

Italiani alle urne: come si schiera la comunità italiana di Twitter riguardo le #ElezioniPolitiche22

Andrea Maver 828725, Eleonora Palomba 876479, Oscar Zanotti 873763

9 settembre 2022

Indice

1	Introduzione	1
1.1	Aspetti metodologici	2
2	Ottenimento e struttura dei dati	2
2.1	Raccolta dei dati	2
2.2	Preprocessing	3
2.3	Dataset	4
3	Social network analysis	5
3.1	Metriche e analisi esplorativa	5
3.2	Visualizzazione del network	6
3.3	Analisi degli hub	6
3.4	Community detection	10
3.5	Valutazione semantica delle comunità	11
3.6	Profili BOT	11
4	Social Content analysis	12
4.1	Sentiment analysis	12
4.2	Emotion analysis	13
5	Conclusioni e sviluppi futuri	16

1 Introduzione

Dopo la caduta del governo Draghi, avvenuta il 21 luglio 2022 dopo le dimissioni dello stesso, si è aperto in Italia un periodo di incertezza politica che culminerà con le elezioni del 25 settembre 2022, nelle quali i cittadini saranno chiamati alle urne per decidere chi salirà al potere. L'attenzione dei media tradizionali come televisione e quotidiani ricade fortemente verso queste discussioni, in cui si dibattono ideali, nuove riforme e modalità per portare avanti il Paese come ognuno ritiene più corretto.

Anche per quanto riguarda i media sociali, in cui ogni utente è libero di esprimere la propria opinione, sono nate diverse piazze digitali in cui discutere e far valere le proprie idee, cercando seguaci e avversari con cui dibattere. Twitter rappresenta la piattaforma principale per questi scopi, in quanto permette di esprimersi con solo pochi caratteri e di interagire facilmente con gli altri utenti, siano essi d'accordo o meno.

Analizzando questi messaggi è possibile cercare delle comunità che condividono lo stesso pensiero e capire se i messaggi rivolti verso i maggiori personaggi di spicco siano favorevoli o critiche. Un altro tipo di analisi riguarda invece la parte più strutturale della comunità, in cui si cercheranno nodi che polarizzano l'attenzione e che risultano perciò particolarmente influenti, mentre altri assumeranno dei comportamenti gregari.

Le domande di ricerca alla base dello sviluppo del progetto sono state le seguenti: la divisione politica della società si rispecchia anche nell'utenza di Twitter? È possibile riconoscere delle comunità nette in base alle preferenze politiche? Quali sono i sentimenti dei messaggi che vengono rivolti ai

principali attori? C'è differenza tra le emozioni che suscitano i messaggi dei leader politici e quelli di chi invece ne parla?

1.1 Aspetti metodologici

Per lo svolgimento del progetto sono stati eseguiti eseguiti dei notebook Python tramite il sito Google Colab e si sono sfruttate librerie come Tweepy, NetworkX, Pyvis, Tree Tagger, Sentix e FEEL-IT.

La presentazione si articola nel seguente modo: come primo step viene esposta la fase di raccolta dei tweet; una volta eseguito del pre-processing dei dati si passa alla costruzione del grafo e all'analisi del network; successivamente si creeranno le comunità andandole a valutare da un punto di vista semantico e di contenuti; l'ultima analisi riguarda i sentimenti e le emozioni espresse sia dai messaggi di alcuni segretari di partito che degli utenti che li citano.

2 Ottenimento e struttura dei dati

In questa sezione verranno presentati il processo di raccolta dei dati e verrà descritto il dataset finale utilizzato per le analisi successive.

2.1 Raccolta dei dati

Per scaricare i tweet si è utilizzata la libreria Tweepy, che permette di interfacciarsi alle API fornite da Twitter direttamente da Python. A causa delle limitazioni imposte dalle API stesse non si è potuti andare particolarmente indietro nel tempo per cercare giorni con avvenimenti o dichiarazioni di spicco, ma la mole di tweet raccolti è stata comunque corposa. L'operazione di download è stata eseguita in due giorni diversi, cioè 1 e 3 settembre, in modo da aggirare parzialmente tali limitazioni e poter ottenere un gran numero di dati. È stato scelto per entrambi i giorni di limitare la ricerca a ventimila messaggi, per poi concatenare i due dataset e eliminare potenziali duplicati. La ricerca si basa su alcuni parametri che permettono di filtrare solamente i messaggi voluti; in particolare sono state utilizzate la lingua e alcuni hashtag.

Per quanto riguarda la lingua si è deciso di scaricare solo tweet in lingua italiana, questo perché, nonostante le elezioni italiane siano un argomento discusso anche dai media internazionali, per raccogliere le opinioni dei cittadini si è ritenuto più adeguato scegliere solo quelle nella lingua natale. Anche per le successive analisi del sentiment scegliere una sola lingua facilita le operazioni, in quanto questo tipo di ricerche vengono sviluppate con diverse eccezioni in base agli idiomi in cui vengono svolte.

Gli hashtag sono una componente fondamentale nella ricerca su Twitter, in quanto permettono di isolare solo messaggi riguardanti un particolare argomento. Questi infatti generano delle sorte di piazze virtuali all'interno delle quali ogni utente può condividere i propri pensieri e interagire con quelli degli altri; ne esistono su qualsiasi argomento, perciò sono un buon mezzo per poter scegliere solo i tweet necessari. In particolare, quelli scelti per il progetto sono stati i seguenti:

- #elezioniPolitiche22
- #elezioniPolitiche2022
- #elezioni2022
- #elezioni22
- #elezionipolitiche

Nonostante siano molto simili tra loro, è stato scelto di utilizzarli tutti perché, non essendoci uno standard e essendo tutte nate dall'utenza, è possibile che diversi individui si riferiscano allo stesso argomento in modo diverso. Inoltre, si è scelto di non utilizzare riferimenti a particolari leader politici per cercare se questi fossero comunque rintracciabili all'interno di una grande comunità più generica.

Un'ulteriore operazione eseguita in questa fase è la creazione di un secondo dataset che verrà utilizzato solamente per sentiment e emotion recognition. Questo si basa sulla ricerca degli ultimi duecento tweet dei principali esponenti politici, per valutare se le emozioni che veicolano sono le stesse di quelle dell'utenza. Per crearlo viene utilizzata un'altra funzione della libreria Tweepy, che permette

di cercare i post di un utente specificandone il nome. Osservando le distribuzioni si nota come questi duecento messaggi si distribuiscano in archi di tempo differenti, avendo come inizio sempre la stessa data ma andando indietro fino a metà giugno, luglio o addirittura fine agosto, a significare che alcuni esponenti utilizzano molto la piattaforma social.

A entrambi i dataset verranno applicate diverse operazioni di pre-processing, con delle leggere differenze in base al tipo analisi per i quali si vogliono utilizzare.

2.2 Preprocessing

Come primo step vengono create delle nuove colonne non presenti direttamente dalle API, cioè hashtag, menzioni e gli autori di un messaggio retweettato. Per fare ciò vengono utilizzate delle espressioni regolari che isolano e separano solo le parti di interesse.

Successivamente viene manipolato il testo del tweet, in modo da renderlo il più fruibile possibile per gli algoritmi. A tale scopo sono state effettuate una serie di operazioni, condotte principalmente attraverso l'utilizzo di regole regex, fra le quali:

- Rimozione delle emoji
- Rimozione delle menzioni e riferimenti a nomi di altri utenti (preceduti dal carattere @)
- Rimozione degli hashtag (preceduti dal carattere #)
- Rimozione degli url
- Rimozione di eventuali spazi extra
- Trasformazione del testo in lowercase
- Rimozione di simboli contenuti nel testo
- Normalizzazione dei caratteri accentati

Avendo ottenuto dei testi puliti si può procedere con le fasi successive del pre-processing, che verranno però effettuate su un subset del dataset originale.

