Twitter Users Characterization durante il Black Friday 2019

Francesco D'Auria[1]
[1]Università degli Studi di Salerno, Fisciano (SA) f.dauria37@studenti.unisa.it

Abstract
Tutti quanti avremo sentito parlare, più di una volta, del Black Friday, soprattutto di recente: ma cos'è e da dove proviene? Il fenomeno del Black Friday arriva direttamente dall'America e, letteralmente, significa "Venerdì nero". Così fu definito il giorno successivo a quello del Ringraziamento, celebrato tipicamente nel quarto giovedì di novembre, tradizionalmente dedicato, negli Stati Uniti, allo shopping natalizio. Secondo alcuni, l'origine di tale nome proviene dalle annotazioni in nero sui libri contabili dei commercianti, che rappresentavano i ricavi, in netta contrapposizione con il rosso delle annotazioni, riguardanti le perdite. Dal punto di vista economico, questo è un giorno molto importante, poiché costituisce un valido indicatore sia sulla predisposizione dei consumatori agli acquisti, sia, indirettamente, sulla capacità di spesa dei suddetti, tanto da essere attentamente osservato dagli analisti finanziari di tutto il mondo. In Italia il fenomeno si è diffuso solo ultimamente ma è già in aumento. Complice la massiccia campagna pubblicitaria messa a punto dai brand di moda, dai grandi colossi dell'e-commerce come Amazon, eBay. I social sono diventati il mezzo per veicolare promozioni e notizie sugli articoli e sempre più vengono utilizzati per vere e proprie manovre di marketing. Gli utenti sono sempre più utilizzatori e partecipi di questi avvenimenti e vogliono dire la propria su prodotti in vendita e recensioni sugli e-commerce. I social stanno diventando lo specchio della società e studiarli significa in qualche modo studiare gli aspetti che cambiano la società e come gli utenti cambiano in base ai periodi dell'anno, ad esempio. Studiare l'andamento, vuol dire anche poi prevederne lo stato tra dieci o venti anni. Per questa ricerca, sono stati raccolti i tweets con hashtag #blackfriday associandolo di volta in volta con l'hashtag ufficiale dell'e-commerce che si voleva studiare. C'è stata quindi la raccolta dei dati e poi un'attenta analisi sulla caratterizzazione degli utenti del social network Twitter che hanno interagito nel periodo in questione e poi si è analizzata la distribuzione di tweet per utente, e non solo, e se fosse possibile associarla e descriverla come una distribuzione Power Law.

Keywords
Twitter, tweet, black friday, amazon, ebay, walmart, wish, gearbest, ecommerce.

1 INTRODUZIONE

La data del Black Friday di quest'anno è venerdì 29 novembre 2019: un giorno da segnarsi in agenda per non perdere neppure un'occasione delle tante proposte nell' arco della giornata. Il Black Friday, tradotto letteralmente venerdì nero, è un giorno interamente dedicato agli acquisti, sia online che offline, che cade sempre il giorno successivo al Ringraziamento, meglio conosciuto come Thanksgiving. Si tratta di una giornata speciale pensata per gli acquisti che nasce in America, ma che ormai gode di fortuna anche da noi in Italia, oltre che nel resto del mondo. Il venerdì nero, soprattutto in America, è un giorno a cui ci si prepara in anticipo per non farsi trovare impreparati, e per arrivare agguerriti al momento in cui si apriranno le porte dei tanti negozi di shopping presi letteralmente d'assalto dai clienti che non vedono l'ora di fare affari d'oro!

I social sempre più vengono studiati perché riflettono (1) la società e le tendenze, e vengono anche utilizzati per fare previsioni in ambito marketing. In questa attività di ricerca, si è studiato il social network più aperto e libero in ambito developer, Twitter, che mette a disposizione API ed endpoint per ottenere dati sui tweets. Si è voluto dare una caratterizzazione agli utenti che hanno 'digitato parole' in quei giorni in particolare dal 18 novembre 2019 al 7 dicembre 2019, in termini di genere, provenienza degli utenti, numero di tweets inviati in quel periodo, numeri di followers, numero di amici, numero di stati pubblicati e quando l'utente ha creato l'account. Soprattutto il primo dato servirà per stabilire se anche il numero di tweets per utente ed altri dati come numero di followers e seguiti seguono la famigerata distribuzione Power Law[1].

2 IDENTIFICAZIONE DEL DATA SET

2.1 Twitter API

Per raccogliere i dati relativi ai tweets, Twitter mette a disposizione API ed endpoints, utilizzabili attraverso linguaggi di programmazione. Twitter è ciò che sta accadendo nel mondo e ciò di cui la gente sta parlando in questo momento. Si può accedere a Twitter tramite il Web o un dispositivo mobile. Per condividere le informazioni su Twitter nel modo più ampio

[1] Le legge di potenza (o Power Law) altri non è che una funzione matematica con un andamento esponenziale caratterizzato da un parametro che può essere sia negativo che positivo e che nella sua forma più generale si esprime come $f(x) = ax^k$

possibile, il social network fornisce anche ad aziende, sviluppatori e utenti un accesso programmatico ai dati di Twitter attraverso le API (interfacce di programmazione delle applicazioni).
Ad alto livello, le API sono il modo in cui i programmi per computer "dialogano" tra loro in modo che possano richiedere e fornire informazioni. Questo viene fatto consentendo a un'applicazione software di invocare ciò che è noto come endpoint: un indirizzo che corrisponde a un tipo specifico di informazioni che forniamo (gli endpoint sono generalmente unici come i numeri di telefono). Twitter consente l'accesso a parti del servizio tramite API per consentire alle persone di creare software che si integra con Twitter, come una soluzione che aiuta un'azienda a rispondere al feedback dei clienti su Twitter.

I dati di Twitter sono unici rispetto ai dati condivisi dalla maggior parte delle altre piattaforme social perché riflettono le informazioni che gli utenti scelgono di condividere pubblicamente. La piattaforma API offre un ampio accesso ai dati pubblici di Twitter che gli utenti hanno scelto di condividere con il mondo. Supportano anche API che consentono agli utenti di gestire le proprie informazioni su Twitter non pubbliche (ad es. Messaggi diretti) e di fornire queste informazioni agli sviluppatori che hanno autorizzato a farlo (2).

2.2 Data Collection

Quando qualcuno vuole accedere alle API, è tenuto a registrare un'applicazione. Per impostazione predefinita, le applicazioni possono accedere solo alle informazioni pubbliche su Twitter. Alcuni endpoint, come quelli responsabili dell'invio o della ricezione di messaggi diretti, richiedono ulteriori autorizzazioni prima che possano accedere alle informazioni dell'utente. Queste autorizzazioni non sono concesse per impostazione predefinita.

Per l'uso di queste API, bisogna scrivere frammenti di codice con l'uso di uno dei tanti linguaggi di programmazione che possono utilizzare queste interfacce, come Python oppure R. I tweets vengono gestiti come se fossero realmente degli oggetti, così come anche gli utenti, e quindi vengono registrati e salvati in una sorta di base di dati, che viene utilizzata proprio come tale, attraverso i Cursor.

L'oggetto Tweet è composto dai campi in **Tabella 1**.

Attributo	Tipo	Descrizione
created_at	String	Ora e giorno in cui il tweet viene creato
id	Int64	Identificatore univoco del tweet
id_str	String	Rappresentazione in stringa dell'identificatore univoco
text	String	Testo del tweet in formato UTF-8
all_hashtag	List	Lista degli hashtag associati al tweet
user	User object	Utente che ha postato il tweet
retweeted_status	Tweet	Indica se *text* è troncato oppure no
followers_count	Int	Numero di followers del tweet

Tabella 1 Lista degli attributi di un oggetto Tweet identificati per questa ricerca

Per raccogliere effettivamente i dati dei tweets è stato utilizzato il linguaggio Python e la libreria *tweepy*. Tweepy è una libreria Python per l'accesso all'API di Twitter. È ottimo per la semplice automazione e la creazione di bot per Twitter. Tweepy viene utilizzato in molte situazioni:

- Ricevere tweet dalla sequenza temporale.
- Creazione ed eliminazione di tweet.
- Seguire e non seguire più gli utenti.

Prima di poter fare qualsiasi cosa, dobbiamo parlare di autenticazione. Prima di poter utilizzare Tweepy (3) e più in generale qualsiasi libreria che permette di raccogliere i dati, bisogna far creare alla dashboard di Twitter delle credenziali e prima ancora configurare un ambiente in cui eseguire un'app configurata ad hoc, il tutto seguendo dei semplicissimi passaggi.

Una volta avute queste credenziali utilizziamole per autenticarci prima di richiedere i dati.

```
#create authentication for accessing Twitter
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

#initialize Tweepy API
api = tweepy.API(auth, wait_on_rate_limit=True)
```

Richiediamo e salviamo i dati dei tweets filtrandoli sia per data che per hashtag, concatenando con l'operatore logico 'AND' tutti gli hashtag. Per questa ricerca, all'hashtag #blackfriday, è stato di volta in volta concatenato con uno tra i seguenti hashtag

- #amazon
- #ebay
- #gearbest
- #walmart
- #wish

per effettuare un'analisi approfondita sul Black Friday e su quale fosse l'e-commerce più utilizzato per fare compere[2], prima, durante e dopo il periodo di acquisti caratterizzato dal venerdì nero. Sono stati presi in considerazione soprattutto i tweets ed i retweets in lingua inglese, e poi raccolti e suddivisi in diversi file.

```
with open('%s.csv' % (fname), 'wb') as file:

        w = csv.writer(file)

        #write header row to spreadsheet
        w.writerow(['timestamp', 'tweet_text', 'username', 'all_hashtags', 'followers_count'])

        #for each tweet matching our hashtags, write relevant info to the spreadsheet
        for tweet in tweepy.Cursor(api.search, q=hashtag_phrase+'', fromDate="2019-11-
18", toDate="2019-12-08",
                                   lang='en', tweet_mode='extended').items(100000000000000000):
                                   if 'retweeted_status' in tweet._json:
                                     retweet_text = 'RT @ ' + api.get_user(tweet.retweeted_status.us
er.id_str).screen_name+' '+tweet._json['retweeted_status']['full_text'].replace('\n',' ')
                                      w.writerow([tweet.created_at, retweet_text, tweet.user.screen_n
ame, [e['text'] for e in tweet._json['entities']['hashtags']], tweet.user.followers_count])
                                   else:
                                    w.writerow([tweet.created_at, tweet.full_text.replace('\n',' '),
 tweet.user.screen_name, [e['text'] for e in tweet._json['entities']['hashtags']], tweet.user.follow
ers_count])
```

Alla fine dell'esecuzione avremo un file con estensione .csv contenente i dati dei tweets differenziando i retweets dai tweets concatenando *'RT @ username'* al testo del retweet.
Avuto a disposizione i dati dei tweets, si dovranno raccogliere informazioni circa gli utenti che hanno digitato quel testo. È stato costruito un frammento di codice che fa anche questo, lo stesso scritto in Python. Prima però bisogna definire gli attributi dell'oggetto User.
Così come i tweets vengono modellati da oggetti di tipo Tweet, anche gli utenti vengono modellati da oggetti di tipo User, costituiti da attributi che delineano le caratteristiche di un utente. Nella **Tabella 2** vengono riportati gli attributi considerati in questa ricerca.

