IN5526 Web Intelligence

Chapter 5 Part 5 - Web Opinion Mining - Case Study II

Spring 2024

Diego Cornejo B., Felipe Hernández M. and Juan D. Velásquez

University of Chile
Department of Industrial Engineering
http://www.wic.uchile.cl/

diego.cornejo@wic.uchile.cl, felipehm.eng@gmail.com, jvelasqu@dii.uchile.cl

Outline

Case Study

About

Design and implementation of a system to monitor consumption and opinion about marijuana on Twitter

Case Study

Sources

Based on ”Employing online social networks in precision-medicine approach using information fusion predictive model to improve substance use surveillance” MF Guiñazú, V Cortés, CF Ibáñez, JD Velásquez Information Fusion 55, 150-163, 2020.
`https://www.sciencedirect.com/science/article/pii/S1566253519304038`

Problem

Context

- Since 2010, there has been a systematic increase in the prevalence rates of drug and alcohol consumption in Chile (Thirteenth National Drug Study in the General Population of Chile, 2018)

	Marijuana			**Cocaine**			**Alcohol**		
Year	Low	Medium	High	Low	Medium	High	Low	Medium	High
2010	6,1	3,3	5,1	1	0,3	0,8	35,2	37,3	47,1
2012	7,4	6,9	7,1	1,5	0,6	0,7	39,9	37,7	44
2014	10,6	10,4	12,5	2,1	1,3	1,1	46,5	46,1	52,5
2016	14,4	12,1	16,6	2	0,9	0,7	41	42,5	51,3
2018	12,4	12,3	13,3	1,6	1	0,7	39,7	42,7	46,7

Problem

Context

- The greatest increase in drug and alcohol consumption has been among students aged 13 to 18 years.
- Chilean secondary school students are the ones that consume more cocaine, cocaine paste, marijuana, tobacco, and tranquilizers at the American level (Report on drug consumption in the Americas 2019).

Third higher consumer marijuana on the world (UNODC)

Mayor consumer alcohol on Latin America (World Health Organization)

Problem

Context

- In 2019, a new preventive model is implemented: “Elige Vivir sin Drogas”.
- This new model involves families, schools, private sector and society in general, to prevent drug consumption among children and adolecents.
- This is based on “Planet Youth” prevention model, known as Icelandic model.

Problem

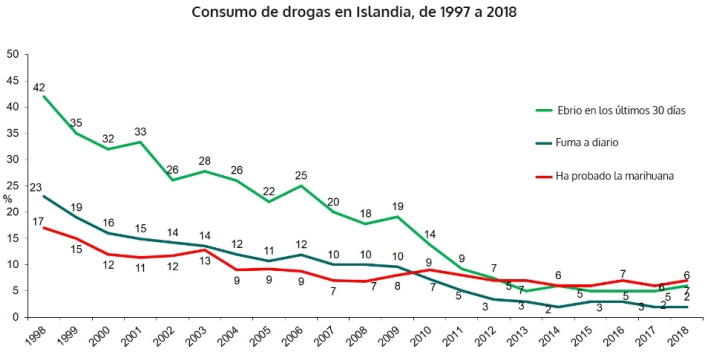
Context

- **Every two years**, a profile of young people is developed through surveys and censuses in educational institutions, in order to create specific action plans by district.
- As well as to identify the main risk factors influencing drug and alcohol consumption and the prevention factors.

Problem

Context

Evolución positiva durante 20 años (estudiantes de 10mo grado)



Problem

Context

- How Web Opinion Mining be useful here?

SONAMA

Objective and Hypothesis

- **Main goal**. Design and implement a system that collects data entered by users chilean Twitter and allow you to track consumption and opinion about marijuana
- **Research Hypothesis**. *"It is possible to extract and process information from Twitter to represent a complex phenomenon such as consumption and opinion about marijuana"*.

SONAMA

Metrics Selection

- Original metrics: Prevalence, risk perception, perception of ease of access, recent drug offering, place of last marijuana offer, drug perception in the home environment, etc.
- Selection criteria: Feasibility, importance, and scope.

