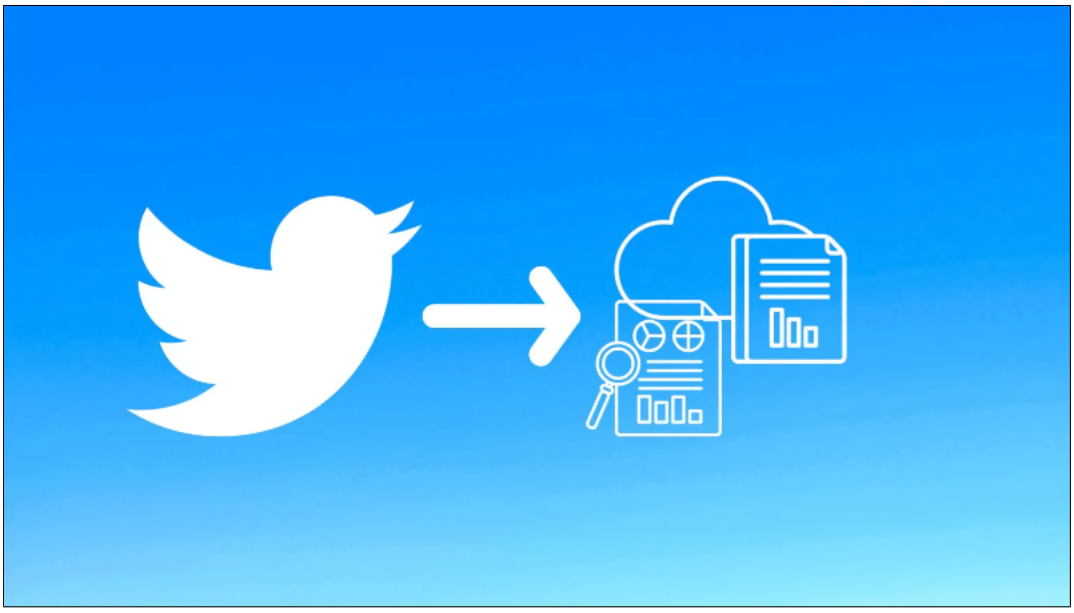
UNIVERSITA' DEGLI STUDI DI TRIESTE

Statistica e Informatica per l'azienda, la finanza e l'assicurazione

TWITTER API, WEB SCRAPING E MODELLO LOGIT

Analisi statistica sull'engagement rate dei tweets sul cambiamento climatico



Progetto a cura di:

Collautti Francesca

Fantuzzi Giulio

Ferfoglia Špela

Regonaschi Margherita

Zaccaron Damiano

Descrizione e obiettivi del progetto

Il progetto realizzato si basa su tecniche di web scraping per estrarre dati dalla piattaforma di Twitter. L'obiettivo del lavoro sarà quello di raccogliere informazioni sui tweets e sui loro utenti, al fine di analizzare e comprendere meglio come un tweet possa essere considerato "di tendenza"/influente. Per fare questo sfrutteremo l'API (*Application Programming Interface*) di Twitter che, insieme ad opportune librerie di R, permette di effettuare la raccolta dei dati. In seguito, dopo un'accurata fase di Data Preparation, applicheremo le nostre conoscenze per effettuare un'analisi quantitativa dei dati. In particolare, l'obiettivo del progetto sarà la costruzione di un modello logit per valutare l'engagement rate dei tweets.

API e Web scraping

Scelta dell'hashtag

Per scaricare i dati da Twitter, è necessario scegliere innanzitutto una parola chiave su cui basare la query per poter effettivamente estrarre i dati dal database di Twitter. La nostra scelta, dopo un'accurata valutazione, è ricaduta sull'hashtag **#climatechange**. Proponiamo in seguito le motivazioni principali:

- La versione gratuita dell'API di Twitter permette di accedere solamente ai dati delle ultime due settimane. Per questo motivo, è stato cercato un hashtag che fosse di attualità e popolare, così da poter disporre di abbastanza dati su cui costruire il modello;
- Si è voluto evitare il rischio che l'hashtag riguardasse un avvenimento troppo recente, al fine di prevenire problemi di bias nei dati raccolti;
- Si è cercato un argomento che potesse interessare sia personaggi influenti, sia l'utente medio di Twitter;
- Si è scelto un hashtag in lingua inglese per garantire una quantità elevata di dati. Inoltre, #climatechange è un termine standard condiviso a livello internazionale, quindi l'audience degli utenti che utilizzano questo hashtag è sicuramente ampia e eterogenea;
- Nella scelta dell'hashtag sono state valutate anche altre opzioni. Ad esempio, è stato pensato e testato l'hashtag #GretaThunberg che, essendo un nome proprio, è un termine universale e indipendente dalle lingue dei vari Paesi. Purtroppo, la query su questo hashtag non forniva abbastanza dati per costruire un modello adeguato.

```
#LIBRARIES
library(rtweet)
library(httpuv)
library(ROAuth)
library(httr)
library(twitteR)
library(graphTweets)
library(igraph)
library(tidyverse)
#API AUTENTICATION
auth_setup_default()
#1) Tweets download
multilang_tweets<- search_tweets("#climatechange", n=10000)
#2) Users data
dati_utenti<- users_data(multilang_tweets)
```

