THE MOOD OF TWITTERERS IN MEXICO

(EL ESTADO DE ÁNIMO DE LOS TUITEROS EN MÉXICO)

Gerardo Leyva

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

October, 2018

The three pillars of official statistics

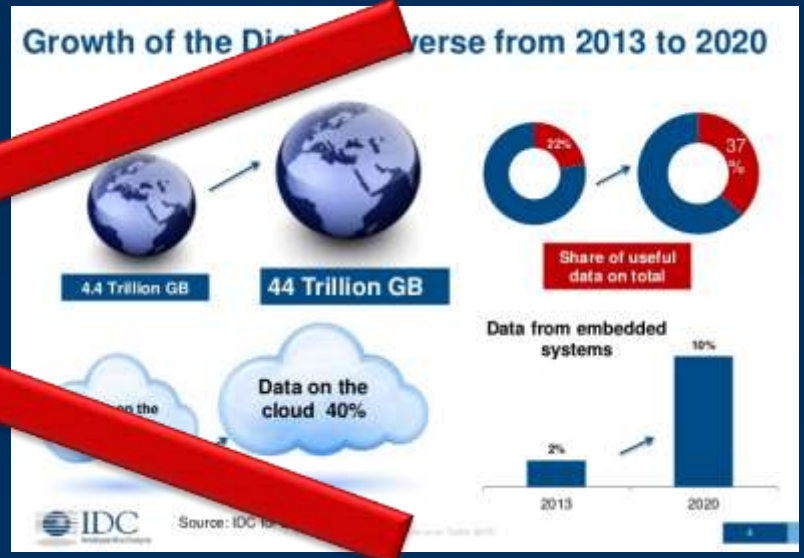
The ~~three~~ four pillars of official statistics

The Big Data definition evolves

- Initially it was about…
  - *Volume*
  - *Velocity*
  - *Variety*

  3 V's

  - *Veracity*
  - *Valu*



- Instead…

Big Data is a flexible approach to use and re-use the totality of a data set, structured or not, in a diversity of possible purposes, normally different to those that originated the information set in the first place.

BIG DATA

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."

Dan Ariely

Big Data (Google trends)

https://www.google.com.mx/trends/

@abxda

PARADIGMS

- Small data
- Big data

Convergence of two agendas

- Big data.
- Subjective Well Being (Martin Seligman).

General idea

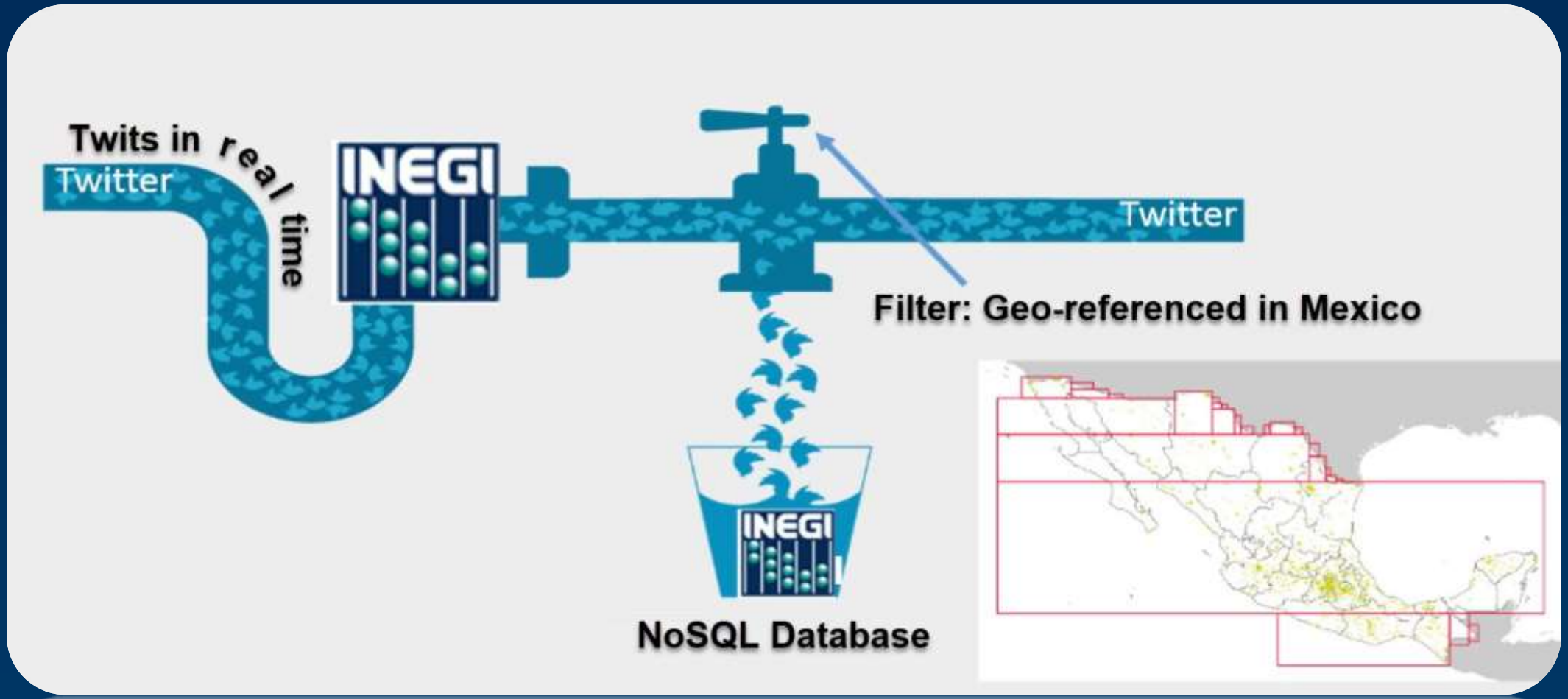
Goal: Automatically measure and report the mood of twitterers in México.

Method: supervised learning

- Humans tag a training set of tweets: 
- The system learns to automatically tag (classify) tweets as close as possible to the way humans would have done it.

