UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

EXTRACCION DE CONOCIMIENTO NUEVO DESDE LOS RECLAMOS RECIBIDOS EN EL SERVICIO NACIONAL DEL CONSUMIDOR MEDIANTE TECNICAS DE TEXT MINING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

CONSTANZA DANIELA CONTRERAS PIÑA

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
FELIPE AGUILERA VALENZUELA
DENIS SAURÉ VALENZUELA

SANTIAGO DE CHILE
2014

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: CONSTANZA DANIELA CONTRERAS PIÑA
FECHA: 05/11/2014
PROF. GUÍA: SEBASTIÁN RÍOS PÉREZ

EXTRACCION DE CONOCIMIENTO NUEVO DESDE LOS RECLAMOS RECIBIDOS EN EL SERVICIO NACIONAL DEL CONSUMIDOR MEDIANTE TECNICAS DE TEXT MINING

El Servicio Nacional del Consumidor (SERNAC) es el organismo estatal que se encarga de que se cumpla la Ley del Consumidor. Para esto, media los conflictos entre consumidores y proveedores tramitando los reclamos de los primeros. Desde el 2010 a la fecha posee más de 1 millón de reclamos, los cuales son utilizados para realizar estudios que establecen las políticas institucionales de los próximos años.

Se considera que SERNAC tiene valiosa información para analizar profundamente permitiéndole mejorar sus funciones y disminuir los tiempos que tardan los estudios. Dado esto, el objetivo de este trabajo es extraer conocimiento nuevo y específico de los reclamos de SERNAC utilizando técnicas de minería de textos.

En la literatura existen muchas técnicas para realizar minería de textos. En este trabajo se utilizaron modelos de tópicos por su capacidad de encontrar semántica subyacente dentro de una colección de documentos. Dado que no queda claro cuál modelo de tópicos es mejor, se compararon 4 de ellos: Latent Dirichlet Allocation (LDA), Pitman-Yor Topic Model (PYTM), Latent Semantic Analysis (LSA) y Non-Negative Matrix Factorization (NMF).

Primero se validó el uso de modelo de tópicos experimentando con LDA. Se logró extraer problemas comunes entre los consumidores, temas de contingencia nacional, problemas específicos de productos o servicios y caracterizar el comportamiento de empresas y consumidores frente a ciertas problemáticas. Esto fue validado por los miembros de SERNAC, definiendo que un tema agrega valor si entrega información específica o revela información no observada.

Después, se fijó un conjunto de datos para trabajar con los modelos (reclamos de tarjetas de multitiendas). Los temas encontrados por cada modelo fueron evaluados por SERNAC en términos de valor. Además, fueron encuestados a opinión popular para ver si eran fáciles de entender y se les calculó su grado de coherencia con respecto a los otros temas del modelo.

Comparando los resultados por modelo se concluye que tanto LSA como NMF son modelos difíciles de interpretar debido a las palabras que utilizan para caracterizar los temas. Los modelos bayesianos en cambio (LDA y PYTM) no poseen ese inconveniente. En particular PYTM logro extraer temas valiosos más específicos para SERNAC que LDA, por lo cual fue el modelo finalmente escogido. Sin embargo, se considera que las métricas utilizadas en este trabajo no son suficientes para realizar una buena comparación del valor (y calidad) de un modelo. Se propone el continuar la investigación en encontrar métricas que logren este objetivo.

A mi hermosa familia. Los *Contreras Piña*

Agradecimientos

En esta pequeña sección quiero agradecer al CEINE por todo su apoyo en la elaboración de este trabajo. Su apoyo fue fundamental en mi desarrollo como profesional al final de mi carrera.

También agradecer a los empleados de SERNAC. Por facilitarme la información para realizar mi memoria y por su disposición a recibirme en contadas ocasiones para explicarme como funciona la institución. Además agradecer a los ejecutivos que a diario tramitan los reclamos. Por esas externas jornadas clasificando tópicos y contandome su experiencia con los reclamos que leen todos los días.

Por supuesto que estoy muy agradecida de mi familia, amigos y de todas las experiencias vividas en mi paso por la Universidad. Pero eso es personal por lo cual le haré llegar a cada uno de ellos por otro canal mis agradecimientos. Una página no basta para la lista enorme de personas a las cuales les tengo que agradecer.

Contents

List of Tables

List of Figures

Introduction

The Federal Trade Commission in USA[1] defined “to prevent business practices that are anticompetitive or deceptive or unfair to consumers” as its mission. Office of Fair Trading in UK[2] and European Consumer Centres Network in Europe[3] have similar functions. In Chile exists the Servicio Nacional del Consumidor (SERNAC). SERNAC intercedes or mediates consumer disputes between businesses and consumers.

Consumers can establish a complaint in SERNAC if they have a problem with a company (eg. wrong charges or have bought defective products). SERNAC contacts the company and negotiates the demands put forward by the consumers according consumers law. It is not authorized to penalize companies with fines or in other legal forms.

SERNAC performed more than 360 collective mediations[4] and 56 demands until 2013, positioning it as an institution with high impact in Chile. In 2010, 2011, 2012 and 2013 it received 207,000, 300,000, 324,000 and 321,000 complaints which are statistically analyzed generating feedback to SERNAC functions[5]. For example, the best known report is the study “Ranking of companies”[6] (e.g. Most claimed banks). All of them are done using aggregate data.

For the increasing amount of complaints (data accumulated until 2013: 1,143,000 claims) and the type of analysis (aggregate data analysis), It considered that SERNAC has valuable information that has to be analyzed further. Indeed, current studies are the support for establishing the next institutional policies for the short, medium and long term. This is the reason why these studies should be increasingly deeper, looking to maximize the information held. They may prove to be a major support in decision-making.

Furthermore, SERNAC wants to obtain more information -so consumers can take better decisions and to improve judicial proceedings increasing fines. Nowadays, it’s difficult to SERNAC to observe if companies change their behavior for good or just in short term (in order to create better regulations) and actions taken are usually one year (at least) after several hundred of people suffer an abuse.

Our work is to extract valuable knowledge from complaints to help SERNAC achieve its goals.

[1] http://www.ftc.gov

[2] http://www.oft.gov.uk

[3] http://ec.europa.eu

[4] a group of consumers supported by SERNAC that present a collective complaint against a company

[5] http://www.sernac.cl/acerca/cuentas-publicas

[6] http://www.sernac.cl/category/rankings-de-empresas-y-servicios/

Afterwards, SERNAC can manage citizen’s complaints farther and take opportune action to defend consumer’s rights.

Chapter 1

Thesis Presentation

In this chapter we present the overview about our work: objectives, expected results and methodology.

