«Big data» versus «small data»: the case of “gripe” (flu) in Spanish

Antonio Moreno-Sandoval (Department of Linguistics UAM & IIC)

Esteban Moro (Department of Mathematics UC3M & IIC)

The starting point

“ ‘Big data hubris’ is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.”

Lazer, Kennedy, King and Vespignani: The Parable of Google Flu: Traps in Big Data Analysis. *Science*, vol. 343, 14 March 2014

The Google Flu Tool (GFT) failure

→ *Nature* reported in Feb 2013 that GFT was predicting more than double of doctor visits for influenza than the USA medical authorities.

- Method: **counting flu-related searches**
- GFT 2009:
  - "big data were overfitting the small number of cases"
  - "GFT was part flu detector, part winter detector"
- GFT 2013, again, "has been persistently overestimating flu prevalence for a much longer time"

The error cause, according to Google

1. **Big data overestimation**

2. **Algorithm dynamics**, which pollute and manipulate data by expanding rumors and trending topics.

What is Big Data?

- The Four V's of Big Data by IBM:
  1. **Volume**: scale of data (2.3 trillion Gb created each day)
  2. **Velocity**: analysis of streaming data
  3. **Variety**: different forms of data (text, video, healthcare records, etc.)
  4. **Veracity**: uncertainty of data

Wikipedia definition

Big data is a broad term for **data sets** so **large and complex** that traditional data processing applications are inadequate.

The term often refers simply to the use of **predictive analytics** to extract value from the **(unstructured) data**.

What is ‘Big LANGUAGE Data’

- 30 billion pieces of content are shared on Facebook every month
- 4 billion hours of video on YouTube each month
- 400 million tweets are sent per day by about 200 million monthly active users.
- In dozens of different languages.

Can we linguistically process big language data?

- Our experiment: to replicate Google's using Twitter messages in Spanish that were:
  1. geolocalized = Spain
  2. included the word 'gripe'
  3. time span: from Jan 2012 to Aug 2014

  → we collected our Spanish Flu Corpus on Twitter

Spanish Flu Corpus in Twitter

- 2759 tweets (including RT)
- 327072 ‘words’
- 140 weeks
- sent from locations in Spain

