Analisi sulle elezioni del Presidente della Repubblica italiana nel 2022

tramite Twitter

Francesco Fustini[13] (matr. 830697) - Francesca Motta[23] (matr. 830107)

[1] *CdLM Data Science* [2] *CdLM Scienze Statistiche ed Economiche* [3] *Università degli Studi di Milano-Bicocca*

Febbraio 2022

Dopo sette anni dalla nomina di Sergio Mattarella a capo di stato nel 2015, il Parlamento torna in seduta comune per eleggere il Presidente della Repubblica. A causa della situazione pandemica, si è tenuta un'operazione di voto al giorno dal 24 al 27 gennaio 2022 mentre si sono tenute due operazioni di voto dal 28 al 29 gennaio 2022. Questo studio è concentrato sull'analisi della community di Twitter e dei loro sentiment riguardo le elezioni. Con l'attività di scraping, sono stati raccolti 94'944 tweet relativi agli hashtag in tendenza dal 23 al 31 gennaio inerenti alle elezioni. Dopo la fase di preprocessing basata su tecniche di Text Mining, è stata effettuata una Social Network Analysis basata su un grafo mezioni-retweet e si è valutata la Community Detection. Infine, sono stati analizzati i singoli tweet avendo come obiettivo la Sentiment Analysis.

1 Introduzione

Allo scadere del mandato presidenziale di Sergio Mattarella, il 24 gennaio 2022 sono cominciate le votazioni per eleggere il nuovo Presidente della Repubblica italiana. In questo lavoro si è voluto analizzare il dibattito sulle elezioni del capo di stato su Twitter e la relativa community.

2 Domande di ricerca

Le domande di ricerca prese in esame sono le seguenti:

1. Come è rappresentato il panorama politico italiano su Twitter?
2. Quali sono i personaggi più influenti all'interno del panorama politico italiano su Twitter?
3. Qual è stato il sentimento generale esternato su Twitter nel periodo di elezioni?
4. Quali sono state le diverse reazioni nel panorama politico rispetto alle elezioni del Presidente della Repubblica?

3 Ottenimento dati e creazione dataset

I dati sono stati scaricati dal social network Twitter[1] con la libreria di Python[2] Tweepy[3]. Essa consente di scaricare tweet risalenti fino a sette giorni precedenti al momento della richiesta. Il processo di scraping è stato inoltre automatizzato impostando un time out di 15 minuti tra una estrazione e l'altra affinchè si rispettasse il limite di tweet scaricati insieme previsto dall'API di Twitter[4].

I topic ricercati sono stati scelti esplorando gli hashtag di tendenza nei giorni delle elezioni e si sono scelti i seguenti: #Quirinale, #Quirinale2022, #PdR, #PresidenteDellaRepubblica e #PresidenzadellaRepubblica. Si è scelto di considerare anche qualche giorno precedente e posteriore ai giorni di votazioni per poter analizzare se il sentiment degli utenti fosse cambiato tra l'aspettativa e l'effettiva elezione del Presidente. Quindi il periodo

temporale d'interesse va dal 23 al 31 gennaio 2022.

Il risultato finale è un dataset composto da 94'944 tweet.

Giorno	tweet scaricati
23 feb	706
24 feb	7883
25 feb	12052
26 feb	5177
27 feb	14860
28 feb	12718
29 feb	25293
30 feb	10314
31 feb	5941

Tabella 1: Distribuzione dei tweet

Oltre alla data, le variabili del dataset sono: id del tweet, testo completo, numero di like e di retweet. Inoltre c'è una colonna dove vengono estratte le menzioni ed una per gli hashtag.

4 Pre-processing

Dopo l'eliminazioni di record duplicati, per ogni documento presente nel corpus sono state svolte le operazioni di preprocessing necessarie per analizzare il testo. In particolare:

1. Tokenization: ciascuna frase è stata tokenizzata in unigrammi
2. Rimozione tag HTML
3. Rimozione URL: ciascun collegamento ipertestuale è stato rimosso in quanto i link non sono significativi ai fini dei nostri obiettivi
4. Punctuation: sono stati rimossi i caratteri di punteggiatura
5. Rimozione Stopwords: alla lista di parole identificate come stopword nella libreria nltk in lingua italiana, sono stati aggiunti alcuni tra i token più frequenti non significativi secondo la nostra opinione
6. Rimozione delle parole con caratteri numerici
7. Rimozione menzioni: sono state rimosse per i fini di analisi testuale ma comunque conservate in una lista per la community detection

Non è stato necessario convertire il testo in lower case poiché il modello di sentiment utilizzato già gestisce questo aspetto. Per lo stesso motivo, i dati non hanno subito un processo di stemming o lemmatization.

