«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.”

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
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

![Graph showing incidence of cases over weeks from January 2012 to December 2014. The graph compares Twitter and Sistema Salud data, with peaks indicating increased activity.](image-url)
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
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…)

Bad data
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
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:
  ‘epidemia’, ‘vacuna’ ‘temporada’
  ‘padecer’ ‘sufrir’ ‘tratar’ ‘cambiar’ ‘superar’
### ‘gripe’ Word Sketch

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

<table>
<thead>
<tr>
<th>object of</th>
<th>5</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sospechar+se</td>
<td>1</td>
<td>8.21</td>
</tr>
<tr>
<td>padecer</td>
<td>1</td>
<td>4.99</td>
</tr>
<tr>
<td>tratar</td>
<td>1</td>
<td>4.92</td>
</tr>
<tr>
<td>sufrir</td>
<td>1</td>
<td>4.27</td>
</tr>
<tr>
<td>tener</td>
<td>1</td>
<td>1.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n modifier</th>
<th>44</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>2</td>
<td>10.58</td>
</tr>
<tr>
<td>20082009</td>
<td>2</td>
<td>10.48</td>
</tr>
<tr>
<td>1918</td>
<td>2</td>
<td>10.3</td>
</tr>
<tr>
<td>porcino</td>
<td>2</td>
<td>9.91</td>
</tr>
<tr>
<td>antigénicamente nuevo</td>
<td>1</td>
<td>9.51</td>
</tr>
<tr>
<td>1973</td>
<td>1</td>
<td>9.51</td>
</tr>
<tr>
<td>20072008</td>
<td>1</td>
<td>9.51</td>
</tr>
<tr>
<td>2008</td>
<td>1</td>
<td>9.39</td>
</tr>
<tr>
<td>pandémico</td>
<td>1</td>
<td>9.25</td>
</tr>
<tr>
<td>verdad</td>
<td>1</td>
<td>9.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>201</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>temporada</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>estación</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>virus</td>
<td>94</td>
<td>101</td>
</tr>
<tr>
<td>antigeno</td>
<td>3</td>
<td>cepa</td>
</tr>
<tr>
<td>pandemia</td>
<td>8</td>
<td>9.85</td>
</tr>
<tr>
<td>brote</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>epidemia</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>epidemiologia</td>
<td>3</td>
<td>8.45</td>
</tr>
<tr>
<td>vacuna</td>
<td>14</td>
<td>8.26</td>
</tr>
<tr>
<td>azote</td>
<td>1</td>
<td>7.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>y o</th>
<th>11</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>gbs</td>
<td>2</td>
<td>10.71</td>
</tr>
<tr>
<td>escara</td>
<td>1</td>
<td>9.14</td>
</tr>
<tr>
<td>resfriado</td>
<td>1</td>
<td>8.89</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>8.33</td>
</tr>
<tr>
<td>sifilis</td>
<td>1</td>
<td>6.44</td>
</tr>
<tr>
<td>vacuna</td>
<td>2</td>
<td>4.64</td>
</tr>
<tr>
<td>fiebre</td>
<td>1</td>
<td>4.64</td>
</tr>
<tr>
<td>infección</td>
<td>1</td>
<td>mortandad</td>
</tr>
</tbody>
</table>
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.”
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’.