Attributo	Tipo	Descrizione
id	Int64	Identificatore univoco dell'utente
id_str	String	Rappresentazione in stringa dell'identificatore univoco
name	String	Il nome dell'utente, non necessariamente un nome di persona
screen_name	String	Un alias con cui si identifica l'utente, spesso considerato come username
location	String	Posizione definita dall'utente, non necessariamente una posizione.
url	String	Un URL fornito dall'utente in associazione con il suo profilo

[2] La lista degli e-commerce più popolari è stata presa da https://disfold.com/top-e-commerce-sites-us/

description	String	La stringa UTF-8 definita dall'utente che descrive il proprio account.
protected	Boolean	Quando è *true*, indica che questo utente ha scelto di proteggere i propri tweets.
verified	Boolean	Se è *true*, indica che l'utente ha un account verificato.
followers_count	Int	Numero di seguaci che questo account ha attualmente
friends_count	Int	Numero di utenti che questo account segue (AKA i "followings")
listed_count	Int	Numero di liste pubbliche in cui questo utente è membro
favourites_count	Int	Numero di tweet che questo utente ha apprezzato nella vita dell'account.
statuses_count	Int	Numero di tweets (compresi i retweets) emessi dall'utente
created_at	String	Data e ora in cui l'account utente è stato creato su Twitter.
profile_image_url_https	String	URL basato su HTTPS che punta all'immagine del profilo dell'utente
default_profile	Boolean	Se è *true*, indica che l'utente non ha alterato il tema o lo sfondo del proprio profilo utente.
default_profile_image	Boolean	Se è *true*, indica che l'utente non ha caricato la propria immagine di profilo e viene utilizzata invece un'immagine di default

Tabella 2 Lista degli attributi di un oggetto User identificati per questa ricerca

Per avere una visione più completa delle caratteristiche degli utenti, bisogna sapere il genere. Twitter non ha previsto un campo del genere nella struttura dell'oggetto User. Per differenziare gli utenti, quindi, si è optato per l'utilizzo delle librerie Python, che dato in input una stringa di testo, restituiscono, con una buona probabilità, il genere della persona che ha scritto quel testo, oppure capiscono da nome se una persona è di sesso femminile o maschile, confrontando con il proprio dataset di nomi americani ed inglesi.
Le due librerie utilizzate sono *gender_guesser* (4) e *Genderizer* (5).
La prima prende in input il nome della persona, preso in questo caso dall'attributo del tweet, e restituisce il genere tra maschile, femminile, quasi-maschile, quasi-femminile ed androgino.

```
d = gender.Detector()
print(d.get_gender(tweet.name))
```

Il funzionamento della seconda libreria è leggermente diverso, ha anche essa il riconoscimento del genere attraverso il nome, ma cosa più interessante, il riconoscimento attraverso il testo. Quindi ogni utente ha scritto una descrizione nel proprio profilo e verrà analizzata con questa libreria affinché si possa risalire al genere della persona che l'ha scritta. Al metodo *detect* diamo in input il nome, preso in questo caso dall'utente, così come viene preso anche l'attributo *description*.
Le due librerie restituiscono il genere della persona sottoforma di stringa oppure "unknown", se non si è riuscito ad indentificare, con una buona probabilità il genere.

```
print(Genderizer.detect(text=user.name))
print(Genderizer.detect(text=user.description))
```

Uniamo queste tre informazioni, e consideriamo

- il genere ottenuto dal *gender_guesser* se è diverso da "unknown"
- il genere ottenuto dal *Genderizer* in riferimento al nome se è diverso da "uknown"
- il genere ottenuto dall'analisi della descrizione, attraverso *Genderizer*, dell'utente anche se è uguale ad "unknown"

3 ANALISI DEI RISULTATI

3.1 Analisi dei Tweets

Una prima analisi che è stata effettuata è quella dei tweets. Vengono presi in considerazione gli attributi testo del tweet per il conteggio di tweets per user e gli hashtag associati. I dati raccolti sono stati raggruppati in grafici per una semplice visualizzazione.

3.1.1 Conteggio dei tweets e retweets per utente

Il conteggio dei tweets per utente è servito per studiare il fenomeno della scrittura di stati virtuali e se fosse possibile descrivere la distribuzione del numero di tweet per utente con la distribuzione Power Law, in cui ci sono poco utenti che scrivono tanti tweets e retweets e tanti utenti che scrivono poco o quasi per niente.

La **Figura 1** riassume questi dati.

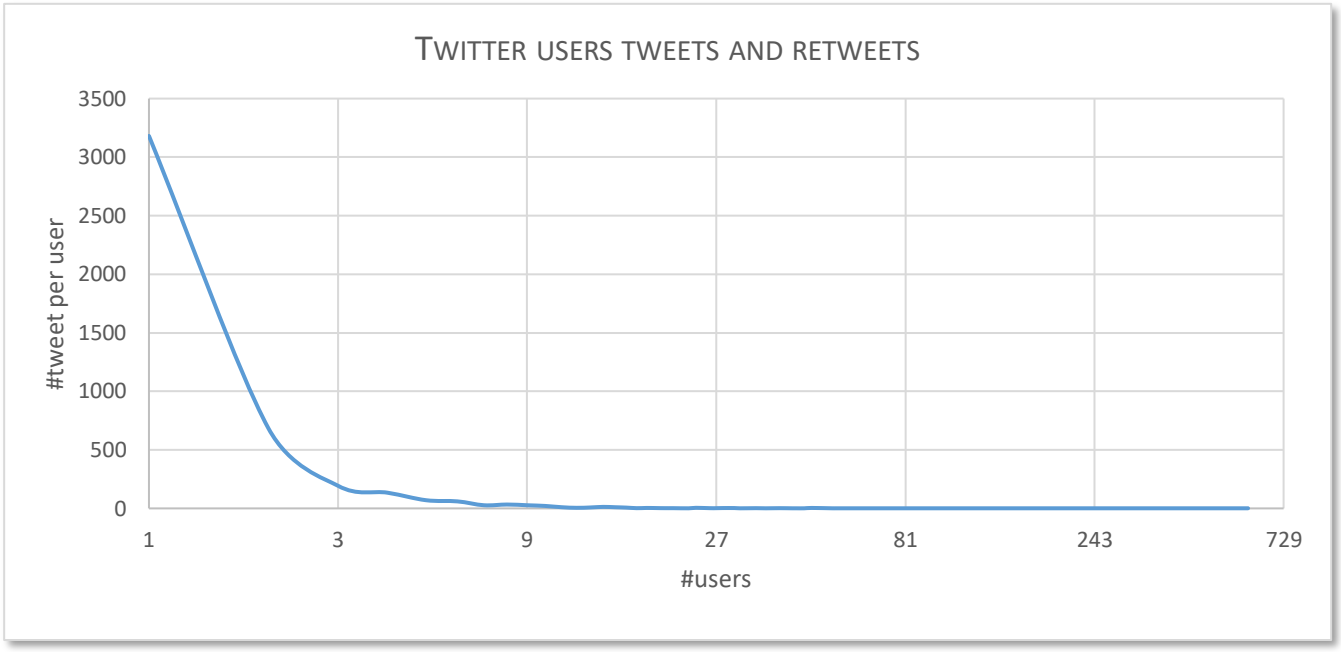

Figura 1 Distribuzione del numero di tweets e retweets per utente.

Anche il numero di retweets ed il numero di tweets segue lo stesso andamento come mostrato nella **Figura 2** e nella **Figura 3**.