1.1 Objectives

1.1.1 General Objective

The main objective of this work to achieve SERNAC interest is “ to extract new and specific knowledge from SERNAC complaints to detect frequent and important problems between consumers and companies”. The underlying idea is extract latent variables which are not directly observed on data.

1.1.2 Specific Objectives

To reach the final goal of this work, 5 specific objectives were established:

1. To identify information needs about complaints in SERNAC (in which aspects SERNAC needs more information to deal with current issues)
2. To define information that will be obtained (what SERNAC needs or expects to find in its data)
3. To extract valuable information using text mining techniques (add major value to SERNAC using a model that fit better its data)
4. To analyze results and evaluates the quality of them (which type of model is best for SERNAC interests)
5. To create a data visualization for an easy interpretation of results

1.2 Expected Results

At the final of this work we expect:

- To obtain insights of most common consumers problems and its trend during the year
- To discover hidden problems between companies and consumers
- To identify best model for SERNAC according SERNAC interests
- To characterize each model used in terms of coherence, human interpretability and expert judgment

1.3 Methodology

The process of transform data in new knowledge its used to be done by experts. Nowadays the vast amount of data available turns difficult that manual analysis. Thus, process as Knowledge Discovery in Database (KDD)[11] and CRISP-DM[8] were created.

1.3.1 Knowledge Discovery in Databases (KDD)

KDD refers to the overall process of discovering useful knowledge from data. KDD is seen as a way to model broader statistics. It aims to provide tools to automate the process of data analysis and hypothesis selection. With this process, the data is a set of cases and the pattern generated with the model, is an expression that describes a subset of them [11]. It consists of the following steps:

1. Selection: Defining the information to be use in the analysis (selection of variables and / or sample data).
2. Pre-Processing: Preparing and cleaning up selected data to use it in next steps.
3. Transformation: Transforming data to a manageable format that enables proper handling with data mining techniques.
4. Data Mining: To extract unknown patterns of interest by modeling the data.
5. Interpretation/Evaluation: To evaluate the results and to interpret them for their use in business.

The complete process and its steps are in Figure 1.1[1]

1.3.2 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology that uses KDD process. The difference is that CRISP-DM proposes a previous step: Business Understanding.

[1] Source: http://bit.uchile.com

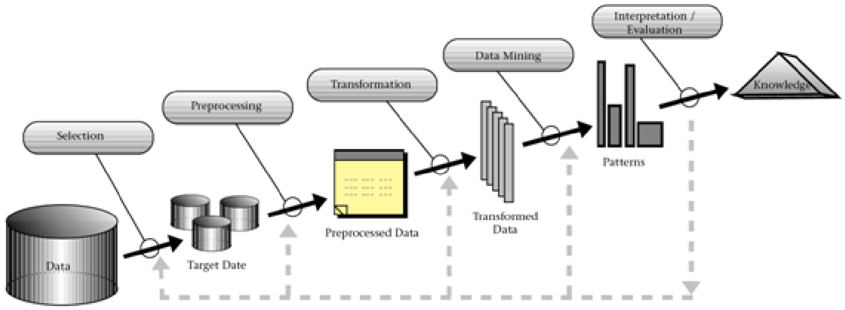


Figure 1.1: KDD process

These step extend to analysts a comprehension of whole the business. Thus, it is for define analysis objectives considering the business context. Additionally, CRISP-DM is presented as a cyclic process (not linear as KDD). Therefore, going back to other phases is completely reasonable if it's needed. Thus, the 6 steps of CRISP-DM methodology are [8]:

1. Business Understanding: Understanding non-technical objectives and business requirements.
2. Data Understanding: Exploring data and understanding it considering business objectives.
3. Data Preparation: Cleaning, sorting and formatting data.
4. Modeling: To select and apply modeling techniques.
5. Evaluation: To evaluate the results and to review the previous steps that created it.
6. Deployment: To present and organize the knowledge gained in a way that the customer can use it. This phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

This cyclic methodology is presented in Figure 1.2 [8].

For the reasons described above we used CRISP-DM Methodology.

1.4 Scope

The work involves the comprehension of the business, the information analysis, the experiments with models and a final report (data visualization). It is not in our working plan to develop software or any kind of platform for SERNAC to use the final model. We just want to prove that our techniques are suitable to make complaints analysis and obtain valuable information to improve SERNAC functions.

Also, we limit to deliver the information found. But we don't want to go farther analyzing laws and proposing changes in them. Our research field it is not public policies. So, we provide the information and it is SERNAC option to take actions or take decisions with it.

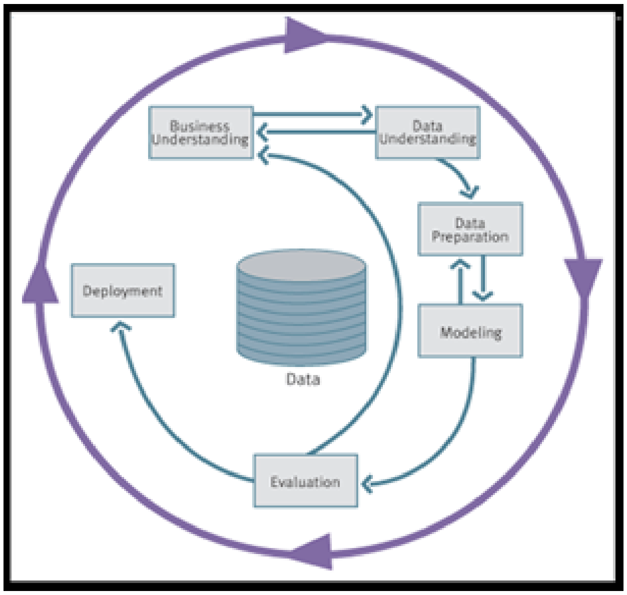


Figure 1.2: CRISP-DM Methodology

1.5 Thesis Structure

On Chapter 2 we present the state of art. Text Mining techniques and previous work are described. Also, it explains the models applied in this work and the metrics used to characterize them.

The experiments and discussion done in each phase of our methodology is presented in Chapter 3. It shows main difficulties on each part of the process and how we faced it. Chapter 4 presents results and their analysis. We explain what we found and the impact of it. We summarize our benchmark and characterize each model according to it. Also, we chose the best suitable model for SERNAC.

Final Chapter concludes our research and discusses future work for us.

Chapter 2

State of art

In literature, extracting knowledge from text is not an uncommon task. It have been developed many models as text mining techniques.

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help to develop new ways to search, browse and summarize large archives of texts[1]. The intuition behind these models is that documents have multiples topics in different proportions.

In practice, researchers suppose that the data was actually generated by a model in question and try to design algorithms that probably find the model that was used to create the data.

We choose apply topic models because our main goal is not to classify complaints (like traditional clustering methods), so this approach allows us to reach SERNAC main goal. But which topic model is the best model to obtain valuable information for SERNAC? Even more what we understand for “valuable information”?

