

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

SOCIAL MEDIA ANALYTICS

PROGETTO FINALE

---

## Zoom su ToloTolo:

analisi della rete sociale e Sentiment Analysis dei tweet  
inerenti all'ultimo film di Checco Zalone

---

*Autori:*

Massimiliano Calani - 838723 - m.calani@campus.unimib.it

Matteo Gaverini - 808101 - m.gaverini1@campus.unimib.it

Matteo Mazzola - 838004 - m.mazzola29@campus.unimib.it

Giugno 2020



## Sommario

I social media oggi, sono diventati degli strumenti essenziali per rimanere informati in qualsiasi momento su eventi che accadono nel mondo: meeting internazionali, anticipazione di film, accordi firmati da potenze mondiali, tutto viene discusso all'interno delle piattaforme digitali. Su Twitter in particolare, i post possono essere anche dei validi indicatori del sentiment degli utenti prima e dopo un particolare evento, come ad esempio l'uscita di un film atteso. Il progetto si focalizza sull'analisi dei tweet, registrati in un periodo temporale di 13 giorni, riguardanti l'ultimo film di Checco Zalone che si chiama *ToloTolo*. I dati raccolti vengono da una parte utilizzati per creare la rete sociale su cui poi verrà svolta la *Community Detection* mentre dall'altra, parallelamente, per effettuare la *Sentiment Analysis*.

## 1 Introduzione

Alcuni film, soprattutto quelli che trattano dei temi di attualità, possono suscitare opinioni contrastanti da parte delle persone. Tra tutti il tema dell'immigrazione è quello che negli ultimi anni in Italia, oltre ad essere teatro di accesi dibattiti politici, sta dividendo l'opinione pubblica in due correnti. La prima fazione rappresentata da un partito populista, sostiene che il Paese non può più accogliere e fornire assistenza agli immigrati perché prima bisogna aiutare i cittadini italiani; la seconda corrente invece, rappresentata da un movimento democratico, ritiene che uno Stato non può negare lo sbarco e l'accoglienza a persone che scappano dalla guerra o dalla povertà. Non è la prima volta che personaggi dello *show-business* (musicisti, conduttori televisivi etc.) esprimano un proprio punto di vista su questo delicato argomento con post pubblicati sui social o interviste sui quotidiani. Anche il cinema risulta essere un valido strumento per trasmettere un messaggio più o meno indiretto, però non tutte le persone riescono a percepire fedelmente ciò che vuole comunicare il regista. Un film che parla del tema dell'immigrazione in modo ironico, può essere letto da uno spettatore sotto diversi punti di vista; alcuni potrebbero apprezzarlo perché indurrebbe ad una profonda riflessione, altri potrebbero rimanere turbati in quanto contrari al fatto che un film possa scherzare su questa tematica, altri ancora potrebbero leggerlo in chiave puramente politica. Un modo per capire quale siano queste sensazioni, opinioni è scaricare i post dai social network, filtrarli per hashtag specifici al film e poi come ultima operazione analizzarne il contenuto testuale. Nel progetto questa strategia è stata adottata in merito all'ultimo film di Checco Zalone che si chiama *ToloTolo*, con l'obiettivo di rispondere ad alcune domande di ricerca che verranno definite successivamente.

## 2 Caso di studio

Il progetto prevede l'analisi di tutti i tweet pubblicati dal 27/12/19 al 8/01/20 riguardanti *ToloTolo*, il film uscito nelle sale 1 gennaio di quest'anno, scritto, diretto e interpretato da Luca Medici in arte Checco Zalone. La pellicola narra le vicende di Checco, un italiano che, deluso dalla madrepatria, decide di emigrare in Africa a cercare fortuna ma, per una serie di vicissitudini, è costretto a far ritorno in Italia percorrendo la tortuosa rotta dei migranti.

L'analisi dei tweet è avvenuta con l'obiettivo di rispondere alle seguenti domande di ricerca:

- Come evolve la rete sociale attorno al film nel tempo?
- Si distinguono utenti più rilevanti all'interno della rete?
- Quali e quante sono le comunità presenti nella rete sociale?
- Come cambia il sentiment degli utenti riguardo al film prima e dopo la sua uscita?
- Esistono dei personaggi famosi italiani maggiormente citati nei tweet?

### 3 Dati

I dati sono stati ottenuti in *batch* utilizzando le API di Twitter e la loro raccolta è avvenuta per un periodo temporale di 13 giorni, nello specifico dal 27/12/19 al 08/01/20. I tweet sono stati filtrati in base a due parametri che risultano essere lingua e hashtag. Per quanto riguarda la lingua si è scelto l'italiano visto che la pellicola è uscita solamente nelle sale italiane mentre per gli hashtag, si sono utilizzati quelli ufficiali del film, ovvero #CheccoZalone e #ToloTolo. Ogni tweet è costituito da 8 campi così definiti:

1. *id\_tweet*: id univoco tweet
2. *time\_tweet*: ora pubblicazione tweet (formato HH:MM:SS)
3. *date\_tweet*: data pubblicazione tweet (formato YYYY-MM-DD)
4. *retweet\_count*: numero di retweet del post
5. *user\_screen\_name*: nome utente che ha pubblicato il tweet
6. *text*: testo tweet
7. *retweet\_screen\_name*: nome utente che ha ripostato il tweet (presente solo se il post corrispondente è un retweet altrimenti settato a Nan)
8. *type*: tipo tweet (ORIGINAL se è un tweet originale, RT\_QUOTE se è un retweet commentato o RT\_NOQUOTE se è un retweet non commentato)

In tutto si sono ottenuti 165147 tweet di cui solamente 22263 risultano essere non duplicati; considerando invece il numero di utenti univoci, se ne sono individuati 9959. I dati sono stati salvati in due file csv separati: il primo chiamato *tolotolo.csv*, contiene tutti i tweet che includono tra gli hashtag #ToloTolo mentre l'altro, chiamato *checcozalone.csv*, i post che presentano tra gli hashtag #CheccoZalone.

### 4 Preprocessing generale

Per quanto riguarda il preprocessing, le uniche due operazioni effettuate sono state l'unione dei due csv creati (*tolotolo.csv* e *checcozalone.csv*) e la rimozione dei duplicati.

## 4.1 Unione

Questa operazione è stata compiuta perché si vuole effettuare un'analisi sulla totalità dei tweet raccolti in *batch*, usando i due diversi hashtag.

## 4.2 Rimozione duplicati

Questa operazione risulta necessaria perché nei singoli csv sono contenuti parecchi tweet duplicati così come sono presenti tweet identici in entrambi i due file. Il motivo per cui si verifica questa situazione è che, durante la fase di *Data Ingestion*, avvenuta giornalmente in *batch*, si è sempre definito come punto di partenza (parametro *since* Tweepy) il giorno 27/12/19. La rimozione, avvenuta dopo l'unione, è stata effettuata in due step: prima si sono rimossi i tweet che avessero i valori di tutte le feature identiche tenendo solo per ognuno una copia, dopo si sono rimossi i post con valore *text*, *user\_screen\_name*, *retweet\_screen\_name* e *date\_tweet* identici tenendo solo il tweet più recente, ovvero quello con il *retweet\_count* maggiore.

Alla fine del preprocessing si ottiene un unico csv chiamato *uniti.csv* che contiene 22263 tweet. Da questo file si ottengono poi altri due csv; il primo (si veda Figura 1.a) chiamato *relazioni.csv*, viene utilizzato per creare la rete sociale mentre l'altro *tweet.csv* (si veda Figura 1.b) per effettuare la *Sentiment Analysis*.