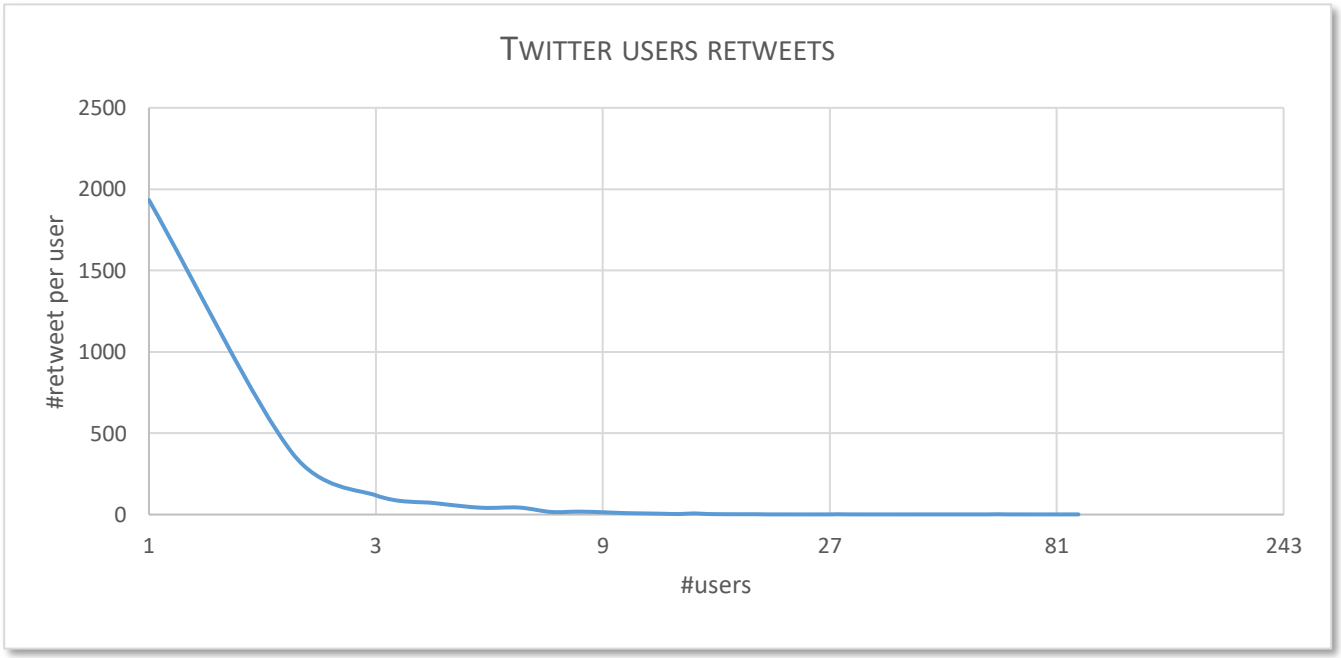

Figura 2 Distribuzione del numero di retweets per utente

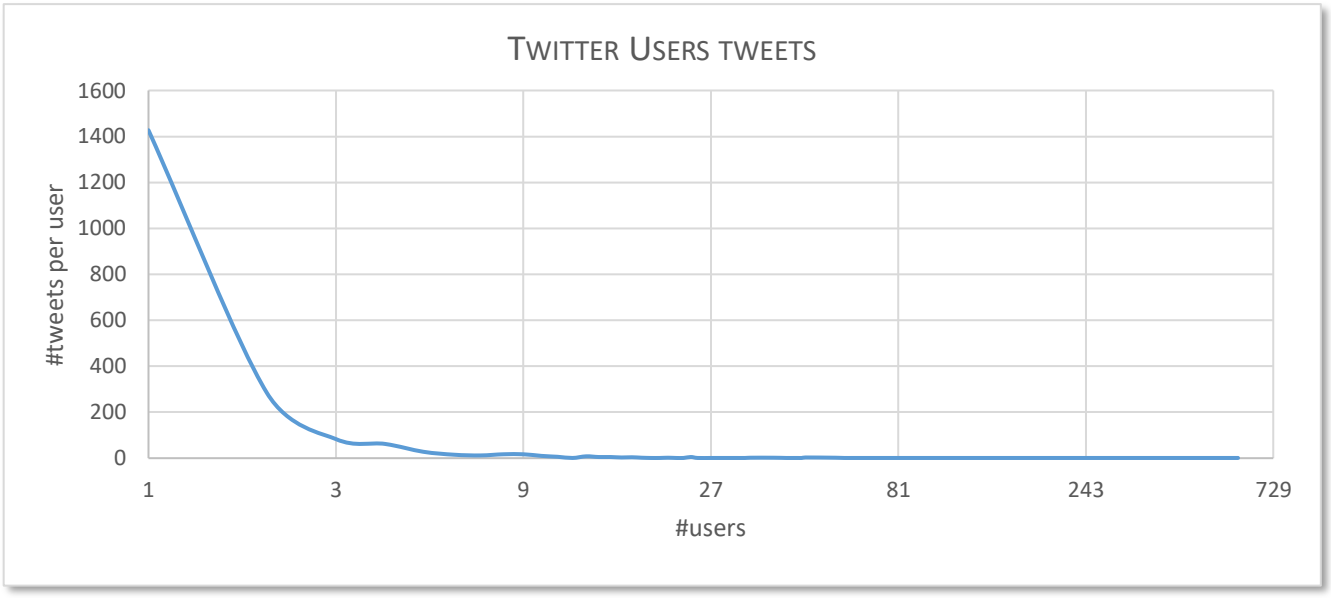

Figura 3 Distribuzione del numero di tweets per utente

3.1.2 Conteggio dei tweets e retweets per utente su scala log-log

Il modo più "naturale" in cui ci aspettiamo di misurare una grandezza è una distribuzione centrale, dove la maggior parte delle osservazioni si concentrano attorno a un valore tipico, il valore medio, e decadono molto velocemente in modo simmetrico ai due estremi. Molte distribuzioni, sia presenti in natura che costruite dall'uomo, hanno queste caratteristiche, che possono essere descritte, con un livello più o meno accettabile di approssimazione, con tutta una famiglia di modelli statistici, come la distribuzione gaussiana, la distribuzione binomiale e la quella di Poisson che, dati una serie di parametri, è in grado di spiegare e descrivere la popolazione.
Non è detto, però, che le osservazioni si distribuiscano in modo simmetrico attorno ad un valore tipico. Alcune grandezze, infatti, si esprimono in popolazioni in cui si ha la maggioranza delle osservazioni su valori bassi delle ascisse, ma presentano anche diverse osservazioni molto a destra, in una lunga coda, con osservazioni che possono raggiungere valori della ascissa molto grandi, anche di diversi ordini di grandezza più grandi della media, proprio come le distribuzioni raffigurate nelle **Figure 1-2-3**.

Il grafico in scala log-log mostra chiaramente un andamento assimilabile a quello di una retta per buona parte del suo dominio. Non lo è, apparentemente, nella parte più a destra, dove, però, vige per tutti e tre i grafici seguenti.

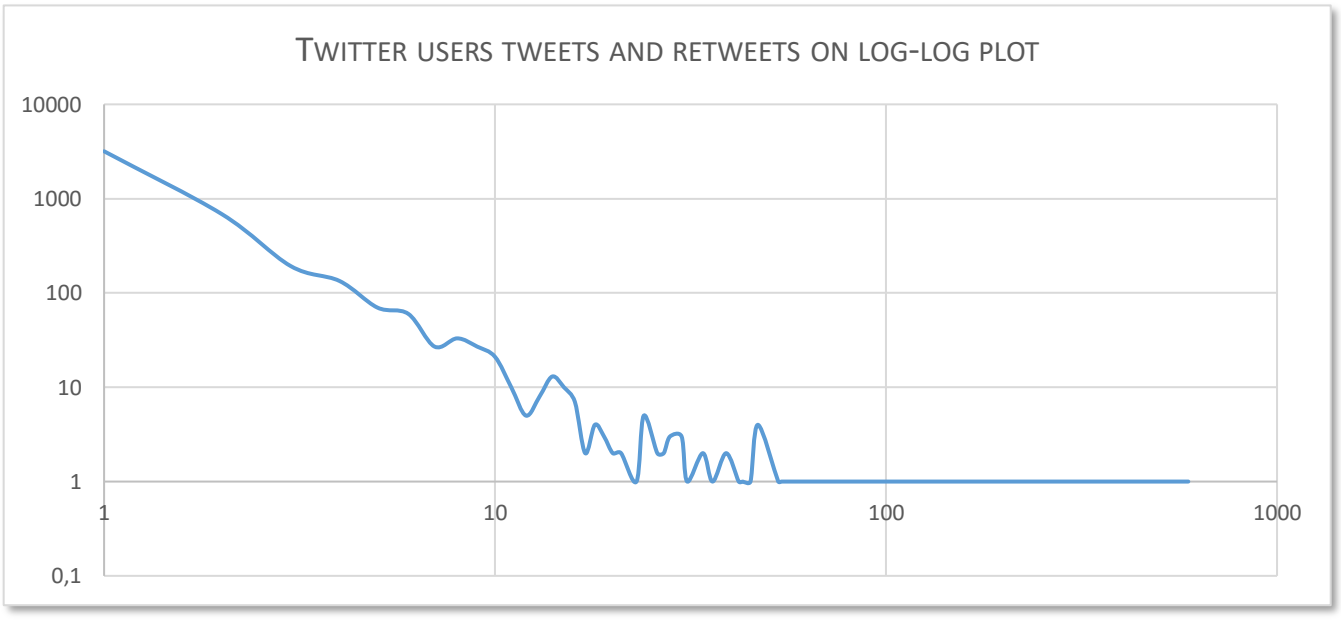

Figura 4 Distribuzione del numero di tweets e retweets per utente su log-log plot.

