Social media *fingerprints* of unemployment

Alejandro Llorente (IIC, Spain)
Manuel García Herranz (UAM and United Nations, NY)
Manuel Cebrián (NICTA, Australia)
Esteban Moro (UC3M, Spain)

PLoS ONE 10(5): e0128692
We are what we repeatedly do.

-Aristotle
Computational Social Science

You are what you repeatedly do [Aristóteles]

Using BigData to infer behavior or society situation

**Situation**
- Demographics
- Health
- Economy
- Unemployment
- Transportation
- Geography
- Politics

**Behavior**
- Social
- Mobility
- Activity
- Content

**Observation**
- Surveys
- Credit card
- Mobile phone
- Social media
- Searches
- ...

Individual - Group - City
Computational Social Science

Health -> Content -> Social Media

Incidences (by 100k users)

- FLU
- Allergy
- Headache
- Fever

Weeks since Jan

Incidences

- alta
- media
- baja

Incidence

@estebanmoro
Computational Social Science

Economy -> Behavior -> Observation

Eagle et al., Science 2010

Development -> Social/Mobility (city) diversity -> Mobile phones

Soto, V. et al., Prediction of Socioeconomic Levels using Cell Phone Records 2011.

Socio economical level -> Social/Mobility/Activity -> Mobile phones


Unemployment -> Content -> Social Media

@estebanmoro
Computational Social Science

Wealth -> Behavior -> Credit card

Krumme, C. et al., 2013. The predictability of consumer visitation patterns. Scientific Reports

Including population patterns we predict (with 30% accuracy) your
Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14 Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827

Economical Impact ~ $1 Billion

@estebanmoro
Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14 Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827

Economical Impact ~ $1 Billion

@estebanmoro
Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14 Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827

Correlation with economic damage

Economical Impact ~ $1 Billion

@estebanmoro
Our objective

Unemployment -> Behavior -> Social media (Twitter)

- It's not another economical problem in Spain. It is the BIG problem
Our objective

Unemployment -> Behavior -> Social media (Twitter)

- Twitter allows us to study social/mobility/content behaviors together
  - But, which behavior(s) are the more relevant with respect to unemployment? (360° view)
- Twitter is BigSparseData: some demographic and geographic areas are not well represented, thus
  - How much of the observed unemployment can be described by Twitter-detected behaviors?
SPAIN

Total Population: 47,370,542
Urban: 77%
Rural: 23%

Internet Users: 33,870,948
Internet Penetration: 72%

Active Facebook Users: 19,600,000
Facebook Penetration: 41%

Active Mobile Subscriptions: 55,740,000
Mobile Subscription Penetration: 118%
SPAIN: SOCIAL MEDIA USE

- Any Social Network: 93%
- Facebook: 87%
- Google+: 57%
- Twitter: 54%
- Tuenti: 29%
- LinkedIn: 26%

OWN AN ACCOUNT

USED IN THE PAST MONTH
TIME SPENT ON SOCIAL MEDIA
Average number of hours per day spent by social media users on all social channels

- ITALY: 2.0
- RUSSIA: 1.9
- FRANCE: 1.7
- IRELAND: 1.7
- POLAND: 1.6
- UK: 1.6
- SPAIN: 1.5
- SWEDEN: 1.4
- NETHERLANDS: 1.3
- GERMANY: 1.3
Our database

- Geo-localized tweets in Spain
- From 29th Nov 2012 to 30th June 2013
- 19.6 million tweets
- 0.57 million unique users
Our database

- Geo-localized tweets in Spain
- From 29th Nov 2012 to 30th June 2013
- 19.6 million tweets
- 0.57 million unique users
Geographical areas in Spain

- Municipalities:
  - ~8200 municipalities in Spain
  - Very heterogeneous: population ranging from 7 to 3.2

- We propose a functional approach to the definition of the areas based on daily mobility

- Users’ municipality home is where they tweet most
Geographical areas in Spain

- We take all trips between two municipalities to construct the flow $T_{ij}$ the number of trips between municipalities
- Flow is well described by the gravity model
  \[ T_{ij} \approx T_{ij}^{grav} = \frac{P_i^{\alpha_i} P_j^{\alpha_j}}{d_{ij}^\beta} \]
  \[ \alpha_i \approx \alpha_j = 0.42, \beta = 0.89 \]
  \[ R^2 = 0.69 \]

(Functional) geographical areas in Spain

- We look for communities in the flow graph
- Infomap algorithm
- 340 functional areas were detected:
  - They are cohesive
  - Statistically robust
  - Modularity is high
  - High overlap with “comarcas”
(Functional) geographic areas in Spain

“The piece is absolutely useless, even ridiculous, outside Spain, because the audience cannot hope to understand its significance, nor the performers to play it as it should be played.”
Twitter penetration

- Is Twitter penetration related to economical development of areas?
  - At country scale twitter penetration ~ GDP
    Hawelka, B. et al., 2013. Geo-located Twitter as the proxy for global mobility patterns.
  - At small scale is the opposite! twitter penetration ~ unemployment

\[ \rho = 0.70 \ [0.6, 0.77] \]
Twitter social interactions

- Granovetter: diversity of interactions yields to more opportunities
- Diversity of interactions between cities is correlated with economical development Eagle et al, Science 2010
- We construct the graph of social interactions
  \[ w_{ij} = \text{number of @ between areas } i \text{ and } j \]
  \[ p_{ij} = w_{ij} / \sum_{j=1}^{k_i} w_{ij} \]
- Measure diversity with entropy
  \[ S_i = - \sum_{j=1}^{k_i} p_{ij} \log p_{ij} \]

\[ \rho = -0.21[-0.37, -0.04] \]
Twitter geographical interactions

- Diversity of geographical mobility is correlated with development

- We use the graph of flows
  \[ \tilde{p}_{ij} = \frac{T_{ij}}{\sum_{j=1}^{k_i} T_{ij}} \]

- Measure diversity with entropy
  \[ S_i = -\sum_{j=1}^{k_i} \tilde{p}_{ij} \log \tilde{p}_{ij} \]
Twitter content

- Two different approaches
  - Classical approach: NLP applied to detect mentions to “unemployment”, “job”, “economy”, ...


\[ \rho = -0.33 \ [ -0.17, -0.47 ] \]
Twitter content

- Our approach: NLP applied to detect **lexical complexity**
  (as a proxy for educational level)

  - Readability (Gunning index)
  

  **Serious misspellings**

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Correct spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alguien se viene <strong>con migo aver</strong> la vida de Pl??</td>
<td>Alguien se viene <strong>conmigo a ver</strong> la vida de Pl??</td>
</tr>
<tr>
<td>La quiero mucho y la <strong>hecho de menos</strong></td>
<td>La quiero mucho y la <strong>echo de menos</strong></td>
</tr>
<tr>
<td>Yo <strong>llendo</strong> a trabajar con este tiempo</td>
<td>Yo <strong>yendo</strong> a trabajar con este tiempo</td>
</tr>
</tbody>
</table>

- We construct a list of more than 600 incorrect expressions of this type validated by Spanish language linguistic experts.

- We do not take into account misspellings due to different Spanish accents and IM abbreviations.

- We compute for each area the fraction of users that make a number of serious misspellings

@estebanmoro
Twitter activity

• Is unemployment reflected in twitter daily patterns?

Just arrived to work, mondays are too hard…

\[ \rho = -0.48 \ [\ -0.34, -0.60] \]
Summary of the variables

Social/geo variables have low correlation
Penetration rate/activity and content are highly correlated with unemployment
Explanatory power of Twitter variables

- Simple linear regression

\[ R^2 = 0.64 \]
Explanatory power as a function of time

- At what time of the year the model has more explanatory power?
Are we really wrong?

Model Error = Model[variables] - Official unemployment

% Shadow Economy *

(* GESTHA report 2012)
Summary
Summary

• Economical development -> Behavior -> Social Media
Summary

• **Economical development -> Behavior -> Social Media**

• Can Twitter be used to infer economical development? YES
  
  • Areas with different behavior in Twitter show different levels of unemployment
  
  • Applications to marketing, media, planning
  
  • Twitter is everywhere! and the API is free! Prove us wrong, please!
Summary

• **Economical development -> Behavior -> Social Media**

• Can Twitter be used to infer economical development? YES
  • Areas with different behavior in Twitter show different levels of unemployment
  • Applications to marketing, media, planning
  • Twitter is everywhere! and the API is free! Prove us wrong, please!

• Can Twitter variables explain the unemployment per area? YES with R² = 0.64
  • Activity, penetration and content account for 80% of the variance explained
  • Diversity on geographical and social interaction amount only for the 20%
Summary

- **Economical development -> Behavior -> Social Media**

- Can Twitter be used to infer economical development? YES
  - Areas with different behavior in Twitter show different levels of unemployment
  - Applications to marketing, media, planning
  - Twitter is everywhere! and the API is free! Prove us wrong, please!

- Can Twitter variables explain the unemployment per area? YES with $R^2 = 0.64$
  - Activity, penetration and content account for 80% of the variance explained
  - Diversity on geographical and social interaction amount only for the 20%

- **What we are doing now:**
  - Can we use the model to forecast future unemployment? Or now-casting?
  - What are the behavioral changes behind socio-economical changes?
  - Use the model in under-developed countries.

@estebanmoro
Behavioral changes behind socio-economical changes

(Geolocalized tweets)
Behavioral changes behind socio-economical changes

(Geolocalized tweets)
Behavioral changes behind socio-economical changes

(Geolocalized tweets)

[Graph showing geolocalized tweets for Working and Unemployed]
Behavioral changes behind socio-economical changes

(Geolocalized tweets)

Less geographical mobility, more probability to be unemployed