N.B. Per motivi di leggibilità vengono visualizzati solo alcuni attributi presenti nei file *relazioni.csv* e *tweet.csv*

	<b>user_screen_name</b>	<b>retweet_screen_name</b>	<b>month</b>	<b>day</b>
51	riccioale19	fabio_falzone	Dec	27
175	RosellinaMarian	RosellinaMarian	Dec	27
335	PoliticallyUn11	PoliticallyUn11	Dec	27
337	leo_caselli	La_manina__	Dec	27
429	Paola30502511	Vince7914	Dec	27

(a) *relazioni.csv*

	<b>id_tweet</b>	<b>text</b>	<b>month</b>	<b>day</b>
43	1210526929922777088	#checcozalone e il suo cast now #tolotolo htt...	Dec	27
158	1210621744970653698	@fabio_falzone ...e allora bisogna correre a v...	Dec	27
314	1210482976687087621	Ciò che difendiamo è la libertà di #CheccoZalo...	Dec	27
316	1210671729561030656	#CheccoZalone presenta 'Tolo Tolo': "Un mostro...	Dec	27
436	1210615527242190848	Siamo tutti migranti: con #ToloTolo Zalone leg...	Dec	27

(b) *tweet.csv*

Figura 1: Header file ottenuti dal preprocessing

## 5 Rete sociale

Una rete sociale o *social network* rappresenta una struttura sociale costituita da attori (quali persone o cose) che interagiscono tra di loro attraverso dei legami diadici (mono o bidirezionali) [1]. Esistono diverse relazioni che si possono rappresentare in una rete (amicizia, like, retweet, etc.) e alcune di esse sono discriminanti rispetto al social media che si monitora. Per esempio la relazione di amicizia in Facebook è diversa rispetto a quella in Twitter; nel primo caso il legame è bidirezionale (se A diventa amico di B anche B è amico di A) mentre nell'altro è unidirezionale (A può seguire B ma B può essere che non segua A). Per quanto riguarda la prima parte del progetto, si è deciso di realizzare come *social network* una rete di retweet dove gli attori sono gli utenti che hanno ripostato un tweet contenente hashtag relativi al film. Il risultato è un

grafo non orientato e non pesato (si veda Figura 2); non orientato perché si assume che il retweet rappresenti un legame bidirezionale tra due persone, non pesato perché l'attenzione è rivolta solo al fatto che tra due utenti ci sia stato un retweet e non quante volte siano avvenuti legami di questo tipo tra i due. Si è scelto di considerare il retweet per realizzare la rete perché rispetto ad altre relazioni come il like o il commento, si è ritenuto più utile e significativo per poter rispondere in maniera esaustiva alle prime due domande di ricerca (si veda Capitolo 2). L'analisi del grafo è avvenuta effettuando due tipi di monitoraggio: giorno per giorno e cumulativo (dal primo giorno si estende progressivamente la rete con utenti e retweet nelle date successive). Il primo monitoraggio si è rivelato utile per osservare l'evoluzione della rete mentre il secondo è servito sia per individuare gli attori rilevanti che riconoscere le comunità all'interno del grafo.

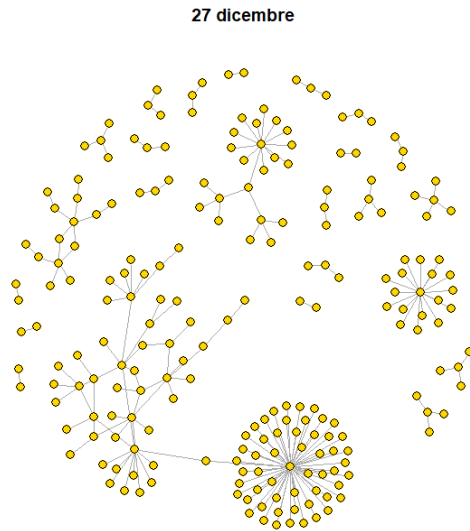
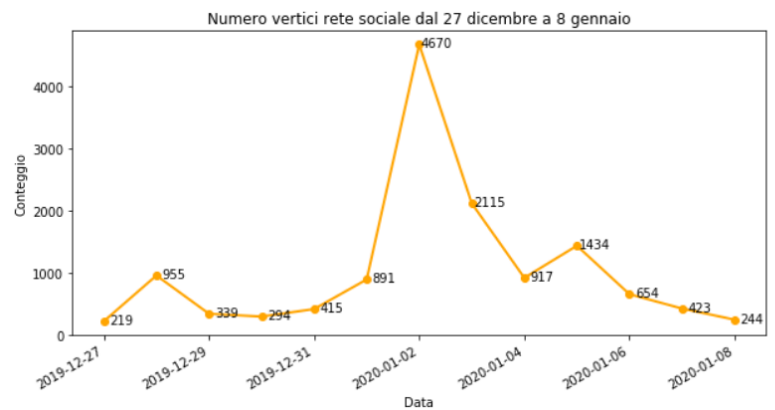


Figura 2: Esempio rete sociale

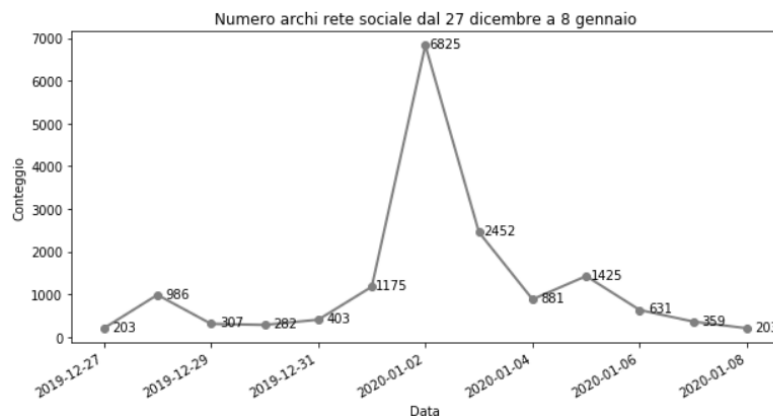
### 5.1 Evoluzione rete sociale

Osservando il grafico giornaliero (si veda Figura 3) emergono diversi aspetti interessanti. Il primo è che la rete aumenta la sua grandezza (intesa come ordine e dimensione del grafo) con il prosieguo del tempo fino al 2 gennaio, giorno in cui si registra l'estensione massima con 4670 utenti e 6825 retweet. Il secondo fattore interessante che emerge è che dopo il picco massimo, si inizia a registrare una flessione del numero di retweet e quindi anche di tweet nelle date successive, ad eccezione del 5 gennaio dove si ha un aumento del 60% circa del numero di utenti e relazioni rispetto al giorno prima. Tale flessione non sorprende perché, una volta terminatosi un evento ampiamente discusso su Twitter, gli utenti con il passare del tempo iniziano a parlarne di meno fino a quando l'argomento stesso non compare più nella lista dei *trending topic*. Questo avviene esattamente per il film *ToloTolo*: nei giorni antecedenti all'uscita gli utenti sono in fervore, dopodiché quando la pellicola è presente in tutte le sale, l'argomento inizia ad essere meno discusso

su Twitter. Bisogna sottolineare però, che attorno al film si era già creato un clamore mediatico molto tempo prima che la pellicola uscisse grazie al *real time marketing*, ossia la capacità di prendere spunto dai fatti di cronaca, dai temi di attualità per promuovere un prodotto [2]. In questo caso l'azione di marketing effettuata è stato rilasciare un trailer i primi di dicembre dove Checco Zalone impersona l'italiano medio che, con una canzone ironica, si lamenta degli immigrati generando così discussioni, polemiche sui social [3]. Questo alimenta l'hype in anticipo, attesa che normalmente si sviluppa solo all'avvicinarsi di un evento. Detto ciò, si sottolinea inoltre che l'aspettativa che si è creata intorno al film è stata favorita anche dall'assenza dal palcoscenico per 3 anni del comico pugliese che ha provocato nelle persone un senso di vuoto, mancanza: gli italiani volevano ridere di nuovo. Per questo motivo la rete già dal primo giorno di monitoraggio, seppur non sia una data prossima al 1 gennaio, contiene 219 utenti e 203 retweet. Il grafo poi aumenta la sua grandezza fino a raggiungere il picco massimo in termini di ordine e dimensione i primi giorni dopo l'uscita del film. Questo può essere legato al fatto che molte persone, vedendo la pellicola, siano rimasti sorprese dalla diversità della trama rispetto a quella intuita dal trailer ed è proprio questa sorpresa, meraviglia che ha generato progressivamente un aumento considerevole di tweet facendo diventare in pochi giorni gli hashtag #Checco Zalone e #ToloTolo *trending topic*.