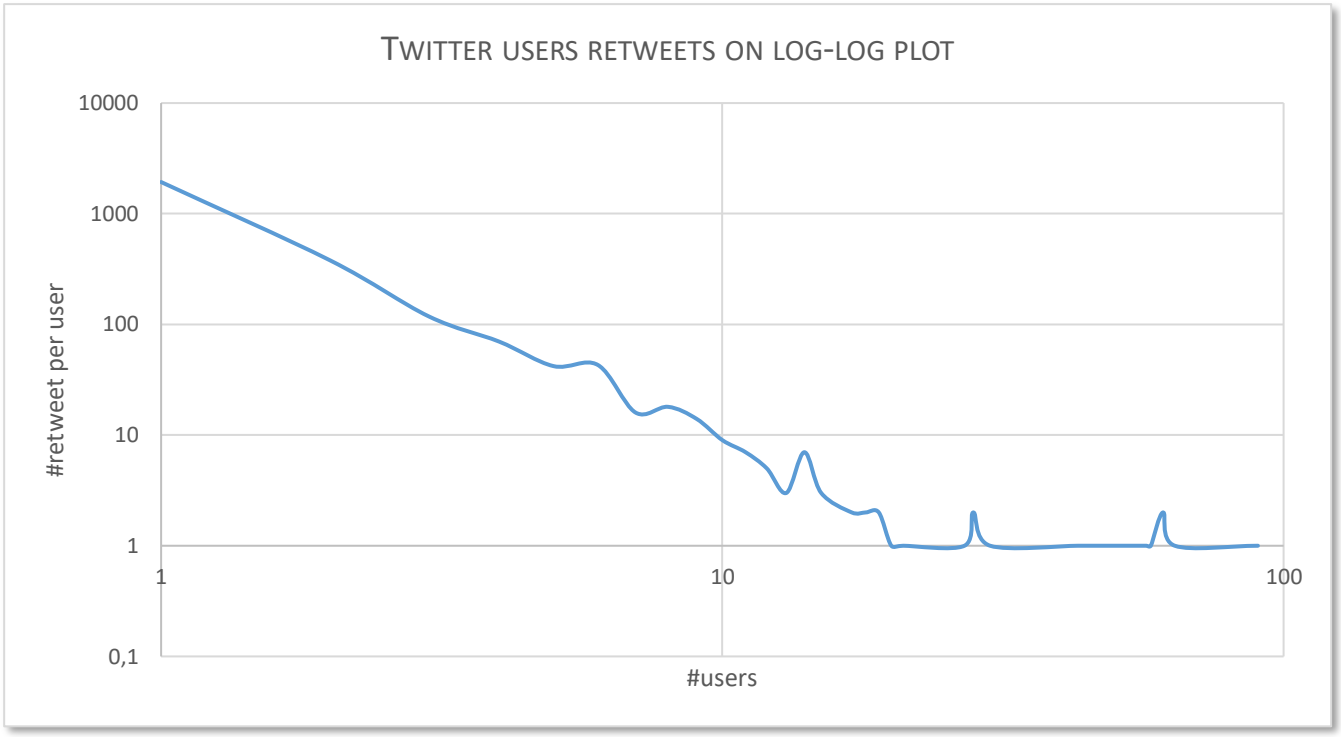

Figura 5 Distribuzione del numero di retweets per utente su log-log plot

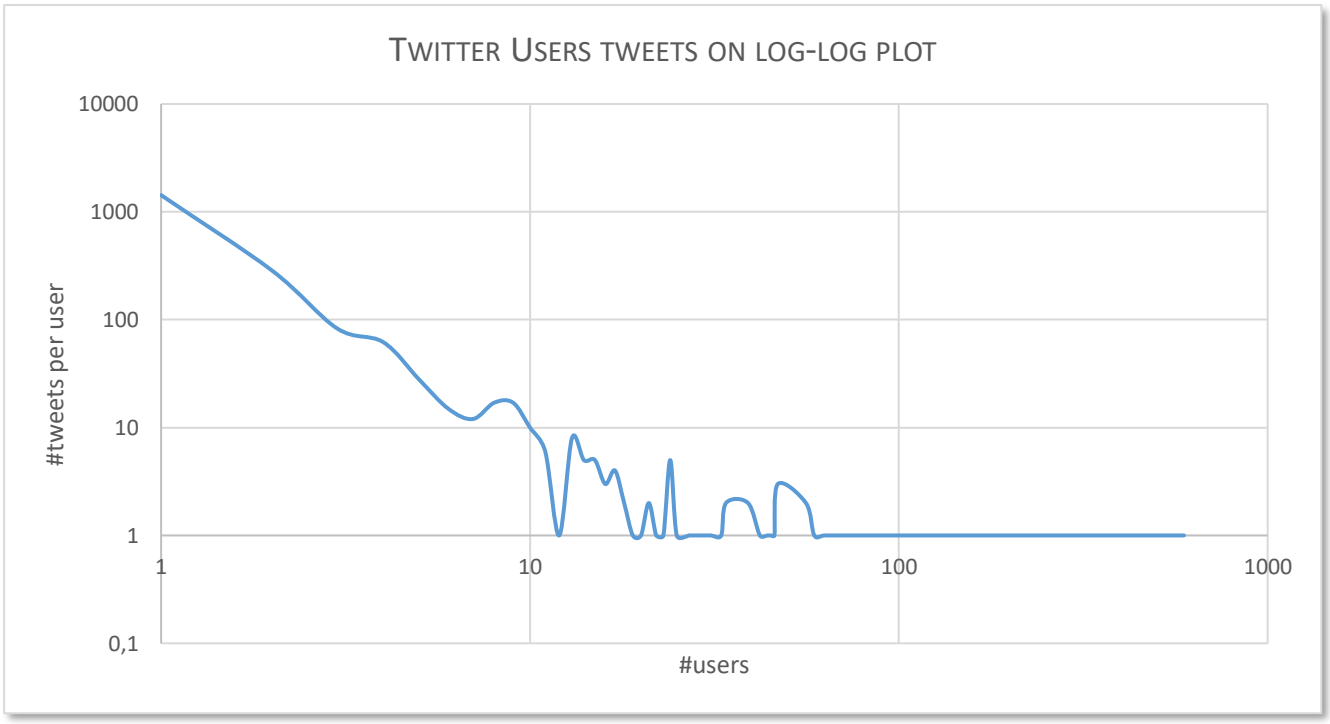

Figura 6 Distribuzione del numero di tweets per utente su log-log plot

3.1.3 Conteggio dei tweets e retweets per e-commerce

Lo studio che si svolto è incentrato anche sull'individuazione dell'e-commerce più popolare prima durante e dopo il Black Friday.

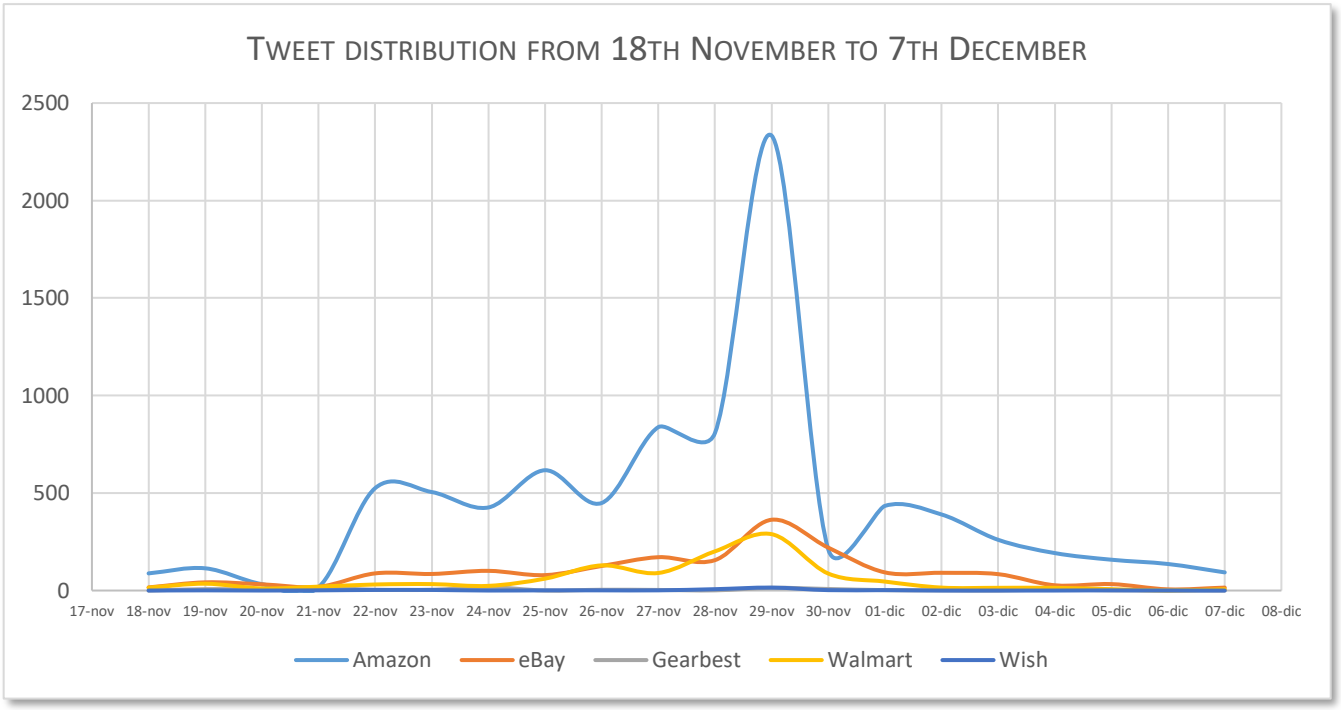

Figura 7 Distribuzione dei tweets dal 18 novembre al 7 dicembre suddivisi per e-commerce

Come si nota nella **Figura 7**, vi è un certo distacco tra gli altri e-commerce che rendono, in una visione complessiva, Amazon l'e-commerce più popolare tra i cinque studiati.

3.2 Analisi degli utenti

Attraverso la raccolta dei tweets, oltre all'analisi degli stessi, sono state individuati anche gli utenti che hanno interagito. Come già discusso nel paragrafo 3.1, il numero di tweets per utente segue una distribuzione Power Law. Per l'analisi degli utenti, sono stati presi in considerazione il genere, la posizione dell'utente registrata nell'account, la verifica dell'utente da parte di Twitter, il numero di seguaci, il numero di seguiti, il numero di tweets totali che ha apprezzato, il numero di tweets e retweets emessi durante la vita dell'account, la data di creazione dell'account, se l'utente ha o meno il profilo di default (cioè se non è stato cambiato tema e sfondo del profilo) e se il profilo ha l'immagine di default.
Twitter non mette a disposizione molti dati degli utenti tipo il genere e la data di nascita. In questo lavoro di ricerca, il genere è stato dedotto da altri attributi, mentre la data di nascita e quindi l'età non è stata presa in considerazione, però sarebbe stato il dato più considerevole per la classificazione degli utenti.

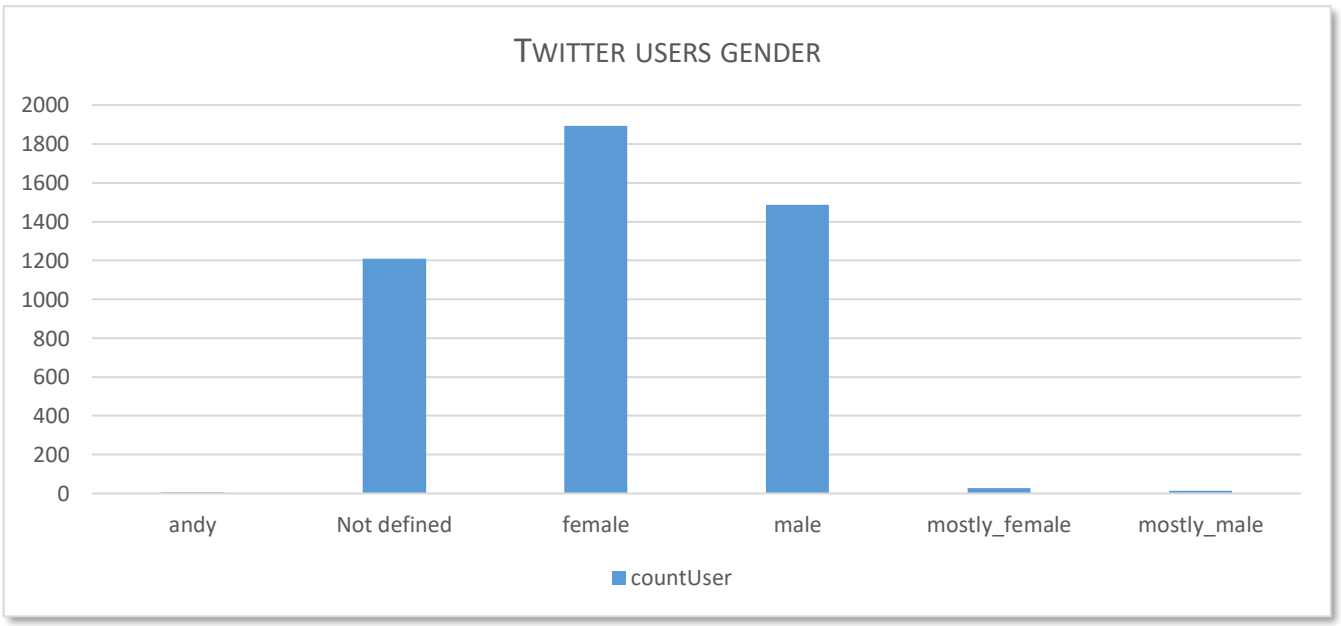

Figura 8 Genere degli utenti

Dopo l'identificazione del genere, descritta nel paragrafo 2.2, dalla **Figura 8** si evince che la maggioranza degli utenti che hanno interagito nel periodo di riferimento sono per lo più di sesso femminile.