Funzione di engagement

Nella scelta della variabile risposta, ci è sembrato più interessante analizzare l'engagement rate dei tweets, una vera e propria chiave per comprendere i social media. Si tratta di una metrica per misurare come le persone interagiscono con i contenuti che l'utente pubblica, basandosi su likes, retweets, replies e followers. In realtà, l'engagement è una misura che viene definita a livello di user. La sua formula è la seguente:

$$ENG_{user} = \left(\frac{\frac{(\text{likes}+\text{retweets}+\text{replies})}{\text{tot tweet}}}{\text{followers}} \right) \cdot 100$$

Come da specifiche, tale funzione è stata riadattata a livello di singolo tweet. In altri termini, l'engagement rate che abbiamo considerato nel nostro progetto è una misura di quanto in voga è andato il tweet.

$$ENG_{tweet} = \left(\frac{\text{likes} + \text{retweets} + \text{replies}}{\text{followers}} \right) \cdot 100$$

NOTE: nella prima formula likes, retweets e replies sono grandezze complessive del profilo, mentre nella seconda si riferiscono al singolo tweet. L'engagement rate, inoltre, è una variabile di tipo quantitativo continuo, ma per poter costruire un modello logit è necessario avere una variabile risposta binaria. Abbiamo scelto una soglia con cui dicotomizzare in maniera opportuna la variabile, ma le considerazioni per effettuare questa scelta verranno spiegate nel seguito. Infine, si noti che la presenza del numero di followers a denominatore è fondamentale per assicurarsi validità e comparabilità della misura ottenuta.

Data preparation

Con la funzione *search_tweets* (si veda il codice a pag.2) vengono scaricati 10000 tweets, ciascuno dei quali è caratterizzato da 43 variabili di vario tipo (quantitative e qualitative). Alcune di queste non sono disponibili con la versione gratuita dell'API, mentre altre non sono ancora in un formato adeguato per poter essere utilizzate nel modello. In generale, è servita una fase preliminare di Data Preparation e Feature Selection.

Selezione delle variabili

Come anticipato, è stata effettuata una selezione delle variabili più significative. Tra queste:

1. A livello di tweet

    - **full_text**: è il testo completo del tweet. Nonostante lo scopo del progetto fosse un modello e non un'analisi testuale, abbiamo ritenuto utile tenere a disposizione il testo del tweet. Infatti, la possibilità di leggere l'intero tweet ci ha permesso di orientarci meglio nella costruzione del modello, oltre ad essere uno strumento prezioso in ottica di debugging. Inoltre, questa variabile è stata utilizzata per ricavarne altre due: *n_hashtags* e *tweet_length*;
    - **display_text_range**: fornisce una misura del numero di caratteri del tweet. Tuttavia, ci siamo accorti che tale valore conta anche caratteri speciali non inerenti al testo (*ad esempio /n per mandare a capo il testo*). Per essere più accurati, abbiamo deciso di sfruttare la funzione built-in di R *nchar()* per contare il numero effettivo di caratteri del tweet;
    - **retweet_count**: si tratta del numero di volte che il tweet è stato retwittato (ricondiviso sulla piattaforma);

- **favorite_count**: si tratta del numero di likes che ha ricevuto il tweet;
- **lang**: è la lingua del tweet, e viene categorizzata tramite etichette (*"en","it", "es",...*).
  Come spiegheremo nel seguito, essa è stata trasformata nella variabile dicotomica *en_tweet*, la quale assume valore TRUE/FALSE a seconda che il tweet sia scritto in lingua inglese o meno;

2. A livello di user (l'utente che ha pubblicato il tweet)

- **followers_count**: è il numero di seguaci (follower);
- **friends_count**: è il numero di seguiti;
- **url**: si tratta di un url/link fornito dall'utente in associazione al proprio profilo;
- **description**: si tratta di una descrizione del profilo da parte dell'utente;
- **verified**: variabile binaria che associa VERO/FALSO a seconda che l'utente sia verificato o meno. In particolare, un account risulta verificato nel momento in cui Twitter certifica che si tratta del profilo ufficiale della persona o del marchio che rappresenta. Poichè l'account verificato viene assegnato generalmente ad account famosi e influenti, abbiamo ritenuto potesse essere particolarmente correlata alla variabile risposta.

```
TweetsSubset <- c("full_text","display_text_range","retweet_count", "favorite_count","lang")
UsersSubset <- c("followers_count","friends_count","url","description","verified")
Tweets <- data.frame(multilang_tweets[TweetsSubset])
Users <- data.frame(dati_utenti[UsersSubset])
#Creo ed esporto un dataframe con le sole variabili selezionate
dataset_CC<- data.frame(Tweets, Users)
write.csv(dataset_CC, file = "../datasets/dataset_CC.csv")
```

Nella fase di selezione delle variabili è stata inoltre valutata la possibilità di includere anche la variabile *location*. Tuttavia, in seguito ad un'analisi del contenuto del dataframe, tale variabile ha rivelato numerose anomalie che la rendevano inconsistente e poco informativa. Ad esempio, alcune locations contenevano il nome di uno Stato, altre di una città, altre ancora di una regione (grandezze non omogenee). Avevamo anche valutato l'idea di dicotomizzare tale variabile in base alla presenza o meno di una location, ma ci siamo accorti che svariati record presentavano location inventate o assurde (*es: Moon, In my mind,...*). Pertanto, abbiamo concluso che la variabile non fosse informativa e abbiamo deciso di non includerla nell'analisi.