2.1 Previous work

In 1988, Deerwester et al. created the Latent Semantic Analysis (LSA)[9]. Also it is known as Latent Semantic Indexing (LSI) when it’s used in Information Retrieval. Those authors find and fit a useful model of the relationship between terms and documents in a document collection. Transforming text in a term-document matrix, they extract a latent semantic space using Singular Value Decomposition (SVD) and reducing the matrix dimensionality. A decade later, Hofmann (1999) presented an improved model called Probabilistic Latent Semantic Indexing (pLSI)[12] that models each word in a document as a sample from a mixture model.

Blei et al. proposed a new model called Latent Dirichlet Allocation (LDA)[5] arguing it is not clear why one should adopt LSI methodology instead of Bayesian methods. LDA is a three level hierarchical Bayesian model, in which each document of the collection is modeled as a finite mixture over

[1] http://www.cs.princeton.edu/ blei/topicmodeling.html

an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities[3]. Finally, the topic probabilities provide an explicit representation of a document.

LDA go beyond of some LSI and pLSI limitations. For example, it can capture significant intra-document statistical structure via the mixing distribution, considering the exchangeability assumption.

Based on those 2 models, many researchers have been proposing new models such as Author-Topic Model[21], Hierarchical Dirichlet processes[23], Nonparametric Sparse Topic Model[10], Correlated Topic Model[4], Pitman-Yor Topic Model[22] and Non-Negative Matrix Factorization[19].

We focused on 2 of these models. Non-negative Matrix Factorization[19], is an algorithm that propose an optimization problem over LSA. All the elements in the matrixes are required to be non negative. If LSA using SVD extracts 3 matrixes, NMF proposes extract 2 matrixes but only with positive elements. This approach is interesting because when we want to model documents based on words, we only have 2 options: the word is related or not to the document.

Sato presents a change in the prior used in LDA model creating the Pytman-Yor Topic Model (PYTM)[22]. The PYTM captures the power-law phenomenon of a word distribution and the presence of multiples topics.

It is usual to apply a metric to evaluate the performance of these models. Sato shows that PYTM outperform LDA in terms of perplexity. Authors in [2] encourage the use of NMF over other topics based on interchangeability property. But Akira Utsumi in [24] demonstrated that NMF is less effective in constructing semantic spaces than LSA, using 2 test and cosine similarity.

It seems that each model outperformed the other one depending on the dimension analyzed by the researcher (coherence, similarity, accuracy, among others). Even some authors created their own metrics to measure topics quality (as Topic Coherence metric[18]).

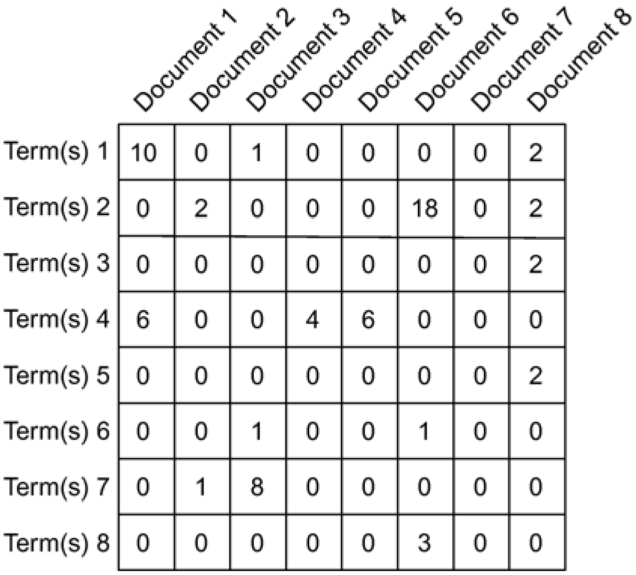
Chang et al.[7] propose a measure of model quality that tries to evaluate if a topic has human-identifiable semantic coherence. Their goal was to measure the success of interpreting topic models across number of topics and modeling assumptions. This work is more related to our desire.

2.2 Handling texts

We introduce traditional representation of a collection of documents or texts for apply text mining techniques. They are used in 2 models which are explained in section 2.3.1 and section 2.3.4.

2.2.1 Term-Document Matrix

First, each document is represented as a set of significant words from it (previously selected to describe it), regardless the order of them. Each row of the term-document matrix corresponds



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Figure 2.1: Representation of Term-Document Matrix

to one of these words appearing in any document (text) of the collection. In turn, each column identifies each document that is part of the collection. Finally, the value of each cell in the matrix refers to the relevance of the word in that row in the document mentioned in that column, being 0 if not listed or is not important.

Thus, given a collection of documents and a vocabulary of words that generated them, the term-document matrix is created as its observed in Figure2.1[2]

Counting the number of appearances (term frequency) is the easiest scheme to define the relevance of each word in each document. However, if the word is too used or too common in the collection, it will emphasize and retrieve documents with high frequency of this word. Meaningful words would not have enough weight in documents representation. It is needed to balance the weight of words to solve this problem. Authors in [25] applied three different term weighting schemes in large datasets. They proved that Term Frequency - Inverse Document Frequency (TF-IDF) scheme performs better than log-entropy and raw term frequency weighting schemes when the text collection becomes very large. Since we have a very large dataset, we applied TF-IDF scheme. It is described in next section.

2.2.2 Term Frequency -Inverse Document Frequency (TF-IDF)

As we explained above, Term Frequency - Inverse Document Frequency (TF-IDF) [15] is a term weighting scheme. It proposed to weight the simplest term frequency with a “term specificity”. This statistic called Inverse Document Frequency (IDF) provides a measure of how rare or common is a term along documents collection. Thus, Term Frequency is defined as the number of times that a term *t* occurs in a document *d* (Equation 2.1).

[2] Source: http://www.emeraldinsight.com

$$tf(t, \mathrm{d}) = \begin{cases} 1 & \text{if } t \text{ occurs in d} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

In turn, IDF is defined dividing the total number of documents N by the number of documents containing the term, and then taking the logarithm of that quotient (Equation 2.2).

$$\mathrm{id}f(t, D) = \log \frac{N}{|\{\mathrm{d} \in D : t \in \mathrm{d}\}|} \tag{2.2}$$

Then, TF-IDF is calculated as is shown in Equation 2.3

$$tf\mathrm{id}f(t, \mathrm{d}, D) = tf(t, \mathrm{d}) \times \mathrm{id}f(t, D) \tag{2.3}$$

2.3 Models

2.3.1 Latent Semantic Analysis (LSA)

LSA constructed a semantic space were terms and documents related are placed near one another. Transforming text in a term-document matrix, they extracted a latent semantic space using Singular Value Decomposition (SVD) and reducing the matrix dimensionality.