SONAMA

Metrics Selection

- Variable to develop:
  - Polarity of tweets: Average of the opinions recorded in the tweets marijuana-related that express or imply positive or negative feelings.
  - Marijuana offer: Percentage of users who issued tweets related to the sale of marijuana.
  - Words used: Distribution of words used to tweets related to marijuana.
  - Prevalence: Percentage of twitter users who used marijuana in the last year.
  - “Friends” consumers: Average percentage of followed users who have published tweets related to consumption
  - Policy tweet polarity: Average of positive or negative feelings in tweets linked to marijuana and politics.

SONAMA

Requirements

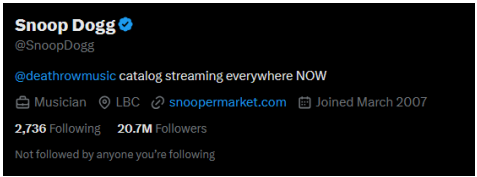
- From tweets:
  - Relationship with marijuana
  - Consumption
  - Policies
  - Sale
- From users:
  - Age
  - Marijuana use

SONAMA

Requirements

- Available:
  - Tweets:

  - User information:



  - Connections between users: Who follows whom?

SONAMA

Requirements

- Necessary:
  - Tweets:

	Variables					Etiquetas		
Tweet	1	...	k	...	m	Consumo	Política	Venta
1	1	...	1	...	0	1	0	0
2	0	...	1	...	0	1	1	0
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮
j	0	...	0	...	1	0	0	1
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮
n	1	...	1	...	0	0	0	1

SONAMA

Requirements

- Necessary:
  - User information:

	Variables					Etiquetas	
Usuario	1	...	k	...	m	Consumo	Edad
1	1.3	...	11	...	0	1	18
2	0.2	...	20	...	0	1	35
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
i	2.2	...	5	...	1	0	22
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
n	-0.1	...	4	...	0	0	15

SONAMA

Requirements

- Necessary:
  - Connections between users:

$$A_{ij} = \begin{cases} 1, \text{if (i,j) in E} \\ 0, \text{otherwise} \end{cases}$$

SONAMA

Data Collection

- Tweet Crawler
  - Tweets were selected if there is a relationship with marijuana.
  - This relationship was based on keywords.
  - Keywords source were experts, bibliography and a poll of related-marijuana words.
  - They were selected according a manual review and topic modeling.

SONAMA

Data Collection

- User Crawler

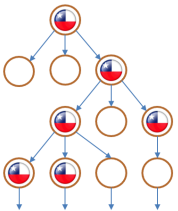
List of followers

```
{
followers: [
12310120310,
9089891,
9120391203903,
61237123,
71237182390,
112831238
]
}
```

Search of keywords on field "location"

```
{
id: 1283713,
screenname: "frame",
description: "Hello there",
locations: "Santiago, Chile"
}
```

Search in levels of followers



SONAMA

Data Labeling

- Data were label to generated a dataset about users location and age.
- Revalidation of the algorithm about marijuana use and age.
- A supervised machine learning approach were used to train a classifier for marijuana use in Twitter account holders.

SONAMA

Data Preprocessing

- Chilean data filtering: by tweet location, last tweet location and/or user location.
- Marijuana use of account holders: if it has marijuana-relationship.
- Age: Under the hypothesis that individuals change the lexicon used throughout their life.
- Localization: By region.
- Gender: Using a pool of names.

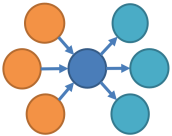
SONAMA

Data Preprocessing

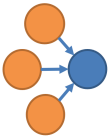
- Tweets to Vector-space representation.
- (1,3) n-grams were used.
- TF-IDF matrix construction.