Inoltre, tutte le osservazioni sono state raggruppate per giorno perdendo così l'informazione oraria non necessaria.

5 Social Network Analysis

Per la costruzione del social network sono state considerate le menzioni e i retweet. Infatti, ci si aspetta che per il dibattito politico su Twitter mezionare qualcuno nei propri tweet e retweettare il pensiero di un altro utente siano di fondamentale importanza per delineare il network.
Il grafo costruito è direzionato in quanto descrive il fenomeno di menzione o retweet che è direzionato per sua natura. Costruita la lista di autori-mezionati(o retweettati), grazie all'utilizzo del software Gephi[5] è stato ottenuto il grafo ed è stato visualizzato con il layout ForceAtlas 2.

Metrica	Valore
Average degree	1.789
Network diameter	28
Average clustering coefficient	0.012

Tabella 2: Metriche del grafo

Community Detection

Con l'obiettivo di analizzare la community d'interesse, è stato impiegato l'algoritmo gerarchico di Louvain, che massimizza la modularità entro i gruppi, cioè massimizza la densità di connessioni entro il cluster rispetto a quella verso l'esterno del cluster.
Sono risultati 736 gruppi con una modularità di 0.638. Osservando la distribuzione dei nodi nei gruppi, si nota che solamente un numero ridotto di cluster è popolato da un numero considerevole di unità. Molti dei cluster trovati hanno una dimensione non significativa e probabilmente includono utenti disposti agli estremi del grafo e poco centrali rispetto alla community.
Di seguito si mostra la rappresentazione grafica della Community Detection, in cui sono state evidenziate le sei community più popolose. La dimensione di ogni nodo è proporzionale al valore dell'in-degree.