Since February 2014

Collecting tweets



More than 300 million tweets

Set of tagged tweets

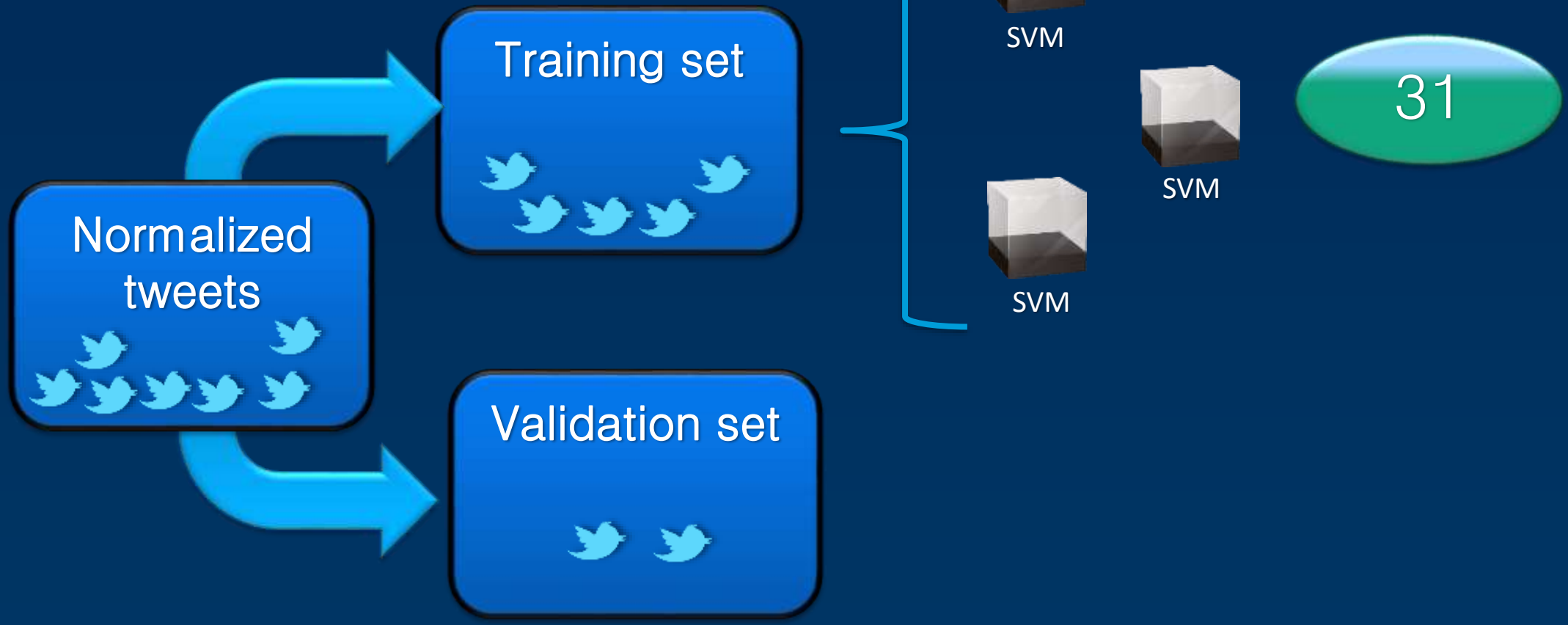
- 9 330 people from Universidad Tecmilenio and INEGI.
- Manually tagged 54 131 tweets.
- Multiple tagging of each tweet.

- Classification system:

https://cienciadedatos.inegi.org.mx/pioanalisis/

Estar enamorada es como ir en un Ferrari a 240 kms/h. Se siente CHINGON pero sabes que en cualquier momento viene el putazo (:

Final solution



Optimal results
(Assamble of SVM)

Goal: Automatically classifying tweets

Unclassified tweets

Normalization
Vector representation
Classification

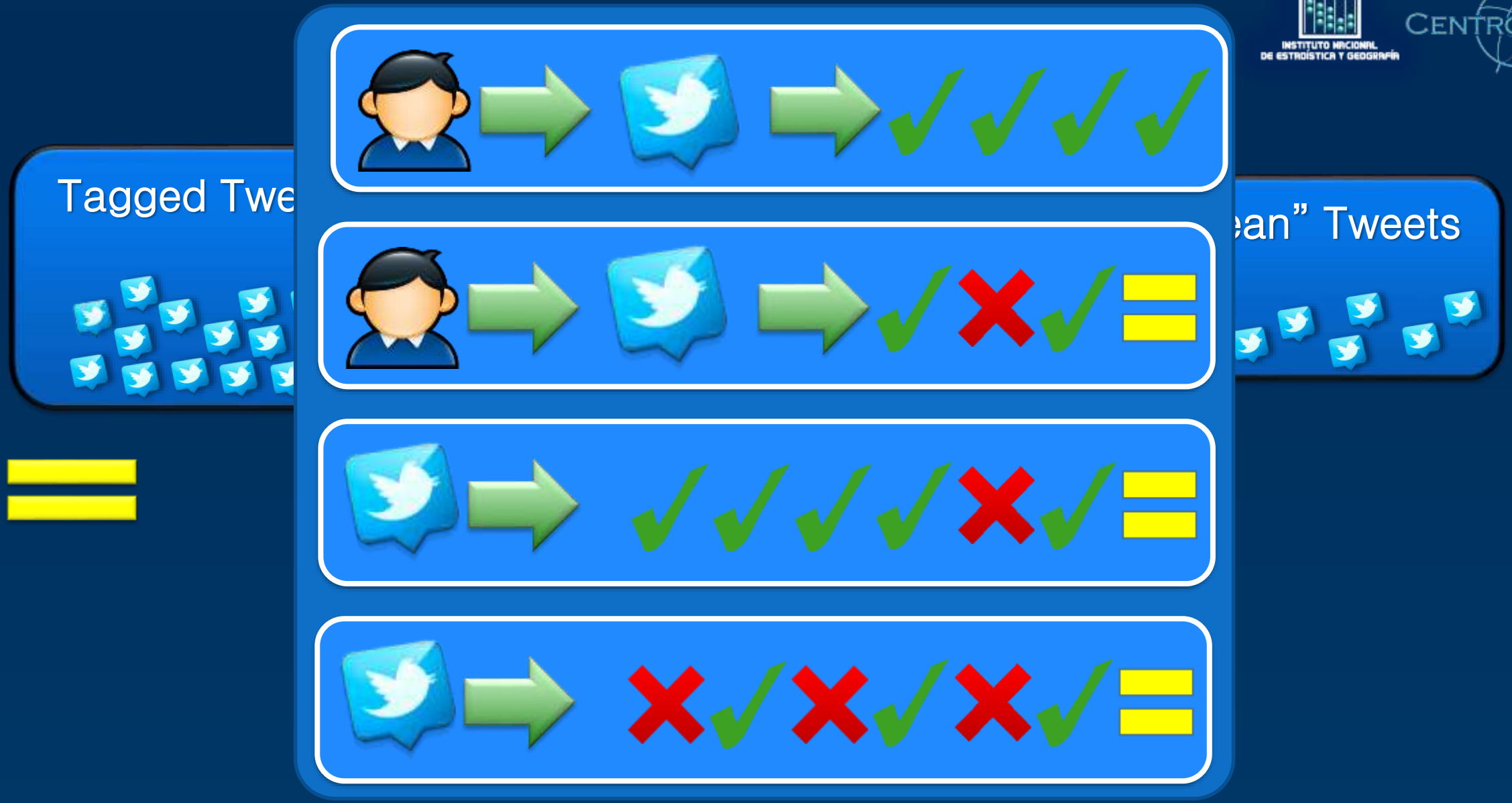
Hundreds of millions of tagged tweets

The process for sentiment classification

- Cleaning
- Text normalization
- Vector representation of text
- Training of the *Machine Learning algorithm*
- Text classification on the fly