(a) Numero vertici



(b) Numero archi

Figura 3: Evoluzione rete sociale

## 5.2 Individuazione hub

Una volta osservata l'evoluzione della rete, si sono individuati gli hub, nello specifico si sono determinati i 3 vertici più importanti rilevati ogni periodo temporale diverso in cui è definito il grafo cumulativo. A tal proposito si è utilizzata come metrica la *degree centrality* normalizzata (si veda Formula 1); si è deciso di normalizzare perché in questo modo si è potuto effettuare un confronto globale degli hub considerando la dimensione della rete in ogni istante temporale diverso. Le altre misure invece, quali *closeness* e *betweenness*, sono state escluse ognuna per un motivo diverso. La prima metrica è stata scartata perché il grafo, per entrambi i tipi di monitoraggio, risulta essere disconnesso e quindi la distanza tra due nodi che fanno parte di due componenti connesse diverse è pari ad un valore infinito. Per questo motivo come suggerisce la letteratura non si può utilizzare la *closeness* quando il grafo presenta più di una componente connessa [4]. La *betweenness* invece è stata scartata in seguito a come è stato interpretato il concetto di importanza: un nodo viene considerato importante solo se risulta essere popolare e non se è un nodo *ponte*, ovvero un vertice capace di influenzare il flusso all'interno della rete svolgendo il ruolo di "mediatore".

$$\overline{C_D}(v) = \frac{d(v)}{n-1} \quad \text{dove } n \text{ indica numero vertici del grafo} \quad (1)$$

Analizzando gli hub individuati (si veda Figura 4) emergono diversi aspetti interessanti. Il primo è che i nodi rilevanti, nella maggior parte dei casi, risultano essere identici tra un istante temporale e un altro in cui è definito il grafo cumulativo. Per esempio la rete dal 27 dicembre al 31 e quella dal 27 dicembre al 1 gennaio, risulta avere gli stessi hub che sono *Capezzone*, *PBerizzi* e *NicolaPorro*. La seconda cosa interessante che emerge è che gran parte degli hub sono giornalisti; questo può essere motivato dal fatto che il film, trattando un tema d'attualità, abbia incentivato gli inviati delle varie testate giornalistiche ad esprimere un'opinione mediante dei tweet. I cronisti individuati come hub risultano essere *Capezzone*, *PBerizzi*, *NicolaPorro* e *Giorgiolaporta*. Il primo è un giornalista che scrive su *La Verità* e *Atlantico*, il secondo è uno scrittore e inviato de *La Repubblica*, il terzo è vicedirettore de *Il Giornale* oltre che essere un conduttore televisivo infine il quarto è un giornalista e portavoce presso la Presidenza della Camera. Gli altri hub invece, ad eccezione di *guffanti\_marco* che non risulta essere un personaggio pubblico, possono essere classificati in 3 gruppi: politici, attivisti e comici. Il gruppo dei politici è rappresentato da *distefanoTW*, vicepresidente di Casa Pound, gli attivisti da *catlatorre* e *OizaQueensday* che sono rispettivamente un'avvocata e una collaboratrice di una rivista online chiamata *The Vision*, entrambe attiviste per i diritti civili, ultimo il gruppo dei comici rappresentato da *Fiorello*. La presenza di un politico come hub non sorprende soprattutto perché il tema dell'immigrazione è uno degli argomenti più discussi e dibattuti dai vari partiti in particolare modo Casa Pound, ha sempre espresso dei pensieri forti su tale tematica anche sui social. La presenza di *Fiorello* invece può essere legata al fatto che essendo un personaggio televisivo famoso, i suoi tweet vengano condivisi da parecchie persone. Questo rende evidente il fenomeno dell'attaccamento preferenziale ovvero la capacità degli hub di essere, in modi diversi, "attrattori" all'interno della rete. Analizzando gli hub individuati, emergono tutte e tre le tipologie di attaccamenti preferenziali che sono *popolarità*, *qualità* e *mix*. Il primo tipo è rappresentato da *Fiorello* visto che è uno showman conosciuto da tutti gli italiani, il secondo dai giornalisti perché il loro lavoro è raccontare la realtà, la verità

dei fatti quindi i loro contenuti devono essere qualitativi, ultimo tipo è rappresentato da *NicolaPorro* in quanto, oltre che essere esperto di politica, è anche conduttore di programmi televisivi di successo come *Matrix* e *Stasera Italia*. Confrontando gli hub da un punto di vista globale, il nodo più rilevante risulta essere *Capezzone* nella rete creata dal 27 al 29 dicembre. Questo può essere motivato dal fatto che tale utente, essendo molto attivo e popolare su Twitter (in media pubblica 8 tweet al giorno), riceve ad ogni suo post un numero considerevole di retweet. Analizzando invece solo l'ultima rete creata, ovvero quella dal 27 dicembre al 8 gennaio, l'hub più rilevante risulta essere *Giorgiolaporta*, a seguire in ordine di importanza *OizaQueensday* e *Fiorello*.

	hub1	grado1	hub2	grado2	hub3	grado3
da 27 a 28	Capezzone	0.26594789	catlatorre	0.09074573	guffanti_marco	0.08805031
da 27 a 29	Capezzone	0.32616487	catlatorre	0.09677419	guffanti_marco	0.09534050
da 27 a 30	Capezzone	0.3032015	NicolaPorro	0.1117389	catlatorre	0.0866290
da 27 a 31	Capezzone	0.2476923	PBerizzi	0.1671795	NicolaPorro	0.1015385
da 27 a 1	Capezzone	0.17879571	PBerizzi	0.14333210	NicolaPorro	0.07351311
da 27 a 2	Giorgiolaporta	0.08098918	Capezzone	0.07527048	Fiorello	0.06970634
da 27 a 3	Giorgiolaporta	0.08865660	distefanoTW	0.06439541	Fiorello	0.06297587
da 27 a 4	Giorgiolaporta	0.08858315	distefanoTW	0.06203234	Fiorello	0.06070480
da 27 a 5	Giorgiolaporta	0.08014730	distefanoTW	0.05675295	OizaQueensday	0.05642803
da 27 a 6	Giorgiolaporta	0.07743409	OizaQueensday	0.06238323	distefanoTW	0.05470210
da 27 a 7	Giorgiolaporta	0.07614986	OizaQueensday	0.06254442	distefanoTW	0.05381257
da 27 a 8	Giorgiolaporta	0.07541675	OizaQueensday	0.06206065	Fiorello	0.05342438

Figura 4: Risultati hub grafo cumulativo

### 5.3 Community Detection

L'ultima operazione effettuata è stata la *Community Detection*, ovvero individuare le comunità presenti all'interno del grafo. Per raggiungere tale obiettivo si è deciso di utilizzare la rete finale che comprende tutti i retweet dal 27 dicembre al 8 gennaio. Il motivo di tale scelta è che si è voluto individuare le comunità non solo in un preciso istante temporale, ma considerando tutto il periodo di monitoraggio della rete. A tal proposito si sono utilizzati due algoritmi della libreria *Networkx* di Python che sono *greedy\_modularity\_communities* e *best\_partition*. Si è effettuata questa scelta perché tra tutti gli algoritmi disponibili, risultano essere gli unici che non richiedono a priori di definire il parametro  $k$ , cioè il numero di comunità da ottenere. Questi due algoritmi, entrambi appartenenti alla famiglia *hierarchy-centric*, aggregano i nodi in gruppi basandosi su uno stesso criterio, ovvero massimizzare la modularità. Tale metrica indica quanto è forte la suddivisione della rete in comunità [5]; per far ciò si misura la connettività all'interno del grafo e la si confronta con il valore atteso delle connessioni che possono essersi create casualmente all'interno di un gruppo di pari dimensione. I valori che può assumere la modularità, nel caso in cui il grafo sia non orientato e non pesato è  $[-\frac{1}{2}, 1]$  mentre in tutti gli altri casi è  $[-1, 1]$  [5].



La formula della modularità  $Q$  è così definita:

$$Q = \frac{1}{2E} \sum_C \sum_{i \in C, j \in C} (A_{ij} - \frac{d_i d_j}{2E}) z_{ij} \quad (2)$$

dove  $C$  indica la comunità assegnata ai nodi  $i$  e  $j$ ,  $A_{ij}$  elemento  $ij$  della matrice di adiacenza,  $d$  il grado del nodo corrispondente,  $E$  la dimensione del grafo infine  $z_{ij}$  rappresenta una variabile decisionale che assume valore 1 se nodo  $i$  e  $j$  fanno parte della stessa comunità, 0 altrimenti.