Un'altra variabile che non abbiamo potuto includere è *replies*. Essa sarebbe servita per il calcolo dell'engagement rate, ma la versione gratuita dell'API di Twitter non ne rende disponibili i valori (N.A.).

Come anticipato, a partire da *full_text* abbiamo ricavato altre due variabili:

- **tweet_length**: fornisce una misura del numero di caratteri effettivi del tweet. Nello specifico, si è ottenuta applicando a *full_text* la funzione *nchar()* di R.
- **n_hastags**: è il numero di hashtags contenuti su ogni tweet. Per calcolarlo, abbiamo implementato una semplice funzione. Di seguito il codice:

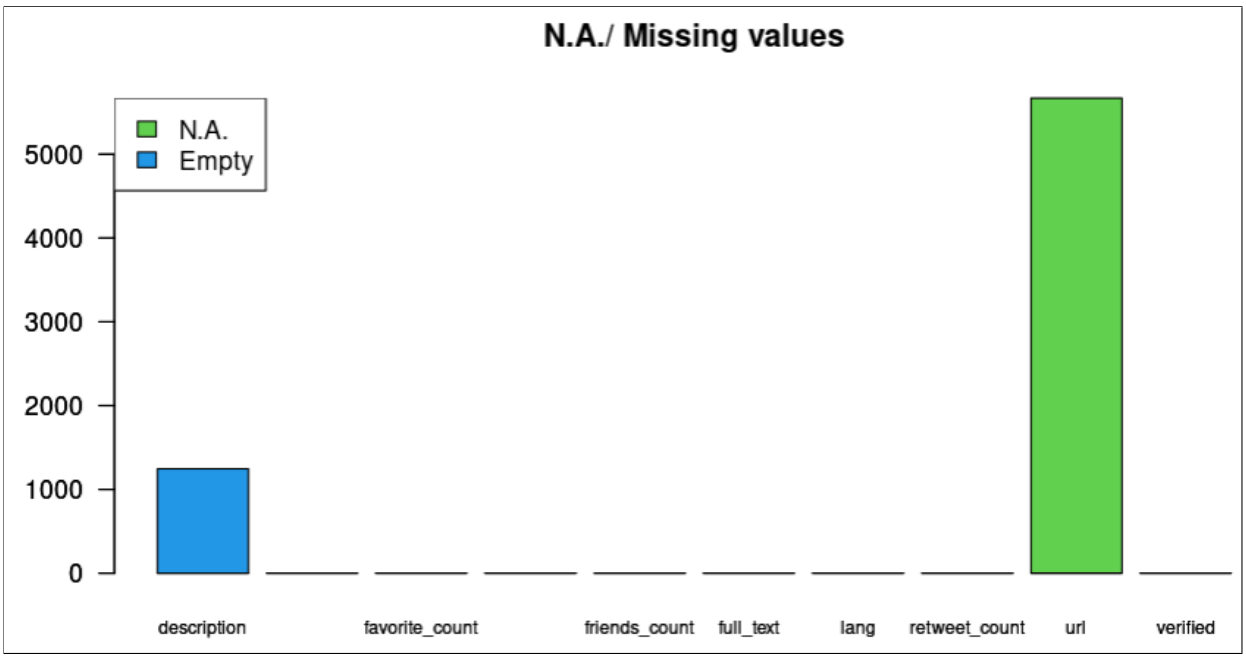
```
count_hashtags <- function(tweet) {
    hashtags <- str_extract_all(tweet, "#\\w+")
    return(length(hashtags[[1]]))
}
```

NOTA: tale funzione richiede il caricamento della libreria *stringr* ed è stata applicata nel modo seguente:

```
library(stringr)
n_hashtags<- vector(mode="numeric", length = length(data[,1]))
for(i in 1: length(data[,1])){
    n_hashtags[i]= count_hashtags(data[i,"full_text"])
}
```

Data cleaning, ricodifica e trasformazioni

Una volta ottenuto il dataframe con le variabili selezionate, siamo poi passati alla fase di pulizia dei dati. Come punto di partenza, abbiamo scelto di gestire un potenziale problema riguardante il calcolo dell'engagement rate. Infatti, essendoci a denominatore il numero di followers, nel caso in cui *followers_count* fosse 0 risulterebbe un engagement infinito. Analizzando i dati, abbiamo riscontrato che solo 39 records su 10000 presentavano un numero di followers nullo. Trattandosi di una percentuale pressoché ininfluente (0.39%) abbiamo deciso di rimuoverli dal dataset. Inoltre, è stata presa in analisi la presenza di N.A. o dati mancanti.

