici.

2.2 SENTIMENT ANALYSIS

Il secondo metodo statistico che si è utilizzato è la sentiment analysis. Questa analisi ci permette di rilevare i sentimenti trasmessi da un input testuale, nel nostro caso sono stati analizzati i testi dei tweet dei politici, al fine di comprendere se è avvenuto un mutamento nella comunicazione degli stessi. Dal punto di vista operativo la sentiment analysis si basa sull'utilizzo di lexicon, particolari dizionari che permettono di associare ad

ogni parola un valore emotivo. In questa analisi ci si è serviti di tre lexicon:

- **NRC Word Emotion Association Lexicon**: lexicon ampiamente usato in letteratura, creato tramite crowdsourching, che permette di associare ad ogni parola 2 sentimenti (positive, negative) e 8 emozioni diverse(anger, anticipation, fear, disgust, joy, sadness, surprise, trust) .
- **Sentix(Sentiment Italian Lexicon)**: lexicon italiano, utilizzato già per l'analisi dei Tweet nel progetto "Twita", basato su SentyWordNet. Questo dizionario permette di associare ad ogni parola 4 diversi valori: positive, score, negative score, polarity, intensity.
- **OpeNer Sentiment Lexicon Italian**: sviluppato in modo semi-automatico da ItalWordNet v.2 partendo da una lista di 1.000 parole chiave controllate manualmente. Contiene 24.293 entrate lessicali annotate con polarità positiva/negativa/neutra.

Tutti e tre lexicon elencati sopra soffrivano di tre grandi difetti, infatti essi non tenevano conto della declinazione di verbi, nomi e aggettivi, risultando pertanto incompleti.

In primo luogo il problema relativo ai verbi era dovuto al fatto che veniva riportata solo la forma all'infinito presente. Si è pertanto proceduto alla coniugazione di tutti i verbi in base all'appartenenza ai gruppi are, ere ed ire. Infine si è associato ad ogni verbo declinato i valori emotivi posseduti dal verbo di partenza. In secondo luogo i nomi comparivano solo ed esclusivamente al singolare, pertanto si è provveduto ad aggiungere i plurali seguendo le regole grammaticali italiane:

- I nomi maschili che finiscono in a al plurale finiscono in i
- I nomi femminili che finiscono in a al plurale finiscono in e
- I nomi che finiscono in o al plurale finiscono in i
- I nomi che finiscono in e al plurale finiscono in i
- I nomi che finiscono in i anche al plurale finiscono in i
- I nomi che finiscono con una lettera accentata al plurale non cambiano

Come in precedenza si è associato ad ogni plurale derivato i valori emotivi posseduti dal singolare.

Infine si è rivolta l'attenzione agli aggettivi, essi presentavano solo la forma al singolare maschile, da esso abbiamo ricavato per ogni aggettivo il femminile singolare e plurale ed il maschile plurale, seguendo le diverse declinazioni in base all'appartenenza dell'aggettivo alle tre diverse classi previste dalla grammatica italiana. Analogamente ai casi precedenti sono stati associati ad ogni forma dell'aggettivo i medesimi valori emotivi.

Dopo aver resi più completi i lexicon a nostra disposizione si è proceduto alla creazione di una funzione computazionalmente efficiente che ricevendo come input una sequenza di parole, restituisse come output le stesse associate con il valore emotivo, con la possibilità di scelta del lexicon da utilizzare.

Nell'applicazione della sentiment analysis si è deciso di utlizzare esclusivamente i lexicon NRC e Sentix, data la loro maggiore popolarità in letteratura. I risultati ottenuti con i due diversi lexicon si sono rilevati coerenti tra di loro ed hanno portato alla luce un cambiamento nella comunicazione dei politici negli ultimi due anni. Esso più che seguire fedelmente il consenso elettorale del partito di appartenenza, avviene in risposta alla variazione del ruolo istituzionale del politico preso in considerazione (principalmente il passaggio da governo a opposizione e viceversa).

3 Lavori Correlati e Spunti Futuri

Ciò che ci ha spinto a fare questa analisi è senza dubbio stato anche il desiderio nell'affrontare un tema totalmente nuovo per noi quale text mining e sentiment analysis. Abbiamo voluto adattarlo al panorama politico dato l'interesse che ci accomuna. Inoltre è importante precisare che esistono numerosi articoli o blogpost sull'analisi dei social dei politici americani(Trump in particolare) ma dei Politici italiani poco o niente. Siamo dell'idea che il nostro studio possa essere usato anche come fonte di ispirazione per analisi più approfondite, ad esempio sulle serie storiche. Sarebbe anche interessante concentrarsi unicamente sul profilo social di Matteo Salvini. Lui, infatti, usa tantissimo i social ed in questo può essere accostato a Donald Trump, come è riuscito a sfruttare i social network nella sua scalata a personaggio politico più acclamato attualmente.

4 Pacchetti R

Nella nostra analisi ci siamo serviti dei seguenti pacchetti:

- rtweet (raccolta di dati di twitter)
- httr (permette di lavorare con url e http)
- base64enc (codifica)
- lubridate (gestione delle variabili temporali)
- dplyr (working with data frames)
- tidyverse (is an opinionated collection of R packages designed for data science)
- readr (legge diversi tipi di dati come "csv", "tsv")
- tidytext (Text mining for word processing and sentiment analysis)
- stringr (operazioni sulle stringhe)
- tokenizers (Fast, consistent tokenization of natural language text)
- broom (Convert statistical analysis objects into tidy tibbles)
- plotly (Create interactive web graphics)

- tweenr (Interpolate data for smooth animations)
- gganimate (create animations with ggplot2)
- wordcloud (crea nuvole di parole)
- readxl (read excel files)
- dygraphs (is an R interface to the dygraphs JavaScript charting library. It provides rich facilities for charting time-series data.)
- tm (text mining)
- ggraph (versione più estesa di ggplot2)
- igraph (grafici. Gestisce bene grafici di grandi dimensioni)
- widyr (Encapsulates the pattern of untidying data into a wide matrix, performing some processing, then turning it back into a tidy form)
- ROAuth (Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice)
- ggplot2 (grafici di analisi esplorativa)
- syuzhet (Extracts sentiment and sentiment-derived plot arcs from text)
- fansi
- devtools
- httpuv
- rlang
- xts
- purrr

Riferimenti bibliografici

[1] NRC Lexicon:
`https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`.

[2] Sentix Lexicon:
`http://valeriobasile.github.io/twita/sentix.html`.

[3] OpeNer Sentiment Lexicon Italian:
`https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-73`

[4] Sentiment Analysis:
`https://en.wikipedia.org/wiki/Sentiment_analysis`

[5] Text Mining with R Julia Silge and David Robinson:
`https://www.tidytextmining.com/`

[6] Intro to rtweet:
`https://mkearney.github.io/blog/2017/06/01/intro-to-rtweet/`

[7] Premium API:
`https://twittercommunity.com/t/premium-apis-with-r/111592`