LSA model construction

Given a term-document matrix (explained in Section 2.2.1), *X*, for example a *t*x*d* matrix of *t* terms and *d* documents, it can be decomposed into the product of three other matrices (Equation 2.4).

$$X = TSD \tag{2.4}$$

This is called Singular Value Decomposition (SVD). *T* and *D* are matrices of singular vectors and *S* is the matrix of singular values. If *m* is the rank of *X*, the dimension of *T* and *D* are *t*x*m* and *m*x*d* respectively. SVD can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors.

SVD gets an optimal approximate fit using smaller matrices. Thus, if *S* is ordered by the size of its singular values, the first *k* largest may be kept and the remaining smaller ones set to zero. Then, the product of the three matrices results in an approximate matrix of X with *k* rank (Equation 2.5)

$$X \approx \hat{X} = T'S'D' \tag{2.5}$$

Therefore, *T'* and *D'* (reduced matrices) are dimension *t*x*k* and *k*x*d*. Rows of singular vectors are taken as coordinates of points representing the documents and terms in a *k* dimensional space.

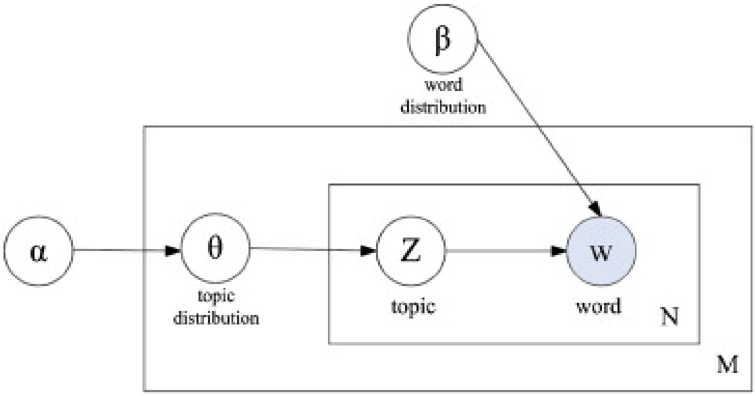


Figure 2.2: Representation of Latent Dirichlet Allocation

2.3.2 Latent Dirichlet Allocation (LDA)

As we introduced before, LDA is a three level hierarchical Bayesian model, in which each document of the collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities[3]. Finally, the topic probabilities provide an explicit representation of a document. The representation of LDA is shown in Figure 2.2[3]

LDA model construction

Given the smoothing parameters β and α and a joint distribution of a topic mixture θ, the idea is to determine the probability distribution to generate - from a set of topics $\mathcal{K}$ - a message composed by a set of N words w ($\mathbf{w} = (w^1, \ldots, w^N)$),

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w^n|z_n, \beta) \tag{2.6}$$

Where $p(z_n|\theta)$ can be represented by the random variable θ_{i}; such topic z_n is presented in document i ($z_n^{\mathrm{i}} = 1$). A final expression can be deduced by integrating Equation **??** over the random variable θ and summing over topics $z \in \mathcal{K}$.

Defining:

- A document as a sequence of N words *w*=(w_1,w_2,...,w_N), where w_n is the nth word in the sequence
- A corpus as a collection of M documents denoted by D={*w*$_1$,*w*$_2$,...,*w*$_M$}

The generative process of LDA to calculate equation 2.6 for each document *w* in a corpus D is:

[3] Source: victorfang.wordpress.com

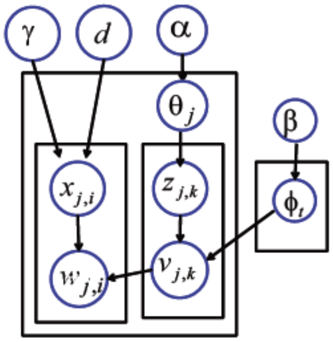


Figure 2.3: Graphical Representation of Pytman-Yor Topic Model

1. Choose $N \sim$ Poisson(ξ)
2. Choose $\theta \sim$ Dir(α)
3. For each of the N words w_n:

    Choose a topic $z_n \sim$ Multinomial(θ)

    Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

2.3.3 Pitman-Yor Topic Model (PYTM)

The Pitman-Yor Topic Model (PYTM) proposed by Sato[22] is based in LDA model and changed the prior used in it. This author uses the Pitman-Yor process to generate the prior.

The called PY process used in this model is a distribution over distributions over a probability space. It use 3 parameters: a concentration parameter γ, a discount parameter d ($0 \leq \mathrm{d} \leq 1$) and a base distribution G_0 that is understood as a mean of draws from PY process. The discount parameter controls the power-law property and generalizes the Dirichlet process when d is 0.

The representation of PYTM is shown in Figure 2.3[4] The PY document model has its perspective based on Chinese Restaurant Process, described below.

Chinese Restaurant Process (CRP)

The Chinese Restaurant Process (CRP) is a process in which n customers sit down in a Chinese Restaurant with an infinite number of tables. Always first customer sits at the first table[5]. The next customers can sit at an occupied table or at a new (unoccupied) table. The probability of sit at any table t is (Equation 2.7)[1]:

[4] Source: Original PYTM publication[22]

[5] videolectures.net/icml05_jordan_dpcrp/

$$\begin{cases} \frac{|b_t|}{n+1} & \text{Probability of sit at occupied table} \\ \frac{1}{n+1} & \text{Probability of sit at new table} \end{cases} \tag{2.7}$$

Where $|b_t|$ is the size of customers in t table. This process defines an exchangeable distribution on partitions of customers. Moreover, it defines a prior on number of tables and on the parameters associated with each tables.

PYTM model construction

For the PY topic model the CRP representation is composed by 4 elements: a customer, a table, a dish and a restaurant. The customer is represented by a word in a document. The table is represented by a latent variable. The dish is represented by a word type. The restaurant is represented by a document. Thus, it is constructed a multiple topics distribution over each document with this representation as a prior.

As it's observed in CRP, if n increases, the number of occupied tables increases, giving a long tale to the distribution. In other words, it is useful when the trend is having many frequency-1 words. This is how CRP in this context captures the power-law distribution over words. Sato adapted this process modifying the 2 probabilities as is shown in equation 2.8:

$$\begin{cases} \text{The k-th occupied table with probability} & \frac{N^c_{j,k}-\mathrm{d}}{\gamma+N^c_{j,\cdot}} \\ \text{A new unoccupied table with probability} & \frac{\gamma+\mathrm{d}K_j}{\gamma+N^c_{j,\cdot}} \end{cases} \tag{2.8}$$

$N^c_{j,k}$ is the number of customers sitting at k-th table, $N^c_{j,\cdot} = \sum_t N^c_{j,k}$ indicates the document length N_j and K_j denotes the total number of tables in restaurant j.

