Universidad Internacional de Valencia

Master en Big Data y Data Science

Práctica 2: Credibilidad de Twitter

Fundamentos de la tecnología Big Data

Autor:
Adrián Hernández Padrón
Mayo 2022

Universidad Internacional de Valencia

Índice

1. Descripción de las métricas usadas para el cálculo de credibilidad

Siguiendo las explicaciones de clases y el paper 'Web Scraping versus Twitter API-A Comparison dor a Credibility Analysis' se planteó realizar un análisis de estas siguiendo 3 criterios: Credibilidad textual, credibilidad de usuario y credibilidad social.

Tanto en la clase como en el paper se plantea un reparto equitativo entre los 3 criterios en los cuales cada uno de ellos tendría un peso equitativo, el 33.3 % en cada caso. Sin embargo en este ejercicio vamos a plantear unos porcentajes de peso distintos:

Criterio	Peso(porcentaje)
Credibilidad Textual	20
Credibilidad Usuario	40
Credibilidad Social	40

1.1. Credibilidad Textual

Este criterio de credibilidad varía en función de las palabras mal sonantes y los errores ortográficos que se encuentren en los twits, la bajada de este criterío se debe a que no se van a analizar los twits palabra por palabra debido a que no poseo los conocimientos ni herramientas necesarias para realizar esta tarea. Hay que tener en cuenta que para este ejercicio las 10 cuentas elegidas son cuentas libres de spam ya que son cuentas que yo personalmente conozco, sin embargo vamos a aplicar unos porcentajes asociados a la credibilidad textual en función de la cuenta que tratemos:

- Lenguaje coloquial: 50 %
- Lenguaje formal: 100 %

Es decir, si una cuenta de twitter suele usar un lenguaje coloquial esta obtendrá la mitad del peso asociada a la credibilidad textual, mientras que si otra cuenta posee un lenguaje mas formal obtendrá el total del peso. Al cubrir ela credibilidad textual de esta manera pensé que sería apropiado disminuir el peso al 20 %.

1.2. Credibilidad de Usuario

Esta credibilidad sigue la extructura explicada en el paper y en clase:

- Verificada: 0-50 %
- $\frac{Edad\_de\_la\_cuenta}{Edad\_maxima}$: 50 %(max)

Estos valores siguen la explicación de que una cuenta verífícada va a poseer mas credibilidad que una que no está verificada y que cuanto mas longeva sea la cuenta mas credibilidad tendrá. El valor de la *Edad_maxima* aparece en el paper y corresponde a 2006.

1.3. Credibilidad Social

Esta credibilidad también sigue los valores explicados en clase que se pueden encontrar en el paper:

- $\frac{Followers}{Followers\_maximos}$: 50 %(max)
- Proporción de Follower/Following: 50 %

Aquí valorizamos el número de followers que tiene una cuenta, puesto esto recae directamente en la credibilidad, a cuantos más followers mas creíble tiende a ser dicha cuenta.

2. Herramientas usadas para el cálculo

Todo el cálculo de la credibilidad se realizó en un programa creado por mi en Jupyter, en dicho programa se fue calculando en distintas celdas los distintos valores de credibilidad. La razón de el uso de Jupyter y Python3 para el cálculo de la credibilidad es porque, además de que estoy familiarizado con esta herramienta me parece que es bastante prácita para este ejercicio. La obtención de los datos para el cálculo se hizo a través del fichero csv del primer ejercicio ,

```python
import pandas as pd
from twython import Twython



APP_KEY = 'kTbYPP7inrfVIdsNYuCdLfbLs' # API Key
APP_SECRET = 'LTQ0Dp7GedTXmAmSdajFcnm2V6TTZjxeqjHxDnoObZmrbtzszm' # API Secret Key
OAUTH_TOKEN = '152150597879305011 2-qelhMJdfv1Czl78ebk4heIxcRBxp1Z' # Access Token
OAUTH_TOKEN_SECRET = 'YhOpm1mo78nBO4BUT8HGQp66ctFQRqzQ586df1JxpLoKZ' # Access Token Secret



twitter = Twython(APP_KEY, APP_SECRET, OAUTH_TOKEN, OAUTH_TOKEN_SECRET)



def follows_twit(twitter,user):
    twitt_info = twitter.show_user(screen_name = user)
    following = twitt_info['friends_count']
    followers = twitt_info['followers_count']
    created_at = twitt_info['created_at']
    verified= twitt_info['verified']
    return (followers,following,created_at,verified)



df = pd.read_csv('/Users/adrihp/Master/MBID01/Practica1/accountsEMBS.csv', encoding='latin-1')



Followers, Following , Created_at , Verified= [], [], [], []
for element in df['Twitter_handle']:
    miau = follows_twit(twitter,element)
    Followers.append(miau[0])
    Following.append(miau[1])
    Created_at.append(miau[2][-4:])
    Verified.append(miau[3])



df['Followers'] = Followers
df['Following'] = Following
df['Created_at'] = Created_at
df['Verified'] = Verified



#df.to_csv('/Users/adrihp/Master/MBID01/accountsEMBS.csv')
```