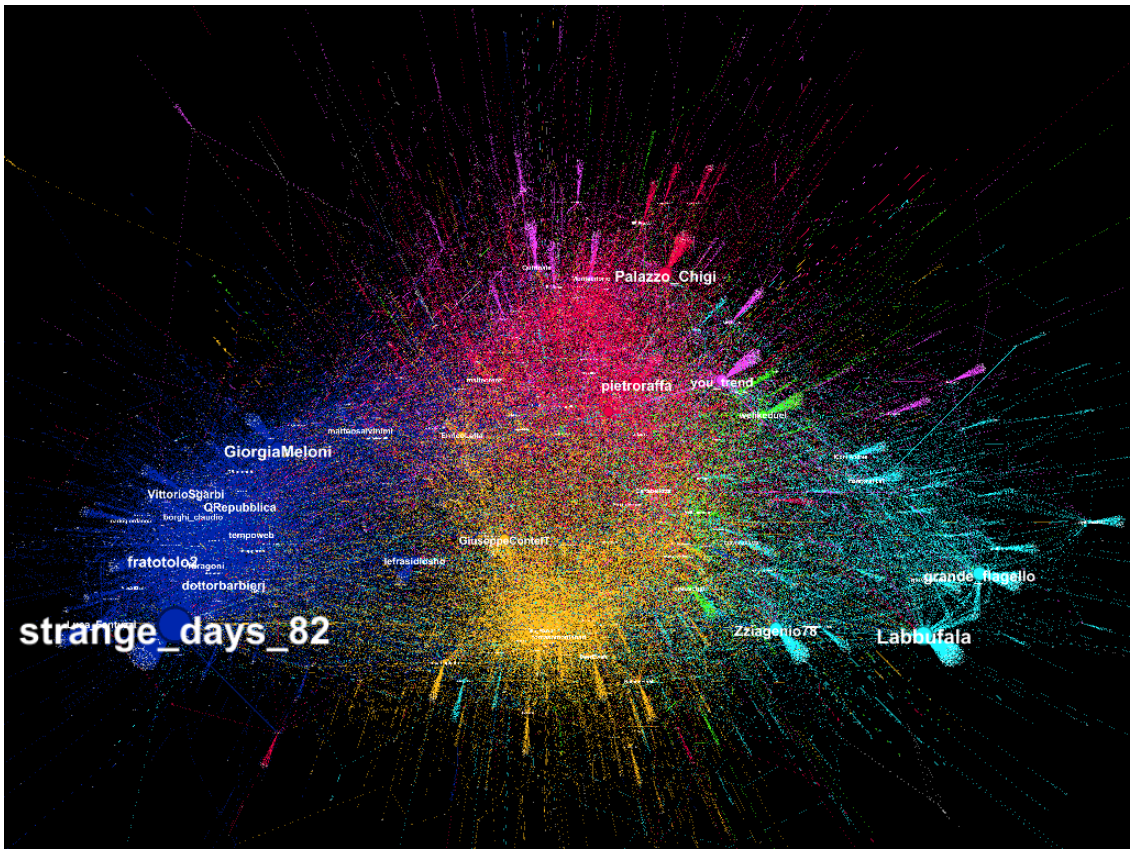

Figura 1: Community detection

Da un'analisi esplorativa congiunta all'analisi dell'in-degree degli utenti appartenenti ai diversi cluster, si è ottenuta una sintesi delle community e si è colto il senso interpretativo del clustering effettuato. Segue l'interpretazione dei sei cluster con dimensioni maggiori partendo dal più popoloso.

Community	Colore	Popolosità
1	Blue	7077
2	Azzurro	4523
3	Rosso	4519
4	Giallo	4448
5	Viola	2117
6	Verde	1194

Tabella 3: Metriche del grafo

- Il cluster 1 comprende gli utenti politicamente schierati a destra. Infatti, analizzando l'in-degree, si individuano gli influencer di questo cluster. Per citare i più importanti: l'opinionista *strange_days_82*, la politica *GiorgiaMeloni* e la giornalista di Il primato Nazionale *fratotolo2*.
- Nel cluster 2, i nodi considerati influencer del gruppo sono utenti di satira. I primi tre per in-degree risultano *Labbufala*, *grande_flagello* e *Zziagenio78*.
- Il cluster 3 include gli utenti neutri come quelli istituzionali e utenti di centro sinistra. I tre influencer più importanti sono l'account istituzionale *Palazzo_Chigi*, il giornalista dell'HuffPost *pietroraffa* e l'opinionista politico *lorepregliasco*.
- Il cluster 4 è composto dagli utenti vicini ai partiti di sinistra o dei 5 stelle. Gli influencer individuati sono il politico *GiuseppeConteIT*, il politico *EnricoLetta* e il rettore dell'università per stranieri di Siena *tomasomontanari*.
- Nel cluster 5 i nodi considerati come influencer sono profili di istituzioni e d'informazione tra cui il canale di news online *you_trend* ed i profili istituzionali *Quirinale* e *Montecitorio*. Come ci si aspetta e si nota anche dal grafo, questi sono utenti menzionati da tutti e quindi di riferimento per il topic analizzato.
- Nell'ultimo cluster preso in analisi, i nodi con in-degree più elevato sono profili d'informazione di sinistra come il profilo del programma Propaganda Live *welikeduel*, la giornalista *creuscher* ed il comico *M49liberorso*. Si consideri l'utente *welikeduel* di riferimento per questo gruppo. Infatti, rappresenta il nodo tra i 1194 presenti nel cluster con in-degree più alto,pari a 381. Dal grafo sottostante risulta evidente che questo nodo non è mai citato dagli utenti appartenenti al cluster 1 (ovvero quello dei personaggi vicini alla destra), a conferma dell'interpretazione data a questi cluster. Al contrario, si osservano diversi collegamenti con gli utenti dei cluster vicini alla sinistra, come ci si può aspettare.