Cleaning of the tagged set

Cleaning of the tagged set (cleaning)



Text normalization

Q-Grams 3,4,5,7
(q=4)

romper → romp

perder → perd

amigs → amig

Polarity of Emoticons
(polarity tag)

+1 0 −1

Others

Number sustitution

URL sustitution

User sustitution

Lower case transformation

Example of text normalization

ORIGINAL TEXT:

pésiiiimo auto :(@autoX fallan frenos y sistema de entretenimiento; no lo compren

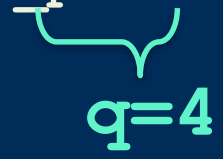
NORMALIZED TEXT:

pesiiiimo auto **_negativo** **_user** fallan frenos y sistema de entretenimiento ; lo **no_compren**

Example of text normalization with q-grams

{_pes, pesi, esii, siii, iiii, iiim, iimo, imo_, mo_a, o_au, _aut, auto, uto_, to__, o__n, __ne, _neg, nega, egat, gati, ativ, tivo, ivo_, vo__, o__u, __us, _use, user, ser_, er_f, r_fa, _fal, fall, alla, llan, lan_, an_f, n_fr, _fre, fren, reno, enos, nos_, os_y, s_y_, _y_s, y_si, _sis, sist, iste, stem, tema, ema_, ma_d, a_de, _de_, de_e, e_en, _ent, entr, ntre, tret, rete, eten, teni, enim, nimi, imie, mien, ient, ento, nto_, to_;, o_;_, _;_l, ;_lo, _lo_, lo_n, o_no, _no_, no_c, o_co, _com, comp, ompr, mpre, pren, ren_ }

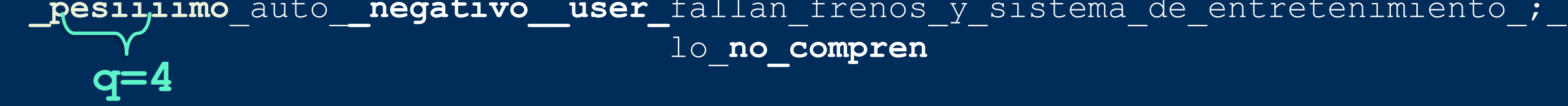
Example of text normalization with q-grams



_pesiiiimo_auto__**negativo__user_**fallan_frenos_y_sistema_de_entretenimiento_;_lo_**no_compren**

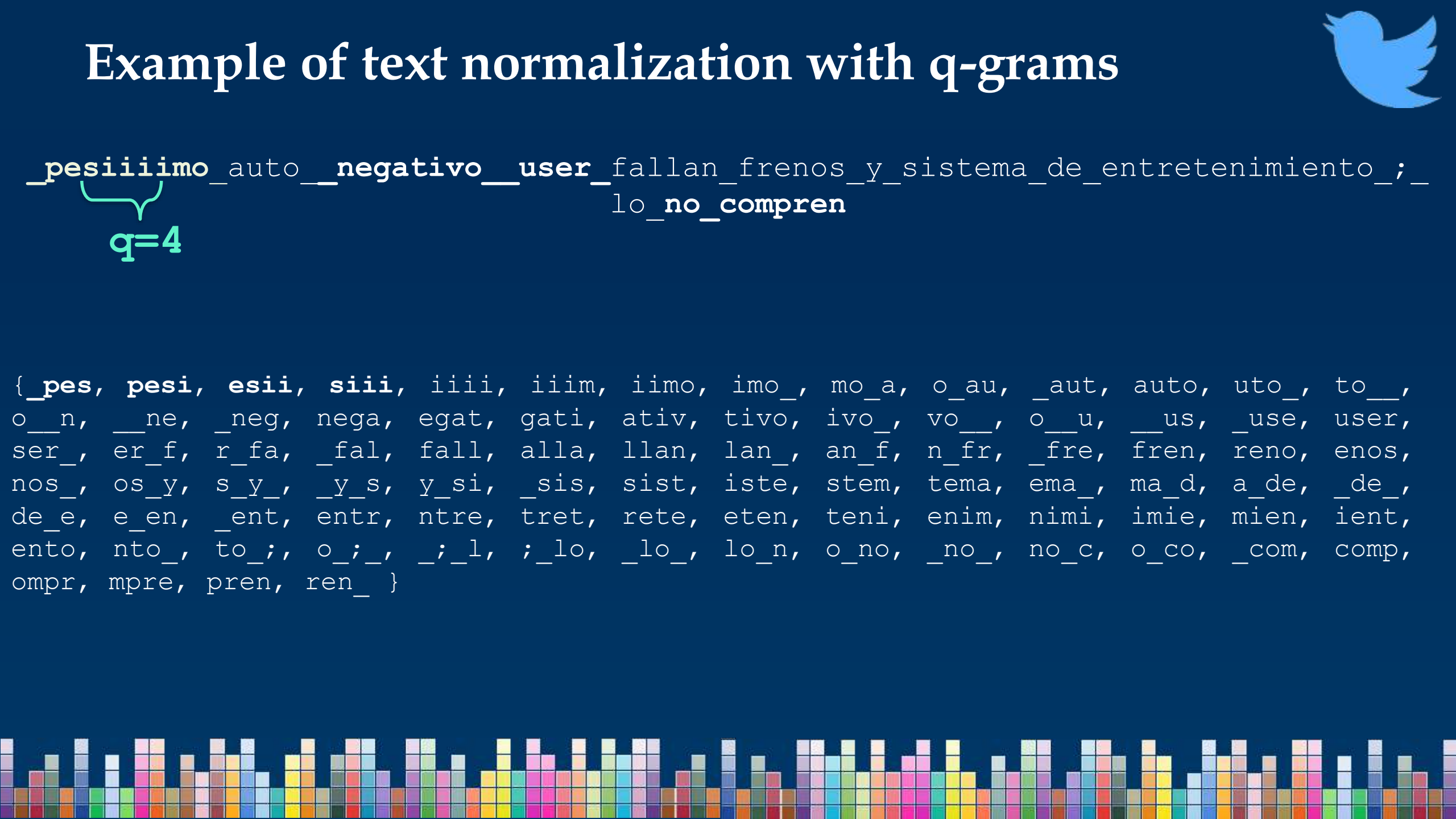
{**_pes**, **pesi**, esii, siii, iiii, iiim, iimo, imo_, mo_a, o_au, _aut, auto, uto_, to__, o__n, __ne, _neg, nega, egat, gati, ativ, tivo, ivo_, vo__, o__u, __us, _use, user, ser_, er_f, r_fa, _fal, fall, alla, llan, lan_, an_f, n_fr, _fre, fren, reno, enos, nos_, os_y, s_y_, _y_s, y_si, _sis, sist, iste, stem, tema, ema_, ma_d, a_de, _de_, de_e, e_en, _ent, entr, ntre, tret, rete, eten, teni, enim, nimi, imie, mien, ient, ento, nto_, to_;, o_;_, _;_l, ;_lo, _lo_, lo_n, o_no, _no_, no_c, o_co, _com, comp, ompr, mpre, pren, ren_ }