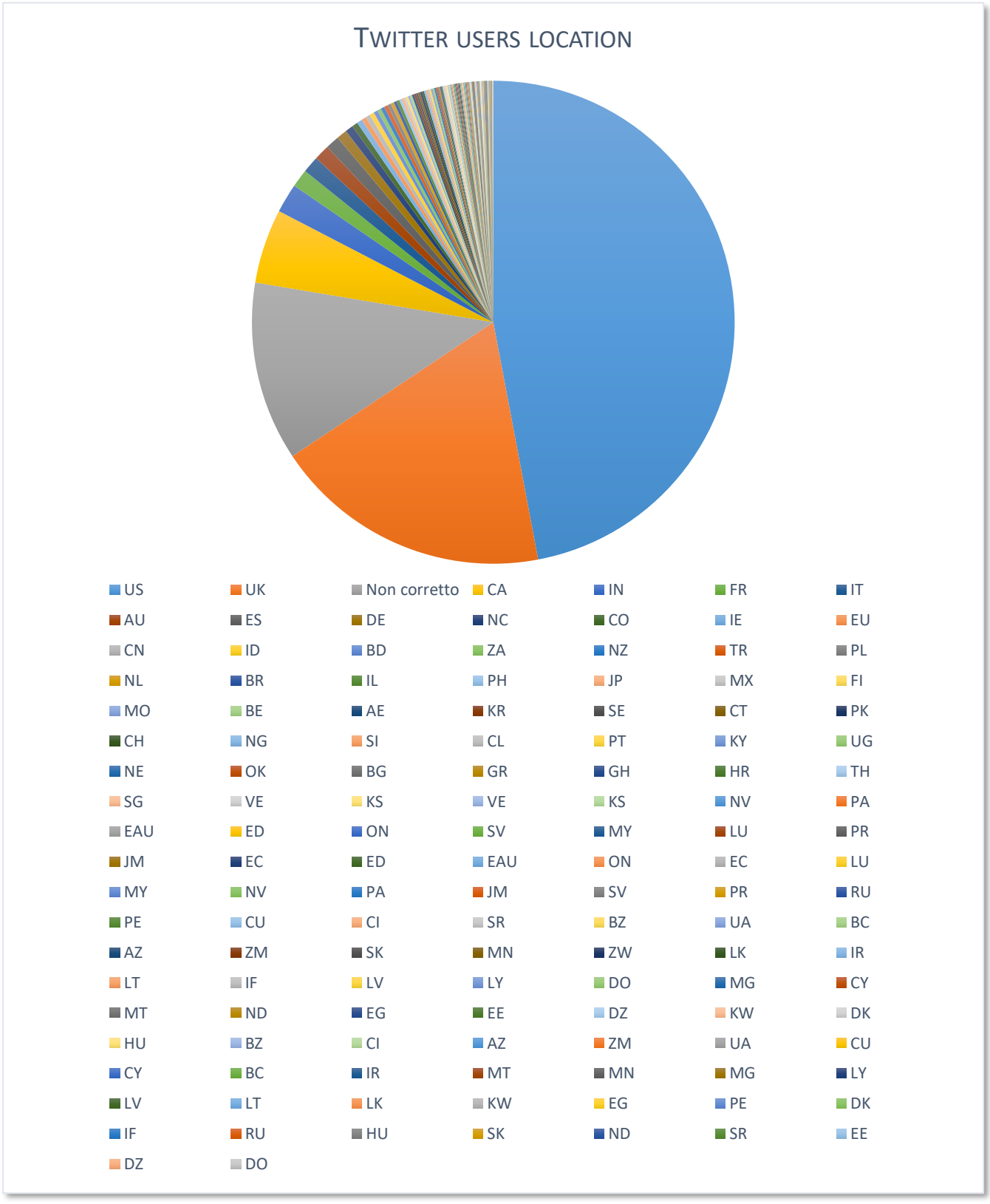

Figura 9 Distribuzione degli utenti negli stati

Dalla **Figura 9** possiamo dedurre che la maggior parte degli utenti proviene dagli Stati Uniti, al secondo posto troviamo il Regno Unito, al terzo posto troviamo il Canada. Per una buona parte degli utenti, non si è riusciti ad identificare il paese di appartenenza poiché l'attributo *location* non prevede limitazioni.

Seguono grafici che riassumono le analisi effettuate e descritte in precedenza.