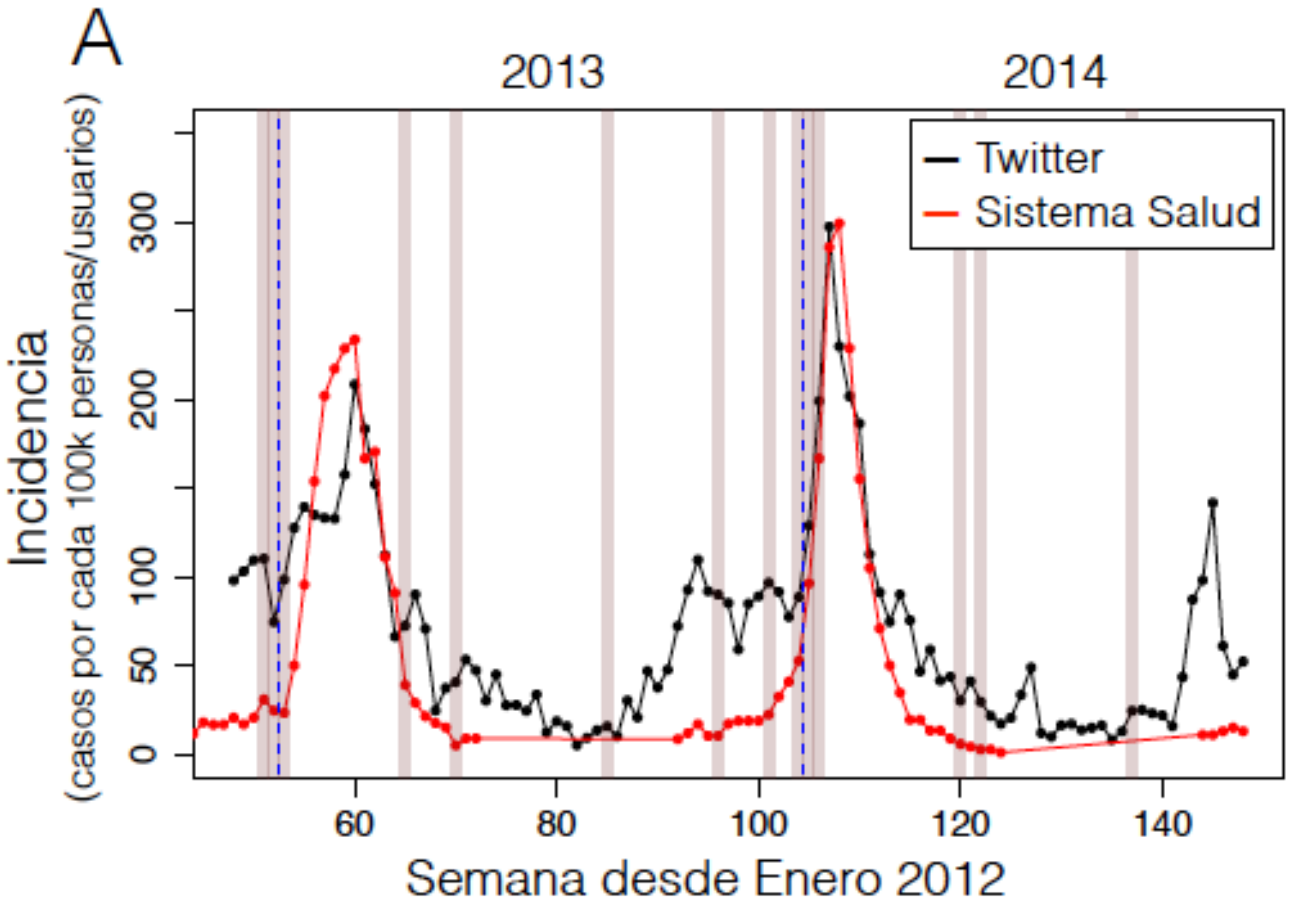
Procedure

- **Hypothesis**: the increased number of messages with **GRIPE** is a predictor of an approaching peak of cases
- **Verification**: to check the prediction against the reported cases by the Spanish Health System (real data sent by doctor in health centers and hospitals)

Elimination of noise in data

- noise: institutional or press messages:
  - “100.000 personas aún no han podido vacunarse contra la gripe”
- → we removed all messages with a URL
- → ‘good data’:
  - → “estoy en la cama con gripe”

Results


A
2013
2014
Twitter
Sistema Salud
Incidencia
(casos por cada 100k personas/usuarios)
0
50
100
200
300
60
80
100
120
140
Semana desde Enero 2012

Discussion

- The messages on Twitter in Spanish also magnify the real cases of flu, as Google (http://www.google.org/flutrends/es/#ES.)
- GFP is based on **flu-related searches**. They compared the query counts with the flu surveillance systems

The problem: VERACITY in data

- Analysis of the Flu Corpus to discover what factors contribute negatively to the prediction.

→ Distinguishing **Good Data** from **Bad Data**

Bad data

- Figurative and humorous language:

a joke:

“No sé si es un constipado **virulento** o una **GRIPE virurápida**.”

‘ I don’t know if this is a virulent cold or a virufast flu’

→ separating real cases from figurative ones requires a good management of pragmatic aspects (intention, irony, metaphor…)

More Bad Data

- when the text contains ‘gripe’ but the speaker doesn’t suffers from it:

“Acabo de ver un anuncio de Gelocatil gripe y me he acordado de @...”

‘I just saw an ad of XXX flu and I have remembered @’

→ all those wrong cases contribute to overestimation.

Some refinements

- Finding synonyms and variants of ‘gripe’ in the Twitter corpus: ‘gripazo’ ‘griposo’
- Detecting sequences that indicate that someone has contracted this disease:
  - “con la gripe en casa”
  - “menudo gripazo he pillado”
  - “con tos y moqueando”
  - “la gripe me mata”

→ THERE ARE VERY FEW LEXICAL AND SYNTACTIC PATTERNS IN THE TWITTER CORPUS

collocations ‘tener’ & ‘gripe’

- Si **tienes** <gripe > lo mejor es descansar.
- además **tengo** un < gripazo > de cuidado
- Seguramente **tenga** <gripe >
- Odio **tener** <gripe > en temporada de calor
- **tengo** un < gripazo > flipante
- que tengo el < gripazo > padre
- Creo que **tengo** <gripe > post-estrés

Now, the Small Data

- Analysing ‘gripe’ in a medical corpus (MultiMedica) with The Sketch Engine:

  just **341** occurrences against **2759**

‘gripe’ collocates with:

‘virus’, ‘brote’, ‘caso’ ‘estación’, ‘azote’, ‘epidemia’, ‘vacuna’ ‘temporada’

‘padecer’ ‘sufrir’ ‘tratar’ ‘cambiar’ ‘superar’

‘gripe’ Word Sketch

gripe *(noun)*
MULTIMEDICA en Español freq = 341 (74.4 per million)

object_of	5	0.2
sospechar+se	1	8.21
padecer	1	4.99
tratar	1	4.92
sufrir	1	4.27
tener	1	1.15

subject_of	50	2.6
ingresar	1	8.85
ee	1	8.13
influir *1*	2	6.77

n_modifier	44	0.5
2007	3	10.58
20082009	2	10.48
1918	2	10.3
porcino	2	9.91
antigénicamentenuevo	1	9.51
1973	1	9.51
20072008	1	9.51
2008	1	9.39
pandémico	1	9.25
verdad	1	9.14

modifies	201	2.2
temporada *16*	17	11.13
estación 1		
virus *94*	101	10.33
antígeno 3 cepa 4		
pandemia	8	9.85
brote *9*	13	8.69
epidemia 4		
epidemiología	3	8.45
vacuna	14	8.26
azote	1	7.33

y_o	11	0.7
gbs	2	10.71
escara	1	9.14
resfriado	1	8.89
16	1	8.33
sífilis	1	6.44
vacuna *1*	2	4.64
virus 1		
fiebre *1*	3	4.64
infección 1		
mortalidad 1		

Conclusions

- Our study supports Lazer et al. 2014:

"instead of focusing on a 'big data revolution,' perhaps it is time we were focused on an **'all data revolution,'** where we recognize that the critical change in the world has been innovative analytics, **using data from all traditional and new sources,** and providing a deeper, clearer understanding of our world."

Conclusions (2)

- In terms of corpora, small but well selected collections of linguistic data should be combined with large repositories from internet and social networks, since sometimes **‘small data’ offer information that is not inferred from ‘big data’.**