Example of text normalization with q-grams



{**_pes**, **pesi**, **esii**, siii, iiii, iiim, iimo, imo_, mo_a, o_au, _aut, auto, uto_, to__, o__n, __ne, _neg, nega, egat, gati, ativ, tivo, ivo_, vo__, o__u, __us, _use, user, ser_, er_f, r_fa, _fal, fall, alla, llan, lan_, an_f, n_fr, _fre, fren, reno, enos, nos_, os_y, s_y_, _y_s, y_si, _sis, sist, iste, stem, tema, ema_, ma_d, a_de, _de_, de_e, e_en, _ent, entr, ntre, tret, rete, eten, teni, enim, nimi, imie, mien, ient, ento, nto_, to_;, o_;_, _;_l, ;_lo, _lo_, lo_n, o_no, _no_, no_c, o_co, _com, comp, ompr, mpre, pren, ren_ }

Example of text normalization with q-grams



{_pes, pesi, esii, siii, iiii, iiim, iimo, imo_, mo_a, o_au, _aut, auto, uto_, to__, o__n, __ne, _neg, nega, egat, gati, ativ, tivo, ivo_, vo__, o__u, __us, _use, user, ser_, er_f, r_fa, _fal, fall, alla, llan, lan_, an_f, n_fr, _fre, fren, reno, enos, nos_, os_y, s_y_, _y_s, y_si, _sis, sist, iste, stem, tema, ema_, ma_d, a_de, _de_, de_e, e_en, _ent, entr, ntre, tret, rete, eten, teni, enim, nimi, imie, mien, ient, ento, nto_, to_;, o_;_, _;_l, ;_lo, _lo_, lo_n, o_no, _no_, no_c, o_co, _com, comp, ompr, mpre, pren, ren_ }