Per limitare e rendere più significativo lo spazio di ricerca per quanto riguarda sentimenti e emozioni si è deciso di procedere nel seguente modo: sono stati cercati i sondaggi politici più recenti e si sono considerati i primi cinque partiti risultanti. In particolare si è fatto riferimento al poll del 5 settembre 2022 realizzato da SWG, una s.p.a. specializzata in indagini politiche e ricerche di mercato, per l'emittente televisivo La7. Da questo si scoprono i primi cinque schieramenti, cioè Fratelli d'Italia, Partito Democratico, Lega, Movimento 5 Stelle e Azione - Italia viva. Per l'analisi è stato scelto di considerare solo i tweet riguardanti i leader dei partiti elencati, in modo da poter scoprire come si divide l'opinione della piattaforma intorno ad essi.

Il primo step necessario è la creazione di diversi subset del dataset originale, che verranno poi concatenati per creare il nuovo insieme di dati da analizzare. Per farlo si utilizzano delle espressioni regolari che cercano il nome del leader di partito, sia all'interno del testo che all'interno della colonna rt_authors, in modo da isolare solo i messaggi che ne hanno parlato esplicitamente o che, invece, sono messaggi originali ricondivisi.

Le stringhe ricercate e i rispettivi partiti sono i seguenti:

1. Fratelli d'Italia: GiorgiaMeloni, Meloni, meloni
2. Partito Democratico: EnricoLetta, Letta, letta
3. Lega: matteosalvinimi, Salvini, salvini
4. Movimento 5 Stelle: GiuseppeConteIT, Conte, conte
5. Azione - Italia viva: CarloCalenda, Calenda, calenda

Una volta creati questi cinque dataset è possibile unirli per ottenere il dataset finale che verrà utilizzato per le analisi, composto inizialmente da 19174 righe (ridotte poi a 14829 in quanto, come verrà riportato in seguito, si è provveduto a rimuovere eventuali tweet postati da account falsi o "bot") e 19 colonne. Il numero di tweet per persona può naturalmente variare, ma questo verrà preso in considerazione nelle fasi successive e potrebbe essere sintomo di una maggiore o minore popolarità di un certo individuo.

A questo nuovo dataset vengono eseguite delle nuove operazioni di pre-processing necessarie per rendere più efficaci le analisi del sentiment; in particolare vengono applicate tokenizzazione e lemmatizzazione del testo. Per questo scopo è stato effettuato dapprima un tentativo con la libreria Spacy, ma non riuscendo a portare a risultati soddisfacenti, nonostante questa comprenda la lemmatizzazione in italiano, si è deciso di provare ad implementare la libreria Tree Tagger, che ha raggiunto performance migliori. In questo caso, la lemmatizzazione avviene tramite l'individuazione della Part of Speech (POS) dei termini all'interno di una frase, associando una funzione sintattica alle parole e riconducendole al relativo lemma.

Come ultima fase vengono rimosse le stopword, parole molto frequenti in una lingua che falserebbero i risultati di qualsiasi analisi se fossero lasciate invariate. Alle stopword italiane definite dalla libreria Nltk vengono aggiunti alcuni caratteri speciali che potrebbero comparire all'interno dei tweet.

2.3 Dataset

Il dataset finale, che comprende tutti i tweet raccolti tra l'uno e il tre settembre, è composto da 35714 righe e 16 colonne e verrà utilizzato per la network analysis, è composto dai seguenti attributi:

- date: giorno e ora in cui è stato pubblicato il post
- id: codice identificativo univoco del tweet
- text: testo del tweet; nel caso sia un retweet presenta il testo del tweet originale
- n_rt: numero di retweet ricevuti dal tweet
- n_like: numero di like riscossi dal tweet
- author_name: nome dell'utente che ha pubblicato il messaggio; se fosse un retweet riporterebbe il nome di colui che sta eseguendo questa operazione, non quello dell'autore originale
- author_id: codice identificativo univoco dell'utente
- location: posizione geografica di dove è stato pubblicato il tweet; non sempre disponibile, dipende se un utente l'abbia specificato o abbia attiva la geo-localizzazione
- rt_authors: lista contenente i nomi degli utenti retweettati e/o menzionati nel post; se il tweet è originale sarà vuota. Ha la struttura di un dizionario contenente come chiavi screen name, name, id e id_str
- hashtag: lista degli hashtag nel tweet; possono esserne presenti anche alcuni diversi da quelli specificati sopra, ma sempre in combinazione con essi
- author_follower: numero totale di seguaci dell'autore del post
- author_friends: numero di persone che l'autore segue
- rt_names: nomi degli utenti retweettati; viene ricavato dal contenuto della colonna rt_authors
- hashtag: lista degli hashtag contenuti nel tweet
- mention: lista degli utenti menzionati nel tweet
- text_clean: testo del tweet processato

Al suo interno sono presenti anche i retweet, che saranno fondamentali per rappresentare le relazioni tra utenti e per creare il grafo.

Viene creato il subset descritto in precedenza con solo le menzioni verso i leader politici di 14829 righe. Vengono aggiunte anche tre colonne utili per i task svolti su questo particolare dataset, cioè cluster, tokenized_text e sentiment_text. La prima è un'indicazione della persona a cui le menzioni appartengono, perciò assume come valore il nome dei diversi leader; la seconda rappresenta il testo dopo aver applicato la tokenizzazione con Tree Tagger, quindi una lista di token lemmatizzati; l'ultima è il testo pulito e ricostruito unendo i vari token ottenuti precedentemente.

Si ricorda che è presente anche un terzo dataset, cioè quello composto dai 1000 tweet pubblicati dai vari segretari di partito, che verrà utilizzato come confronto con quello appena descritto. Per questo motivo subisce le stesse fasi di pre-processing, in modo da avere dati strutturalmente consistenti.

3 Social network analysis

Per eseguire questo tipo di analisi è prima di tutto necessario creare un network di interazione tra gli utenti; questo viene sviluppato utilizzando i retweet all'interno del dataset. Nel caso in cui un utente abbia retweettato il post di un altro, tra i nodi rappresentanti i due individui viene creato un arco direzionato dal primo verso il secondo; a questo link viene assegnato un peso in base a quante volte l'utente ha citato i messaggi dell'altro, essendo possibile che questa operazione sia stata eseguita per diversi post. Si creano nodi anche per gli utenti che hanno scritto post originali, senza necessariamente menzionarne di altri. Si nota che, nonostante retweet, menzione e citazione rappresentino tre operazioni leggermente diverse all'interno della piattaforma di Twitter, all'interno del report verranno utilizzati come sinonimi, in quanto l'unico tipo di relazione studiata è stata quella del retweet. Il grafo risultante è pesato e direzionato, e composto da 12947 nodi e 27097 archi;

3.1 Metriche e analisi esplorativa

Analizzando la struttura del grafo attraverso alcune dimensioni, mostrate in tabella 1, se ne possono studiare le caratteristiche.