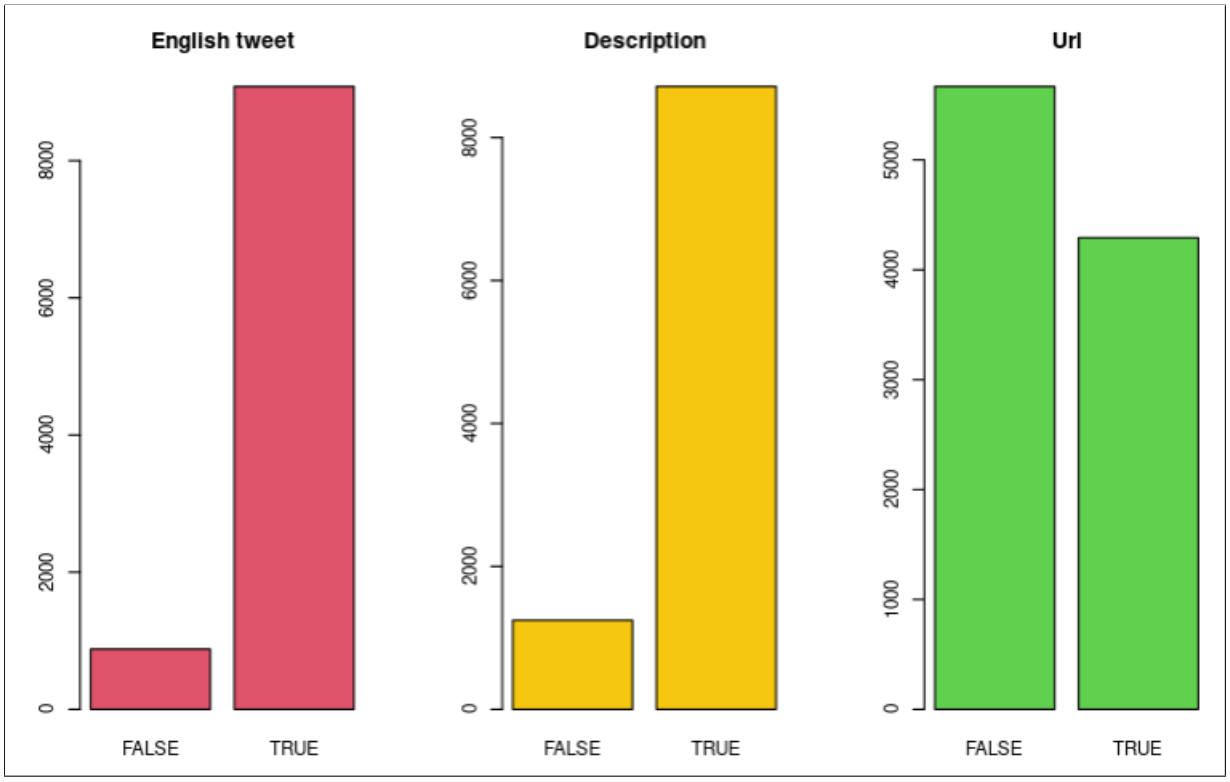


In generale, per gestire la presenza di dati mancanti, si possono adottare varie tecniche di imputazione dei valori. Tuttavia, le due variabili evidenziate dal grafico sopra sono qualitative testuali e, a differenza di dati quantitativi, la presenza di un valore vuoto o mancante non costituisce un problema. Al contrario, abbiamo sfruttato questa caratteristica per dicotomizzare le variabili *description* e *url*

```
# DESCRIPTION
#Assegno true se c'è descrizione; false se campo vuoto
data[,"description"]<- data[,"description"]!=""
# URL
#Assegno true se c'è un url; false se dato N.A.
data[,"url"]<- !is.na(data[,"url"])
```

Oltre a *description* e *url*, anche le variabili *verified* e *en_tweet* sono dicotomiche. Per quanto riguarda la prima era già codificata nel dataset di partenza, mentre la seconda era originariamente codificata come una variabile qualitativa con oltre 30 categorie: *lang*. Tuttavia, siccome oltre il 90% dei tweets risultava scritto in lingua inglese, abbiamo deciso di trasformarla nella variabile binaria *en_tweet* (già introdotta a pag. 4).

Proponiamo di seguito la distribuzione di queste variabili dicotomiche:



Engagement rate: definizione e scelta del valore soglia

La fase di Data Preparation per le variabili esplicative è terminata: esse sono state tutte codificate e/o trasformate in maniera opportuna, e sono pronte per essere usate nella costruzione del modello. A questo punto, è possibile definire la variabile risposta: *eng_rate*. Come introdotto a pagina 3, l'engagement rate è una funzione di likes, retweets e followers. Per calcolarlo, abbiamo implementato la seguente funzione:

```
eng_tweet<- function(data){
    likes<- data$favorite_count
    retweets<- data$retweet_count
    followers<- data$followers_count
    #replies non sono dati disponibili gratuitamente!
    eng_rate<- ((likes+retweets)/followers) *100
    return(eng_rate)
}
```

In questo modo *eng_rate* è una variabile quantitativa che assume valori continui in $\mathbb{R}^+$. Tuttavia, un modello logistico è caratterizzato da una variabile risposta binaria, dunque è stato necessario dicotomizzarla.