Figura 1: Comienzo del código en donde se obtienen los valores necesarios para el cálculo de métricas.

Con esto obtenos los datos y en las siguientes celdas se ve como se van calculando las métricas y se van añadiendo a un DataFrame de manera organizada. Esto genera un nuevo csv mas completo con las métricas añadidas. Este programa posee una ventaja y es que se puede reutilizar para cualquier cuenta de twitter ya que solo depende de los datos del fichero csv de entrada.

En cuanto al tiempo para el calculo de estas métricas se tardó una hora en completar el código en Python.

```python
#Credibilidad del texto

credibilidad_textual = [20,10,10,20,10,20,20,20,10,20]
df['Credibilidad_textual'] = credibilidad_textual
```

Python

```python
df.head(10)
```

Python

	Unique_ID	org_name	org_url	Twitter_URL	Twitter_handle	earliest_tweet_in_db	number_of_tweets_in_db	Followers	Following	Created_at	Verified	Cr
0	1	Pedro Sanchez	NaN	https://twitter.com/sanchezcastejon	sanchezcastejon	NaN	NaN	1667662	6035	2009	True	
1	2	Estopa	NaN	https://twitter.com/estopaoficial	estopaoficial	NaN	NaN	1010695	723	2011	True	
2	3	Billkilgore	NaN	https://twitter.com/billkilgore_	billkilgore_	NaN	NaN	202580	1590	2017	False	
3	4	Antonio	NaN	https://twitter.com/levmauc	levmauc	NaN	NaN	377948	236172	2016	False	
4	5	Knekro	NaN	https://twitter.com/KNekro	KNekro	NaN	NaN	439769	272	2012	True	
5	6	Crespo	NaN	https://twitter.com/QuantumFracture	QuantumFracture	NaN	NaN	370972	695	2012	True	
6	7	Arcane	NaN	https://twitter.com/arcaneshow	arcaneshow	NaN	NaN	540902	14	2021	True	
7	8	LoL Esports	NaN	https://twitter.com/lolesports	lolesports	NaN	NaN	2184932	549	2012	True	
8	9	Illojuan	NaN	https://twitter.com/LMDShow	LMDShow	NaN	NaN	715494	371	2014	True	
9	10	Alex Riveiro	NaN	https://twitter.com/alex_riveiro	alex_riveiro	NaN	NaN	317667	797	2009	True	

Figura 2: Código donde evaluamos la credibilidad textual.

```python
#Credibilidad de usuario
#Verified 0/50
Credibilidad_de_usuario = []
for element in df['Verified']:
    if element == True:
#Añadimos 20 porque una cuenta verificada pondera la mitad de la credibilidad de usuario.
        Credibilidad_de_usuario.append(20)
    else:
        Credibilidad_de_usuario.append(0)
#Sacados de los papers
Max_account_age = 2006
for i in range (0,10):
    Credibilidad_de_usuario[i] = Credibilidad_de_usuario[i]+((2023-int(df['Created_at'][i]))/(2023-Max_account_age))*20
df['Credibilidad_de_usuario'] = Credibilidad_de_usuario
```