Metrica	Valore
Grado medio	2.809
Densità	0.000162
Connettività nodi	0
Connettività archi	0
Assortatività	-0.0616
Reciprocità generale	0.01705
Componenti fortemente connesse	12678

Tabella 1: Metriche di network analysis

Osservando il valore di connettività si deduce che il network è tendenzialmente neutro, ma leggermente disassortativo, infatti il valore è negativo ma molto vicino a zero; questo significa che i nodi si legano tra loro con una distribuzione randomica, senza particolari pattern tra gli hub, i nodi con un maggior numero di connessioni. Dal basso valore di densità si evince che il grafo è molto sparso, confermato anche dalla reciprocità, che indica che gli utenti non interagiscono particolarmente tra loro. La connettività, sia di archi che di nodi, assume valore zero, portando alla conclusione che il grafo sia disconnesso. Inoltre, dal grado medio si può osservare che ogni utente ha effettuato in media quasi tre retweet nell'arco dei giorni presi in considerazione.

La rete può essere definita di tipo ego-centrico, tipologia in cui gli hub riuniscono gran parte delle relazioni, in questo caso i retweet degli utenti, dando vita al fenomeno del 'preferential attachment'.

Si procede con un ulteriore analisi esplorativa dei dati della rete, in particolare la distribuzione di tweet e retweet da parte degli utenti. In particolare, dalla figura 1 si nota che la maggior parte degli utenti non ha ricevuto alcun retweet e per quelli che ne hanno ricevuto uno il numero è circa un quinto del precedente. Anche aumentando la granulometria dell'analisi, cercando perciò dei gruppi più ampi per valori maggiori di connessioni (figura 2), si osserva come pochissimi nodi superino la soglia dei 120

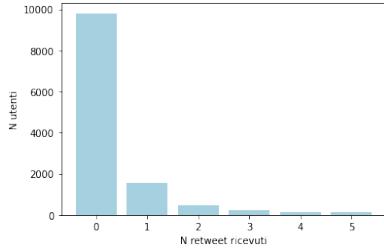


Figura 1: Retweet ricevuti

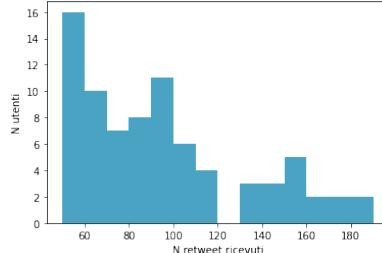


Figura 2: Retweet ricevuti

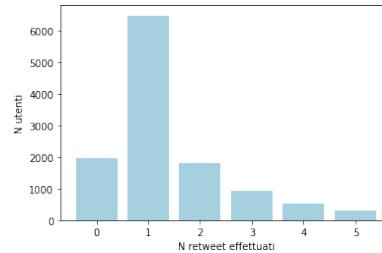


Figura 3: Retweet effettuati

retweet ricevuti. Osservando infine il numero di retweet effettuati, figura 3, si osserva come la maggior parte degli utenti ne abbia effettuato uno e come il numero decresca all'aumentare dei re-post.

3.2 Visualizzazione del network

Una volta effettuate queste prime analisi si può procedere con la visualizzazione della rete; ne sono state effettuate due varianti: la prima, più semplice, utilizzando la libreria NetworkX, che permette di dare un veloce sguardo ai principali attori del grafo; la seconda, creata con Pyvis, fornisce un risultato più definito e permette di interagire con i diversi nodi. Questa sarà anche la base sulla quale verranno visualizzate le comunità nella fase di community detection.

Partendo dalla prima versione in figura 4, si è scelto di ingrandire le dimensioni dei nodi con $in_degree \geq 150$, che rappresentano gli utenti che hanno ricevuto più di centocinquanta retweet da utenti diversi; inoltre sono stati colorati di blu e se ne mostra l'etichetta, cioè il nome utente. Il valore è stato scelto dopo aver eseguito diversi tentativi per cercare di mostrare nodi significativi senza rendere eccessivamente affollata la visualizzazione. I nodi colorati di verde invece non hanno effettuato alcun tweet nell'arco di tempo preso in considerazione, ma sono stati menzionati da altri utenti, perciò risultano all'interno della rete come creatori di post originali.

Da questo si possono trarre alcune conclusioni: è possibile notare come siano presenti i maggiori esponenti dei diversi partiti politici in gara per le elezioni, tra cui Giorgia Meloni, Enrico Letta, Giuseppe Conte e Carlo Calenda. Questi sono tutti presenti come nodi di colore verde a significare che non hanno twettato direttamente nei giorni analizzati (o non l'hanno fatto utilizzando gli hashtag scelti), ma che nonostante questo siano stati molto citati dall'utenza nei giorni successivi. Si nota anche la presenza di diversi account di rappresentanza dei partiti stessi, che siano ufficiali o gestiti da sostenitori. Per quanto riguarda i nodi colorati in blu invece non si osservano nomi noti al grande pubblico, ma sui quali verrà comunque effettuata un'analisi manuale alla ricerca di ruoli particolari.

Il grafo mostrato con Pyvis (figura 5) permette, oltre ad avere una struttura più definita, di confermare alcune deduzioni fatte durante le analisi delle metriche. Si osserva infatti che il grafo è disconnesso, in quanto sono presenti diversi nodi lungo il bordo esterno senza alcun collegamento. Inoltre, si possono iniziare a visualizzare delle sorte di comunità che vengono a crearsi intorno agli hub, con delle particolari strutture a mongolfiera, nonostante la parte centrale sia particolarmente affollata. Osservando nuovamente i nodi con più connessioni si notano corrispondenze con la visualizzazione precedente, per cui spiccano i nodi riferiti ai vari leader dei partiti.

3.3 Analisi degli hub

Si prosegue con l'analisi dei principali hub trovati all'interno della rete. In tabella 2 si possono osservare gli account presenti nel grafo con più follower, tra quelli che hanno postato almeno un tweet; si nota la presenza di diverse testate giornalistiche, oltre a qualche famoso giornalista. Non sono presenti i leader politici perché, come detto, non hanno pubblicato post con gli hashtag ricercati, perciò non è possibile calcolarne il numero di seguaci.

In tabella 3 invece si considerano solo i nodi etichettati all'interno del grafo, ovvero quelli con più di centocinquanta archi entranti, e se ne mostrano i follower; questi rappresentano gli utenti con più collegamenti e quindi alcuni degli hub principali della rete. Si nota come alcuni siano gli stessi della tabella precedente, come Enrico Letta e Maria Elena Boschi (meb), come compaiano alcuni dei