With this prior, Sato modified the generative process of LDA, creating the PYTM. Hence, the generative process for PYTM is:

```
1. Draw $\phi_t \sim$ Dir$(\phi|\beta)(t = 1, ..., T)$
2. for all document $j (= 1, ..., M)$ do:
3.     Draw $\theta_j \sim$ Dir$(\theta|\alpha)$
4.     for all word i$(= 1, ..., N_j)$ do:
5.         Sit at the $k$-th occupied table in proportion to $N^c_{j,k} - \mathrm{d}$
6.         Sit at a new unoccupied table in proportion to $\gamma + \mathrm{d}K_j$, draw a topic $z_{j,k^{new}} \sim Multi(z|\theta_j)$ and draw a word type $\nu^{new} \sim p(w|z_{j,k^{new}}, \phi)$ at the new table
7.     end for
8. end for
```

Where $\phi_t (t = 0, 1, ..., T)$ is the word distribution for each document and θ_j is the topic distribution for each document. PYTM generates topics as much as tables are created, while the number of topics in LDA is equal to number of words.

2.3.4 Non-Negative Matrix Factorization (NMF)

As we explained before, Non-negative Matrix Factorization[19] proposes an optimization problem over SVD extracting 2 matrices instead of 3.

Given a term-document matrix X (refer to section 2.2.1 for term-document matrix) of dimensions nxm it can be decomposed in (equation 2.9):

$$X = GF + E \tag{2.9}$$

Where G is an unknown factor matrix (scores) of dimensions nxp and F is an unknown factor matrix (loadings) of dimensions pxm. E is the matrix of residuals.

NMF model construction

To approximate X matrix, we want to minimizes E so it is necessary estimate residuals, i.e. estimating standard deviation of each element in E. This problem has a weighted least squares sense: G and F are determinate so that the Frobenius norm of E divided (element by element) by standard deviation is minimized.

Defining σ as standard deviation, p as the selected rank and $\|B\|_F$ as Frobenius norm of any matrix B, NMF model is described as:

$$\{G, F\} = \arg\min \|X - GF\|_F \tag{2.10}$$

$$\{G, F\} = \arg\min \|E\|_F \tag{2.11}$$

$$\{G, F\} = \arg\min \sum_{\mathrm{i}=1}^{n} \sum_{j=1}^{m} \frac{E_{\mathrm{i}j}^2}{\sigma_{\mathrm{i}j}^2} \tag{2.12}$$

2.4 Metrics

In this section we described selected metrics to characterize each model in order to choose the model that fits better SERNAC data.

2.4.1 Perplexity

Perplexity[14] is a measurement used for compare probability models. It captures how well a sample is predicted by the model. Based on the entropy of the distribution underlying, perplexity identifies how difficult is for the model to choose a word in the distribution. Using D as the collection of documents, M as the number of documents and N as the number of words (w), perplexity is calculated as is shown in Equation 2.13:

$$perplexity(D) = \exp\{-\frac{\sum_{\mathrm{d}=1}^{M} \log p(w_{\mathrm{d}})}{\sum_{\mathrm{d}=1}^{M} N_{\mathrm{d}}}\} \tag{2.13}$$

Better general performance is indicated by lower perplexity. For us is difficult to find an intuition behind this metric. There are discussions about this on-line[6]. We want to go beyond to theoretical metrics and definitions. Additionally, this metric was applied to compare LDA with pLSA [5] and LDA with PYTM [22]. Thus, perplexity will not be calculated and use it to compare our 4 models.

2.4.2 Topic Coherence

David Mimno et al. created an automatic evaluation metric for identifying semantic coherence in topic models[18]. They argued that the presence of poor quality topics reduce users confidence in the utility of statistical topic models. Thus, authors proposed a new topic coherence score that corresponds well with human coherence judgments. This makes possible to identify specific semantic problems such semantic coherence without human evaluations or external references.

Using human experts, authors classified topics as “good”, “intermediate” or “bad”. Based on the insight that “words belonging to a single concept will co-occur”, they defined *topic coherence* as:

$$C(t, V^t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^t, v_l^t) + 1}{D(v_l^t)} \tag{2.14}$$

where $D(v)$ is the document frequency of word type v (the number of documents with least one token of type v), $D(v, v')$ is the co-document frequency of word types v and v' (the number of documents containing one or more tokens of type v and at least one token of type v') and $V^t = (v_1^t, v_2^t, ..., v_M^t)$ is a list of M most probable words in topic t.

Authors demonstrated that “standard topic models do not fully utilize available co-occurrence information, and that held-out reference corpus is therefore not required for purposes of topic evaluation”. Additionally, they remark that topic size is a good indicator of topic quality; although there is some overlap between bad and good topics (bad topics are smaller than good topics in general).

The result of this metric is a vector with negative numbers, in which topics with numbers closer to zero indicate higher coherence.

To validate the results of this metric, authors also compare it with the work of Chang et al.[7] (*Word Intrusion*) presented in next section.

2.4.3 Word Intrusion

Chang et al.[7] propose a measure of model quality that tries to evaluate if a topic has human-identifiable semantic coherence. Their goal was to measure the success of interpreting topic models across number of topics and modeling assumptions. This work is more related to our desire.

They created the *word intrusion* task which involves to find an “intruder” between a bag of words. The task is constructed as follows:

[6] http://metaoptimize.com/qa/questions/5715/what-is-perplexity

1. Select random topic from the model
2. Select 5 most probable words from that topic
3. Select randomly the *intruder*: a word with low probability in the current topic but high probability in any other topic
4. Shuffle the 6 words selected before. Present them to subject
5. The subject have to choose the word which is out of place or does not belong with the others

Being S the number of subjects, $\mathrm{i}_{k,s}^{m}$ is the intruder selected by the subject s on the bag of words related to topic k from model m and w_k^m is the intruding word among the other words, *word intrusion* is calculated as the number of subjects agreeing with the model as is shown in Equation 2.15

$$MP_k^m = \sum_s \mathbb{1}(\mathrm{i}_{k,s}^{m} = w_k^m)/S \tag{2.15}$$

Comparing MP_k^m with the model's estimate of the likelihood of the intruding word, authors found that higher probabilities and higher predictive likelihood did not have higher interpretability.

2.4.4 Expert Judgment and Value

For reach our main goal, we need to define a measure of value in order to compare it with word intrusion (not expert opinion) and topic coherence (what data say). We consider as "Expert Judgment" the interpretation of executives of SERNAC, what they understand about the topic. This information helps us to label each topic.

Value, instead, is defined as *if the topic is useful to SERNAC*. With "Expert Judgment" topics are valuable if:

- The topic brings an insight that have to be analyzed further in order to take actions (new knowledge)
- The topic brings more information about actual/common problems (specification)

This categorization is given by the statistical Chief and Head of Customer Service in SERNAC. It allowed us to define how useful is each topic to SERNAC functions. Therefore, we can relate usefulness with semantic coherence and people's understanding.