Python

```python
df
```

Python

	Unique_ID	org_name	org_url	Twitter_URL	Twitter_handle	earliest_tweet_in_db	number_of_tweets_in_db	Followers	Following	Created_at	Verified	Cr
0	1	Pedro Sanchez	NaN	https://twitter.com/sanchezcastejon	sanchezcastejon	NaN	NaN	1667662	6035	2009	True	
1	2	Estopa	NaN	https://twitter.com/estopaoficial	estopaoficial	NaN	NaN	1010695	723	2011	True	
2	3	Billkilgore	NaN	https://twitter.com/billkilgore_	billkilgore_	NaN	NaN	202580	1590	2017	False	
3	4	Antonio	NaN	https://twitter.com/levmauc	levmauc	NaN	NaN	377948	236172	2016	False	
4	5	Knekro	NaN	https://twitter.com/KNekro	KNekro	NaN	NaN	439769	272	2012	True	
5	6	Crespo	NaN	https://twitter.com/QuantumFracture	QuantumFracture	NaN	NaN	370972	695	2012	True	
6	7	Arcane	NaN	https://twitter.com/arcaneshow	arcaneshow	NaN	NaN	540902	14	2021	True	
7	8	LoL Esports	NaN	https://twitter.com/lolesports	lolesports	NaN	NaN	2184932	549	2012	True	
8	9	Illojuan	NaN	https://twitter.com/LMDShow	LMDShow	NaN	NaN	715494	371	2014	True	
9	10	Alex Riveiro	NaN	https://twitter.com/alex_riveiro	alex_riveiro	NaN	NaN	317667	797	2009	True	

Figura 3: Código encargado del cálculo de la credibilidad de usuario.

```python
#Credibilidad social
#Followers impact

Credibilidad_social = []
#Sacado del paper
max_followers = 2200000

for element in df['Followers']:
    Credibilidad_social.append((element/max_followers)*20)

#Followers/following

for i in range(0,10):
    Credibilidad_social[i] = Credibilidad_social[i] + (df['Followers'][i]/(df['Followers'][i]+df['Following'][i]))*20

df['Credibilidad_social'] = Credibilidad_social
```

✓ 0.2s Python

```python
df
```

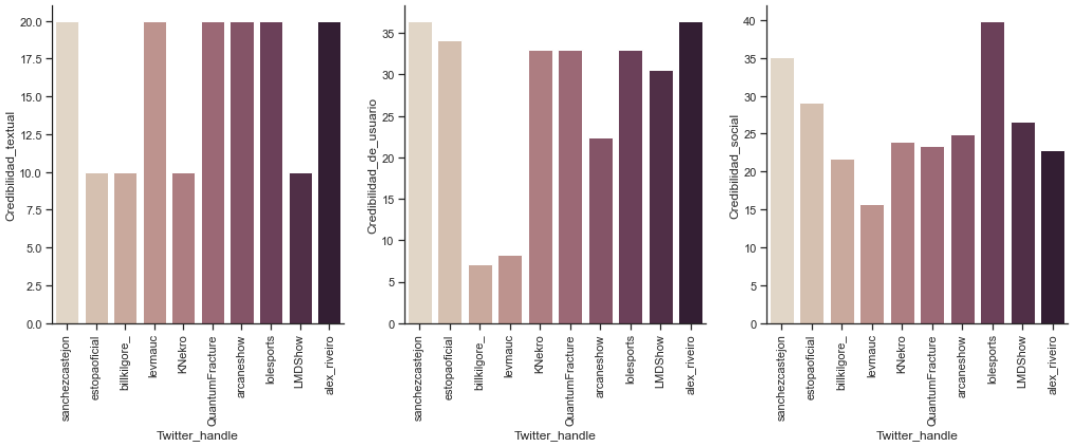
✓ 0.2s Python

	Unique_ID	org_name	org_url	Twitter_URL	Twitter_handle	earliest_tweet_in_db	number_of_tweets_in_db	Followers	Following	Created_at	Verified	Cr
0	1	Pedro Sanchez	NaN	https://twitter.com/sanchezcastejon	sanchezcastejon	NaN	NaN	1668260	6035	2009	True	
1	2	Estopa	NaN	https://twitter.com/estopaoficial	estopaoficial	NaN	NaN	1010811	723	2011	True	
2	3	Billkilgore	NaN	https://twitter.com/billkilgore_	billkilgore_	NaN	NaN	203150	1619	2017	False	
3	4	Antonio	NaN	https://twitter.com/levmauc	levmauc	NaN	NaN	378028	236126	2016	False	
4	5	Knekro	NaN	https://twitter.com/KNekro	KNekro	NaN	NaN	440036	274	2012	True	
5	6	Crespo	NaN	https://twitter.com/QuantumFracture	QuantumFracture	NaN	NaN	371803	695	2012	True	
6	7	Arcane	NaN	https://twitter.com/arcaneshow	arcaneshow	NaN	NaN	541408	14	2021	True	
7	8	LoL Esports	NaN	https://twitter.com/lolesports	lolesports	NaN	NaN	2189198	549	2012	True	
8	9	Illojuan	NaN	https://twitter.com/LMDShow	LMDShow	NaN	NaN	720339	371	2014	True	
9	10	Alex Riveiro	NaN	https://twitter.com/alex_riveiro	alex_riveiro	NaN	NaN	317713	799	2009	True	