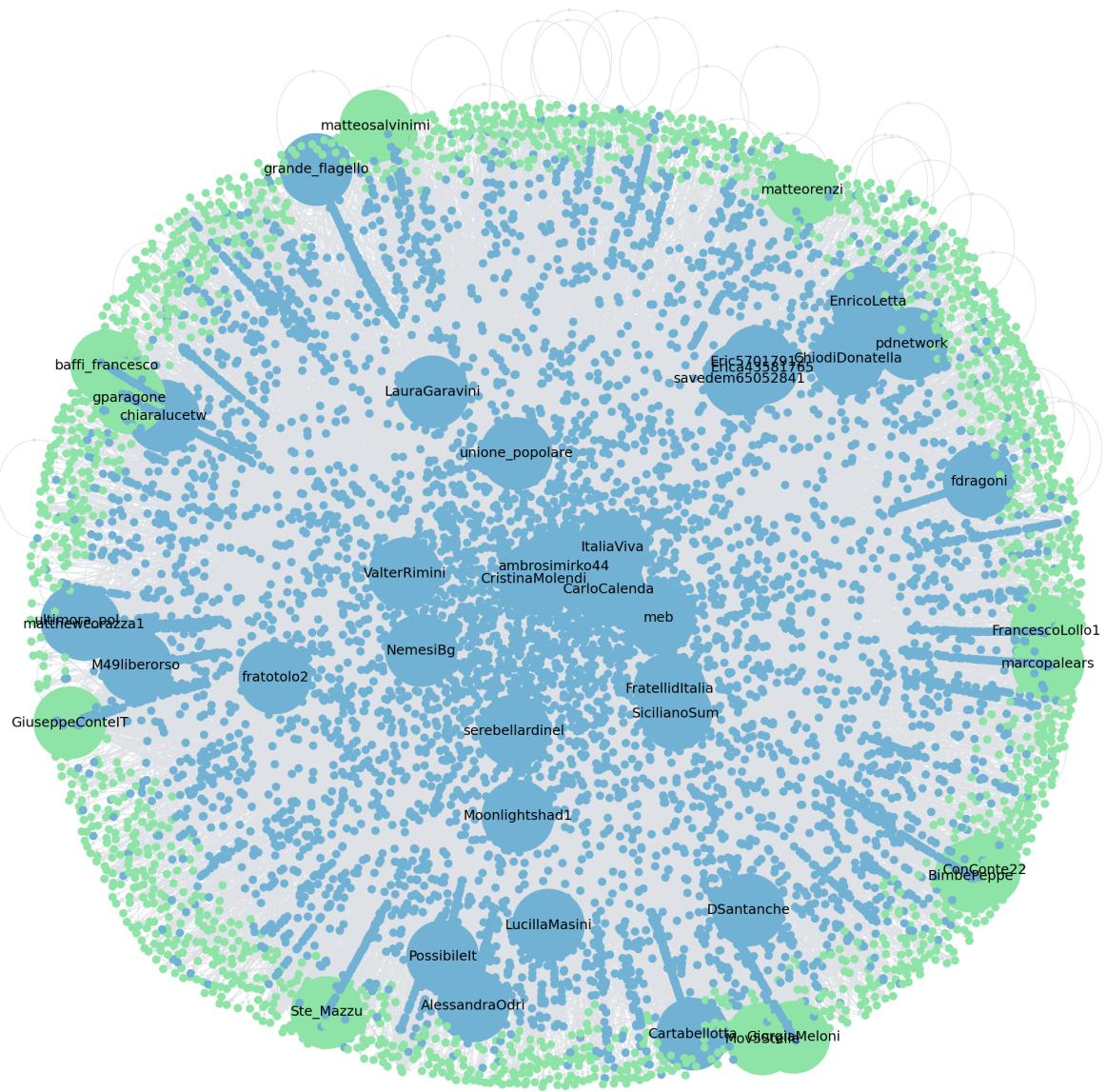


Figura 4: Visualizzazione del network con NetworkX

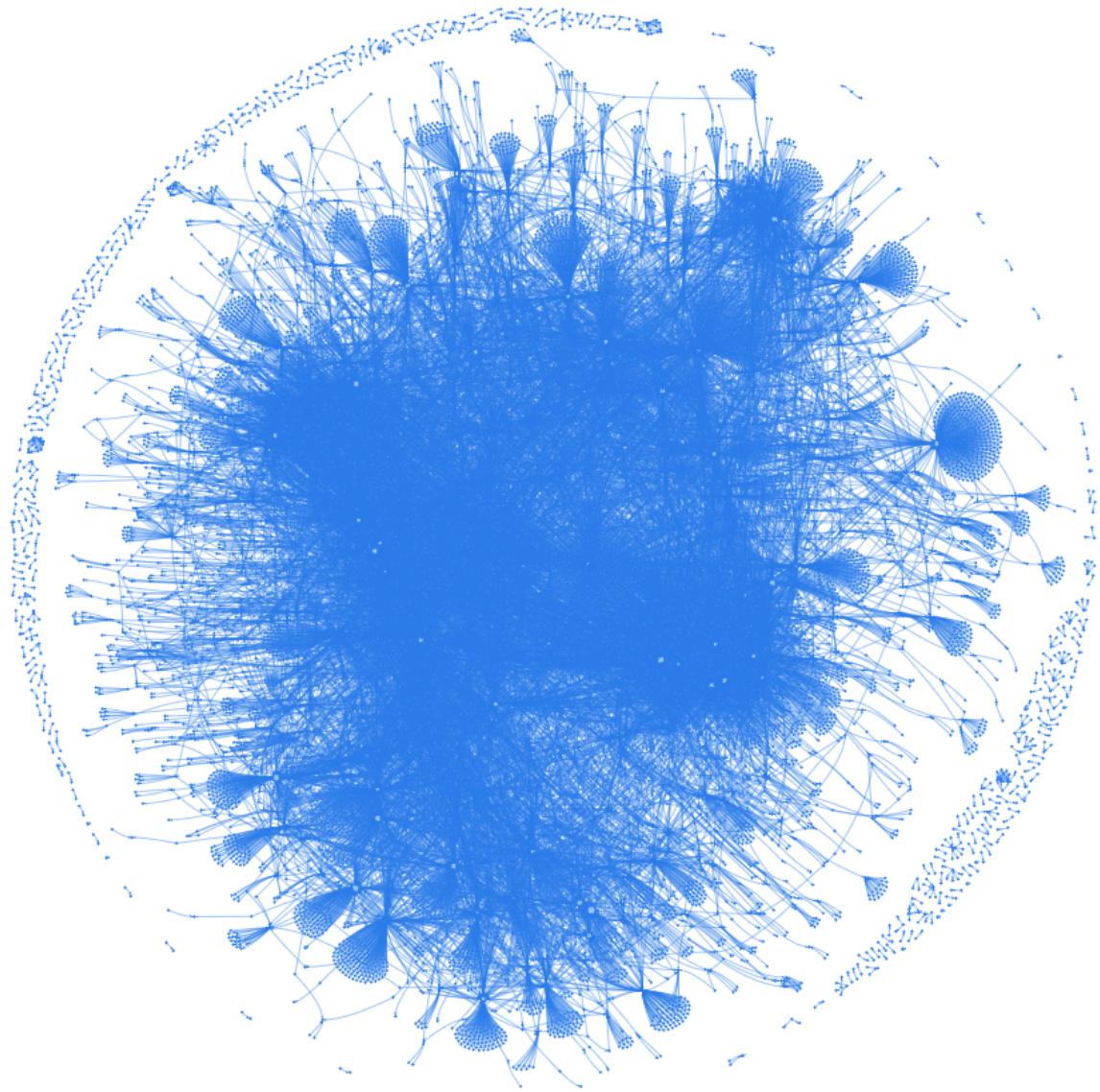


Figura 5: Visualizzazione del network con NetworkX

author_name	author_follower
SkyTG24	3'618'782
TwitterGov	2'371'841
fattoquotidiano	2'179'848
sole24ore	1'741'955
marcotravaglio	1'652'386
MediasetTgcom24	1'407'494
TgLa7	770'899
EnricoLetta	716'015
ilpost	713'621
meb	654'442
Adnkronos	593'177
La7tv	589'914
ilgiornale	567'333
Roma	506'153
civati	408'480

Tabella 2: Top 15 account più seguiti

author_name	author_follower
EnricoLetta	716'015
meb	654'442
pdnetwork	400'377
CarloCalenda	396597
DSantanche	191'623
FratellidItalia	182'065
gparagone	172'004
Cartabellotta	76'223
PossibileIt	58'214
fdragoni	54'974
ItaliaViva	50'143
grande_flagello	25'756
M49liberorso	25'744
Moonlightshad1	21'249
ChiodiDonatella	20'074

Tabella 3: Top 15 hub più seguiti

deputati e deputate dei diversi partiti, come Daniela Santanchè e Gianluigi Paragone, e altri account umoristici di satira.