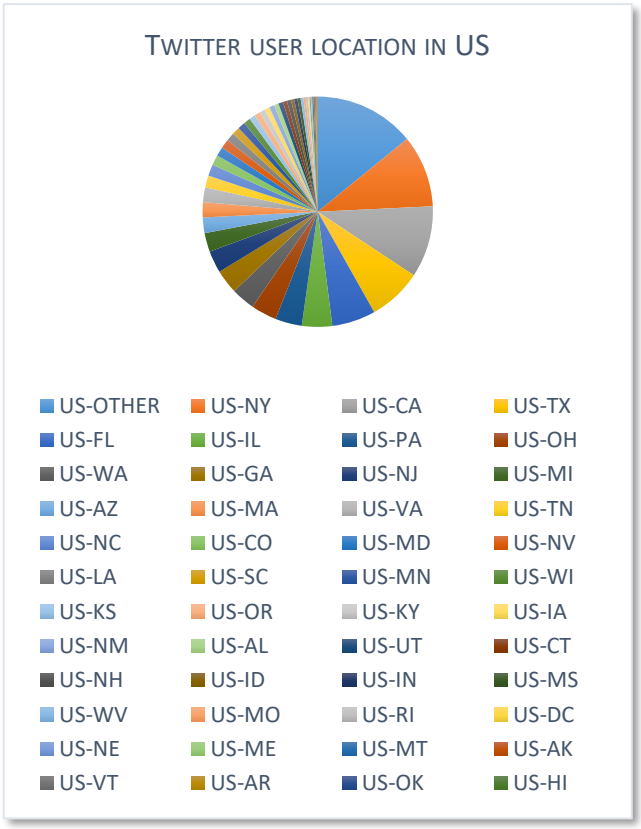

Figura 10 Utenti dagli Stati Uniti suddivisi per stati americani

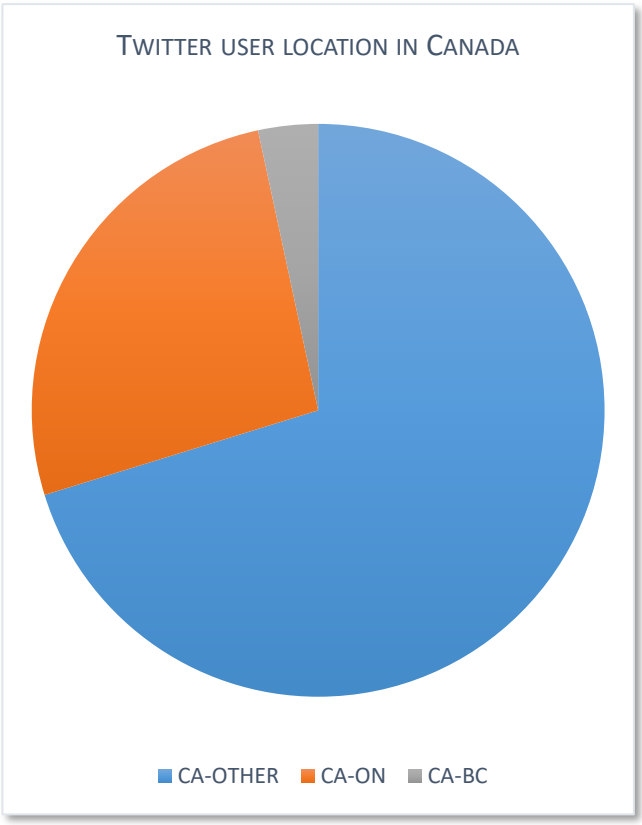

Figura 11 Utenti del Canada suddivisi per regioni canadesi

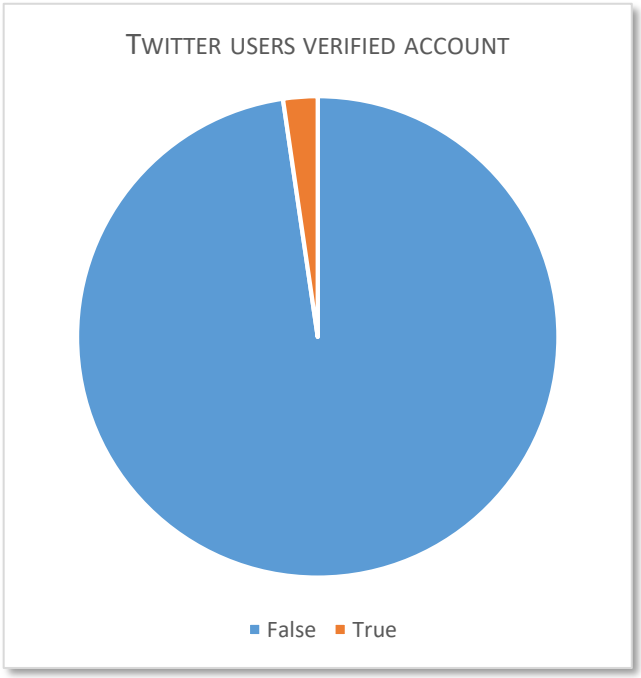

Figura 12 Numero di utenti verificati

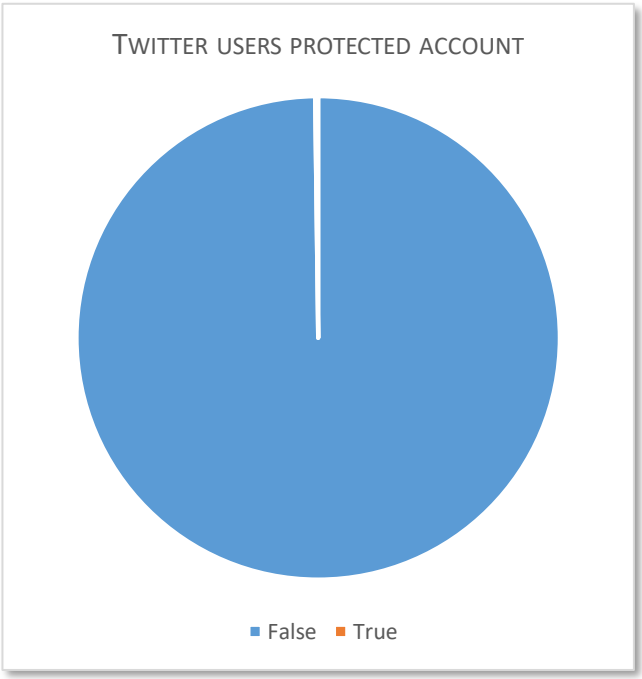

Figura 13 Numero di utenti protetti

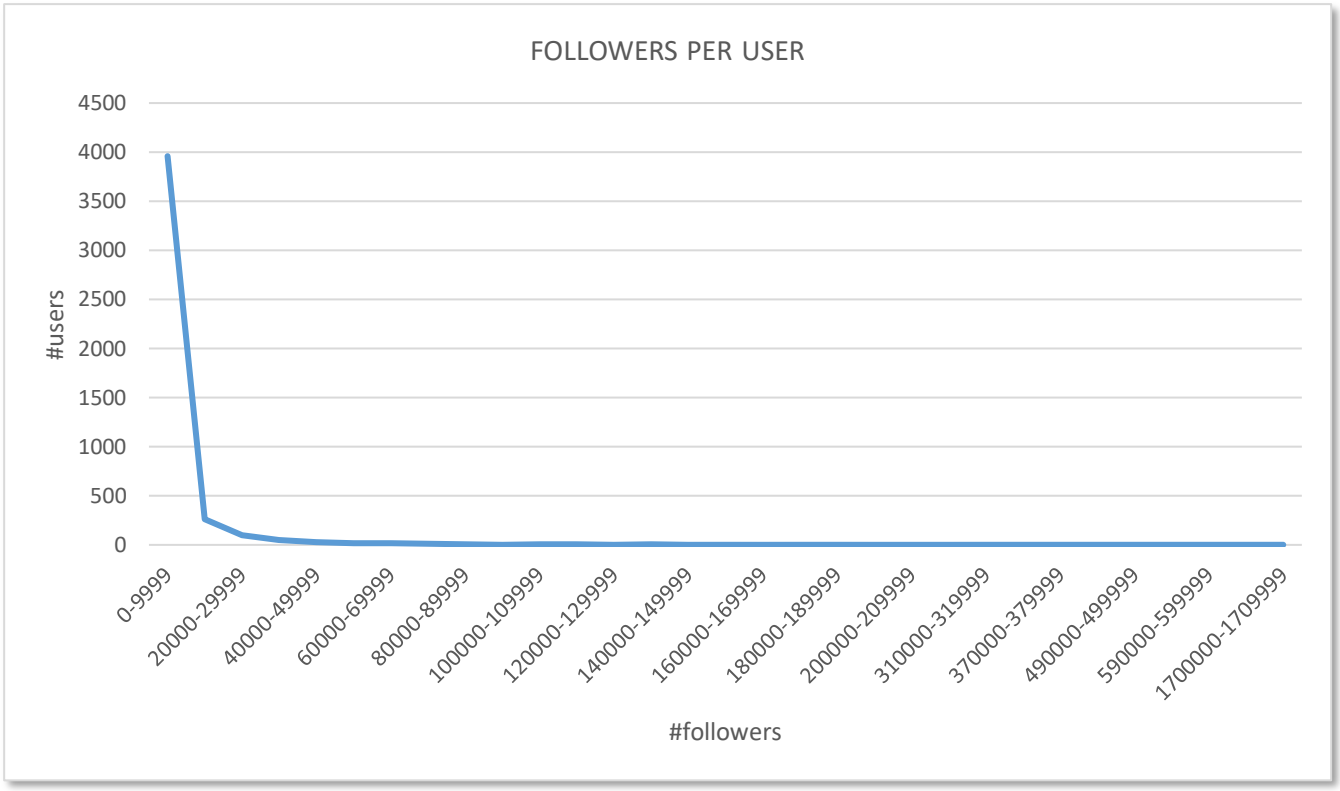

Figura 14 Numero di utenti suddivisi per numero di followers

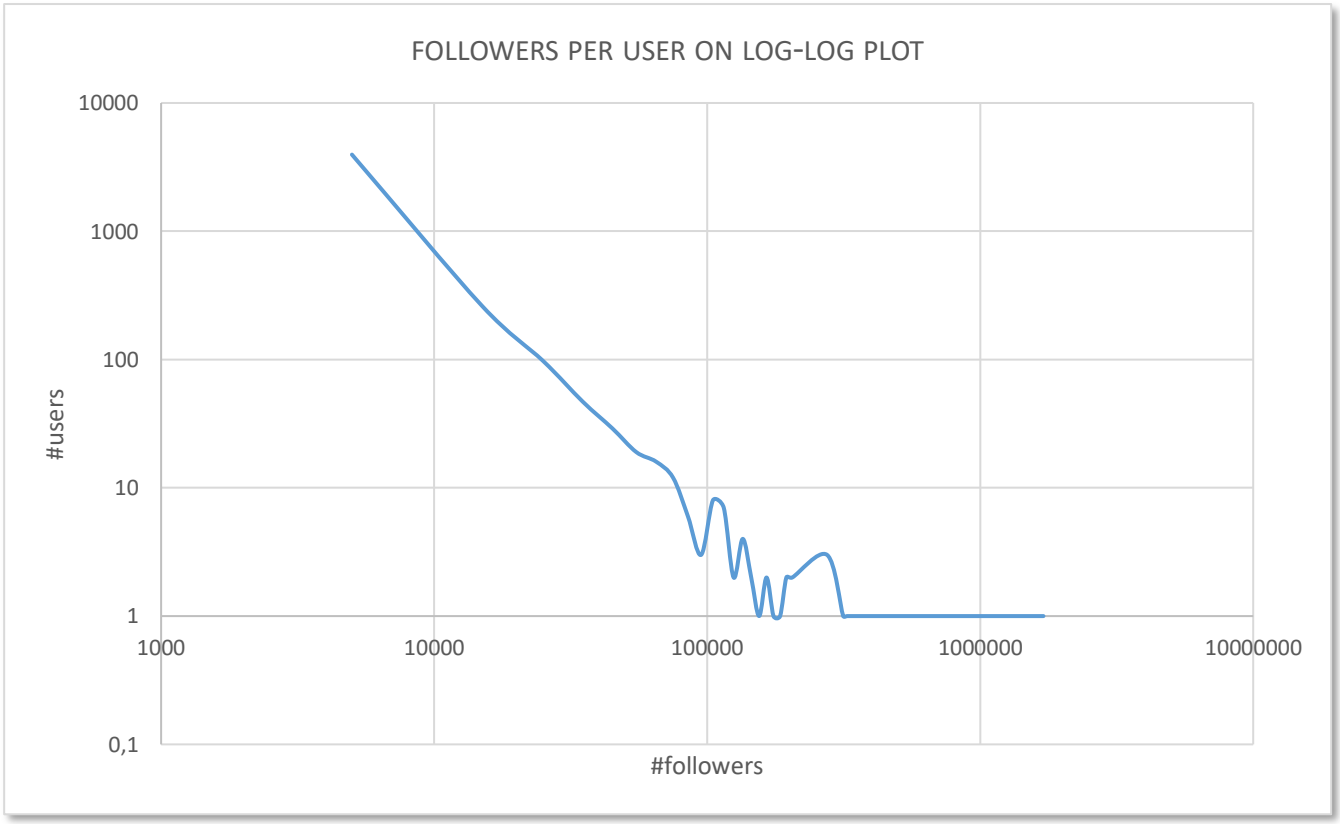