Gli algoritmi utilizzati, anche se entrambi massimizzano la modularità, seguono due strade diverse per perseguire il loro obiettivo. Il primo utilizza l'algoritmo di *Clauset-Newman-Moore*; esso prevede, tramite un approccio *greedy*, di assegnare inizialmente un gruppo ad ogni singolo nodo e poi aggregare ripetutamente due comunità con l'obiettivo di massimizzare la modularità intergruppo fino a convergenza, cioè quando non esistono più raggruppamenti che migliorino tale criterio [6]. Il *best\_partition* invece, individua le comunità utilizzando l'algoritmo di *Louvain* il quale, a differenza del primo, massimizza la modularità adottando un *local moving heuristic*. Tale approccio consiste nel spostare ripetutamente un nodo, preso in modo casuale dal grafo, da una comunità ad un'altra in modo tale che ad ogni cambiamento corrisponda un incremento della modularità [7]. L'algoritmo quindi, partendo da una situazione iniziale dove ogni nodo è associato a una comunità, individua i gruppi adottando *local moving heuristic* fino a quando non esiste più un qualsiasi spostamento che incrementi la modularità.

### 5.3.1 Risultati

Una volta eseguiti i due algoritmi di *Community Detection*, si sono valutati i risultati ottenuti. A tal proposito, visto che nel caso di studio non è presente una *ground truth* e che entrambi gli algoritmi massimizzano la modularità, si è effettuato un confronto dei valori delle loro funzioni obiettivo. Da qui è emerso che l'algoritmo *greedy-modularity-communities* risulta essere migliore rispetto a *best\_partition* avendo ottenuto un valore di modularità di 0.67 contro 0.66. Bisogna comunque dire che globalmente i risultati si rivelano in tutti e due i casi molto buoni in quanto il valore di modularità è prossimo a 1; ciò implica che gli algoritmi *hierarchy centric* che si basano su tale criterio, risultano essere una strategia vincente nel contesto in esame. A questo punto, una volta decretato l'algoritmo ottimo da cui si ottengono 20 comunità (si veda Figura 5), si è cercato di etichettarle in base ai suoi componenti. Per far ciò, si è dapprima considerato l'utente con grado più alto per ogni comunità (si veda Figura 6), successivamente si è ricercato sul web chi fosse (giornalista, comico etc.) infine si sono analizzati gli argomenti discussi all'interno di ogni gruppo qualora la ricerca precedente non fosse andata a buon fine (zero informazioni ottenute dal motore di ricerca). Si ottengono così 8 label che verranno poi utilizzate per raggruppare le comunità, tali etichette sono: "giornalismo", "comicità", "varietà", "antisalvinismo", "notizie dal blog", "religione", "beni culturali" e "haters".

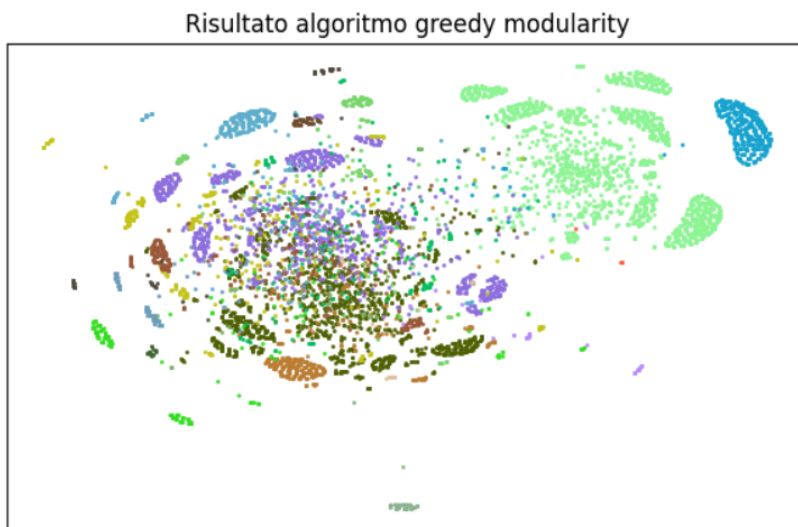


Figura 5: Risultato algoritmo migliore

La prima categoria "giornalismo" è rappresentata dalle comunità 0, 4 e 16; nei primi due gruppi i rappresentanti sono dei giornalisti (*Giorgiolaporta*, *PBerizzi*) mentre nell'altro l'esponente è uno scrittore palermitano (*fulvioabbate*). La seconda categoria "comicità" è rappresentata da 5 comunità che sono 1, 8, 12, 14 e 17. Si sono identificati questi gruppi perchè tutti contengono come rappresentanti o dei comici (*Fiorello*, *0DanieleFabbri*, *GliAutogol*) o account di pagine satiriche (*segnanti*); inoltre si registra all'interno di una comunità anche un vignettista (*NatangeloM*). La terza categoria "varietà" è rappresentata da un'unica comunità che è la 11; essa include come rappresentante, l'account ufficiale di un programma televisivo di genere varietà (*chetempochefa*). La quarta categoria "antisalvinismo" è identificata da 3 comunità che sono 2, 10 e 18; rispetto agli altri gruppi l'assegnamento della corrispondente label è avvenuta non in base alla professione ma rispetto a quello che pubblicano i rappresentanti di ogni *community*, ovvero per la maggior parte tweet contro Matteo Salvini (si veda Figura 7), segretario federale e leader della Lega. La quinta categoria "notizie dal blog" è rappresentata da 4 comunità che sono 3, 5, 7, 9; si sono identificate queste *community* perchè tutti i rappresentanti sono dei blogger. La sesta categoria "religione" è rappresentata da due comunità che sono 6 e 13; nella prima *community* il rappresentante è *fam\_cristiana*, settimanale di ispirazione cattolica che parla di attualità, nella seconda invece *donTonio66*, parroco di Capurso, un comune in provincia di Bari. Le ultime due categorie sono identificate ognuna da un solo gruppo; la prima "beni culturali" è rappresentata dalla comunità 15 in quanto il rappresentante è l'account ufficiale di un parco archeologico (*appia\_day*) mentre l'altra label "haters", identificata dalla comunità 19, contiene come nodo rappresentante un utente che critica pesantemente il personaggio Checco Zalone.

Dai risultati ottenuti emergono due aspetti interessanti; il primo è l'assenza di gruppi che parlano di cinema o di film in generale, scenario atteso visto che si stanno monitorando tweet inerenti ad una pellicola, il secondo invece è la presenza di comunità che criticano Matteo Salvini. Quest'ultimo aspetto può essere interpretato come in realtà non ci sia stata una vera e propria discussione sul film, ma il focus, l'interesse fosse più che altro sul tema d'attualità affrontato nella pellicola, argomento di cui il leader della Lega non ha mai nascosto il suo punto di vista.

	Community	Degree
0	0	Giorgiolaporta
1	1	Fiorello
2	2	catlatorre
3	3	OizaQueensday
4	4	PBerizzi
5	5	mchiarissima
6	6	fam_cristiana
7	7	KelleddaMurgia
8	8	GliAutogol
9	9	manginobrioches
10	10	O_Strunz
11	11	chetempocheffa
12	12	segnanti
13	13	donTonio66
14	14	0DanieleFabbri
15	15	appia_day
16	16	fulvioabbate
17	17	NatangeloM
18	18	senzaproblemi1
19	19	SaverioSara