Per dicotomizzare *eng_rate* abbiamo dovuto scegliere una soglia, così da assegnare alla variabile il valore 1 se superiore alla soglia, 0 altrimenti. In altri termini, abbiamo scelto un valore α tale che:

$$\text{eng\_rate} = \begin{cases} 1 \text{ se eng\_rate} > \alpha \\ 0 \text{ se eng\_rate} \leq \alpha \end{cases}$$

Per la scelta del valore soglia (α) abbiamo considerato varie opzioni. Inizialmente ci siamo basati sulla distribuzione di *eng_rate*, pensando che la media potesse essere un buon criterio. Prima di procedere, però, abbiamo ritenuto opportuno analizzare come si distribuisse:

```
> summary(eng_data$eng_rate)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     0.0      0.1      1.4    498.7     43.6 342550.0
```

Dal *summary* emerge una forte asimmetria verso destra. Come si può notare la media assume un valore molto più elevato (498.7) rispetto alla mediana (1.4), e ciò è causato dalla presenza di outliers; abbiamo così deciso di non utilizzare la media come soglia. Poiché la mediana aggira il problema dei valori estremi (non è influenzata da essi) ci era parsa un buon sostituto. Allo stesso tempo, per definizione la mediana divide la distribuzione in due parti uguali. Utilizzarla per una classificazione binaria non ci ha dato l'impressione che potesse fornire dei risultati rilevanti.

Non trovando soglie significative tra le caratteristiche della distribuzione, abbiamo spostato il nostro focus verso il significato intrinseco di engagement rate. Poiché questo misura il successo di un tweet in relazione al numero di interazioni rispetto ai followers dell'utente, abbiamo ritenuto sensato fissare la soglia $\alpha = 200$.

Per comprendere meglio il motivo di tale scelta, riconsideriamo la formula dell'engagement rate:

$$ENG_{tweet} = \left(\frac{\text{likes} + \text{retweets}}{\text{followers}} \right) \cdot 100$$

Siamo partiti dal seguente ragionamento: ogni utente può interagire con un tweet al massimo in due modi, mettendo like e ricondividendo il post. Possiamo dunque considerare un tweet di successo nel momento in cui il numero delle interazioni ricevute supera quello delle interazioni che potrebbero garantire i soli follower.

Considerato un tweet di un utente, la situazione limite si ha quando ciascuno dei suoi follower interagisce con il post sia mettendo like sia rewtittandolo. In questo caso il rapporto $\frac{\text{likes+retweets}}{\text{followers}}$ sarebbe pari a 2 e, di conseguenza, risulterebbe un engagement rate pari a 200. Nel momento in cui il tweet riceve più interazioni di quelle che potrebbero garantirgli i suoi seguaci, ecco che l'engagement rate supererebbe il valore di 200. In sintesi, un engagement rate maggiore della soglia è una valida misura per stabilire se un tweet è stato influente o meno, poiché in questo caso vuol dire che è "arrivato" anche a utenti esterni al cerchio dei suoi followers.

Costruzione del modello logit

Il modello logistico (o *modello logit*) si basa sulle relazioni tra una variabile dipendente binaria (ad esempio, successo/fallimento) e una o più variabili indipendenti. Sotto questo punto di vista, risulta uno strumento utile a modellizzare il fenomeno analizzato nel progetto. Come spiegato in precedenza infatti, l'engagement rate di un tweet può essere interpretato come una misura del successo che ha avuto.

Il punto di partenza è stato dicotomizzare la variabile risposta:

```
# Dicotomizziamo la variabile risposta
eng_data$eng_rate<- as.integer(eng_data$eng_rate > 200)
```

Prima di procedere alla costruzione del modello, è stato necessario rimuovere dal dataframe le variabili utilizzate per il calcolo dell'engagement rate. Questo è inevitabile, in quanto essendo componenti della formula dell'engagement, renderebbero automaticamente tutte le altre variabili non significative.

```
#Rimuoviamo dal dataframe le variabili usate per calolare l'engagement:
eng_data <- eng_data[,-which(colnames(eng_data)=="favorites_count")]
eng_data <- eng_data[,-which(colnames(eng_data)=="retweet_count")]
eng_data <- eng_data[,-which(colnames(eng_data)=="followers_count")]
```

Successivamente, abbiamo trasformato le variabili non numeriche in variabili categoriche:

```
eng_data$en_tweet = factor(eng_data$en_tweet)
eng_data$url = factor(eng_data$url)
eng_data$description = factor(eng_data$description)
eng_data$verified = factor(eng_data$verified)
```

Inizialmente, non avendo un'idea di quali variabili fossero significative o meno, abbiamo costruito un primo modello logistico con tutte le esplicative. Di seguito il codice:

```
#1)MODELLO CON TUTTE LE VARIABILI
modello1= glm(eng_rate ~., data=eng_data,
              family = binomial(link = logit))
```