Figura 15 Numero di utenti suddivisi per numero di followers su log-log plot

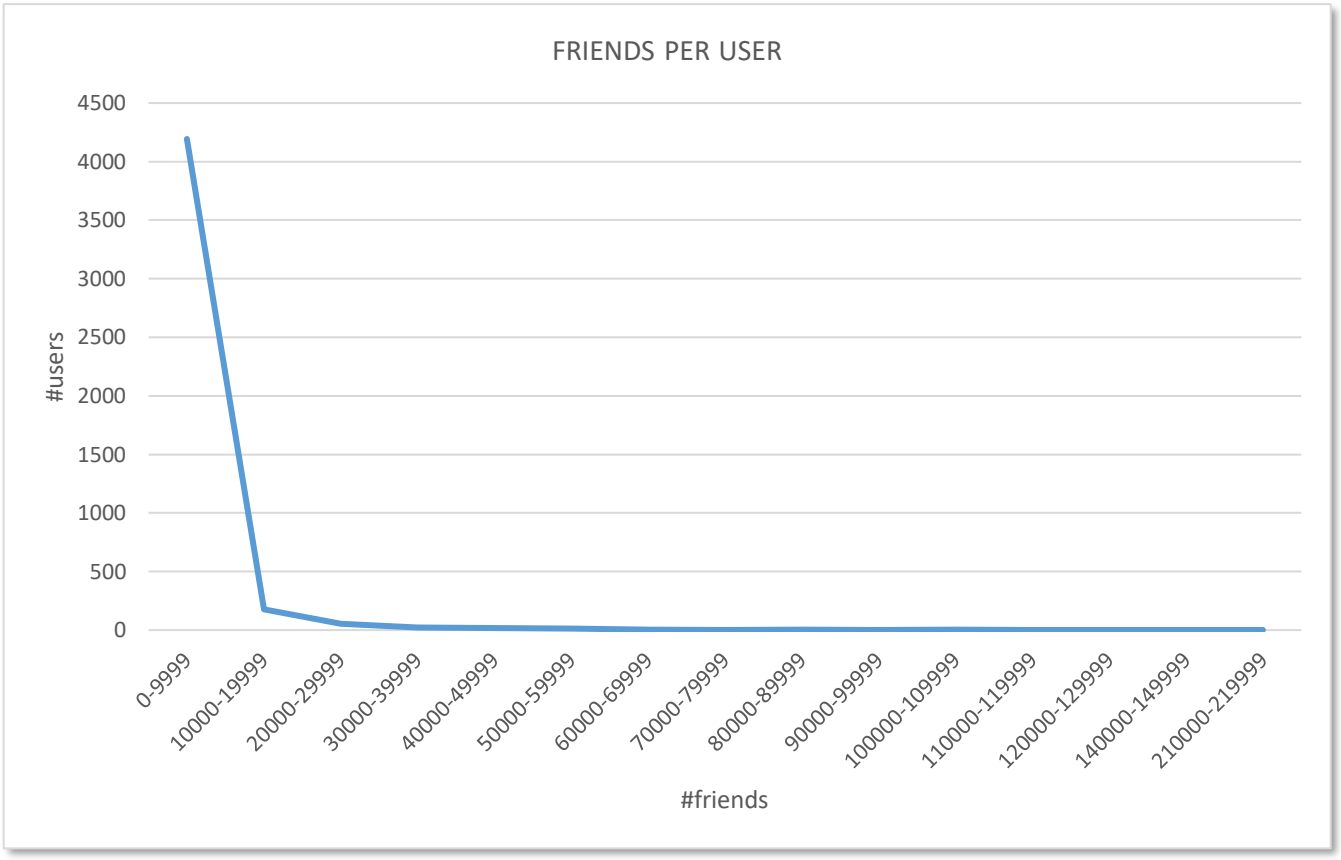

Figura 16 Numero di utenti suddivisi per numeri di seguiti

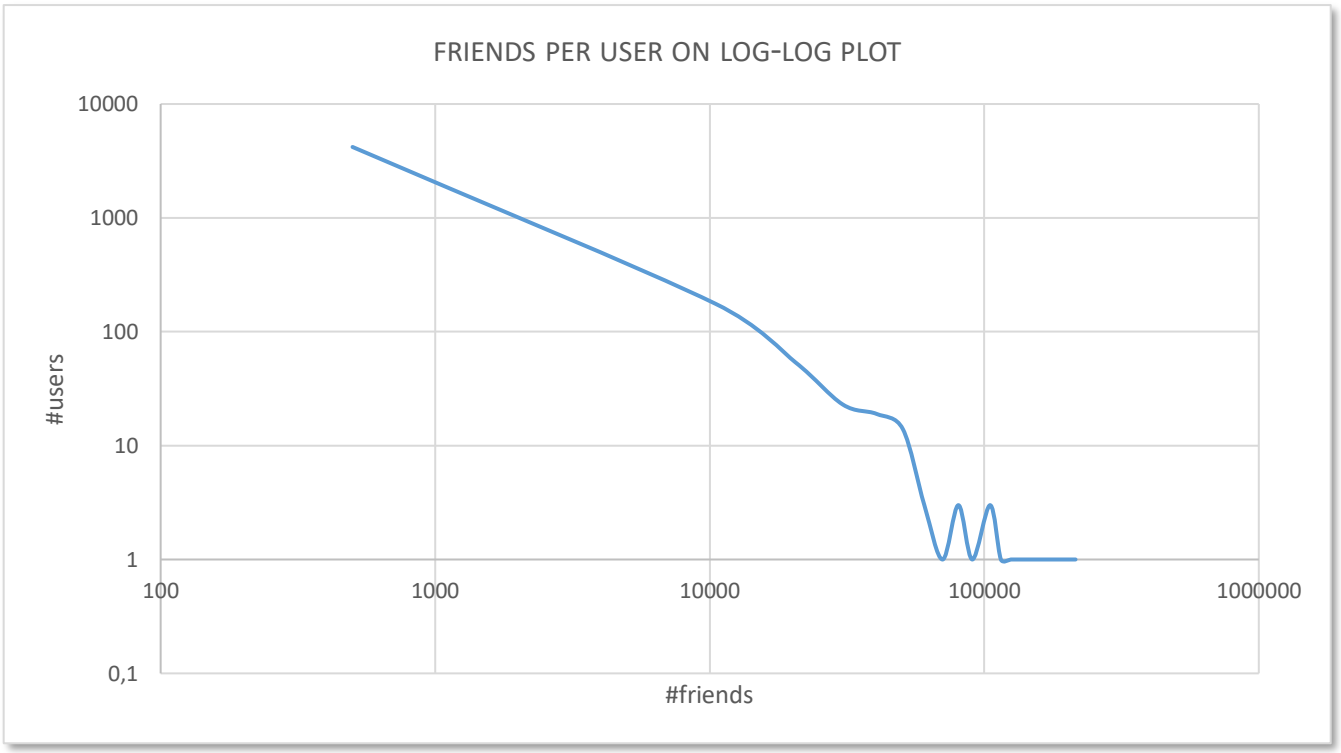

Figura 17 Numero di utenti suddivisi per numeri di seguiti su log-log plot

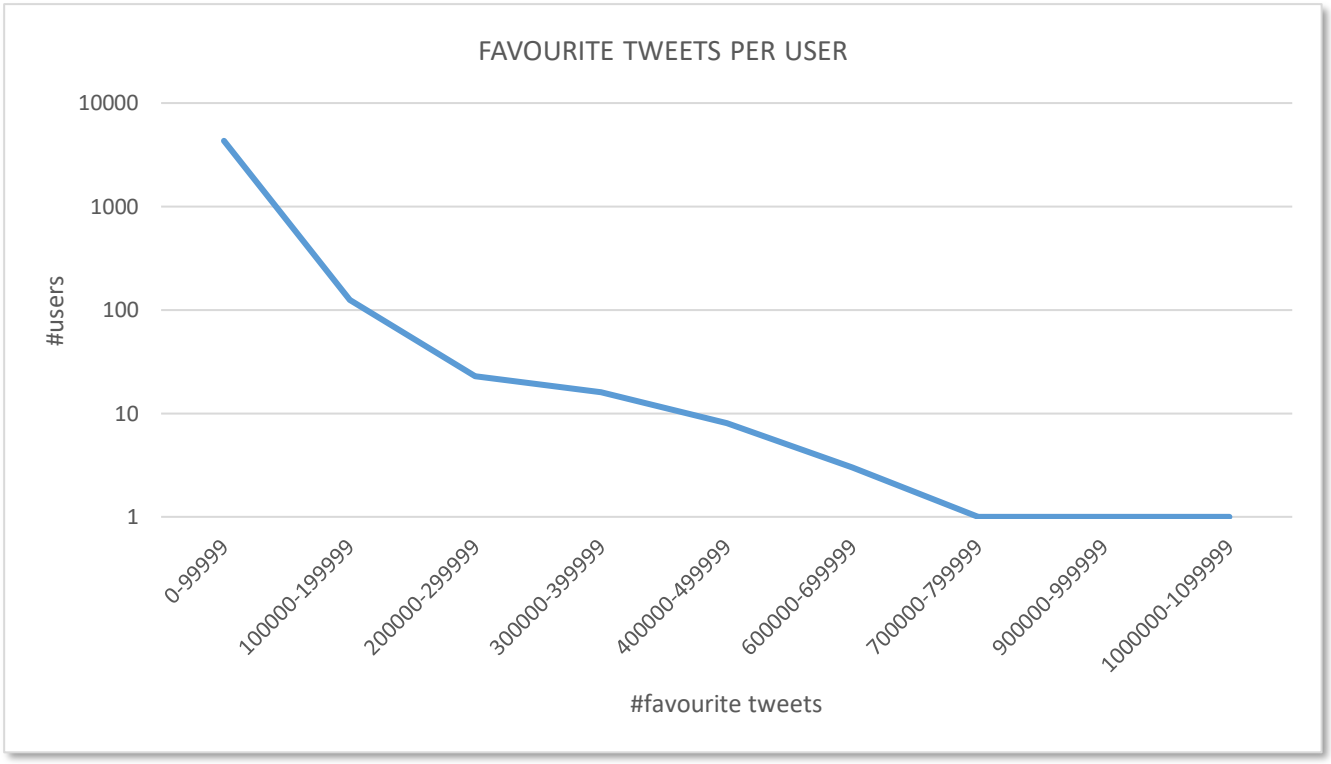

Figura 18 Numero di utenti suddivisi per numero di tweet apprezzati

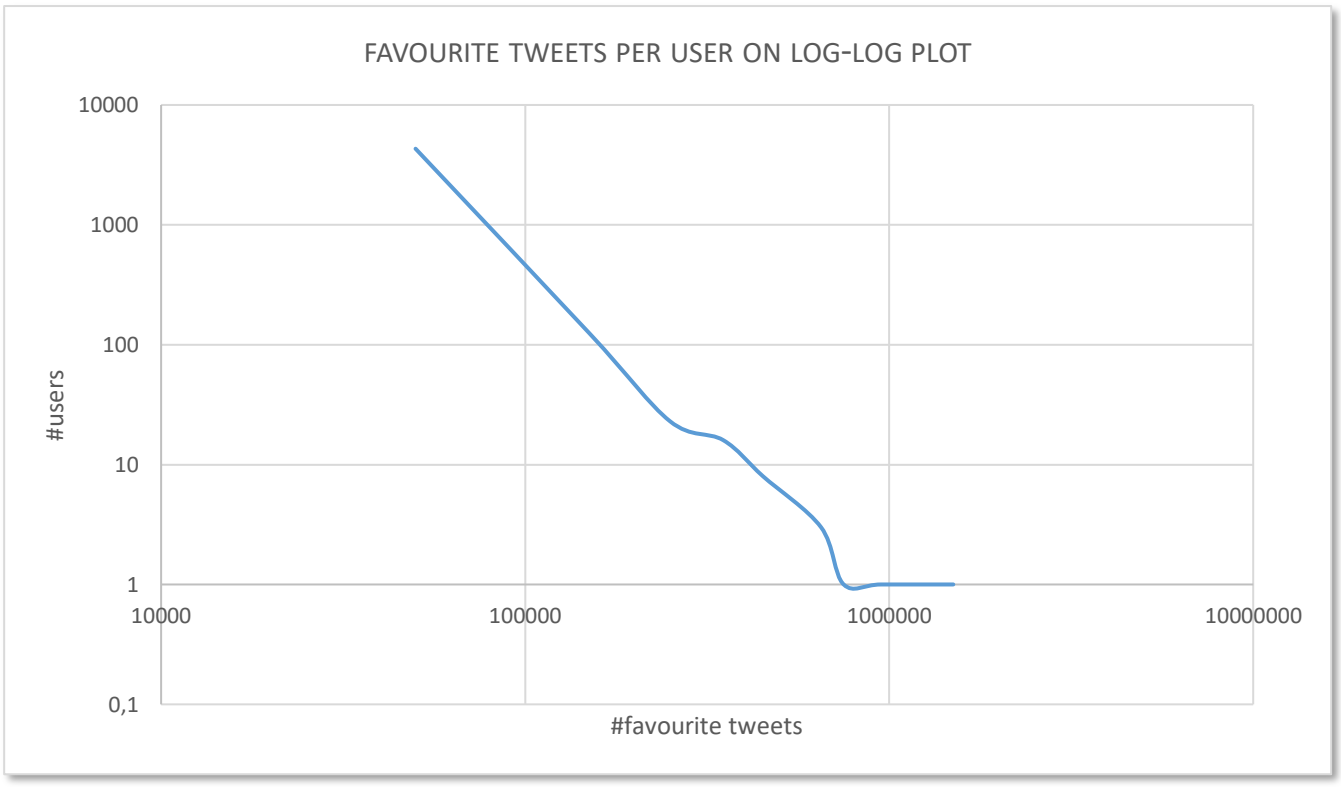

Figura 19Numero di utenti suddivisi per numero di tweet apprezzati su log-log plot