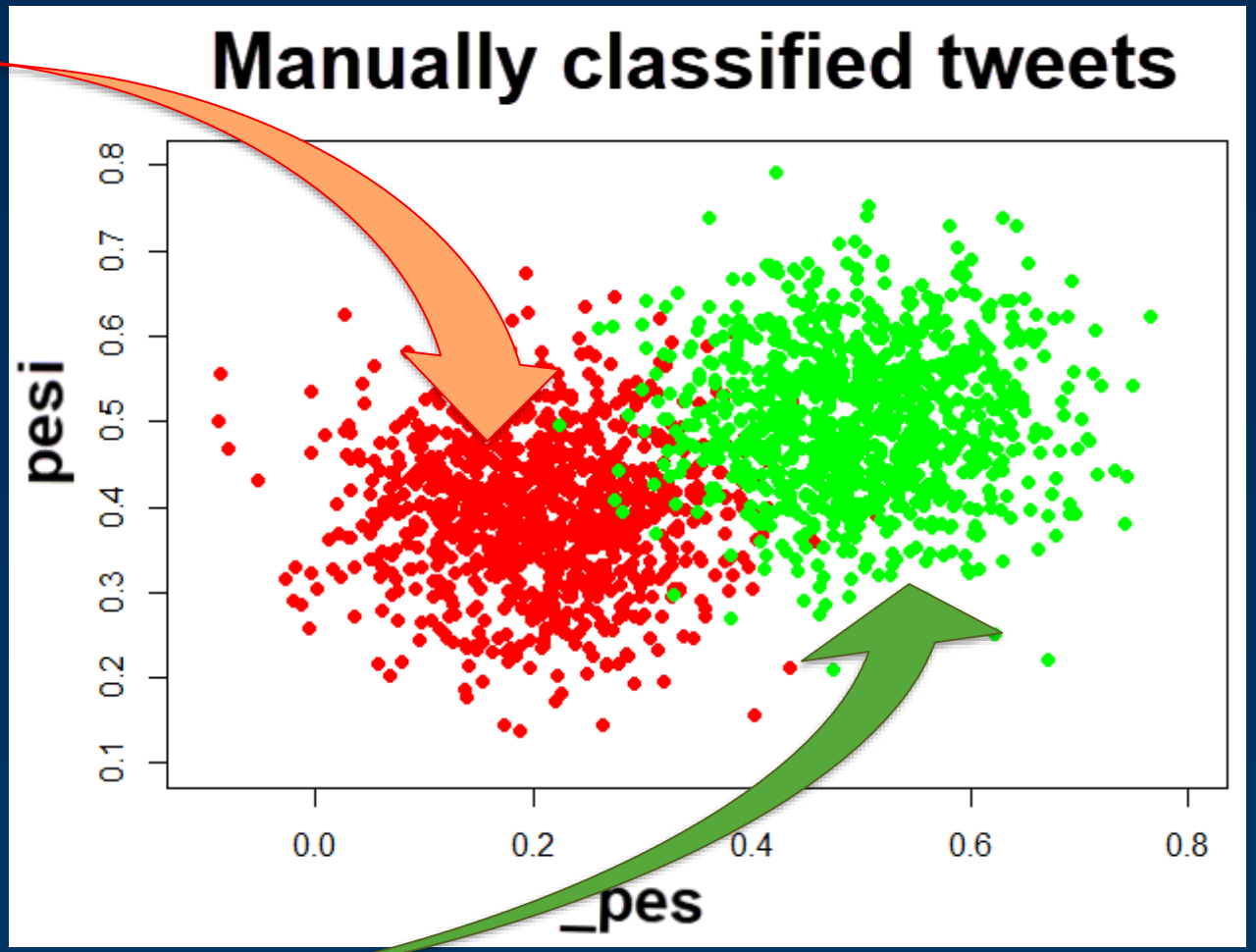
Vectoral representation of the text

Term frequency- inverse document frequency
TF IDF

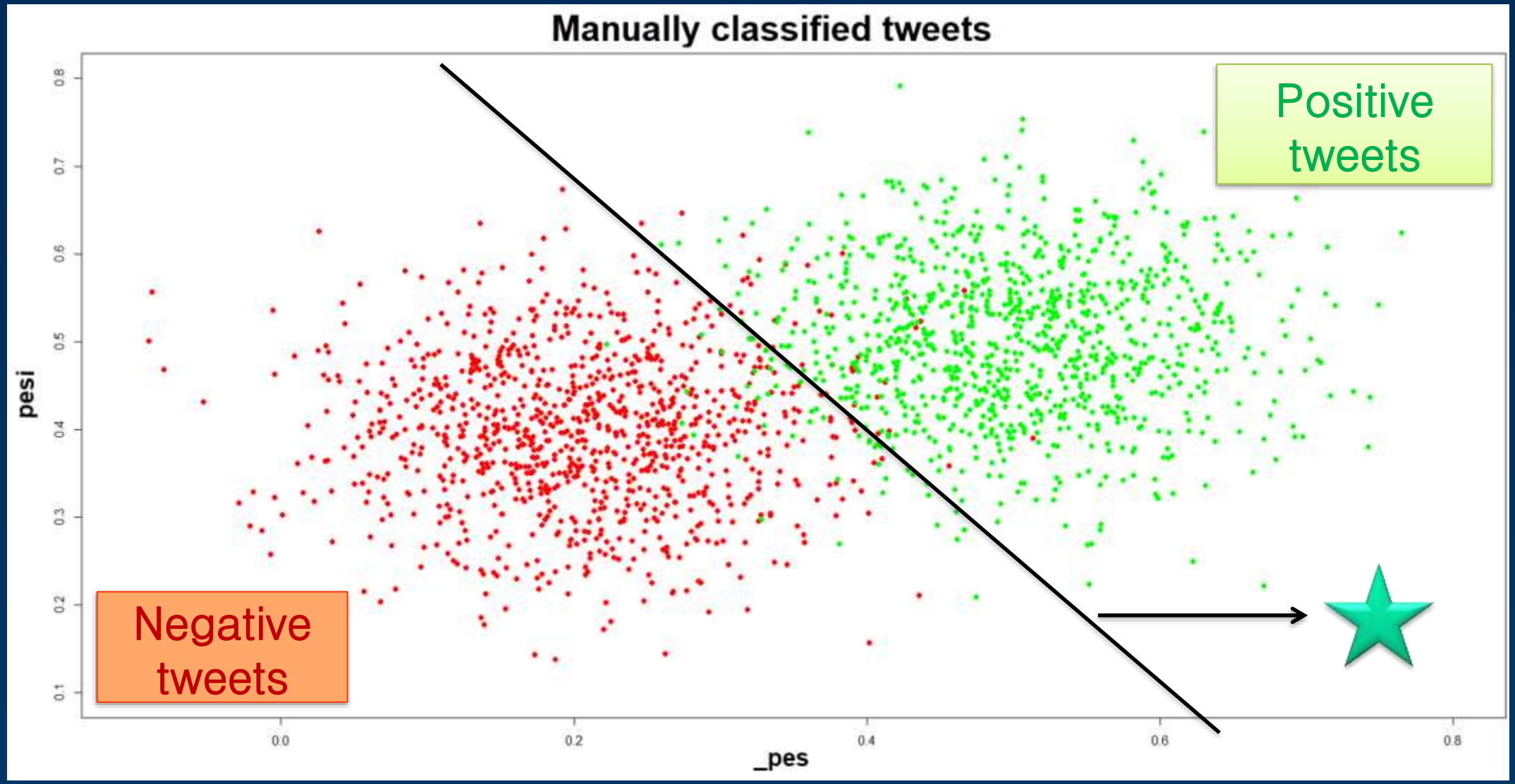
	tag	_pes	pesi	esii	siii	iiii	iiim	iimo	imo_	mo_a	o_au	_aut	auto
Tweet 1	NEGATIVE	0.22	0.48	0.10	0.25	0.21	0.21	0.21	0.21	0.30	0.30	0.10	0.50
Tweet 2	NEGATIVE	0.12	0.55	0.20	0.10	0.30	0.24	0.00	0.00	0.00	0.00	0.00	0.00
Tweet 3	NEGATIVE	0.25	0.39	0.00	0.00	0.00	0.00	0.48	0.00	0.70	0.20	0.30	0.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Tweet n	POSITIVE	0.6	0.35	0.00	0.00	0.00	0.00	0.48	0.00	0.70	0.20	0.30	0.50

Machine learning algorithm SVM

	tag	_pes	pesi
Tweet 1	NEGATIVE	0.22	0.48
Tweet 2	NEGATIVE	0.12	0.55
Tweet 3	NEGATIVE	0.25	0.39
...	...	...	...
Tweet n	POSITIVE	0.6	0.35



Training the SVM algorithm



The task of text classification…in a nutshell:

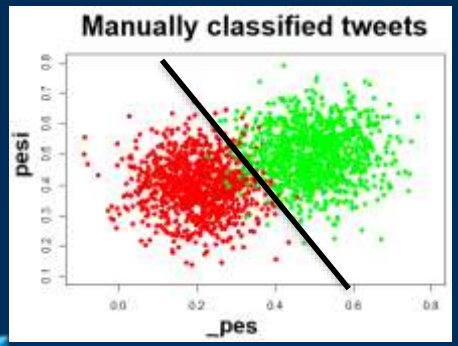
Tagged tweets

Normalization and vector representation

Term frequency- inverse document frequency
TF IDF

	tag	_pes	pesi	esii	siii	iiii	iiim	iimo	imo_	mo_a	o_au	_aut	auto
Tweet 1	NEGATIVE	0.22	0.48	0.10	0.25	0.21	0.21	0.21	0.21	0.30	0.30	0.10	0.50
Tweet 2	NEGATIVE	0.12	0.55	0.20	0.10	0.30	0.24	0.00	0.00	0.00	0.00	0.00	0.00
Tweet 3	NEGATIVE	0.25	0.39	0.00	0.00	0.00	0.00	0.48	0.00	0.70	0.20	0.30	0.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Tweet n	POSITIVE	0.6	0.35	0.00	0.00	0.00	0.00	0.48	0.00	0.70	0.20	0.30	0.50