Figura 6: Rappresentanti di ogni comunità

<p>Ricapitolando : - Volevano # CheccoZalone senatore a vita senza aver visto il film ; - Hanno visto il film ; - Attaccano # CheccoZalone Sovranisti e nazionalisti . Poi si lamentano se li definisci incoerenti . Sic ! # ToloTolo</p> <p>Leghisti che chiedono il rimborso dopo aver visto il film di Checco Zalone . Tranquilli Amici , è arrivato a 23 milioni di incassi . Quando arriva a 49 con calma ve li restituisce . # tolotolo # CheccoZalone # 5gennaio</p> <p>Per settimane lo hanno chiamato genio , Salvini lo ha persino “ eletto ” senatore a vita . Poi si sono accorti che # Tolotolo non parlava dei neri ma di loro , della loro miseria e ipocrisia . E di colpo # CheccoZalone è diventato una “ zecca comunista ” . Che bellezza il sovranismo ...</p>
(a) community 2
<p>Il sovranista è quel soggetto che fa passare Zalone da idolo degno di essere nominato senatore a vita , a zecca sfigata mangiabambini . Il tutto senza avere ancora visto il film . # ToloTolo</p> <p>La Russa bocchia il film di Zalone : “ Sonnacchiavo ” . Doveva prendere il Viagra , magari sarebbe riuscito a tenere la testa dritta . ( @ LucillaMasini ) # 2gennaio # LaRussa # zalone # ToloTolo # viagra</p> <p>Perché è l'unico che sa parlare a tutti gli italiani ( si : tutti ) . E perché nel nuovo film # ToloTolo lancia un messaggio antifascista e antirazzista forte e chiaro . Rischiamo come nessun</p> <p>Zalone svela che la trama di Tolo Tolo è contro i fascisti . Vi spoilerò il film : Salvini fa una figura di merda . ( @ LucillaMasini ) # Zalone # tolotolo # Salvini # 27dicembre</p> <p># CheccoZalone è passato in un attimo da attore comico a Senatore a vita e poi da Senatore a vita a pessimo attore esattamente come i meridionali sono passati da puzzolenti terroni a elettori fraterni . ( Fabio Macchia @ FaxyMac )</p>
(b) community 10

Figura 7: Esempi tweet contro Matteo Salvini

## 6 Analisi testuale dei tweet

Per quanto riguarda l'analisi dei tweet, si sono svolti due task: *Sentiment Analysis* e *Named Entity Recognition (NER)*. La prima operazione consiste nell'esaminare il contenuto testuale dei post al fine di identificare il sentiment, interpretato nel progetto come polarità, espresso dagli utenti [8] (per svolgere tale task si è adottato un approccio *lexicon-based*). La seconda operazione invece, prevede di riconoscere le entità presenti all'interno di un testo come ad esempio persone o organizzazioni [9]. I due task, entrambi problemi di classificazione, si sono rivelati utili per rispondere efficacemente alle ultime due domande che hanno guidato il caso di studio: la *Sentiment Analysis* per verificare il variare delle opinioni degli utenti prima e dopo l'uscita del film mentre la *NER* per individuare i personaggi famosi italiani maggiormente citati all'interno dei post. Oltre a questo la *NER* si è rivelata utile durante il preprocessing specifico dei tweet; tale procedimento, descritto nella sezione successiva, ha avuto come obiettivo quello di uniformare la struttura dei post, operazione fondamentale per effettuare la *Sentiment Analysis*.

### 6.1 Preprocessing testuale

Come prima operazione si è eliminato il "rumore" all'interno dei tweet scartando tutti quelli non pertinenti all'argomento trattato, dopodiché si è effettuato il preprocessing generale dei tweet svolgendo globalmente 7 task. Prima di tutto si è pianificata la gestione degli hashtag, nello specifico si è decretato quali fossero gli hashtag da tenere e quali invece da eliminare; dopodiché si è svolta la *tokenization* ed eliminati gli eventuali allungamenti espressivi presenti all'interno di una parola. A questo punto, una volta effettuate tale operazioni, si sono svolti due task ulteriori: il primo prevede la gestione delle eventuali negazioni delle frasi mentre il secondo l'individuazione e la gestione delle abbreviazioni di parole comuni nei tweet. Dopo aver svolto tali task, si sono poi effettuate come ultime due operazioni la *stop-words removal* e la *lemmatization*.

#### 6.1.1 Filtraggio

Questa operazione si è effettuata perché analizzando la distribuzione degli hashtag (si veda Figura 8), emerge che i tweet che ne contengono un numero superiore ad una soglia uguale a 5, risultano essere *off-topic* ovvero non inerenti con l'oggetto di interesse. Questo può essere legato al fatto che molti utenti includano all'interno dei loro tweet parecchi hashtag virali solo per avere più visibilità su Twitter (si veda Figura 9). Per questo motivo si è deciso di eliminare tutti i post il cui numero di hashtag supera la soglia definita precedentemente ottenendo così 21570 tweet, dati che verranno utilizzati per svolgere la *Sentiment Analysis*.

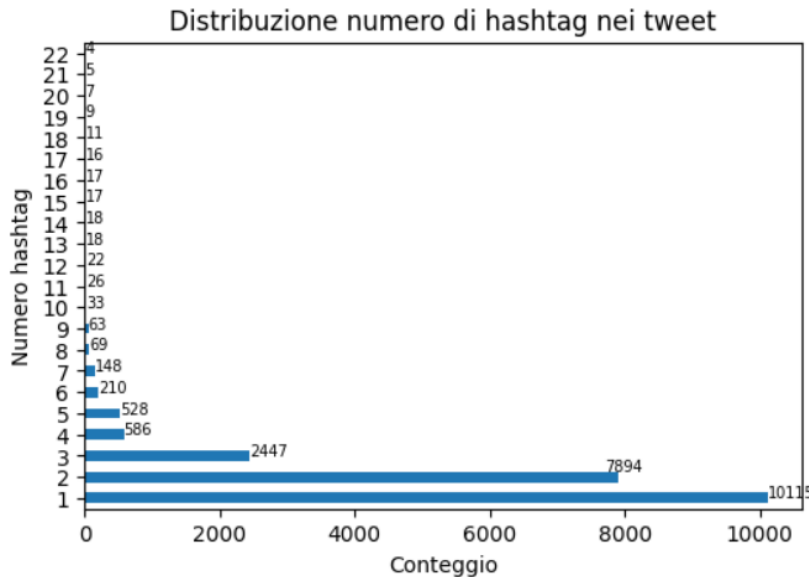


Figura 8: Distribuzione numero hashtag



Figura 9: Esempio tweet *off-topic*

### 6.1.2 Gestione Hashtag

La prima operazione è stata la gestione degli hashtag, ovvero capire se queste particolari parole fossero informative o meno per la *Sentiment Analysis*. Analizzando i tweet è emerso che gli hashtag che si trovano all'inizio o alla fine di un testo sono raramente rilevanti mentre quelli che si collocano in mezzo lo sono quasi sempre. Per questo motivo si è deciso di eliminare gli hashtag che si trovano alla fine di un tweet e tenere tutti gli altri convertendoli in semplici parole.

### 6.1.3 Tokenization

La *tokenization* è la seconda operazione effettuata nonché la fase principale del preprocessing di un testo; esso prevede di suddividere il contenuto testuale dei tweet in token significativi. Questa operazione risulta essere molto importante perché, oltre ad essere la base per effettuare la *lemmatization*, permette di ottenere una rappresentazione *Bag of Word (BoW)* dei tweet nel contesto in esame. La *tokenization* inoltre presenta dei problemi che bisogna gestire, uno fra tutti è la lingua. Questo perché la definizione dei token, oltre che dipendere dal contesto di riferimento, dipende strettamente dal linguaggio. Per questo motivo la suddivisione del testo è avvenuta considerando la struttura grammaticale e le forme della lingua italiana. In questa fase si è inoltre applicata congiuntamente una *Named Entity Recognition* per affinare la definizione dei token. Questa operazione è stata effettuata perché all'interno di alcuni tweet, si sono trovati dei riferimenti a personaggi pubblici i cui cognomi sono costituiti da più di una

parola ("La Russa", "Di Maio" etc.) e quindi, se si dovesse effettuare la *tokenization* senza considerare questi casi particolari, si otterrebbero suddivisioni errate delle parole.

#### 6.1.4 Gestione negazioni

La terza operazione è stata gestire le negazioni in quanto molti tweet presentano delle frasi negate. Questo è un problema che bisogna assolutamente considerare perchè il dizionario usato per la *Sentiment Analysis*, contiene parole le cui polarità positive e negative, possono cambiare in base alla frase in cui si trova tale termine. Per esempio la parola "bello" all'interno del dizionario è associata una polarità positiva però se quel termine fosse all'interno di un tweet negato ("il film non è bello"), a quel punto deve essere associata la polarità negativa. Per questo motivo si è deciso, una volta appurato che il tweet presentasse delle negazioni, di negare tutte le parole che seguono l'avverbio di negazione "non" fino a quando si incontra o un segno di punteggiatura che indica la fine di una parte del testo (.;: ([?!)) oppure una congiunzione avversativa ("però", "ma", "tuttavia").