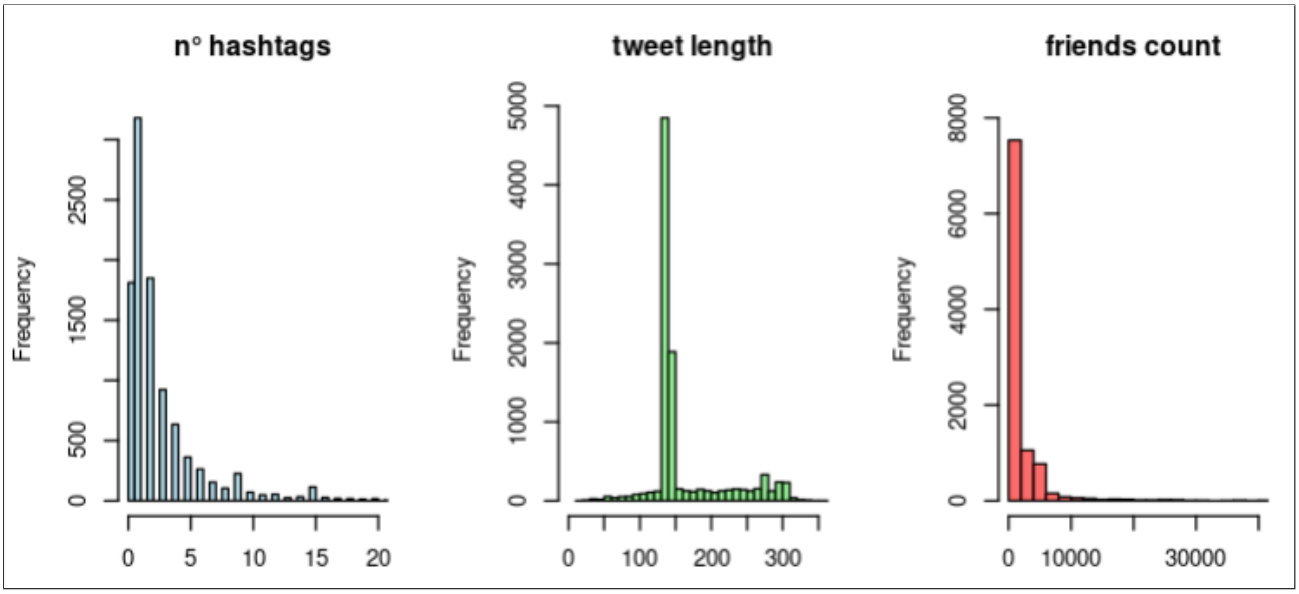
Compilando il codice, è comparso il seguente messaggio di avvertimento:

> *"glm.fit: fitted probabilities numerically 0 or 1 occurred"*

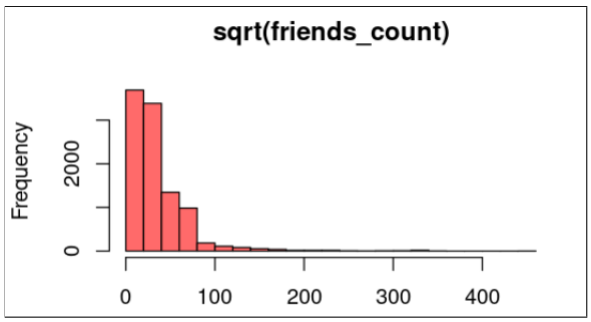
Questo avviso si verifica quando una o più previsioni risultano indistinguibili tra 0 o 1. Due sono le cause principali per questo tipo di errore:

1. un numero di dati eccessivamente basso (decisamente non il nostro caso);
2. la presenza di esplicative con una distribuzione fortemente asimmetrica e su scala molto ampia.

Tra le variabili esplicative del modello, le quantitative sono le seguenti tre:



Come si può notare in figura, le variabili che presentano una distribuzione asimmetrica sono *n_hashtags* e *friends_count*. Tra le due, *n_hashtags* si distribuisce su una scala relativamente ristretta, al contrario di *friends_count* che assume valori su una scala molto ampia. Abbiamo tentando di risolvere il problema applicando a *friends_count* una trasformazione che potesse correggere l'asimmetria. In primis abbiamo pensato alla trasformazione logaritmica, che però non poteva essere adottata poiché alcuni record (59) presentavano un valore di *friends_count* pari a 0 (il logaritmo in zero non è definito!). L'altra opzione che abbiamo valutato è stata la radice quadrata, essendo l'altro caso particolare delle trasformazioni di Box-Cox. Tuttavia, il problema non è stato risolto: nonostante la trasformazione, la distribuzione della variabile presentava ancora forte asimmetria (si veda il grafico sotto)



Date le problematiche dovute alla presenza di tale variabile, abbiamo proseguito con la formulazione di un modello che contenesse tutte le variabili esplicative eccetto *friends_count*

```
#2) MODELLO SENZA FRIENDS_COUNT
modello2= glm(eng_rate ~tweet_length+ n_hashtags +en_tweet+
              description+ url + verified, data=eng_data,
              family = binomial(link = logit))
```

In questo modo il nuovo modello non presenta più messaggi di errore, e tutte le sue componenti sono state stimate correttamente.