La community detection ha rilevato un quadro generale del panorama politico italiano che rispecchia la situazione politica reale. Infatti, la destra è descritta da un unico cluster densamente collegato mentre la sinistra, seppur leggermente più popolosa, è divisa in più cluster. Effettivamente, nello scenario politico italiano di questi anni, più importanti partiti di destra sono più incisivi a livello mediatico rispetto ai partiti di sinistra e questo potrebbe aver portato ad una community sui social più coesa.
I risultati ottenuti dalla community detection sono coerenti con l'analisi dell'in-degree relativa a tutto il social network. Alcuni utenti di destra e di satira sono stati identificati come maggiori influencer del social network e sono infatti i nodi rilevanti per i due cluster più popolosi.
Inoltre, tra i nodi con in-dregree elevato, si osserva che quelli con betweenness centrality maggiore sono profili di informazione come programmi e giornali o direttamente giornalisti neutri sulla scena politica italiana. Il senso interpretativo è che il network abbia come nodi centrali canali d'informazione che risultano influencer del dibattito politico italiano su Twitter come *tempoweb*, *lorepregliasco*, *fdragoni* e *welikedue*.
Si nota anche un'elevata betweenness centrality per nodi con un alto livello di out-degree come *ErmannoKilgore* e la giornalista *DianaLanciotti*. Questi profili menzionano e retweettano post di utenti collocati in tutte le zone del grafo fungendo così da nodi centrali.

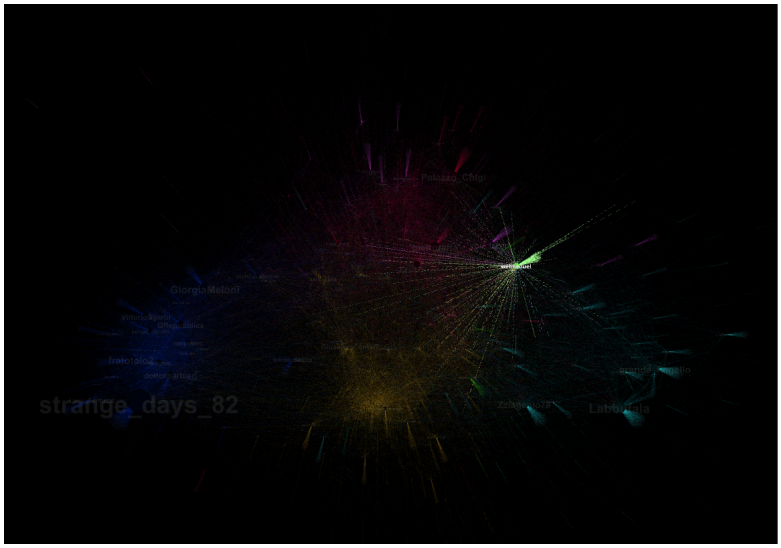

Figura 2: Nodo welikeduel

6 Sentiment Analysis

Il modello scelto è UmBERTo, versione italiana del BERT, ottimizzato poi su tweet italiani grazie al dataset FEEL-IT[6] e su opinioni nel campo semantico della lingua italiana relativo alla politica.
Il problema di classificazione è binario: infatti, il modello classifica i tweet in sentiment positivo o negativo.
I grafici realizzati per lo studio della Sentiment Analysis sono stati realizzati tramite il software Tableau[7] che è una piattaforma per l'analisi visiva. Il classificatore è stato utilizzato per analizzare i tweet: in generale, con riferimento temporale e congiuntamente alla cummunity detection eseguita.
Da una prima analisi generale sui tweet, si evince che il sentiment degli utenti di Twitter riguardo alle elezioni del Presidente della Repubblica del 2022 è prevalentemente negativo.