#### 6.1.5 Eliminazione allungamento espressivo

La quarta operazione risulta essere la rimozione di tutti gli eventuali allungamenti espressivi presenti all'interno di una parola come ad esempio "mooolto" o "immigra-tooooo". Si è scelto di eliminare tali forme espressive e non considerare le parole così come sono perchè il dizionario usato per la *Sentiment Analysis*, contiene termini nella forma base e quindi, se in un tweet ci fosse per esempio la parola "bellooooo", non si riuscirebbe a ottenere la polarità associata anche se il termine nella forma base è presente nel dizionario.

#### 6.1.6 Gestione abbreviazioni

La quinta operazione prevede di individuare le abbreviazioni generalmente usate nei tweet. Per far ciò si è creato manualmente un dizionario che contiene per ogni abbreviazione la parola associata; in totale si ottengono 33 termini che risultano essere quelli più usati nei tweet come ad esempio "cmq" o "dv". Si è effettuata questa operazione perchè in questo modo si possono individuare ulteriori parole di senso compiuto che possono aiutare il processo di *Sentiment Analysis*.

#### 6.1.7 Stop words removal

La sesta operazione è stata rimuovere le *stop word*, ovvero eliminare tutte quelle parole che risultano essere poco significative per capire il contenuto dei tweet. I termini eliminati sono stati estratti da una lista di 279 parole (ottenuti con la libreria Python *nltk*) che rappresentano le principali categorie lessicali della lingua italiana come ad esempio articoli e preposizioni. Oltre a queste parole, si è deciso di aggiungere alla relativa lista anche tutti i termini contenuti nei tweet non rilevanti per la semantica come ad esempio i segni di punteggiatura e i numeri (per l'elenco completo si veda Tabella 1).

#### 6.1.8 Lemmatization

L'ultima operazione riguarda la *lemmatization*, ovvero ridurre le forme flesse delle parole ponendole in una forma base. Per effettuare tale tecnica si è utilizzato *TreeTagger*

Elemento	Famiglia	Esempio
<i>articolo</i>	determinativo	il, lo..
	indeterminativo	uno, una..
	partitivo	della, dei, degli..
<i>preposizione</i>	semplice	di, a, da..
	articolata	nello, sullo, allo..
<i>aggettivo</i>	possessivo	mio, tuo, nostro..
	interrogativo/esclamativo	quali, quanti, quante..
<i>pronomi</i>	personale	mi, ti, ci..
	possessivo	nostro, vostro, loro..
<i>elemento_tweet</i>	link	<a href="https://t.co/LxhQeBwJQE">https://t.co/LxhQeBwJQE</a>
	citazione	@fabio_falzone, @GianniJ08..
	numero	100, 20
	simbolo	€, %..

Tabella 1: Elenco stop word usate

[10], un tool che annota le parti del discorso di un testo ed estrae per ogni parola il corrispondente lemma. Si è scelto di applicare la *lemmatization* e non la principale alternativa che è lo *stemming* perchè il dizionario usato per la *Sentiment Analysis*, contiene tutti i vocaboli nella forma base e quindi, per capire quale è la polarità associata ad una certa parola, è necessario un match identico con il termine contenuto nel glossario. Per questo motivo applicare unicamente lo *stemming* oppure usarlo dopo la *lemmatization* risulta inutile perché tale tecnica riporta solo la radice di una parola (per esempio "pesca", "pescare" diventa "pesc").

## 6.2 Sentiment Analysis

Per quanto riguarda la *Sentiment Analysis*, si è effettuata una *Polarity Analysis* di tipo *lexicon-based*. Si è deciso di considerare la polarità e non l'emozione (gioia, rabbia etc.) per interpretare il sentiment, in seguito alla domanda di ricerca a cui si è voluto rispondere, ovvero se il giudizio delle persone dato al film, fosse cambiato prima e dopo la sua uscita. La tecnica *lexicon-based* invece, si è scelta perchè il dataset di riferimento non contiene alcun tweet etichettato con la label reale della polarità, di conseguenza non è possibile effettuare alcun apprendimento sia supervisionato che semi-supervisionato. Oltre a questo, i dati non consentono di effettuare alcun operazione di *Transfer Learning* perchè gli unici modelli che si possono utilizzare nel dominio di riferimento, sono addestrati usando un dataset di recensioni di film in inglese e quindi, avendo due lingue diverse, si otterrebbero risultati pessimi. Il dizionario usato per la *Sentiment Analysis*, si è ricavato unendo due glossari: *OpenNER Sentiment Lexicon Italian* [11] e *Emoji Sentiment Ranking 1.0* [12]. Il primo (si veda Figura 10) contiene 24293 entrate lessicali italiane a cui è associata la corrispondente polarità; l'etichettatura è avvenuta in modo semi-automatico utilizzando come riferimento *ItalWordNet 2.0*, un database semantico-lessicale. Il secondo glossario (si veda Figura 11) invece, contiene 751 emoji dove ad

ognuna sono registrate, oltre ad altre informazioni, il numero di volte in cui compare tale emoticon come positiva, negativa, neutrale, se presente, in un campione di 70000 tweet (incluse 14 lingue europee diverse); le label in questo caso non sono assegnate in modo semi-automatico ma manualmente da 83 persone di diversa nazionalità. La polarità di ogni emoji, si è ottenuta considerando le sue occorrenze rispetto a quante volte è positiva e negativa nel campione: se il primo valore è maggiore dell'altro, la polarità assegnata è positiva, se invece il secondo valore è maggiore del primo è negativa, altrimenti è neutrale. Una volta ottenuto il dizionario, si è deciso di includere anche le *bad words* comuni della lingua italiana associando ad ognuna la polarità negativa. Si è effettuata questa operazione perché si è riscontrato che alcuni tweet contengono delle parolacce e tutti questi termini vengono utilizzati per offendere. Per calcolare la polarità di ogni tweet, si è proceduto nel seguente modo: prima di tutto si sono contate le occorrenze dei termini positivi e negativi contenuti in ogni messaggio usando come riferimento il dizionario creato, dopodiché si è calcolato il rapporto in percentuale tra il numero di termini della classe minoritaria e maggioritaria. A questo punto, una volta effettuata tale operazione, si verificano delle condizioni che, a seconda di quali vengano soddisfatte, decretano la polarità finale del tweet: se le occorrenze positive superano quelle negative e il rapporto è inferiore o uguale al 20%, il tweet viene considerato positivo, se invece si verifica la stessa condizione però al contrario quindi sono i negativi ad essere in numero maggiore rispetto ai positivi, il tweet viene considerato negativo, altrimenti in tutti gli altri casi è neutrale. Si è deciso di aggiungere una threshold del 20% al calcolo della polarità perché è emerso che la quasi totalità dei tweet raccolti, contengono al loro interno parole di entrambe le polarità e quindi, da teoria, risulterebbero tutti neutrali. Per questo motivo si è deciso di definire un trade-off tra il risultato voluto, ovvero ottenere tweet di diverse polarità e la definizione teorica di neutralità.

	parola	polarità
41	errore	negative
42	dialogo	neutral
43	trucidare	negative
44	espansionista	neutral

Figura 10: Header primo glossario





	Emoji	Unicode codepoint	Occurrences	Position	Negative	Neutral	Positive
0		0x1f602	14622	0.805101	3614	4163	6845
1		0x2764	8050	0.746943	355	1334	6361
2		0x2665	7144	0.753806	252	1942	4950
3		0x1f60d	6359	0.765292	329	1390	4640
4		0x1f62d	5526	0.803352	2412	1218	1896