Interpretazione dei coefficienti

```
> summary(modello2)

Call:
glm(formula = eng_rate ~ tweet_length + n_hashtags + en_tweet +
    description + url + verified, family = binomial(link = logit),
    data = eng_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5884  -0.6712  -0.3207  -0.1116   3.4042

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.071432   0.234104   0.305    0.760
tweet_length     -0.009413   0.001204  -7.815 5.48e-15 ***
n_hashtags       -0.163860   0.022697  -7.219 5.22e-13 ***
en_tweetTRUE      1.576431   0.195296   8.072 6.92e-16 ***
descriptionTRUE  -1.378047   0.070813 -19.460  < 2e-16 ***
urlTRUE          -1.692093   0.096675 -17.503  < 2e-16 ***
verifiedTRUE    -14.124170 211.101258  -0.067    0.947
```

Come emerge dal summary del modello, i coefficienti associati alle variabili sono tutti significativi, tranne quello di *verifiedTRUE*. Il fatto che il coefficiente associato ai profili verificati sia risultato significativamente nullo, a primo impatto, è stato inaspettato. Ritenevamo infatti che i profili di questa tipologia, essendo associati in genere a utenti influenti o celebri, potessero avere un impatto notevole sull'engagement del tweet. Per comprendere meglio questo risultato, abbiamo analizzato meglio i dati:

```
#Quanti tweets hanno avuto successo?
sum(eng_data$eng_rate==1) #1567
#Quanti tweets sono di account verificati?
sum(eng_data$verified==TRUE) #309
#Quanti tweets di account verificati hanno avuto successo?
sum(eng_data$eng_rate==1 & eng_data$verified==TRUE) #0
```

Effettivamente nei nostri dati nessun tweet associato ad un profilo verificato ha avuto successo. Da qui è possibile comprendere il motivo della nullità del coefficiente associato.

Abbiamo dunque creato un terzo (e ultimo) modello, senza includere tra le variabili esplicative *verified*.

```
modello3<- glm(eng_rate ~tweet_length+ n_hashtags
               +en_tweet+ description+ url , data=eng_data,
               family = binomial(link = logit))
```

Il modello in questione presenta coefficienti tutti significativamente diversi da zero, quindi ora ci concentreremo sulla loro interpretazione. In ambito di regressione logistica, un'utile quantità per comprendere il contributo delle singole esplicative (effetto marginale) è l'odds ratio (OR).

Al posto di riproporre la schermata del summary anche per questo modello, abbiamo costruito una tabella riassuntiva comprendente anche gli odds ratio.

	Estimate	Std. Error	z value	Pr(>|z|)	OR
tweet_length	-0,009637	0,001201	-8,027	1,00E-15	0,9904092871
n_hashtags	-0,162048	0,022725	-7,131	9,97E-13	0,8504003843
en_tweetTRUE	1,585868	0,195317	8,119	4,68E-16	4,883528442
descriptionTRUE	-1,383398	0,070853	-19,525	< 2e-16	0,2507251399
urlTRUE	-1,751002	0,096576	-18,131	< 2e-16	0,1735999092

- **tweet_length:** l'odds ratio associato a tale variabile è risultato pari a circa 0.99. Ciò significa che, a parità di tutto il resto, un tweet con un carattere in più rispetto ad un altro ha una probabilità 0.99 volte maggiore di avere successo (la direzione dell'effetto è dunque negativa). Trattandosi di un valore molto prossimo a 1 l'effetto che si registra è quasi ininfluente, e ciò non è affatto sorprendente. Teniamo presente che il numero di caratteri di un tweet ha dei vincoli standard, e può variare da un minimo di 1 carattere a un massimo di 280, dunque l'effetto marginale che può avere sul successo del tweet è pressochè nullo.
- **n_hashtag**: l'odds ratio associato al numero di hashtags è risultato pari a circa 0.85. Ciò significa che, a parità di tutto il resto, un tweet con un hashtag in più rispetto ad un altro ha una probabilità 0.85 volte maggiore di avere successo. Anche in questo caso l'effetto marginale dell'esplicativa verso il successo del tweet ha una direzione negativa. Pur non disponendo dei dati per fare un'analisi più approfondita, abbiamo ipotizzato due diverse motivazioni:
  1. Più sono i caratteri che vengono utilizzati per scrivere gli hashtags, meno possono essere quelli destinati al contenuto del tweet. In altri termini, tweets che hanno un contenuto più povero hanno una minor tendenza ad avere successo;
  2. I tweets con un numero di hashtags elevato possono essere percepiti dagli utenti come "attention seeker" (tweets in cerca di interazioni forzate) e, di conseguenza, tendono ad essere ignorati.
- **en_tweet:** l'odds ratio associato a tale variabile è risultato pari a circa 4.88. Ciò significa che, a parità di tutto il resto, un tweet scritto in inglese ha una probabilità 4.88 volte maggiore di avere successo rispetto ad un tweet scritto in una lingua diversa. L'effetto (positivo) che questa variabile esplicativa ha nei confronti dell'engagement del tweet è tanto notevole quanto prevedibile. Infatti, l'inglese è la lingua che copre l'audience più ampia, ed è sicuramente il mezzo più efficace per condividere un contenuto a livello internazionale.
- **description:** l'odds ratio associato a tale variabile è risultato pari a circa 0.25. Ciò significa che, a parità di tutto il resto, un tweet scritto da un utente che ha una descrizione del proprio profilo ha una probabilità 0.25 volte maggiore di avere successo rispetto ad un tweet scritto da un utente senza descrizione. Tuttavia, riteniamo che non si tratti di una relazione di natura causale, ma casuale.
- **url:** l'odds ratio associato a tale variabile è risultato pari a circa 0.17. Ciò significa che, a parità di tutto il resto, un tweet scritto da un utente che associa un url al proprio profilo ha una probabilità 0.17 volte maggiore di avere successo rispetto ad un utente che non lo fa. Come per *description*, è difficile trovare un'interpretazione fondata, dunque riteniamo si tratti di una relazione non causale, bensì casuale.