Per concludere l’analisi, prima di continuare con la community detection, si effettua un calcolo della *centrality*, in particolare *closeness* e *betweenness*. Queste misure permettono di trovare i nodi che assumono un importante ruolo strutturale all’interno della rete: la prima fornisce informazioni su quali siano gli attori che risultano più centrali, quindi con la distanza minore dal resto dei nodi; la seconda invece mostra quali utenti, se ne sono presenti, sono utili per collegare diverse parti del grafo e mettere potenzialmente in contatto diverse comunità. I risultati per entrambe sono mostrati in tabella 4.

Considerando la *closeness*, non si notano valori particolarmente alti, anche se risultano presenti i leader politici dei due principali schieramenti, con il loro rispettivo partito, cioè Giorgia Meloni con Fratelli d’Italia e Enrico Letta con il Partito Democratico. Per ognuno di essi sono anche presenti alcuni deputati, cioè Roberto Menia per i primi e Peppe Provenzano per i secondi. I rimanenti nodi sono di siti di informazione, come Ultima ora, Il giornale e Dire.it.

Osservando invece i valori di *betweenness* risultano tutti molto bassi, ad indicare che non sono presenti nodi che mettono in contatto diverse parti del grafo quasi esclusivamente. Anche il fatto che non siano presenti attori osservati precedentemente è un buon risultato. Tra gli utenti trovati non sembrano esserci nomi di particolare spicco tranne Luigi de Magistris, deputato di Unione Popolare. Andando a osservare i profili Twitter si nota però un pattern interessante: tutti questi account sono evidenti sostenitori del Movimento 5 Stelle o di Unione Popolare; il primo è un partito centrista molto in voga nei poll, che anche nel grafo sembra quasi posizionarsi tra i due schieramenti di destra e sinistra mettendo in comunicazione le diverse aree della rete.

closeness centrality	betweenness centrality
ultimora_pol	0.127
GiorgiaMeloni	0.0838
pdnetwork	0.0747
peppaprovenzano	0.0600
EnricoLetta	0.0586
FratellidItalia	0.0575
ilgiornale	0.0511
direpuntoit	0.0511
robertomenia	0.0504
DSantanche	0.0493
ambrosimirko44	0.00316
Yoda15271485	0.00176
Rapace781	0.00169
Finiguerra	0.00159
romano_tano	0.00156
BuselliGiulia	0.00154
UPrieste	0.00152
demagistris	0.00119
Marzo531	0.00110
radionowher	0.00109

Tabella 4: Top 10 closeness e betweenness centrality

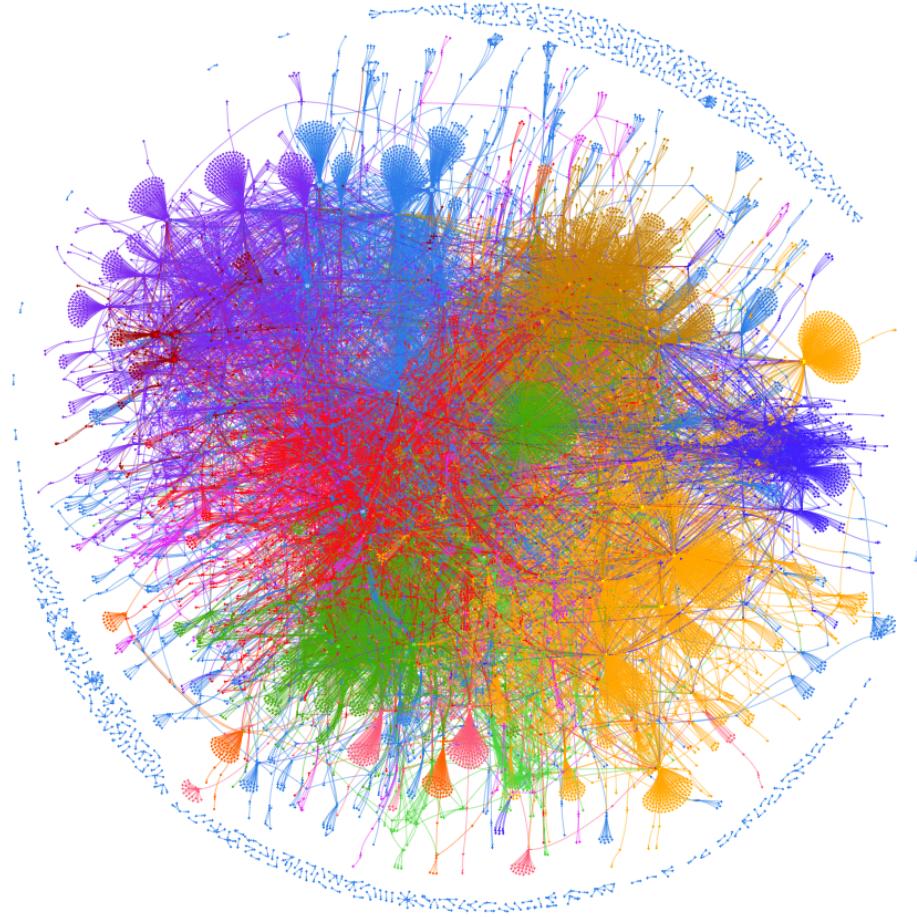


Figura 6: Visualizzazione delle comunità trovate

3.4 Community detection

Lo step successivo nell'analisi della rete riguarda la ricerca di comunità di utenti che interagiscono tra loro e posseggono dei legami particolari. Un social media come Twitter permette di comunicare con altri individui che condividono gli stessi interessi, andando così a creare gruppi di persone che posseggono tratti simili. Diversi algoritmi permettono di cercare e trovare queste comunità all'interno delle reti, ottimizzando diversi parametri, a livello di nodi o di network.

La modalità scelta per il progetto è stata la massimizzazione della modularità, una tipologia *network centric* di community detection. La modularità è la differenza tra il numero potenziale e quello effettivo di archi all'interno di un gruppo; massimizzare questo valore significa cercare cluster che abbiano il maggior numero possibile di archi al loro interno, ma che non comunichino particolarmente con gli altri. Può assumere valori compresi tra -1 e 1; naturalmente, un risultato vicino a 1 porta a dedurre che le comunità trovate sono significative e ben legate al loro interno.

In figura 6 si può osservare il risultato dell'applicazione dell'algoritmo, con diverse colorazioni per ognuna delle comunità con più di 150 elementi; in particolare, sono state identificate 351 comunità con un valore di modularità medio di 0.712. Si nota subito che in ognuna di esse sia presente uno degli hub trovati precedentemente, siano essi persone famose, siti d'informazione o utenti influenti, ad indicare nuovamente come siano fondamentali per la diffusione delle informazioni all'interno delle comunità.

Andando leggermente più nel dettaglio, anche se la navigazione del grafo è difficoltosa, si osserva che Giorgia Meloni, Fratelli d'Italia e Libero sono stati inseriti nella stessa comunità. Questo sembra un buon risultato, chiaramente per i primi due, e potrebbe significare che la testata giornalistica sia schierata verso il centro destra.

3.5 Valutazione semantica delle comunità

Per poter valutare la bontà degli algoritmi e dei risultati della community detection si possono utilizzare diversi metodi in base ai dati che si hanno a disposizione. Non avendo una ground-truth con la quale confrontare i cluster ottenuti è necessario sfruttare degli strumenti diversi: analizzando il contenuto semantico di ogni gruppo è possibile scoprire se al suo interno ci siano dei termini ricorrenti, che possono perciò permettere di identificare le idee e particolari inclinazioni dello stesso. Per fare ciò sono state prodotte delle wordcloud per le dieci comunità più numerose.