Training

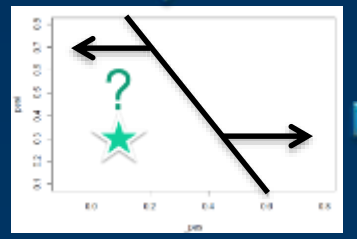


Production

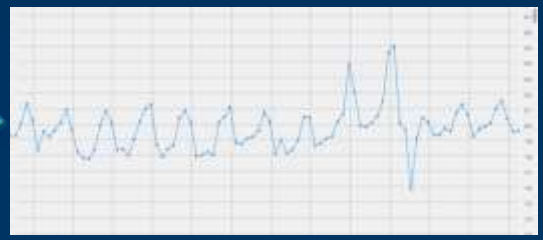
New tweet

Normalization and vector representation

	_pes	pesi	esii	siii	iiii	iiim	iimo	imo_	mo_a	o_a
New tweet	0.1	0.3	0	0.1	0.1	0.16	0.45	0.3	0	0



Decision rule

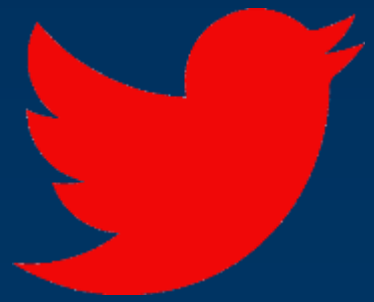