Cambio del valore soglia

Abbiamo voluto valutare come sarebbe cambiato il modello utilizzando un'altra soglia. Se nel modello precedente la soglia era stata fissata sulla base del significato intrinseco di engagement rate, questa volta abbiamo provato a sceglierla a partire dalla distribuzione. Avendo già scartato l'ipotesi di utilizzare la media e la mediana, la scelta è ricaduta sul terzo quartile. Dunque, se prima il successo di un tweet era misurato in base a una definizione puramente teorica, ora esso viene messo in relazione all'andamento generale del dataset. Abbiamo dunque considerato un tweet "di successo" nel caso in cui superasse il 75esimo percentile della distribuzione.

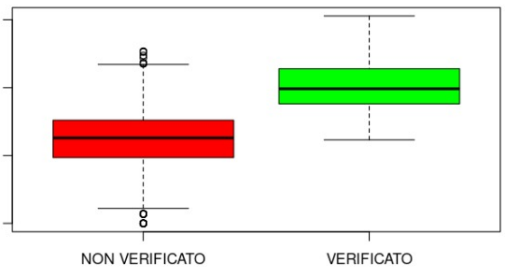
Come effettuato in precedenza, siamo partiti tentando di costruire il modello utilizzando tutte le variabili esplicative. Pur avendo cambiato la soglia, si è comunque presentato lo stesso messaggio di errore: *"glm.fit: fitted probabilities numerically 0 or 1 occurred"*. Anche in questo caso, il modello con la radice quadrata di *friends_count* non ha permesso di migliorare tale situazione. Abbiamo dunque costruito il modello logit senza includere tra le esplicative *friends_count*

```
#2) MODELLO SENZA FRIENDS_COUNT
modello2= glm(eng_rate ~tweet_length+ n_hashtags +en_tweet+
              description+ url + verified, data=eng_data,
              family = binomial(link = logit))
```

I coefficienti sono risultati:

	Estimate	Std. Error	z value	Pr(>|z|)	OR
tweet_length	-0,0100505	0,0009252	-10,863	< 2e-16	0,9899998375
n_hashtags	-0,1574064	0,0175634	-8,962	< 2e-16	0,8543567777
en_tweetTRUE	1,3439677	0,1394054	9,641	< 2e-16	3,834226425
descriptionTRUE	-1,3172652	0,0700142	-18,814	< 2e-16	0,2678668635
urlTRUE	-2,5804108	0,7139411	-3,614	0,000301	0,07574288245
verifiedTRUE	-1,4770554	0,0685586	-21,544	< 2e-16	0,2283089782

A differenza della soglia usata precedentemente il coefficiente associato alla variabile *verified* è risultato significativo. In particolare risulta che i tweet postati da account verificati hanno una probabilità di successo oltre quattro volte minore rispetto a quelli di utenti "standard". La direzione dell'effetto è negativa; questo potrebbe essere spiegato dalla forte relazione tra utenti verificati e numero di follower (come evidenziato dal Box-Plot). Abbiamo infatti constatato che il profilo di un utente con un numero elevato di follower ha difficoltà a raggiungere engagement più alti, questo poichè il numero di followers costituisce il denominatore della formula.



Per quanto riguarda i coefficienti delle altre variabili, la direzione del loro effetto, rispetto all'altro modello, non cambia:

- Le variabili *n_hashtags, description, url* continuano ad avere un effetto negativo sull'engagement;
- La variabile *tweet_length* continua ad avere un effetto pressochè ininfluente (OR è prossimo a 1);
- La variabile *en_tweet* è l'unica che mantiene un effetto positivo nei confronti dell'engagement.

L'unica differenza riguarda l'entità del loro effetto, che tuttavia non è abbastanza accentuata da modificare l'interpretazione del modello.