Figura 7: Wordcloud delle 10 comunità più numerose

Osservando i risultati si notano alcuni cluster con parole frequenti e particolarmente significative. Il gruppo 0 sembra essere focalizzato sulla figura di Giorgia Meloni, anche se dalla visualizzazione non si può evincere se l'opinione sia positiva o meno; al suo interno compare anche Silvio Berlusconi, per cui valgono le stesse considerazioni. Il cluster 2 si riferisce al terzo polo e i nomi Matteo e Carlo sono molto probabilmente quelli di Renzi e Calenda, i principali leader del movimento. I gruppi 3 e 6 sembrano riferirsi al Partito Democratico, anche se composti da termini che spaziano ampiamente nell'arco dei vari temi; da notare come in nessun gruppo compaia Enrico Letta, leader del partito. Il cluster 5 sembra avvicinarsi ideologicamente al Movimento 5 stelle e a Giuseppe Conte. Per i restanti gruppi non sembrano apparire particolari pattern.

Di seguito sono mostrati alcuni esempi di tweet contenuti nei cluster appena descritti:

- Cluster 0: "Giorgia #Meloni sulla sua eventuale ascesa all'ufficio di Presidente del Consiglio: "Ammetto che mi tremano i polsi all'idea" - Sapessi a noi Giorgia"
- Cluster 2: "Di #TerzoPolo si parlava anche nel 2010-2013 con Fini, Casini e successivamente anche Monti. Com'è finita? Che poi sono andati al governo col PD. Cope per calendiani e renziani. #ElezioniPolitiche22 #ItaliaSulSerio"
- Cluster 3: "Bisogna combattere il pensiero unico del PD e della pseudosinistra. Qualcuno chiama ancora il PD di sinistra. Ma il PD è principale azionista del governo Draghi, ha fatto macellerie sociali, ha privatizzato tutto. Che cosa ha di sinistra? #UnionePopolare #ElezioniPolitiche22"
- Cluster 5: "#Conte è l'unico che può guardare dritti negli occhi i suoi avversari politici perché i suoi comportamenti sono sempre stati coerenti e rispettosi e alle parole alle promesse sono sempre seguiti i fatti. #IoVotoM5sConte #IoVotoM5S #DallaParteGiusta #M5S #ElezioniPolitiche2022"

3.6 Profili BOT

In sede di esplorazione manuale dei tweet postati dagli utenti ne sono emersi alcuni che hanno richiamato l'attenzione, come ad esempio:

- "GRANDE RESET, SE LI RIVOTI, CONTINUERANNO L'AGENDA. #USA #NATO #UE #SANZIONI #NAZISTI #UCRAINA #GAS #ENERGIA #PartitoDemocratico #PD #EnricoLetta #CinqueStelle #Salvini #Lega #GiorgiaMeloni #Meloni #FratellidItalia #ElezioniPolitiche2022 #GreenPass #Vaccino #ZELENSKY"

- ”SE LI RIVOTI, LO FARANNO ANCORA. #PartitoDemocratico #PD #EnricoLetta #Meloni #FratellidItalia #GiorgiaMeloni #Salvini #Lega #CinqueStelle #Conte #Grillo #ElezioniPolitiche2022 #GreenPass #Vaccino #Pfizer #Moderna #QuartaDose #TerzaDose #NessunaCorrelazione #Covid_19 #Vita”

Si è notata infatti la presenza di autori che tendevano a postare ripetutamente lo stesso testo più e più volte. Ipotizzando che questi profili possano essere riconducibili a bot, per cui non appartenenti ad utenti reali ma creati probabilmente al solo scopo di sfruttare la popolarità degli hashtag citati per pubblicizzare un qualche slogan ”rivoluzionario”, si è provveduto a ricercare all’interno del grafo le cricche contenenti questi utenti. Per fare ciò è stato necessario trasformare il grafo da diretto a indiretto, in modo da poter utilizzare la funzione `EnumerateAllCliques` fornita dalla libreria NetworkX. Notando la presenza di numerosi utenti ”sospetti” che tendono a retweettarsi a vicenda, si è scelto di rimuoverli dai dataset utilizzati: includere anche questi tweet all’interno delle analisi avrebbe infatti rischiato di comprometterne i risultati.

4 Social Content analysis

In questa sezione verranno analizzati i contenuti dei tweet raccolti, in particolare verranno effettuate sentiment e emotion analysis utilizzando i dataset creati durante le fasi di pre-processing. Verrà eseguito un confronto tra quello che l’utenza pensa dei principali leader e quello che loro pensano o vogliono veicolare attraverso la propria comunicazione. Le analisi verranno portate avanti nello stesso modo e alla fine se ne confronteranno i risultati.

4.1 Sentiment analysis

L’analisi del sentiment degli utenti, come accennato in apertura, è finalizzata all’identificazione e all’estrazione delle opinioni riportate all’interno di un testo. Nel caso in esame, essa è stata eseguita utilizzando un approccio di tipo lexicon-based, dopo aver ricondotto le parole contenute nei tweet ai loro lemmi originari, per poi associarvi un punteggio di sentiment tramite un dizionario (lexicon). Tale metodologia utilizza i testi tokenizzati e lemmatizzati, e ha seguito alcuni passaggi:

- Gestione del lessico italiano tramite Sentix: si tratta di un lexicon che fornisce punteggi di sentiment analysis per lemmi in lingua italiana. In questo modo è possibile salvare all’interno di un dizionario la lista dei lemmi e la relativa polarity, espressa con valori da -1 (negativo) a +1 (positivo)
- Calcolo del sentiment score tramite VADER: VADER è un tool che permette allo stesso tempo di analizzare la polarità (positivo/negativo) e l’intensità delle emozioni. Quest’ultima viene colta analizzando l’enfasi con cui le parole vengono riportate: in questo modo, ad esempio, verrà attribuito un peso maggiore ai concetti accompagnati da punteggiatura (es. ”no!”). La scelta di adottare di un simile strumento è stata dettata dal fatto che il tipo di linguaggio dei testi analizzati si presta particolarmente a questo tipo di trattamento, nonché dalla necessità di integrare diversi strumenti per riuscire ad ottenere dei risultati soddisfacenti per lemmi in lingua italiana. VADER viene aggiornato con il dizionario ottenuto da Sentix, e attraverso l’unione dei diversi elementi contenuti nella stringa di testo viene fornito un compound score normalizzato, ovvero un valore di sentiment complessivo compreso tra -1 ed 1. Questo permette di classificare un tweet come neutro se il suo valore di compound risulta essere compreso tra -0.05 e 0.05, positivo se maggiore di 0.05, e negativo se inferiore a -0.05.

I risultati dell’analisi di sentiment sono mostrati nelle figure 8 e 9, che mettono a confronto i tweet dell’utenza (a sinistra) e quelli postati dai leader (a destra).