The mood of tweeterers

Positivity quotient

POSITIVES

NEGATIVES

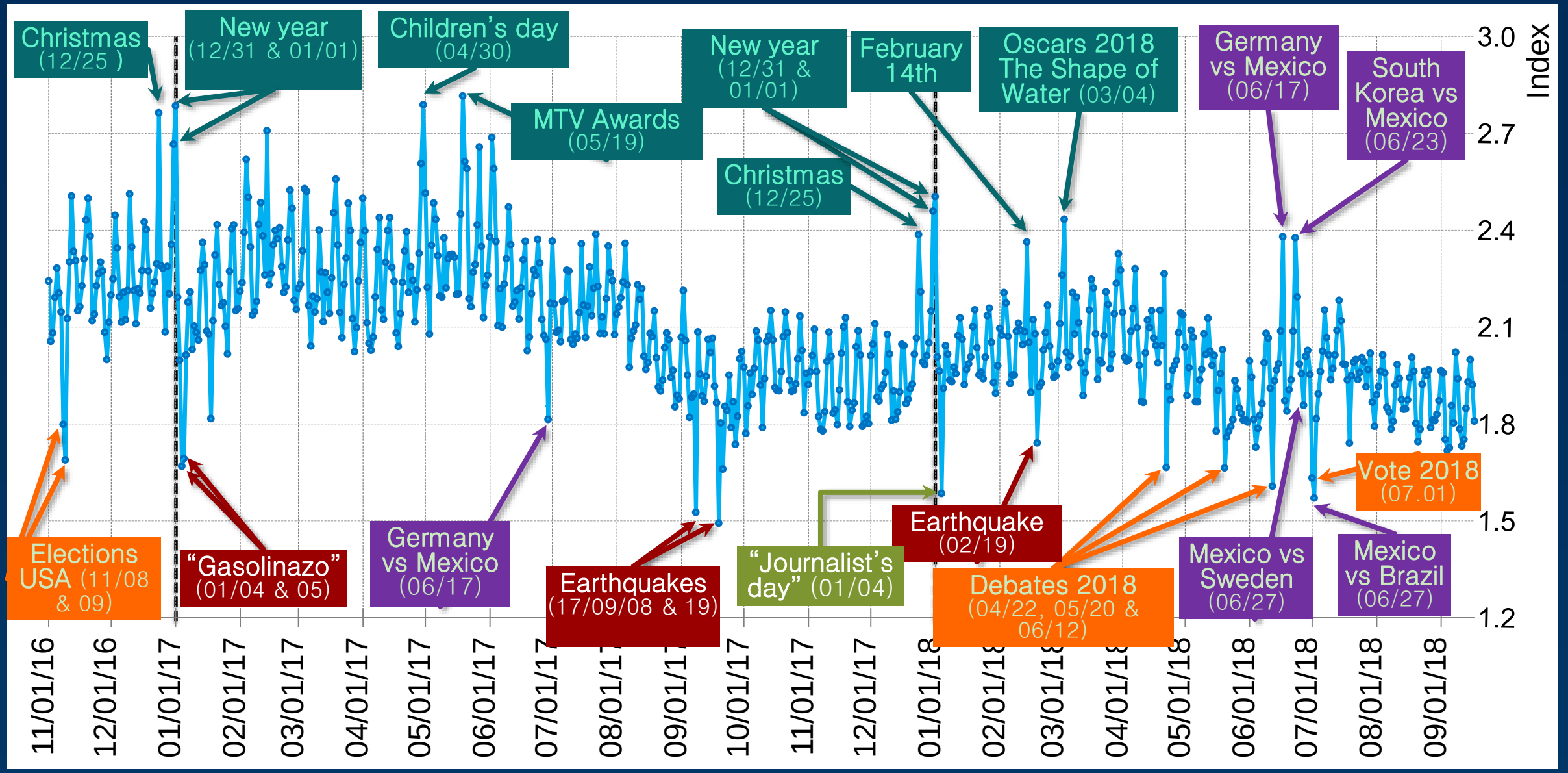


=

Positivity quotient

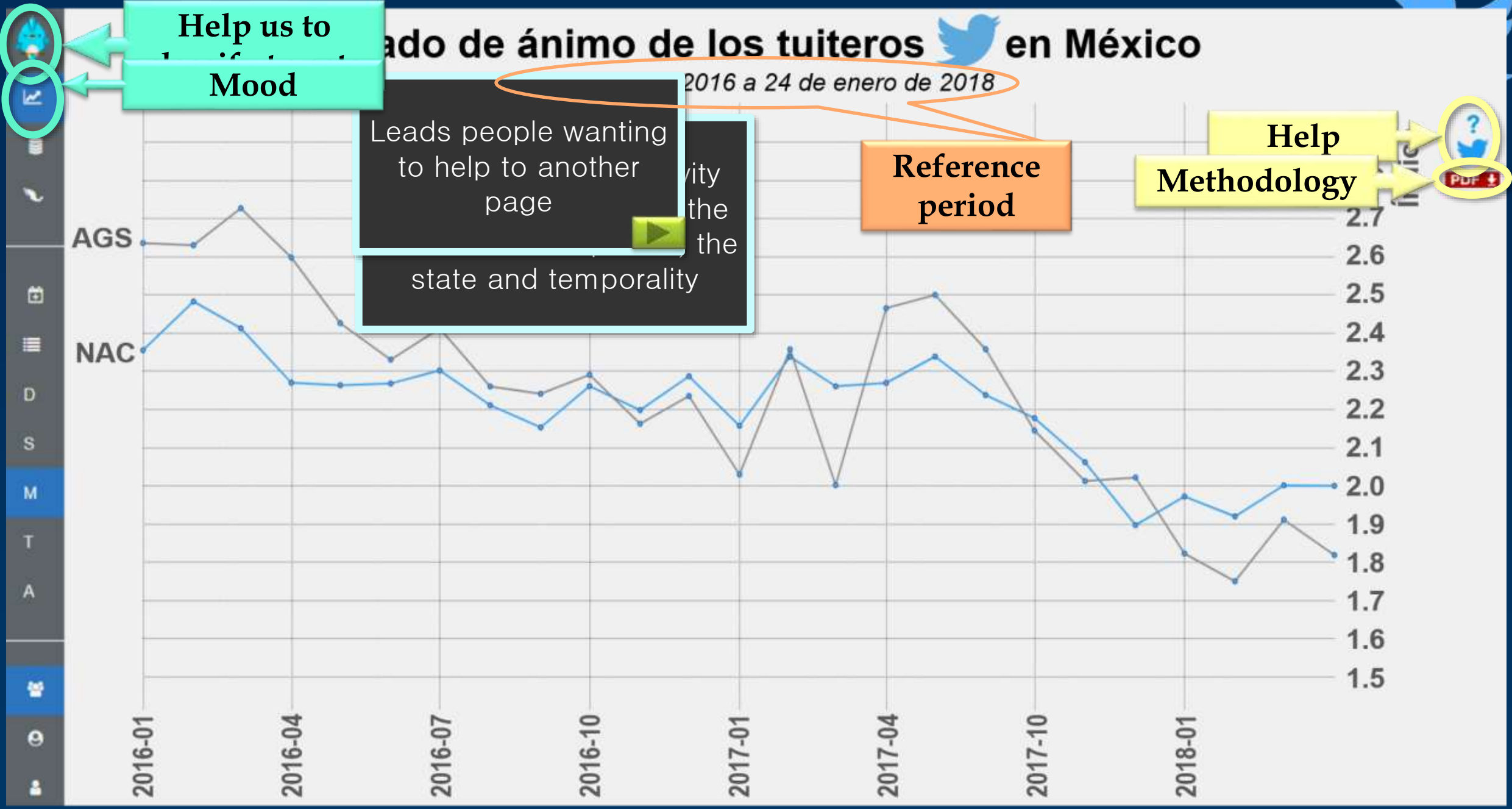
The mood of tweeters in Mexico

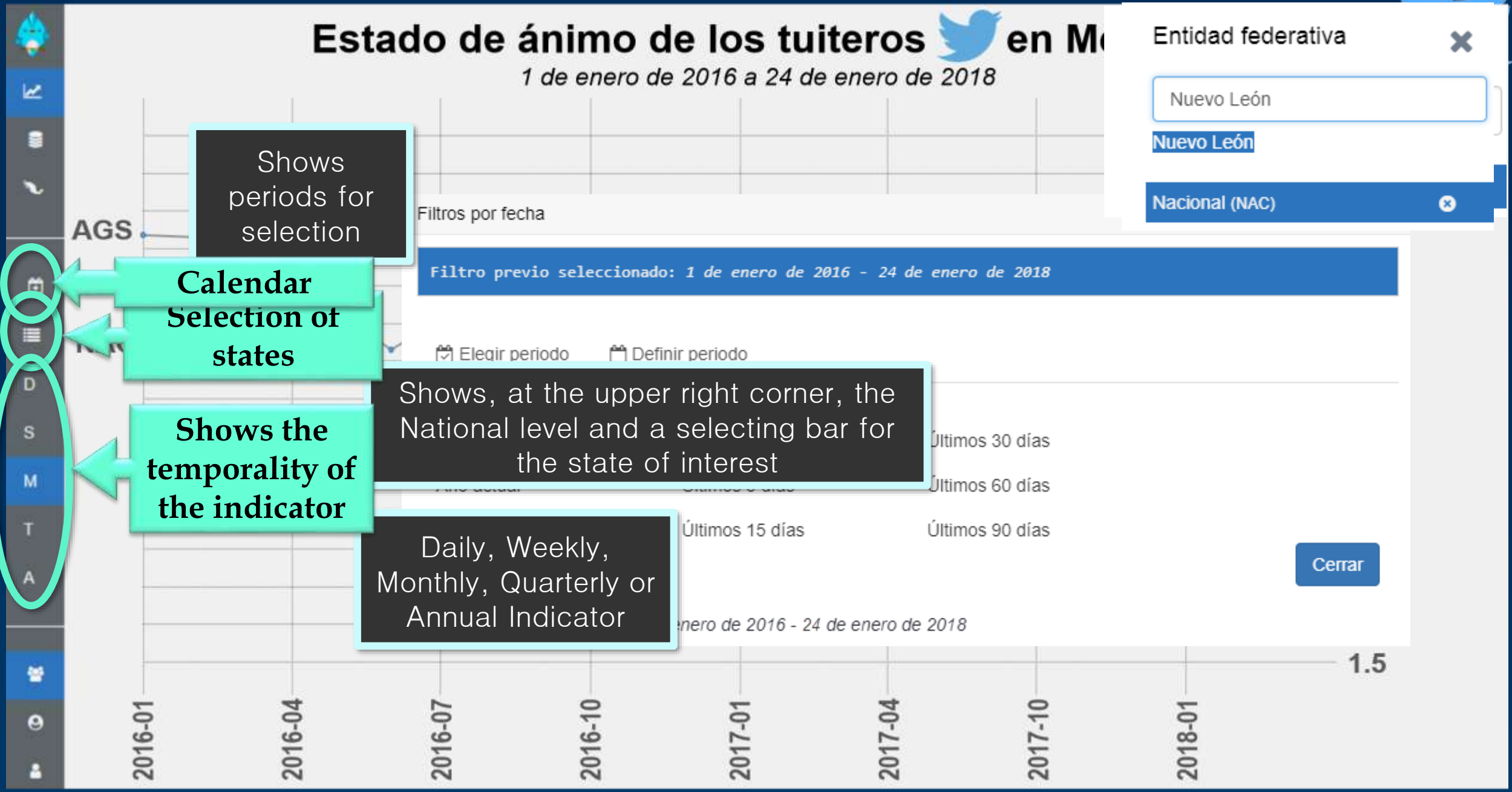
Showing 2016/Nov-2018/Sep (daily)