SONAMA

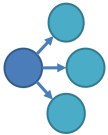
Social Network Analysis



Relations



List of followers



Friends list

$V = \{1, 2, \dots, n\}$ Set of vertices (users)

$F_k = \{1, 2, \dots, n_k\}$ Set of followers of the Username k

$A_k = \{1, 2, \dots, m_k\}$ Set of friends of the Username k

$N_k = F_k \cup \{k\}$ Neighborhood from k

$M = \{1, 2, \dots, w\}$ Set of users that they mention its consumption

Density: $density_k = \dfrac{2 * (|F_k| + \sum_{i \in F_k} \sum_{j \in F_k} b_j^{N_k})}{|N_k| * (|N_k| - 1)}$ *where* $b_j^{N_k} = \begin{cases} 1 \; si \; j \in N_k \\ 0 \; si \; no \end{cases}$

SONAMA

Social Network Analysis

Indegree: $in_k = |F_k|$ Outdegree: $out_k = |A_k|$

Reach Centrality: $reach_k = \frac{\left|\bigcup_{j \in F_k} F_j \cup F_j - \{k\}\right|}{|V|}$

Friends consumers: $m_k = \frac{\sum_{j \in A_k} c_j^M)}{|A_k|}$ *where* $c_j^M = \begin{cases} 1 \; si \; j \in M \\ 0 \; si \; no \end{cases}$

Polarity from neighborhood: $p_k = \frac{\sum_{j \in A_k} d_j)}{|A_k|}$ d_j: *polaridad de usuario* j

Distance: $dist_k = \begin{cases} 1 \; si \; A_k \cap M \neq \emptyset \\ 2 \; si \; A_k \cap M = \emptyset \wedge \bigcup_{j \in A_k} A_j \cap M \neq \emptyset \\ 3 \; si \; no \end{cases}$

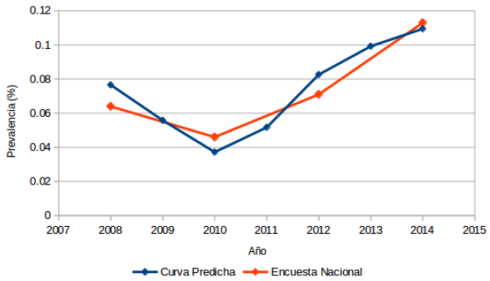
Nominations external: $outnom_k = friends_k - out_k$

SONAMA

Results - Prevalence

Correlationion: 0.933

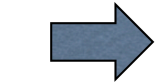
Prevalence by Real



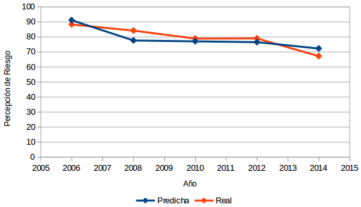
SONAMA

Results - Risk perception and Frequency from consumption

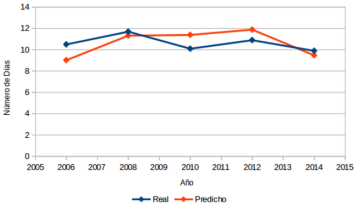
Risk perseption



Correlationion: 0.819



Frequency from consumption



SONAMA

Results - Per region

		PCC	MAE	RMSE
Age Range	[14–18]	0,671	0,0906	0,1085
	[19–25]	0,959	0,0427	0,0508
	[26–34]	0,883	0,0437	0,0587
	[35–44]	0,696	0,0308	0,0382
	[45, 64]	0,676	0,0258	0,0321
Region	1 Tarapacá	0,321	0,0874	0,1124
	2 Antofagasta	0,358	0,1138	0,1325
	3 Atacama	0,602	0,1091	0,1236
	4 Coquimbo	-0,031	0,0988	0,1237
	5 Valparaíso	0,971	0,0575	0,0620
	6 O'Higgins	0,777	0,0946	0,1083
	7 Maule	0,957	0,1224	0,1354
	8 Biobío	0,923	0,0984	0,1116
	9 Araucanía	0,933	0,0952	0,1036
	10 Los Lagos	0,896	0,1029	0,1152
	11 Aysén	-0,017	0,1001	0,1214
	12 Magallanes	0,337	0,0917	0,1179
	13 Metropolitana	0,935	0,0539	0,0638
	14 Los Ríos	0,891	0,0751	0,0848
	15 Arica y Parinacota	0,655	0,1157	0,1392
Gender	Male	0,944	0,0495	0,0612
	Female	0,911	0,1078	0,1199