Dalle figure si nota che i diversi sentimenti si distribuiscono similmente tra i vari attori considerati; Meloni, Salvini e Calenda risultano avere la più alta percentuale di positivi, sia per quanto riguarda l’utenza che per i propri messaggi. Conte appare essere il leader percepito in modo peggiore, nonostante i suoi tweet non lo risultino particolarmente. Un fatto interessante è che per quanto riguarda i messaggi pubblicati dai candidati, la percentuale di quelli classificati come neutri è molto minore rispetto a quelli dell’utenza. Questo potrebbe essere dato dal fatto che nei tweet degli utenti sono presenti anche quelli



Figura 8: Sentimenti dell'utenza

Figura 9: Sentimenti dei leader

di testate giornalistiche che riportano fatti e notizie accadute, mentre i post dei leader cercano di essere più polarizzanti, cercando di influenzare positivamente i loro seguaci o invece attaccando gli avversari.

Vengono mostrati alcuni tweet riferiti a ogni sentimento:

- Negativo: ”Oggi come nel 2018 l’unico argine al populismo deteriore di #Salvini e al sovranismo della #Meloni è il #M5S. Stare #DallaParteGiusta insieme a #Conte vuol dire fermare la restaurazione che #Renzi e il #PD ha avviato portando #Draghi al Governo. #IoVotoM5S #ElezioniPolitiche2022”
- Neutro: ”Tensione in apertura al comizio di Giorgia #Meloni a #Cagliari, quando un attivista ha fatto irruzione del palco della leader di #Fdi impugnando una bandiera del movimento #Lgbtqplus #2settembre #elezioni2022”
- Positivo: ”#Conte è l’unico che può guardare dritti negli occhi i suoi avversari politici perché i suoi comportamenti sono sempre stati coerenti e rispettosi e alle parole alle promesse sono sempre seguiti i fatti. #IoVotoM5SconConte #IoVotoM5S #DallaParteGiusta #M5S #ElezioniPolitiche2022”

4.2 Emotion analysis

Per eseguire questo tipo di analisi viene utilizzata la libreria FEEL-IT, che contiene un modello nato nel 2021 appositamente per assegnare a tweet in lingua italiana quattro emozioni elementari, cioè rabbia, gioia, paura e tristezza. Queste quattro fanno parte delle otto emozioni fondamentali della ruota di Plutchik (figura 10), e rappresentano i punti cardinali da cui poi vengono derivate le altre come loro combinazione.

Il modello è stato creato facendo fine-tuning di UmBERTO, una versione italiana di BERT, famoso modello per svolgere task di NLP. FEEL-IT è stato allenato su un dataset di circa duemila tweet raccolti in giornate diverse in base agli hashtag di tendenza, in modo da coprire argomenti differenti,

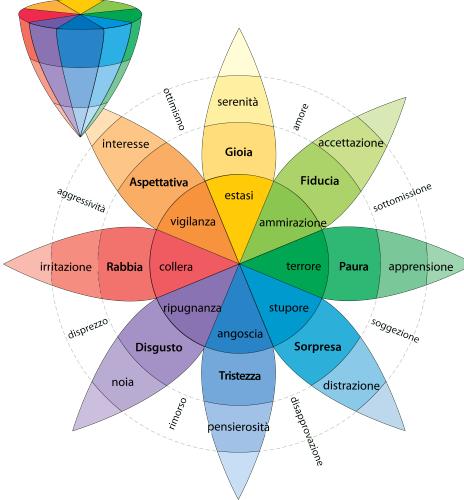


Figura 10: Ruota delle emozioni di Plutchik

poi etichettati manualmente dai ricercatori. Tra gli esempi forniti si notano argomenti come politica, la pandemia del 2019, sport e musica. Permette di fare anche sentiment analysis, ma solo con risultato positivo e negativo; per questo è stato scelto di utilizzare le tecniche esposte in precedenza che permettono di avere anche risultati neutrali.

Una volta eseguito il modello di classificazione sui diversi messaggi si ottiene l'emozione corrispondente; applicando dei raggruppamenti per i diversi cluster si possono ottenere dei valori normalizzati per la dimensionalità di ogni cluster, in modo tale da evitare sproporzioni legate alla quantità di tweet trovati.

Di seguito si possono osservare i risultati, che mettono a confronto le emozioni provate dall'utenza e quelle che invece i leader vogliono far passare (figure 11 e 12).

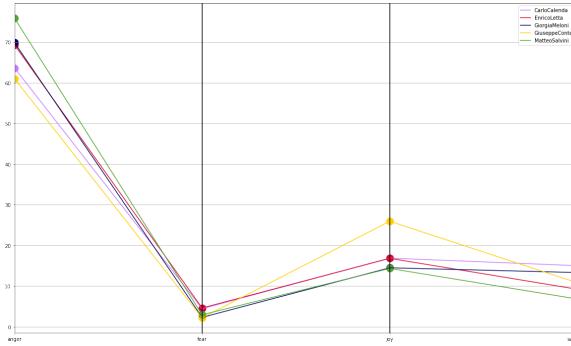


Figura 11: Emozioni percepite dall'utenza

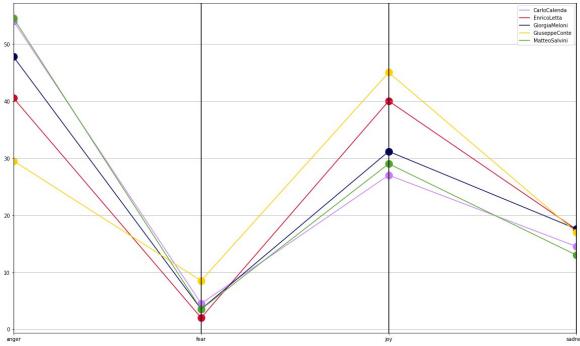


Figura 12: Emozioni veicolate dai leader

Osservando la figura di sinistra si nota che l'emozione rilevata maggiormente sia stata la rabbia, seguita dalla gioia e tristezza. Salvini, Meloni e Letta raggiungono i valori più alti; questo potrebbe significare che gli utenti che parlano di loro lo fanno in maniera aggressiva, probabilmente sperando che il leader del partito opposto non salga al Governo e cercando di screditarsi a vicenda. Detto questo, anche per i restanti due il valore rimane molto alto e nettamente superiore alle altre emozioni. Per quanto riguarda la gioia al primo posto si trova Conte, che doppia quasi tutti i suoi avversari. Questo non si sa se sia dovuto a un effettivo gradimento verso l'ex Presidente del Consiglio (dalla sentimènt è risultato il contrario), a una piccola comunità di seguaci che ne parla bene e molto frequentemente o addirittura alla potenziale influenza di pagine umoristiche. I risultati sono in linea con quelli trovati dai ricercatori nella fase di addestramento e test di FEEL-IT e rispecchiano lo sbilanciamento del dataset su cui è stato allenato. Questo potrebbe aver influito in una così forte disparità tra risultati.

Confrontando invece i messaggi scritti dai vari leader, quindi potendo valutare la loro comunicazione e come vogliono far percepire la propria immagine, si nota una fortissima crescita per quanto riguarda



Figura 13: Wordcloud dei tweet dei leader

la gioia, che in alcuni casi supera la paura. Si possono quindi distinguere messaggi di incoraggiamento e proposte volte a danneggiare l’immagine dei rivali. Questo è un risultato interessante che mostra come ci sia una forte differenza tra quello che l’utenza percepisce e comunica e quello che i vari segretari tentano di veicolare. È però necessario notare che il campione di tweet sui quali è stata eseguita questa seconda analisi è molto minore, perciò aumentando i dati a disposizione queste relazioni potrebbero cambiare, portando magari a risultati più in linea con i primi.

Seguono alcuni esempi di tweet ritenuti significativi per ogni tipo di emozione.

- Rabbia: ”#Elezioni2022 Giorgia #Meloni: ”Non ci interessano le provocazioni, io non vi racconto balle sull’attuale situazione italiana. . . Non siamo messi bene, no. . . Non è un momento facile, ma penso che con un pò di coraggio e buona volontà, si può invertire tendenza”
- Paura: ”sono preoccupato per queste elezioni, per lo più per la Meloni, non tanto per il reinserimento della leva militare, ma più per l’andamento generale e di quello che accadrà a noi giovani, naturalmente ha alcune buone idee ma non ci resta che vedere #Elezioni2022 #GiorgiaMeloni”
- Gioia: ”L’emozione di tornare nel paese natale. Solo 300 abitanti. Il calore di #VolturaraAppula E Vaiii!!!! #conte #giuseppeconte #puglia #CampagnaElettorale #ElezioniPolitiche2022”
- Tristezza: ”L’unica persona che mi rappresenta è #Salvini. Mi spiace che molta gente preferisca sudici come #Letta e #Conte miserabili che ci porteranno alla distruzione. Ieri il grullino ha ribadito che le sanzioni alla Russia sono necessarie. Finirà molto male. #ElezioniPolitiche22 #credo”

L’ultima breve analisi consiste nel cercare, per ognuno dei personaggi osservati fin’ora, quali siano le parole più frequenti all’interno dei loro tweet tramite delle wordcloud, mostrate in figura 13. Questo permette di avere un’idea su quali siano i temi più ricorrenti sui quali i diversi segretari fanno leva.

Esplorandole una ad una, Calenda sembra parlare principalmente dei partiti avversari, per cui si notano ”destra”, ”pd” e ”salvini” probabilmente screditandone le opinioni. Si nota come compaia anche ”draghi”, presidente uscente, le cui idee di riforma e piani vorrebbero essere portati avanti proprio dal partito di Calenda. Interessante come pare che Letta e Meloni non parlino del proprio partito, ma piuttosto dell’opposizione, sicuramente attaccandosi a vicenda. Per quanto riguarda proprio la leader di Fratelli d’Italia si notano anche riferimenti alla difesa dell’idea di ”nazione”, punto importante per il suo partito; sulla stessa lunghezza d’onda si trova Salvini, che parla di ”lavoro” e ”Paese”. Entrambi i partiti di centro-destra sembrano rivolgersi agli ”italiani”, al contrario di Conte che fa appello ai ”cittadini”. Per il segretario del Movimento 5 Stelle non compaiono particolari riferimenti ai rivali ma termini più generici utilizzati in campagna elettorale.

Di seguito vengono mostrati alcuni tweet ritenuti rappresentativi, anche in base alle wordcloud appena descritte.

- Carlo Calenda: ”Posso partecipare anche io al giochino? Se votate PD rischiate di far vincere Fratoianni. Se votate la destra rischiate di far vincere Putin. Se votate 5S rischiate di far vincere il nulla. Enrico ci vediamo al confronto domani e parliamo di cose serie. Che è meglio.”
- Enrico Letta: ”A Catania con CaterinaChinnic conferenza stampa per presentare le nostre proposte su una nuova #PubblicaAmministrazione in grado di spendere bene i soldi europei del #PNRR. Soldi che vanno usati e ben spesi, non rinegoziati come dicono a destra. Li perderemmo in quel caso.”
- Giorgia Meloni: ”La sinistra dei salotti ha dimostrato in questi anni di essere lontana dai temi del lavoro, della sicurezza e dello sviluppo. Gli italiani sanno bene che solo un governo capace di difendere gli interessi nazionali potrà combattere delocalizzazioni e concorrenza sleale. #VotaFdI”

- Giuseppe Conte: "Chiediamo che le urgenze dei cittadini siano anche le urgenze del governo. Né più, né meno. E su questo il Mov5Stelle e il Paese aspettano le risposte di Palazzo Chigi."
- Matteo Salvini: "I libri di scuola sono un salasso per le famiglie. Come aiutare gli italiani? Ecco la proposte della Lega. Fatemi sapere cosa ne pensate. Ti aspetto su TikTok per rimanere aggiornato"

5 Conclusioni e sviluppi futuri

Le analisi effettuate hanno permesso di rispondere adeguatamente alle domande di ricerca poste inizialmente, mettendo in evidenza in tutte le sue sfumature il frammentato panorama politico italiano. Dall'analisi dei tweet pubblicati nel periodo considerato emerge la presenza di community contrapposte, coerenti con la disparità di opinioni che i principali leader politici avanzano nei confronti dei temi che più contraddistinguono la nostra realtà. La sparsità riscontrata nella conformazione del grafo complessivo ha permesso di affermare che gli utenti analizzati non sono molto propensi all'interazione reciproca, e che tra gli attori che risultano più presenti all'interno del dibattito sono presenti i due esponenti dei partiti favoriti, ovvero Giorgia Meloni (con Fratelli d'Italia) ed Enrico Letta (con il Partito Democratico). Inoltre, un aspetto interessante emerso riguarda l'effettiva centralità all'interno del grafo occupata da alcuni sostenitori del Movimento 5 stelle, principale partito centrista candidato. Anche l'analisi semantica delle diverse comunità riscontrate all'interno del grafo ha permesso di evidenziare la presenza di cluster distinti, riuniti intorno ai diversi schieramenti presenti. L'analisi del sentiment degli utenti ha inoltre permesso di osservare una certa coerenza nella distribuzione della polarità (positiva/negativa/neutrale) espressa dai cittadini e veicolata dai 5 leader politici selezionati attraverso i tweet pubblicati, mentre dall'analisi delle emozioni emergono alcune differenze tra la comunicazione adottata dai capipartito e la percezione degli utenti, i cui tweet appaiono frequentemente caratterizzati da un certo malumore.

I principali ostacoli riscontrati hanno riguardato essenzialmente limitazioni dovute ai costi computazionali richiesti per effettuare le analisi, nonché all'impossibilità di reperire un numero maggiore di tweet a causa dei limiti API imposti da Twitter, dovendo dunque circoscrivere le analisi ad un periodo di tempo ridotto e non riconducibile all'intera campagna elettorale. Inoltre, l'argomento di studio ha necessariamente richiesto l'utilizzo di librerie in lingua italiana per effettuare le analisi di opinion mining che, oltre ad essere presenti in numero ridotto, non sono state addestrate su un numero troppo ampio di vocaboli.

Per le motivazioni riportate, sarebbe interessante ripetere le analisi a fini migliorativi utilizzando delle API Premium invece della versione Standard, nonché sperimentare eventuali variazioni nei risultati utilizzando algoritmi differenti per la community detection e la sentiment analysis o ampliando lo spettro di emozioni coinvolte. Inoltre, un possibile sviluppo futuro nell'immediato potrebbe riguardare la replica delle analisi a elezioni concluse, in modo da confrontare i risultati pre e post campagna elettorale.

Riferimenti bibliografici

- [BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O'Reilly Media, Inc.”, 2009.
- [BN13] Valerio Basile and Malvina Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.
- [BNH21] Federico Bianchi, Debora Nozza, and Dirk Hovy. ”FEEL-IT: Emotion and Sentiment Classification for the Italian Language”. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021.
- [HSSC08] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

- [Hut14] E.E. Hutto, C.J. Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". *Eighth International Conference on Weblogs and Social Media*, 2014.
- [Roe20] Joshua Roesslein. Tweepy: Twitter for python! URL: <https://github.com/tweepy/tweepy>, 2020.
- [SWG] SWG. Il sondaggio politico di lunedì 5 settembre 2022. <https://tg.la7.it/politica/il-sondaggio-politico-di-luned%C3%AC-5-settembre-2022-05-09-2022-175480>.