Link:

- http://www.inegi.org.mx/
- http://www.beta.inegi.org.mx/app/animotuitero/#/app/multiline


Help us to
Mood
ado de ánimo de los tuiteros en México
2016 a 24 de enero de 2018
Leads people wanting to help to another page
vity
the
the
state and temporality
Reference period
Help
Methodology
AGS
NAC
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04
2017-10
2018-01

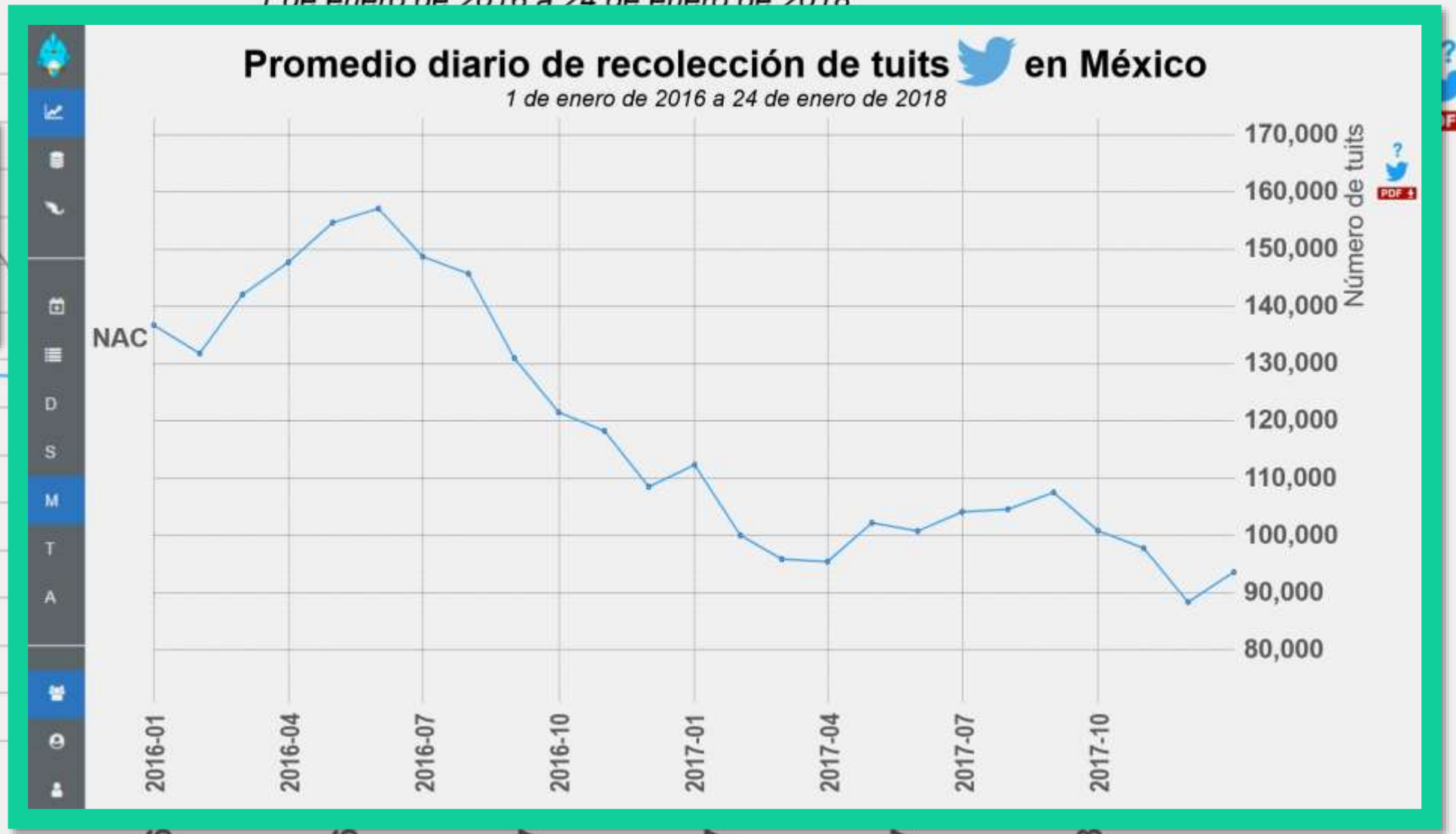

Estado de ánimo de los tuiteros en M
1 de enero de 2016 a 24 de enero de 2018
Entidad federativa
Nuevo León
Nuevo León
Nacional (NAC)
Shows periods for selection
Calendar
Selection of states
Shows the temporality of the indicator
Shows, at the upper right corner, the National level and a selecting bar for the state of interest
Daily, Weekly, Monthly, Quarterly or Annual Indicator
AGS
Filtros por fecha
Filtro previo seleccionado: 1 de enero de 2016 - 24 de enero de 2018
Elegir periodo
Definir periodo
D
S
M
T
A
Últimos 30 días
Últimos 60 días
Últimos 15 días
Últimos 90 días
Cerrar
enero de 2016 - 24 de enero de 2018
1.5
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04
2017-10
2018-01

Estado de ánimo de los tuiteros en México

1 de enero de 2016 a 24 de enero de 2018

Gathering

Shows the number of tweets gathered




Mapa del estado de ánimo de los tuiteros en México
1 de enero de 2016 a 24 de enero de 2018
Índice = Positivos (♥) / Negativos (♥)
Map
Shows, on the map, the states coloured according to the positivity quotien
Shows the tweets of all people in the state or the country
All
Residents
Shows the tweets of people visiting the state
Visitors

Other INEGI projects with Twitter:

- Domestic tourism.
- Mental health.
- Mobility in Mexico City.
- New agglomerations.
- Consumer confidence.
- Insecurity.

Other INEGI projects with big data:

- CFE electricity consumption for nowcasting of industrial activity.
- Use of satellite images for diverse purposes including land cover, agricultural activity and new settlements.
- Cooperation with Telefonica and BBVA-Bancomer to generate a rapid response system to face natural disasters.
- Web scraping and scanner data for prices.

¡Thank you!

Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa

@ INEGI_INFORMA

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA