Use of data from social networks to obtain statistical and geographical information

Global Conference on Big Data for Official Statistics
October 2015, Abu Dhabi, UAE
Purpose of the presentation

Sharing Mexican experiences about the use of tweets to obtain statistical information by the National Institute of Statistics and Geography of Mexico, INEGI.

Pilot Test
CONTEXT
INEGI is a Constitutionally **Autonomous Entity** with:

- **Statistical and geographical** responsibilities
- Around 20,000 employees
- **10 Regional** Offices and **34 State** Offices all around the country
- **5 General Directorates** which generate and integrate statistical data: economic, social, demographic, government, public safety and justice

INEGI is responsible for Mexico’s **National System of Statistical and Geographical Information**
Emergence and Evolution of new information sources
Big Data Sources

- Smart meters and sensors
  - Traffic cameras, GPS devices, price scanners, power monitors, smart watches, smart phones, etc.

- Social Interactions
  - Talks and publications on social networks like Twitter, Facebook, FourSquare, etc.

- Business Transactions
  - Movements of credit cards, electronic cash registers, cell phone records, etc.

- Electronic files
  - Documents which are available in electronic formats such as PDF files, websites, videos, audio, digital media broadcasting

- Broadcast media
  - Digital video and audio produced on real time
Governing Board Commitment

Eduardo Sojo Garza Aldape

Presidente

Enrique de Alba Guerra
Vicepresidente

Félix Vélez Fernández Varela
Vicepresidente

Mario Palma Rojo
Vicepresidente

Rolando Ocampo Alcántar
Vicepresidente
Technological Landscape

**Internet of things, people and ideas**

- Signal processing
- Probability models
- Machine learning
- Statistical learning
- Data mining
- Database
- Data engineering
- Pattern recognition
- Learning patterns
- Predictive analytics

**Advanced statistics, mathematics and data Science**

- Uncertainty modeling
- Data warehousing
- Data compression
- Computer programming
- High-performance computing
- Geolocation
- Geo-referencing
- ...

**Business Knowledge (experts)**

**IT Infrastructure: Robust Computing and Communications, Specialized software tools for processing, analysis, visualization, etc.**
Institutional Cooperation

- National
- International
TWEETS COLLECTION
Stratification: First Efforts

www.inegi.org.mx/est/contenidos/Proyectos/estratificador/
Twitter as a data source

Real time

Twitter

Query: México Georeferenced

INEGI

NoSQL Data Base
Why Twitter?

- Readily available
- Up to 1% of global tweets at no cost
- Around 12 M accounts in Mexico
- Geolocated tweets by 700 thousand accounts
- 150 M plus tweets downloaded since January 2014
- Even though its drawbacks: Not documented, not supported by “traditional” statistics methodologies
Polygon for the collection of tweets
150 Millions of Tweets, August 2015
70+ Millions of Tweets
August 2015
MOBILITY
Tweeter Users Mobility

4,469,550 displacements
347,157 Tweeter users
Tweets behavior on a Long-Weekend

Desplazamientos hacia Guanajuato o Puebla antes, durante y después del fin de semana largo

- GUANAJUATO
- PUEBLA
Use of Twitter Data for Tourism Studies

GUANAJUATO
Domestic Mobility
Research for development of analytical method to measure trans-border mobility through Geo-referenced tweets.

Mexican (red) and US (blue) tweets

Mexican tweets
Frequency of Tweet Generation in the country

882,007 Twitter Users
43’079,312 Geo-Referenced Tweets
Geo-referenced Tweets in the Municipalities
Use of the Mexican Roads Network

- 70 millions of geo-referenced tweets
  - October 2014

- 121 millions of geo-referenced tweets
  - January 2015
Average Tweets on Banks and Bars in a Week
Next studies on mobility…

- “Feria Nacional de San Marcos” visitants
- Internal tourism in all the country
- Mobility in our Borderlines
- Urban-Rural Systems
SUBJECTIVE WELLNESS
Project Objective

Generation of experimental indicators, new or complementary to traditional methods using data science technologies for the extraction, storage, processing and visualization of big data.
Expected benefits

- New indicators obtained from Big Data Sources
- Correlation of results with traditional methodologies information
- Scientific production
Inception

- **Supervised Learning Method**
  - Humans put qualifications on a training set
  - The system uses similarities to qualify other tweets

Knowledge gathering...

http://cienciadatos.inegi.org.mx/pioanalisis

Desarrolló (5000+ students)
Knowledge

Set of tagged tweets from PioAnálisis (Tweet Analysis)

- 5000+ volunteers
- 374 000 tagged tweets
- 40 000 different tagged tweets (each tweet have been revised 9 times on average)
Automatized Analysis and Qualification

With the manually tagged set of tweets we built a training set to teach the system to recognize and use similarities to qualify other tweets
Tagged Tweets (40,136) → Cleaning → Normalization → Training → Validation
Cleaning process

Cleaning
Contradictions and repetitions

Entropy

Tagged tweets

Cleaning
Contradictions and repetitions
Information Pre-Processing (normalization)

- **Word correction** (dictionary, statistics and heuristics)
  - rojooooo jajajaja
  - rojo

- **Coding** (diacritic symbols)
  - caña
  - perdón
  - amigo

- **Lematizing** (FreeLing)
  - rompí
  - llegó
  - subió

- **Filtering** (Suppressing of stop words)
  - adverbs
  - interjections
  - verbs
  - adjectives
  - nouns
  - hashtags

- **Polarity of Emoticons** (tag of polarity)
  - 😊 😐 😞

- **Q-Grams** (q=4)
  - romper
  - perder
  - amig
Training with algorithms from the state of the art

- Normalized Tweets (19,163)
- Tweets Training (17,163)
- Validation of Tweets (2,000)
- ≈ 50% of accuracy

Algorithms:
- Support Vector Machines
- Bayesian
- Genetic Algorithms
- Decision Trees
- Support Vector Machines + Emotive Charge Words
Development of new Classification Algorithms

Normalized Tweets (19,163)

Tweets Training (17,163)

Validation of Tweets (2,000)

Linear Discrimination Analysis + Elio Weight

Graffeticos

Saddinger

≈ 70% of accuracy
Classification Algorithms Assembly to Improve Accuracy

Accuracy Based Assembly

Saddinger
LDA + Elio Weight
Graffeticos

Positive
Negative

≈ 80% of accuracy
Automatic classification of other tweets

Classification of 60 millions of tweets with the Assembly

Not qualified tweets (60’000’000) → Normalization (Sabinization) → Classification (assembly) → 60 millions of classified tweets
On line demo: www.ingeotec.mx
Estadisticas de ánimo de los tuiteros en México

En colaboración con:

INSECT - CENTROGEO
High Level Meeting
INEGI, INFOTEC, Centro-GEO, CIDE and CIMAT
Results: Collaboration INEGI, INFOTEC, Centro-GEO, CIDE y CIMAT

• Collaboration lines:
  – Common research
  – Seminars
  – Internships
  – academic programs
  – Spaces exchange
  – Micro data access
NEW PROJECTS
Social Networks monitoring for INEGI

**Goal:** Publication, following and monitoring of social networks

- Evaluation of new tools: Semantic Web Builder developed by INFOTEC
- Living Lab workshop: Dissemination staff
- Implementing on internal environment
Mental Health of Teenagers (Data2X)

**Objective:** “Generation of information about mental health on teenager women in Mexico from Tweeter messages”

- INEGI-Data2X agreement (Data2X is a ONG supported by UN)
Mental Health of Teenagers (Data2X)
Statistics on Security and Safety

Research on the possibility to use the tweets database to get information about:

- Collection of dues in urban areas
- Information about natural disasters
Thank you!!