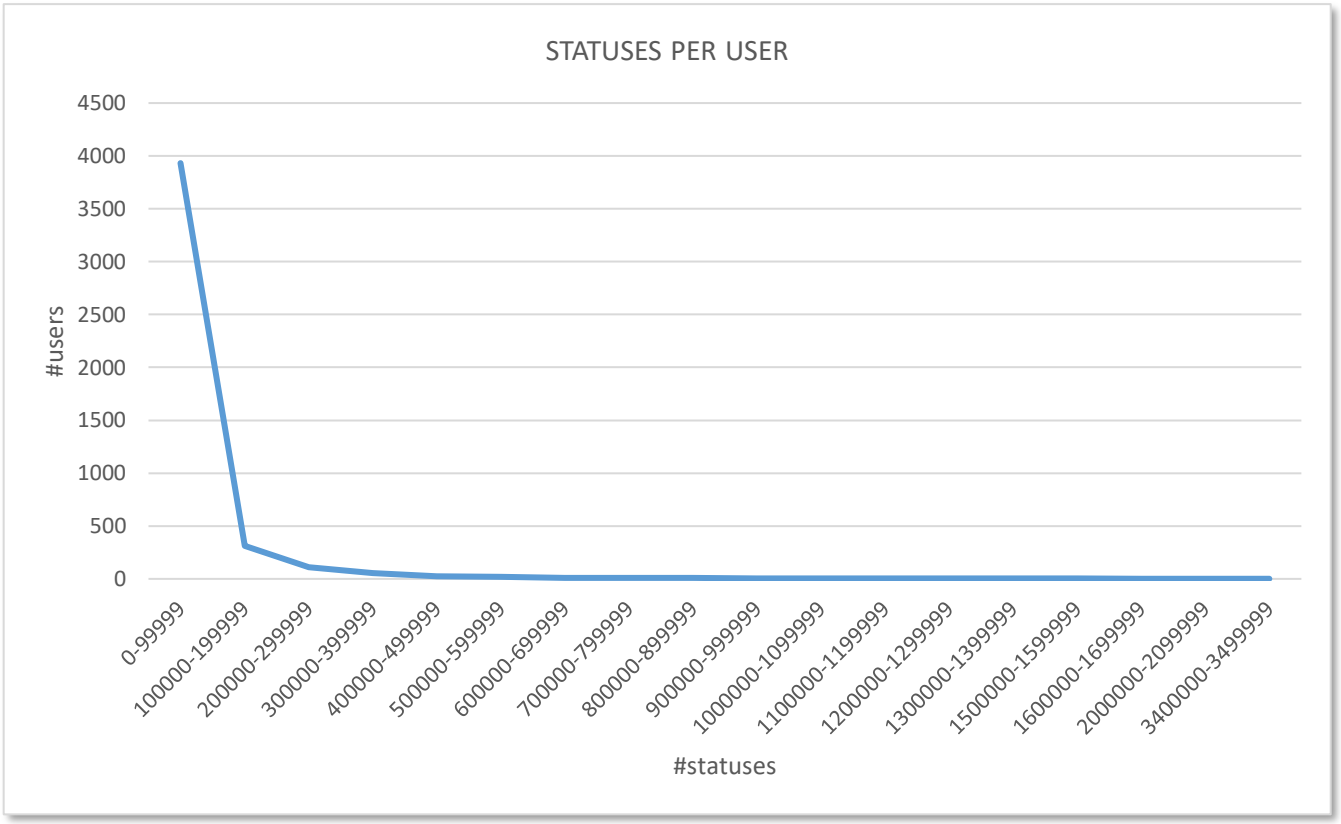

Figura 20 Numero di utenti suddiviso per numeri di stati inviati

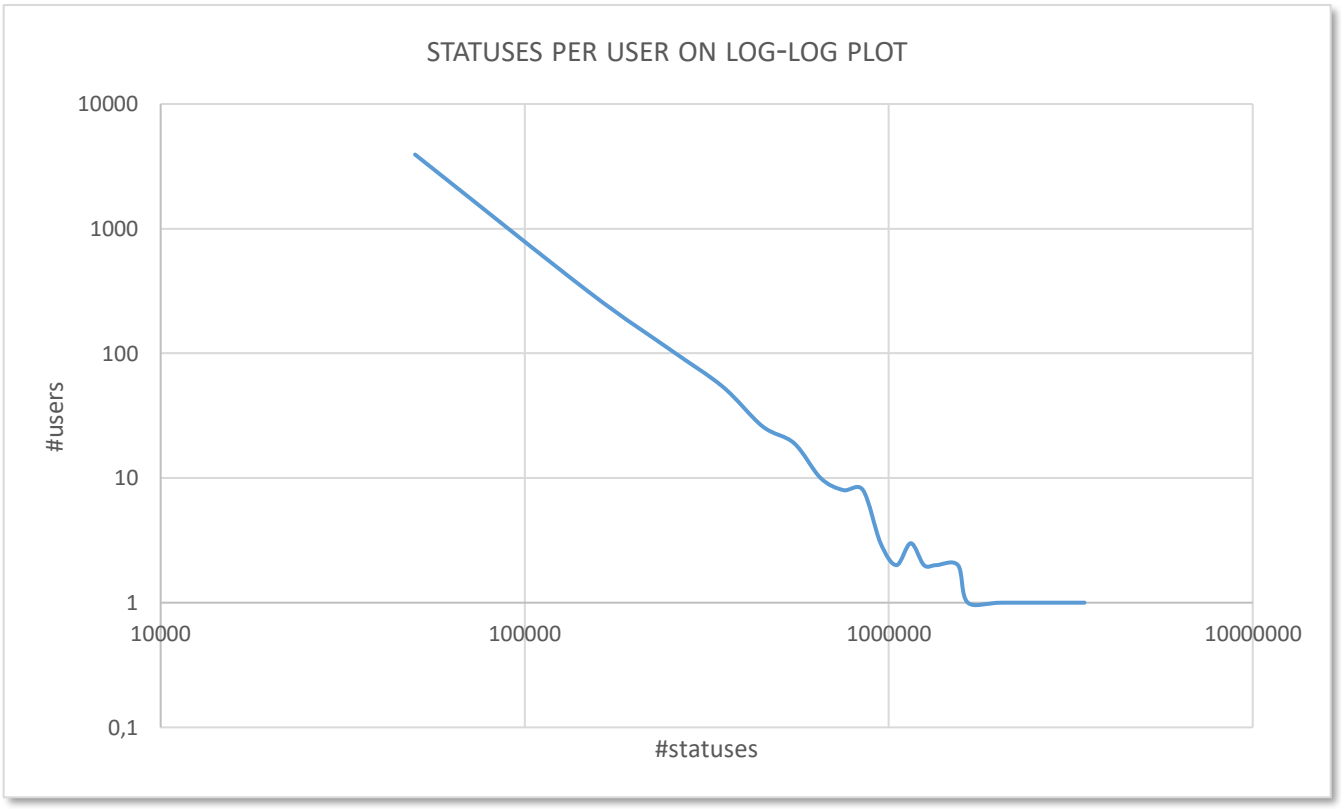

Figura 21 Numero di utenti suddiviso per numeri di stati inviati su log-log plot

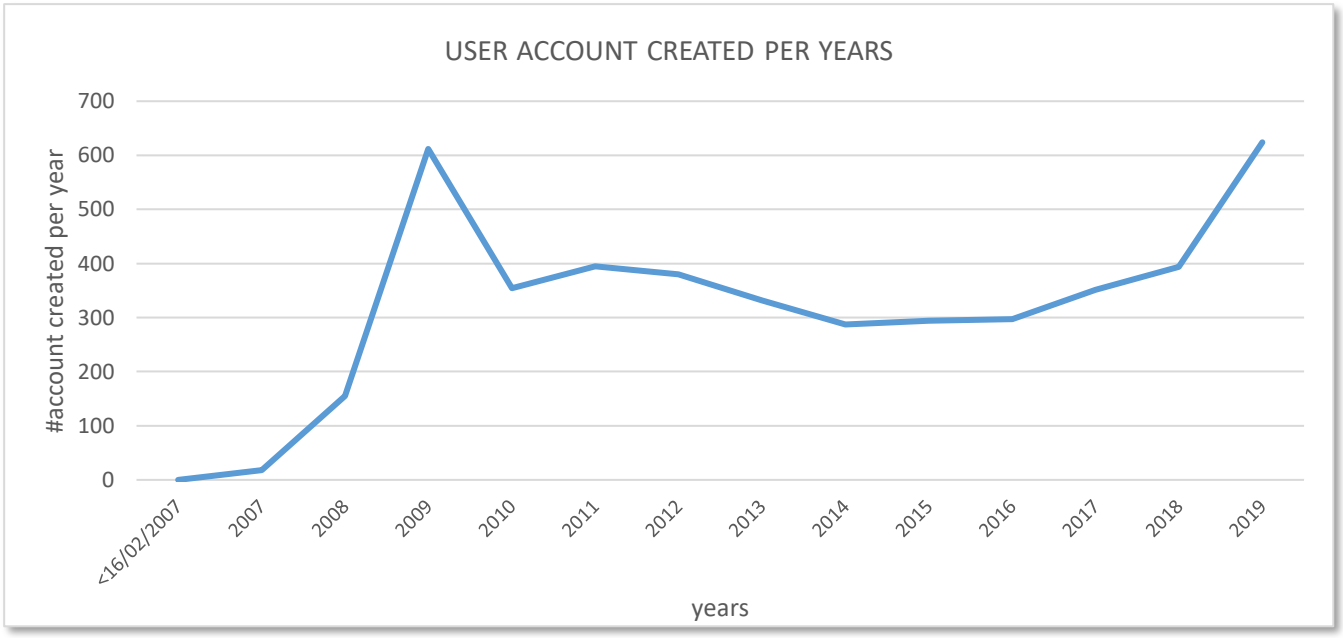

Figura 22 Numeri di utenti suddivisi per anno di creazione dell'account

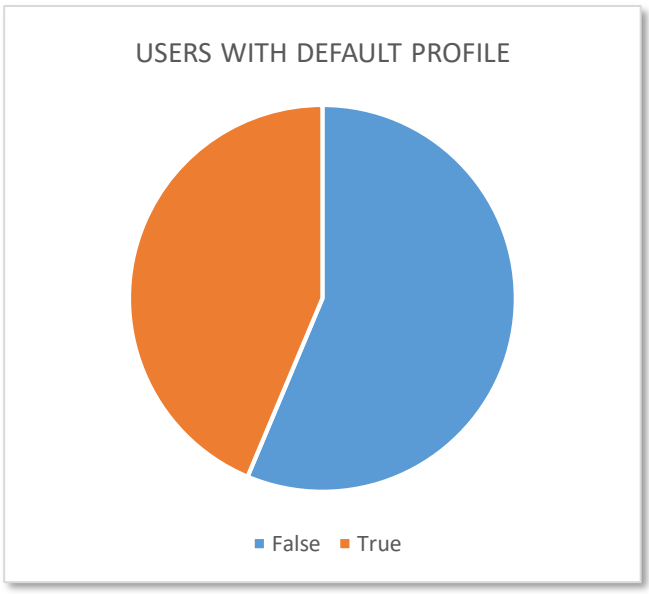

Figura 23 Numero di utenti con profilo di default

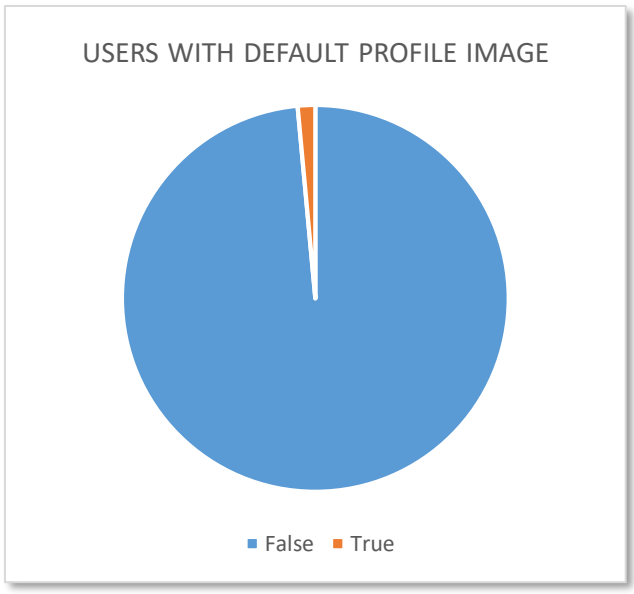

Figura 24 Numero di utenti con imm. del profilo di default

4 CONCLUSIONI

Lo studio effettuato ha mostrato come la distribuzione del numero di tweet per utenti segue la distribuzione Power Law, così come il numero di followers per utente, il numero di seguaci per utente oppure il numero di tweets apprezzati. I tweets raccolti sono in lingua inglese, perciò si spiega la maggioranza di utenti di provenienza anglosassone o statunitense e per di più sono di genere femminile. Ulteriori dettagli è possibile apprezzarli nei grafici, che mostrano una suddivisione e un'organizzazione dei dati raccolti.

5 RIFERIMENTI

(1). Tratto da https://www.tomorrownews.it/i-social-come-specchio-della-societa-o-siamo-noi-lo-specchio/
(2). Tratto da https://help.twitter.com/it/rules-and-policies/twitter-api
(3). https://www.tweepy.org/
(4). https://pypi.org/project/gender-guesser/
(5). https://pypi.org/project/genderizer/