2.5 Measures

2.5.1 Cosine Similarity

In Text Mining, the Cosine Similarity between 2 vectors (usually documents represented as vectors) is a measure of how similar they are in terms of their underlying matter. It calculates the cosine of the angle between them.

Cosine similarity is a measurement of orientation and not magnitude[7]. If they are pointing to the same direction they are similar. Thus, given vectors a and b, the cosine similarity $\cos\theta$ is calculated as is shown in Equation 2.16:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}||||\vec{b}||} \tag{2.16}$$

It's considered that 2 vectors are similar if cosine is near to 1. They are independent if cosine is near to 0. If cosine is near to -1 they are dissimilar.

We want to use this measure to measure how close topics from one model to another are. Instead of documents, we want to compare topics. This can give us an insight of how many different topics from other model are extracted in each model. In other words, for each topic of model A, we want to know if a similar topic was extracted in model B, C and D.

[7] Source:http://pyevolve.sourceforge.net/wordpress/?p=2497

Chapter 3

Discussion and Experiments

In this section we describe all the discussion and the experiments done in each phase of CRISP-DM methodology. Results are presented in next chapter.

3.1 Business Understanding

In the introduction (see chapter) we presented a brief about SERNAC and its context. In this section we want to remark some aspect about SERNAC in order to understand better the institution, its needs and the tasks developed.

As we mentioned before, SERNAC cannot penalize companies if they are breaking the consumer law. Instead of, SERNAC can establish a lawsuit against companies. It is also able to contact consumers with the same problem and support them in a collective negotiation or in a collective lawsuit. Despite that, this legal situation turns difficult to SERNAC to take actions. The institution needs to get information from consumers or others sources because it cannot control directly companies and markets. Additionally, the lack of regulation in some markets (as retail market) has caused that consumers are use to many abusive practices from companies.

SERNAC has 3 functions to help citizens: Education, Citizen Participation and Consumer Protection. Education area is in charge of educating citizens spreading consumer law (by social media, in schools, on TV, among others). So, citizens can notice when a company is breaking the law and alert to SERNAC. The Citizen Participation area support collective lawsuit and negotiations. It encouraging consumers to create associations and look for compensations and finish with abusive practices. Finally, Consumer Protection area it is probably the best known function. Also it is a support of the other 2 areas. It is responsible of receive complaints from consumers and help them contacting companies.

Besides the law facilitates SERNAC to ask for some information to companies, complaints are the principal way in which SERNAC can obtain information about companies and consumers behavior. Since SERNAC cannot be day by day controlling companies, the actual process in which SERNAC can help consumers is educating them about the law, encouraging them to defend their rights and

Field	Description
Fecha_Hecho	Date in which the situation happened
Region_Consumidor	Number of the district in which the consumer lives
Nombre_Mercado	Name of the Market related to the complaint
Categoria_Mercado	Name of the sub-category in the market
Tipo_Prod	Type of product or service in the complaint
Motivo_Legal	Legal motive
Categoria_ml	Specification (sub-category) of the legal motive
Descripcion_inconformidad	Description of the complaint
Solicitud	Consumer's request
Nombre_canal	Channel that received the complaint (on-line or presential)

Table 3.1: Important fields in SERNAC database

warning companies about it. The product of this process is the increasing amount of complaints.

Therefore, complaints have to be analyzed farther. Besides the information extracted of them, it can be a precedent for change the law and ask for more faculties to the institution. We based our work in this analysis about SERNAC and the importance for the institution and the country of the complaints.

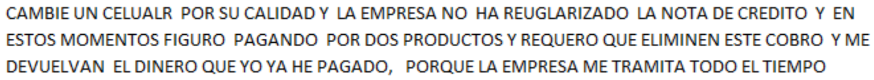
3.2 Data Understanding

Complaints are receiving by 2 channels: in person and on-line. Normally, in the in person channel, an executive of SERNAC write the situation told by the consumer. So, it is written in third person. In the case of online complaints, consumers fill an on-line questionnaire in which describe in their own words their claims.

Normally, people that write complaints do not explain a cohesive story. They abuse of technical vocabulary i.e. use legal words in wrong contexts. Also they have many misspellings and bad typing.

We have a database of 217,300 complaints of 2012. They belong to 3 markets: Financial, Retail and Telecoms. The database has 47 fields with information of each complaint and was given in a excel sheet. The first column is the number of case. In Table 3.1 we describe important fields for the data preparation and the modeling. The complaint itself has 2 parts. The first part is the description of the situation ("Descripción de inconformidad" in Spanish). The consumer explains about what happened and why he/she is complaining about it. The second part is the request ("Solicitud" in Spanish). In this part the consumer request compensation or detail what wants he/she.

The database has also a sub-classification of markets. For example in financial market there are banks, collections agencies and compensation fund, among others. Over each of those categories, a new classification is made by "Tipo_Prod" (type of product): credit cards, ATM, checking accounts, etc. Additionally, every complaint has a legal motive description and an specification of it. In Figure 3.1 we show an example of a complaint in the excel sheet.



CAMBIE UN CELUALR POR SU CALIDAD Y LA EMPRESA NO HA REUGLARIZADO LA NOTA DE CREDITO Y EN ESTOS MOMENTOS FIGURO PAGANDO POR DOS PRODUCTOS Y REQUERO QUE ELIMINEN ESTE COBRO Y ME DEVUELVAN EL DINERO QUE YO YA HE PAGADO, PORQUE LA EMPRESA ME TRAMITA TODO EL TIEMPO

Figure 3.1: Example of a complaint in SERNAC database

3.3 Data Preparation

First we selected which complaints are useful. We found some complaints with no explanation because the explanation is in an “attached letter”, which is not in our database. Thus, we filtered by in person channel (using “nombre_canal” field) and we removed all complaints with texts such as “The situation is described in the attached letter”.

The preliminary preparation of data was to create an SQL database and upload it all complaints with their fields. This facilitates filtering, information crossing and other functions that are slow in excel because the amount of data.

For modeling phase was necessary to adapt the data. We worked with the first part of the complaint: “Descripcion_inconformidad”. We didn’t select the other part because is shorter (less words to explain) and consumers explain the entire situation in the first part. So, consumer’s insights should be hidden in this variable.

The preparation was done in 2 steps: cleaning and transformation.

3.3.1 Cleaning

Cleaning the complaints was done with many functions programmed in python. This functions were added to a framework developed by Camilo Lopez: T-Analyzer [17]. T-Analyzer was created to classify posts of many blogs and predict in which blog should be posted a new entry. It was developed using Django framework[1] and programmed in python. Using this framework for cleaning made the process easier and faster.

We upload our complaints in T-analyzer creating a new database. We just uploaded the field “descripcion_inconformidad”.

The functions of the cleaning process are described below:

- Lower Case: Data was in upper case. So we pass all texts to lower case.
- Remove non-alpha/numeric characters: We removed all characters that are not a-z/0-9. For example, we removed punctuation signs and non-ascii characters.
- Misspelled words: We created a dictionary of common misspelled words. In Figure 3.1 the word “celular” (cell phone in English) is misspelled (“CELUALR”). We identified most common cases and corrected them. Most of them are related to companies, i.e. the names of many companies were misspelled.

[1] https://www.djangoproject.com/

- Remove double spaces: Removing characters leave many spaces between words. We removed them. So, between every word there is just one space.

3.3.2 Transformation

In order to get better results and apply the implementation of models described in Section 3.4, we transform data using the following functions:

Key Entity

There are many words that together have a special meaning in consumer context like “Letra Chica” (abusive clause in English) or the name of some companies such as “Banco de Chile” (Bank of Chile). Thus, “bank” and “Chile” by themselves have a meaning but together mean something different (the name of a bank). Therefore, we identified this words calling them *Key Entities*. We use the first letter of the second word in upper case to recognize them. So now “Banco de Chile” is “bancoChile”. For this function we created a list of keywords that have to be replaced.

Remove Numbers

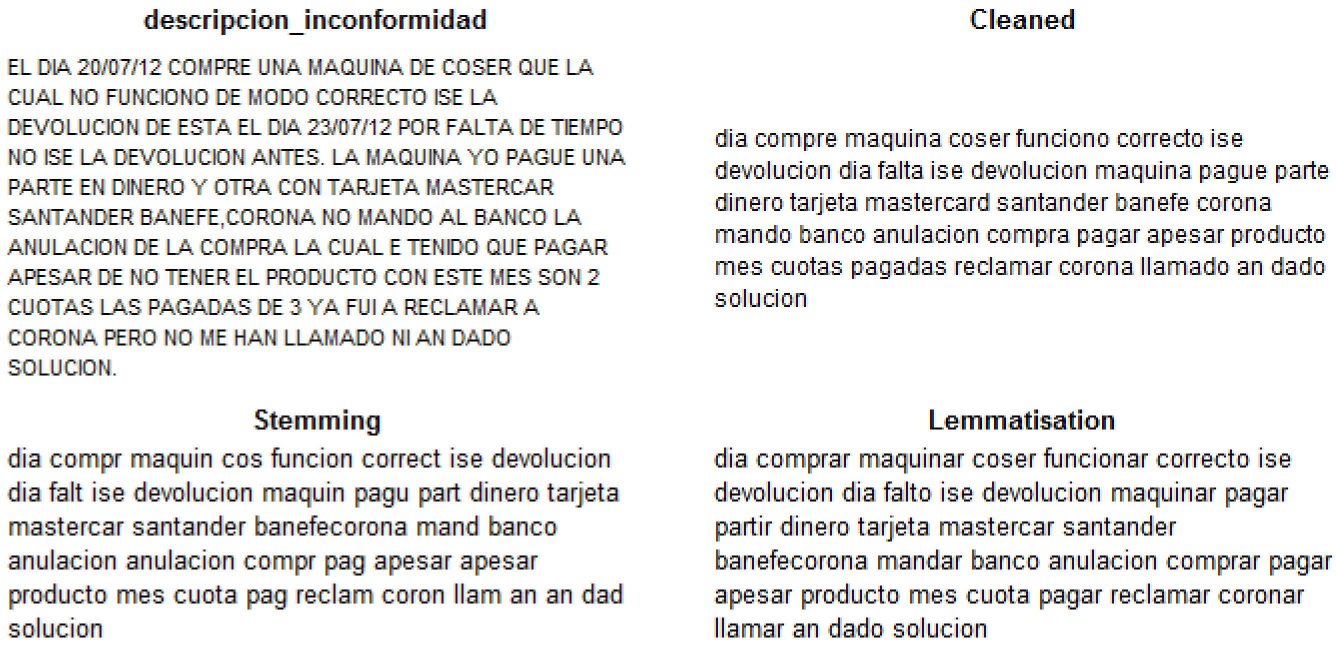
Many consumers disclosure numeric data to support their claims such as dates, number of checking accounts, Chilean id, number of bills, etc. This numbers are not useful for the research and also many of them like dates and ids are presented in other fields of the database. So we remove them to reduce the number of unique words in our complaints collection.

Some numbers that are special were identified as key entities. For example, the name of a financial company in Chile is “Caja 18”.Thus, we replaced for “Caja18” to not remove the number. Therefore, the criterion for remove numbers was that the number cannot be a part of a String. This limited us to not remove all useless numbers in the complaints because some numbers were written as a part of a string like Chilean ids ended with “k”. Also consumers wrote the number of their bills as “bill n0848293942”.

Remove Stop Words

If we consider that each complaint has 100 words (average), 217,300 complaints have 21,730,000 words. However, many of those words are useless for semantic space, such as connectors and short function words. These type of words are called *stop words*. They do not contain important significance to be used in our work.

We use a dictionary with 350 Spanish stop words to remove them from our complaints. We created a new field in our database to store complaints without stop words.



descripcion_inconformidad

EL DIA 20/07/12 COMPRE UNA MAQUINA DE COSER QUE LA CUAL NO FUNCIONO DE MODO CORRECTO ISE LA DEVOLUCION DE ESTA EL DIA 23/07/12 POR FALTA DE TIEMPO NO ISE LA DEVOLUCION ANTES. LA MAQUINA YO PAGUE UNA PARTE EN DINERO Y OTRA CON TARJETA MASTERCAR SANTANDER BANEFE,CORONA NO MANDO AL BANCO LA ANULACION DE LA COMPRA LA CUAL E TENIDO QUE PAGAR APESAR DE NO TENER EL PRODUCTO CON ESTE MES SON 2 CUOTAS LAS PAGADAS DE 3 YA FUI A RECLAMAR A CORONA PERO NO ME HAN LLAMADO NI AN DADO SOLUCION.

Cleaned

dia compre maquina coser funciono correcto ise devolucion dia falta ise devolucion maquina pague parte dinero tarjeta mastercard santander banefe corona mando banco anulacion compra pagar apesar producto mes cuotas pagadas reclamar corona llamado an dado solucion

Stemming

dia compr maquin cos funcion correct ise devolucion dia falt ise devolucion maquin pagu part dinero tarjeta mastercar santander banefecorona mand banco anulacion anulacion compr pag apesar apesar producto mes cuota pag reclam coron llam an an dad solucion

Lemmatisation

dia comprar maquinar coser funcionar correcto ise devolucion dia falto ise devolucion maquinar pagar partir dinero tarjeta mastercar santander banefecorona mandar banco anulacion comprar pagar apesar producto mes cuota pagar reclamar coronar llamar an dado solucion

Figure 3.2: An example of a complaint in its original form (descripcion_inconformidad), cleaned, with stemming and lemmatisation.

