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

@gruppo_1

TWITTER API, WEB SCRAPING E MODELLO LOGIT:

Analisi statistica sull’engagement rate dei tweets sul cambiamento climatico. #ClimateChange

Show this thread

SCELTA DELL'HASHTAG

#ClimateChange

Motivazioni della scelta:

- Popolare
- Di attualità ma privo di *recency bias*
- Utilizzato sia da personaggi influenti, sia dall'utente medio
- In lingua inglese ma termine standard a livello internazionale
- Fornisce una quantità adeguata di dati

```
#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)
```

DATA PREPARATION

Selezione delle variabili:

Viene effettuata una distinzione tra variabili:

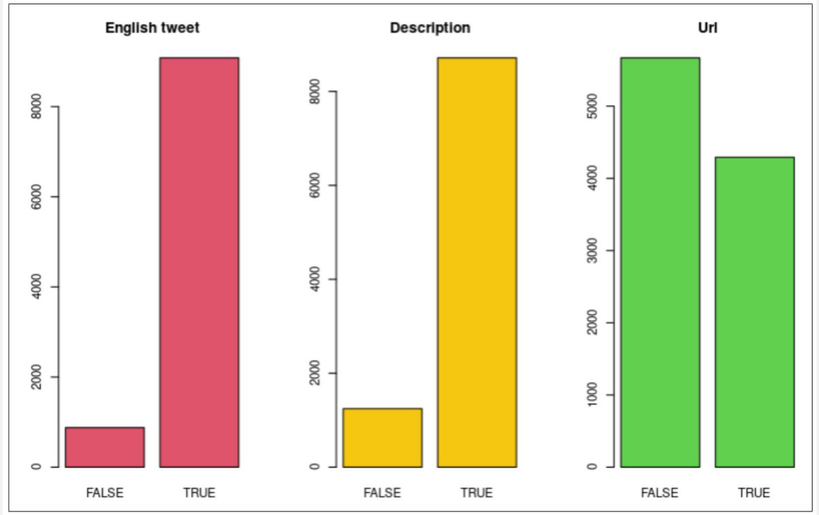
- A livello di **TWEET**: full_text, display_text_range, retweet_count, favorite_count, lang
- A livello di **USER**: followers_count, friend_count, url, description, verified
- **full_text** viene utilizzata per creare **n_hashtags** e **tweet_length**.

Data cleaning:

Si sceglie di dicotomizzare le variabili **verified**, **url**, **en_tweet** (trasformazione di **lang**) e **description**.

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

```
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"])
}
```



FUNZIONE DI ENGAGEMENT

La funzione di engagement, che misura come le persone interagiscono con i contenuti pubblicati, è definita a livello di user. Per lo scopo del progetto, viene adattata a livello di tweet:

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

SCELTA DEL VALORE SOGLIA

Il risultato della funzione è una variabile reale, mentre il modello logit è caratterizzato da una variabile binaria, dunque è stato necessario dicotomizzarla tramite una soglia α.

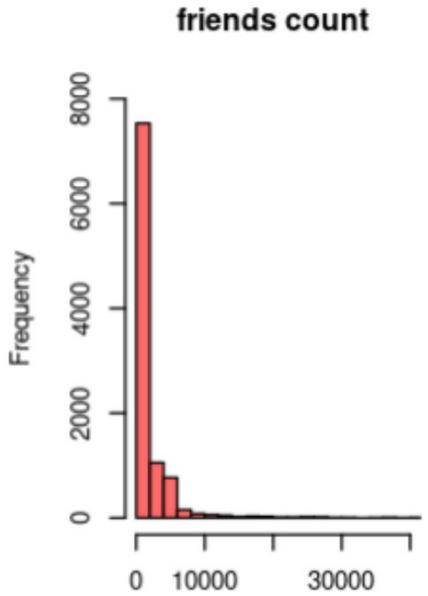
$$\text{eng\_rate} = \begin{cases} 1 \text{ se eng\_rate} > \alpha \\ 0 \text{ se eng\_rate} \leq \alpha \end{cases}$$

Sono state prese in considerazione soglie basate sulla distribuzione, ma data la forte asimmetria che questa presentava, abbiamo optato per il valore 200, derivato dal significato della funzione di engagement.

```
> 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
```

COSTRUZIONE DEL MODELLO

Dopo aver rimosso le variabili contenute nella formula di engagement e categorizzato le variabili qualitative, si è costruito il modello utilizzando tutte le esplicative.



La forte asimmetria di **friends_count** ha causato l'errore: *"glm.fit: fitted probabilities numerically 0 or 1 occurred"*

Non potendo risolvere il problema, si è costruito il modello contenente 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))
```

COSTRUZIONE DEL MODELLO

Analizzando il *summary* è risultato che tutti i coefficienti, eccetto **verified**, sono significativi per il modello. In effetti:

```
#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
```

Si è quindi costruito il modello definitivo escludendo quest'ultima variabile.

INTERPRETAZIONE DEI COEFFICIENTI

	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

Variabili con effetto negativo:	Variabili con effetto positivo:	Variabili con effetto neutro:
• n_hashtags • description • url	• en_tweet	• tweet_length

E CAMBIANDO LA SOGLIA?

Utilizzando come valore soglia il terzo quartile della distribuzione e costruendo il modello con tutte le variabili eccetto **friends_count** si ottengono i seguenti 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

L’unica differenza sostanziale è rappresentata da **verified**, che ora risulta significativo e con effetto negativo sul successo del tweet.

Resources