Figura 11: Header secondo glossario



### 6.2.1 Risultati

Una volta effettuata la *Sentiment Analysis*, si è deciso di realizzare un'infografica interattiva (consultabile al seguente [link](#)) al fine di mostrare i risultati ottenuti. Gli aspetti più rilevanti emergono in particolare modo osservando due plot, rispettivamente un barchart e un linechart che, opportunamente filtrati, mostrano in due modi differenti (totale e giorno per giorno) la polarità dei tweet nei giorni di dicembre antecedenti all'uscita del film e in quelli di gennaio dopo tale uscita. Osservando il barchart (si veda Figura 12) emergono una serie di aspetti interessanti. Il primo è che indipendentemente dal periodo temporale, la quasi totalità dei tweet risultano essere neutrali mentre i positivi sono sempre in numero maggiore dei negativi. Il secondo aspetto rilevante è che i messaggi positivi nel primo periodo rappresentano 17.7% rispetto al totale dei tweet mentre nel secondo periodo 35.2%. Questo implica un evidente aumento di tweet positivi; si può dire quindi che il popolo di Twitter ha valutato positivamente il film, le persone hanno lodato il cambiamento dell'attore pugliese e la sua evoluzione verso temi più impegnativi. Nonostante questo, si evidenzia comunque un aumento dei tweet negativi tra un periodo ed un altro anche se, rispetto ai positivi, l'incremento è minimo: prima dell'uscita del film sono 3.9% mentre dopo 6.2% rispetto al totale. Si può affermare quindi che alcune persone, nonostante la maggioranza abbia apprezzato il film, l'abbiano comunque valutato negativamente. Questo può essere legato da due fattori, il primo strettamente politico mentre l'altro legato allo stile della commedia. Politico perché Checco Zalone, diversamente da quanto si era intuito dal trailer del film, non ritrae l'Italia invasa dagli immigrati ma un Paese visto con gli occhi di uno straniero dove l'attenzione si focalizza sulle storie che ogni migrante potrebbe aver vissuto prima di approdare in Italia. Questo ovviamente viene visto in modo negativo da tutte le persone che sostengono determinati partiti politici di destra (Lega, Casa Pound etc.) perché il regista, così facendo, trasmette indirettamente un messaggio di appoggio e sostegno agli immigrati. Il secondo fattore invece è puramente legato al film la cui comicità, risulta essere completamente diversa rispetto a quella di tutte le altre pellicole in cui emerge lo "zalonismo"; qui lo stile della commedia non è più semplice, elementare dove vengono esasperati gli stereotipi dell'italiano medio ma una comicità che fa riflettere, un equilibrio tra la leggerezza e il dramma vissuto dai migranti. Il cambio di rotta del comico e regista pugliese quindi, può aver contribuito all'aumento di tweet negativi in quanto tutti si aspettavano il classico film ma in realtà hanno assistito a tutt'altro sicché molte persone sono rimaste deluse.

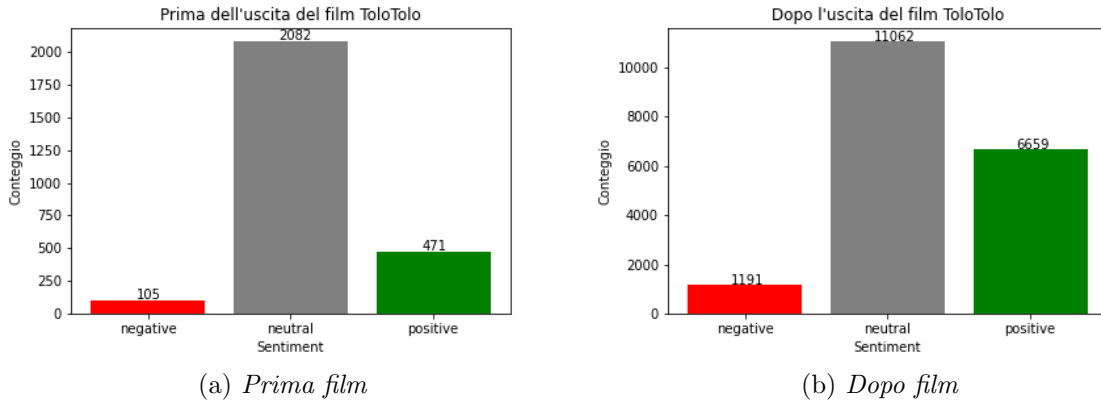


Figura 12: Barchart polarità tweet nei due periodi temporali

Osservando il linechart (si veda Figura 13) invece, emergono tre aspetti interessanti. Il primo è che i tweet negativi assumono trend simili tra un periodo ed un altro, per entrambi si evidenzia un picco e una decrescita dell'andamento fino al raggiungimento di un *plateau*. Il secondo aspetto rilevante è che il trend dei messaggi positivi, sia nei giorni antecedenti che successivi all'uscita del film, risulta essere altalenante: nel primo periodo l'andamento è caratterizzato da un picco il 28 dicembre, una discesa nelle fasi successive infine un ulteriore picco il 30 dicembre seguito da una nuova discesa, nel secondo periodo invece, il trend è caratterizzato da due picchi, uno il 2 gennaio e l'altro il 5 gennaio, seguiti entrambi da un andamento decrescente. L'ultimo aspetto interessante che emerge osservando i linechart, è che il giorno in cui si registra il numero più alto di tweet positivi e negativi risulta essere il 2 gennaio. Questo può essere motivato da due fattori; il primo riguarda l'uscita del film che è avvenuta il giorno prima, di conseguenza molte persone sono andate immediatamente a vederlo e subito dopo hanno twittato le loro impressioni, il secondo fattore invece è legato alle festività natalizie, periodo in cui la gente, essendo a casa dal lavoro o dall'università, è più propensa ad andare a vedere un film tanto atteso come *ToloTolo*.

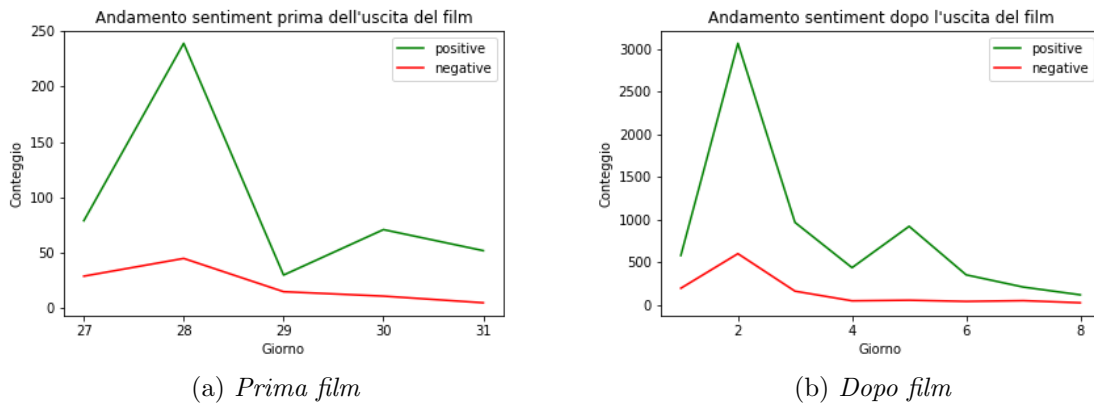


Figura 13: Linechart polarità tweet nei due periodi temporali

## 6.3 Named Entity Recognition

Come si è visto nell'introduzione del Capitolo 6, la *NER* è risultata utile sia per effettuare correttamente la *tokenization* sia per rispondere all'ultima domanda di ricerca, ovvero individuare le persone famose italiane maggiormente citate all'interno dei tweet. Tale task è stato svolto utilizzando una *convolutional neural network* multi-task (*'it-core-news-sms'*) resa disponibile dalla libreria Python *spaCy*. Questo modello, addestrato su un corpus italiano estratto da *WikiNer*, restituisce per ogni tweet i token candidati ad appartenere ad un'entità fornendo anche informazioni riguardo al tipo (persona, luogo etc.). Per lo scopo di questo progetto si è deciso di tenere in considerazione solo output riferiti all'entità di tipo persona. Una volta ottenuti i risultati, si è deciso di filtrarli con un dizionario (si veda Tabella 2) realizzato ad hoc in modo da eliminare i falsi positivi (token individuati che nella realtà non appartengono ad alcuna entità) restituiti dal modello; questo passaggio risulta necessario in quanto la struttura grammaticale dei tweet si rivela diversa dalla struttura grammaticale di riferimento utilizzata per addestrare la rete neurale.