Figura 4: Código encargado del cálculo de la credibilidad social.

3. Gráficas y análisis

Estas gráficas también se hicieron en el mismo programa que el cálculo de métricas usando la librería matplotlib. Vamos a comenzar viendo las gráficas de las tres tipos de métricas usadas:



(a) Credibilidad textual (b) Credibilidad de usuario (c) Credibilidad social

Vamos ahora a ver los resultados de la credibilidad total y hacer un análisis de los resultados.

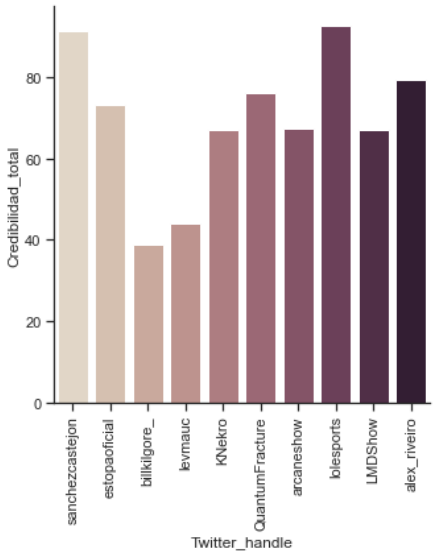


Figura 6: Credibilidad total. Al ser esta credibilidad por tweet, cada barra tendra un total de 200 tweets puesto que la credibilidad de cada tweet no varía en función de la cuenta.

3.1. Discusión de los resultados

Primero que nada vamos a separar los resultados:

- CT >75 % = Buena credibilidad
- 65 % < CT < 75 % = Credibilidad media
- CT < 65 % = Mala credibilidad

Comenzando con las cuentas que poseen una buena credibilidad podemos ver que estas son cuatro: Pedro Sánchez, QuantumFracture, Alex Riveiro y LoLEsports. Estas cuentas pertenecen a organizaciones de deportes electronicos, el presidente de españa y divulgadores ciéntificos, es decir, son cuentas que deben poseer una alta credibilidad sobre todo en el caso de Pedro Sánchez y LoLEsports que por su naturaleza deberían ser las cuentas mas creíbles, cosa que se demuestra con nuestras métricas.

Siguiendo con las cuentas de credibilidad media en este grupo entran las siguientes: Knekro, Illojuan, Estopa y Arcaneshow. Estas cuentas pertenecen a streamers, cantante y la famosa serie de Arcane. En el caso de Knekro, Illojuan y Estopa es normal que estén en este rango ya que son cuentas que usan un lenguaje mas coloquial y aunque si que son cuentas verificadas no dejan de perder credibilidad por esto. En el caso de Arcane cabría esperar que entrara en el rango anterior, de alta credibilidad, sin embargo al ser este tan reciente (2021) hace que, según nuestras métricas, su credibilidad baje.

Para el último grupo solo encontramos dos cuentas: Billkilgore y levmauc. En el caso de Billkilgore es normal que caiga dentro de este grupo, es una cuenta no verificada el cual su uso principal es para poner memes y bromas, eso hace que su credibilidad baje y se demuestra con nuestras métricas. Por otro lado, levmauc es una cuenta muy famosa que se dedica a compartir información sobre peliculas, la razón de que entre en este grupo es que posee una cantidad increible de following, eso hace que su credibilidad social sea baja. Además, su credibilidad de usuario también es bastante baja porque no es una cuenta verificada.

Los resultados son los esperados salvo el caso de levmauc que si considero una cuenta con una credibilidad media-alta que sin embargo por la situación de su cuenta hace que su credibilidad baje bastante.