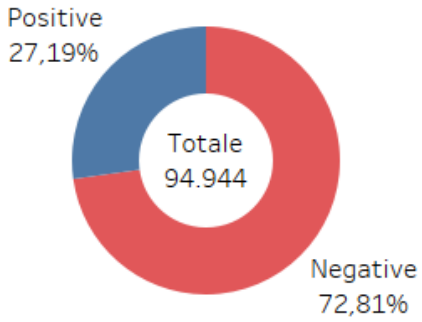

Figura 3: Sentiment generico

Considerando la serie storica degli utenti nei giorni dal 23 al 31 gennaio, non si notano grandi variazioni. Il trend di fondo rimane negativo dal giorno precedente alla prima votazione fino a due giorni dopo la rielezione di Sergio Mattarella.
C'è un leggero miglioramento nel giorno della rielezione di Mattarella a cui segue però una ricrescita significativa del sentiment negativo. Questa è un'informazione importante riguardo al periodo storico attuale della politica italiana. La rielezione del presidente uscente dovrebbe essere un'eccezione straordinaria e non la regola. Tuttavia, per la seconda volta consecutiva, così come accadde a Giorgio Napolitano, il Parlamento ha rieletto il Presidente uscente. Questo potrebbe essere uno dei motivi dei risultati negativi rilevato sui tweet. Un'altra causa che porta

nei primi giorni di elezioni a una percentuale molto alta di tweet negativi potrebbe essere il fatto molto criticato dall'opinione pubblica riguardo ai nomi inaspettati letti durante lo spoglio dei voti.

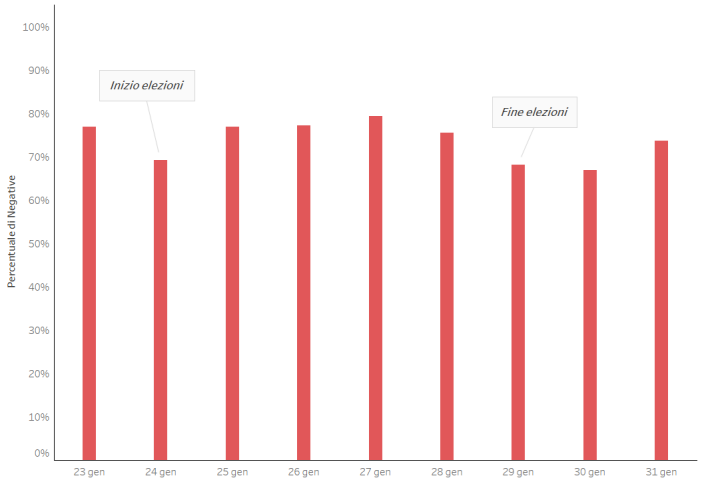

Figura 4: Serie storica

Come ultimo obiettivo, si è voluto analizzare e confrontare il sentiment nei diversi gruppi ottenuti nella community detection.

Figura 5: Sentiment per community

L'analisi differenziata per community riporta risultati simili per i gruppi politicamente schierati. Essi riportano infatti una sentiment molto negativa, mentre i gruppi costituiti da utenti più istituzionali o d'informazione riportano percentuali più equilibrate tra i tweet con sentiment opposti. Un'osservazione da farsi riguardo alla seconda community è che, essendo di satira, difficilmente il modello utilizzato potrà essere in grado di prevedere con accuratezza la sentiment.

7 Conclusioni e sviluppi futuri

Il lavoro svolto è riuscito a rispondere a tutte le domande di ricerca che ci si era posti. Inoltre, dall'analisi delle elezioni del Presidente della Repubblica è stata individuata la community di Twitter relativa al dibattito politico italiano. Questo progetto potrebbe risultare utile a un partito che volesse migliorare la propria comunicazione online poiché scoprirebbe quali sono gli influencer a cui fare riferimento. Uno sviluppo futuro potrebbe essere quello di ampliare la community detection all'intero quadro del dibattito politico italiano e non solo relativamente al topic delle elezioni del Presidente della Repubblica.

Bibliografia e risorse online

[1] *Twitter*. URL: https://twitter.com/.

[2] Guido Van Rossum e Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[3] Joshua Roesslein. ≪Tweepy: Twitter for Python!≫ In: *URL: https://github.com/tweepy/tweepy* (2020).

[4] *Twitter API*. URL: https://developer.twitter.com/en/portal/projects-and-apps.

[5] Mathieu Bastian, Sebastien Heymann e Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[6] Federico Bianchi, Debora Nozza e Dirk Hovy. ≪FEEL-IT: Emotion and Sentiment Classification for the Italian Language≫. In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: Association for Computational Linguistics, apr. 2021. URL: https://aclanthology.org/2021.wassa-1.8.

[7] *Tableau Software, Inc.* URL: https://www.tableau.com/.