Elemento	Fonte
deputati italiani	<a href="http://camera.it">camera.it</a>
ministri e sottosegretari italiani	<a href="http://governo.it">governo.it</a>
attori italiani dal 1910 ad oggi	<a href="http://wikipedia.org_attori">wikipedia.org_attori</a>
registi famosi italiani	<a href="http://wikipedia.org_registi">wikipedia.org_registi</a>
romanzieri, saggisti, poeti italiani	<a href="http://wikipedia.org_scrittori">wikipedia.org_scrittori</a>

Tabella 2: Contenuto dizionario

### 6.3.1 Risultati

Osservando il barchart (si veda Figura 14) emergono una serie di aspetti interessanti. Il primo è che tra i 10 personaggi più citati all'interno dei tweet, 5 sono politici incluso il presidente del Consiglio Giuseppe Conte; ciò sottolinea ancora una volta che la discussione intorno al film è legata soprattutto a questioni politiche. Oltre a questo si evidenzia che la persona più citata in assoluto risulta essere Matteo Salvini, confermando così l'assegnamento della label "antisalvinismo" a una o più comunità. Tale risultato non sorprende visto che il leader della Lega ha sempre manifestato la sua contrarietà a politiche migratorie troppo permissive; inoltre il politico, dopo l'uscita del trailer del film, ha addirittura proposto di far diventare Checco Zalone senatore a vita pensando che la pellicola mostrasse i "problemi" causati dagli immigrati. Un altro aspetto interessante che emerge è che la seconda persona più citata non risulta essere un politico ma una scrittrice e blogger che si chiama Michela Murgia. Il motivo di tale fattore è che molti utenti hanno retweettato un post dove viene citata la scrittrice. L'ultimo aspetto curioso che traspare dai risultati è che tra i 10 personaggi più citati si trovano 4 attori; i primi due fanno parte della storia del cinema italiano, Franco Fantasia e Alberto Sordi mentre gli altri, Massimo Boldi e Roberto Benigni, sono famosi per motivi diversi. Massimo Boldi è conosciuto per i suoi cinepanettoni mentre Roberto Benigni, oltre che per i suoi monologhi teatrali dirompenti, è famoso per aver interpretato e diretto *La vita è bella*. Il motivo per cui ci siano tutti questi attori può essere legato al fatto che molte persone abbiano trovato somiglianze tra il film di Checco Zalone e altre pellicole

appartenenti alla storia del cinema italiano come *Un americano a Roma* e *La vita è bella*; il primo film perché Alberto Sordi così come Checco non ha timore di apparire sgradevole, è sempre diretto con le persone mentre la seconda pellicola perché così come *ToloTolo*, narra una tragedia umana in chiave comica.

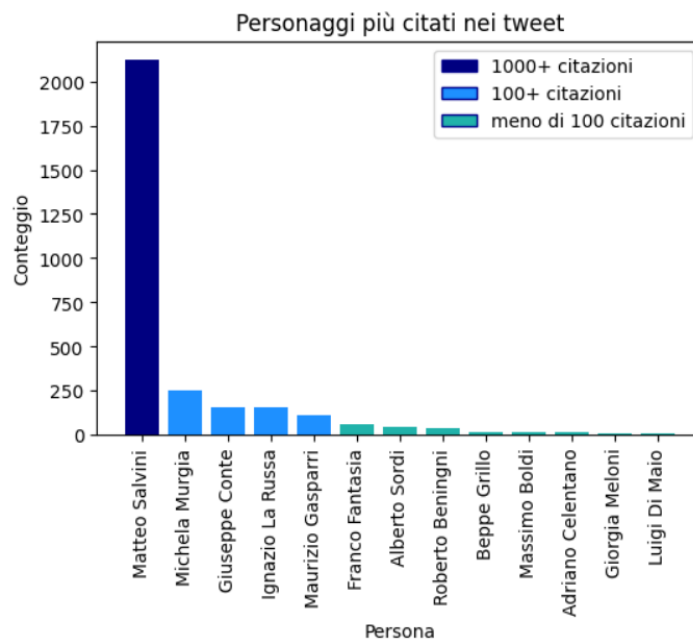


Figura 14: Risultati NER

## 7 Conclusioni

Il progetto si è incentrato sul monitoraggio e l'analisi dei tweet inerenti a *ToloTolo* con l'obiettivo di rispondere ad alcune domande di ricerca definite nella fase iniziale. Per quanto riguarda la prima domanda, i risultati ottenuti evidenziano un'estensione della rete sociale fino al 2 gennaio, giorno in cui si registra l'espansione massima del grafo. Considerando la seconda domanda di ricerca, ovvero individuare gli hub, emerge che la maggior parte di questi nodi risultano essere personaggi pubblici che hanno un certo seguito sui social, in special modo si evidenzia una forte presenza di giornalisti e opinionisti. Oltre a questo si riscontra anche una differenza tra coloro che sono gli hub prima e dopo l'uscita del film, in particolare il nodo più rilevante, dal 27 dicembre al 1 gennaio risulta essere *Capezzone* mentre dopo 1 gennaio è *Giorgiolaporta*. Considerando la terza domanda di ricerca, si individuano 20 comunità ben suddivise tra di loro. Dall'assegnamento di una di 8 label per ciascun gruppo, emerge che le categorie "giornalismo" e "comicità" risultano essere quelle a cui sono assegnate il maggior numero di comunità. Per quanto riguarda la quarta domanda di ricerca, i risultati ottenuti evidenziano, da un punto di vista globale, un non cambiamento del sentiment degli utenti prima e dopo l'uscita del film visto che i tweet neutrali risultano essere in numero maggiore di tutti gli altri nei due periodi. Analizzando invece solo i tweet positivi e negativi, si riscontra una crescita considerevole per entrambe le polarità anche se, l'incremento più evidente, riguarda la polarità positiva. Considerando l'ultima domanda di ricerca, dai risultati emerge che i politici risultano essere le persone più citate nei tweet. Per quanto riguarda gli sviluppi futuri di questo progetto, si registrano in particolar modo operazioni di

*refine* delle tecniche usate sia per effettuare la *Sentiment Analysis* che la *Named Entity Recognition*. Considerando il primo task, si potrebbero distinguere i tweet oggettivi da quelli soggettivi e concentrarsi su questi ultimi per capire quali sono gli aspetti del film (trama, personaggi etc.) piaciuti di più o di meno al pubblico. Considerando la *NER* invece, un'operazione di *refine* potrebbe essere utilizzare un corpus più specifico al caso in esame per addestrare la rete neurale, nello specifico impiegare un insieme di testi provenienti dai social media in modo tale da poter catturare al meglio le entità citate nei vari tweet e quindi migliorare i risultati ottenuti. Alla fine da questo progetto si possono trarre due conclusioni; la prima è che le discussioni sui social si sono incentrate per la maggior parte su questioni di natura politica e non su eventuali critiche cinematografiche alla pellicola, la seconda è che Checco Zalone è riuscito ancora una volta a "centrare il bersaglio" visto che la maggior parte delle persone ha giudicato positivamente il film.

## Riferimenti bibliografici

- [1] S. Wasserman, K. Faust *et al.*, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.
- [2] G. Marchina, "Cinema, checco zalone e il real time marketing: due facce della stessa medaglia."
- [3] G. Canova, "Tolo tolo – il marketing di luca medici."
- [4] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social networks*, vol. 32, no. 3, pp. 245–251, 2010.
- [5] Wikipedia, "Modularity (networks)." [Online]. Available: [https://en.wikipedia.org/wiki/Modularity\\_\(networks\)](https://en.wikipedia.org/wiki/Modularity_(networks))
- [6] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.
- [7] L. Waltman and N. J. Van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *The European physical journal B*, vol. 86, no. 11, p. 471, 2013.
- [8] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [9] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 473–480.
- [10] H. Schmid, M. Baroni, E. Zanchetta, and A. Stein, "The enriched treetagger system," in *proceedings of the EVALITA 2007 workshop*, 2007.
- [11] I. Russo, F. Frontini, and V. Quochi, "OpeNER sentiment lexicon italian - LMF," 2016, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational

Linguistics ”A. Zampolli”, National Research Council, in Pisa. [Online]. Available: <http://hdl.handle.net/20.500.11752/ILC-73>

- [12] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, “Emoji sentiment ranking 1.0,” 2015, slovenian language resource repository CLARIN.SI. [Online]. Available: <http://hdl.handle.net/11356/1048>