Stemming and Lemmatization

Stemming is the process of reduce words to their *stem* (base or root form). In information retrieval is a common technique for web search engines. For example, the words “playing”, “played” and “player” are reduced to the stem “play”. Thus, the 3 words can be treated as same.

Lemmatization is similar to stemming but more complex. This process identifies the primitive form of a word called *lemma*. If we have the words “writing”, “write” and “written”, the stem recognized is “writ”. The lemma instead is “write”.

To apply both process we used 2 dictionaries. One with words and their stem related and another with the same words and their lemma related. Both dictionaries have 1 million Spanish words. Complaints with stemming and lemmatization were stored ir our database in 2 new fields.

In Figure 3.2 it is presented a complaint after cleaning and transformation stages. All of the functions described above were programmed in python. Each of those transformation used more than 24 hours to process all our complaints.

3.3.3 Discussion of Data Preparation phase

We experienced many problems preparing complaints. As we explained before, complaints are far from being a well written text. So we found a large amount of misspelled words that appeared less than 4 times in our collection of complaints (about 20,000 words in a data set of 21,800 complaints).

We identified important words and replaced them for their correct form. However, it was impossible to correct all of them. For this process we used our implementation of LDA model described in section 3.4. This implementation gave us a list with unique words (vocabulary). We reviewed this words, selected important misspelled words, replaced them and then, we run our model again. We observed that some words appear with higher probability in some topics. For example, a word in top 30 of most probable words in a topic was later in top 20. Hence, we repeat this process until most meaningful words were corrected.

Polysemy problem

Spanish has a complex conjugation of verbs. The acute accent is used to distinguish tenses and meaning. For example, the word "río" means river but "rio" means laugh (verb in past tense). Also, many words have their meaning according to context. In our vocabulary the word "cuenta" can means account or count (verb in present tense).

We have each complaint in 3 forms:

1. Cleaned and transformed but without any stemming or lemmatization transformation
2. Cleaned and transformed with stemming
3. Cleaned and transformed with lemmatization

Comparing complaints with stemming and lemmatization we find that lemmatization complaints were easier to interpret and read than stemming complaints. Thus, we dismiss the option of process stemming complaints. Then, we process a data set with LDA model in the other 2 forms (about 38,500 complaints). We presented those results to the executives of SERNAC. Their appreciations were that topics with original words were easier to interpret than words with lemmatization. We agree because lemmatization form increases polysemy problem. As an example, the word "monto" (amount) was transformed in "montar" (ride as a verb) giving the word no meaning in SERNAC context. Also the word "cuenta" from "cuenta corriente" (checking account) was transformed in "contar" which could refer counting (money for example) or telling a story.

To avoid some polysemy problems we used key entities so we could identify some words that have special meaning in SERNAC context. Moreover, we discard the idea of use lemmatization. We found that in some cases the tenses are valuable. As an example, it is different "a consumer paid a bill and then the consumer had a problem" than "when a consumer was paying a debt, the problem happened". In fact, when we were testing, we noticed some topics in which just appear a word in one tense in top 30 most probable words. Other topics have in their top 30 a word in many tenses.

3.4 Modeling

In this section we present experiments done with each model. First of all, as we explain before, we used LDA model in data preparation phase. The reason is that we already have a code in Java for

LDA model[2]. This implementation is easy to run, fast and return many outputs useful for posterior analysis. LDA implementation is presented in section 3.4.1.

Additionally, we use LDA model to fix the number of topics to extract, the number of iterations and the amount of complaints to process. For the number of iterations, we tested from 100 iterations to 10,000 iterations. From 100 to 1,000 we used a step of 100. Then, from 1,000 to 10,000 we used a step of 1,000. Comparing topics we noticed that after 2,000-3,000 iterations topics did not change significantly. Thus, we fixed the number of iterations in 2,000.

To determinate the number of topics we selected a data set of 38,500 complaints and process them extracting from 5 to 100 topics. We used a step of 5 between 5 and 40. Then we used and step of 10. We observed that extracting less than 15 topics generate topics too aggregated. Also, generating more than 40 topics turns tedious the analysis of all of them and many of them were too fine-grained or with no sense. In our appreciation and considering executive's available time, 25 is a good number of topics to extract. Thereby we fixed the number of topics in 25.

For the amount of complaints we proved using SERNAC categories. First, we process the entire data base (217,300 complaints). We found many topics referring to markets: complaints about financial market, complaints about telecom market, etc. As we shown in Section 3.2, markets are a category in sernac database (Nombre_mercado column). So we selected a smaller data set using sub-categories (categoria_mercado column). Then, we worked with complaints classified by products (tipo_prod column).

We observed that modeling complaints classified by products found topics more disaggregate than the other experiments. Moreover, those topics obtained more specific information about products, consumers and markets. This is consistent with SERNAC interest. So we decided to work with data sets splitted by type of product/service.

The data sets used to test models were: 1,000 complaints of collection companies, 21,800 complaints of department store's credit cards and a data set of 38,500 complaints of banks. In some cases we used the data set of supermarkets corresponding to 7,000 complaints.

3.4.1 LDA model implementation

We use JGibbLDA, a Java implementation of Latent Dirichlet Allocation (LDA) using Gibbs sampling for parameter estimation and inference[20][6][3]. The default value of α in this implementation is 50/k, with k is the number of topics. In our case, α is 2. β instead is 0.1 by default.

The input for this implementation is a file in which the first line contains the total number of documents. Then each line is one document represented by its words separated by the blank character. The parameters used for this implementation were:

- $\alpha = 2$
- $\beta = 0.1$

[2] Source:http://jgibblda.sourceforge.net/

[3] Source:http://jgibblda.sourceforge.net/

- number of topics: $k = 25$
- number of iterations = 2,000
- number of most likely words for each topic = 30

Even when time was not a variable to considerate in this research, we used this model for experiment in previous phases because it is fast. It takes less than 5 minutes to process 21,800 complaints in our server.

3.4.2 PYTM model implementation

Victor Chahuneau has repositories in GitHub with his implementations of many models. One of them is "vpyp"[4], an implementation of LDA with a Pitman-Yor prior instead of a Dirichlet prior. The code is in python.

The input is a file in which each line is one document represented by its words separated by the blank character (similar to LDA implementation). The parameters used were the number of topics (25), the number of iterations (2,000) and the number of top words (30).

The main problem with this implementation was that it is too much slow. It takes one week to process 21,800 complaints and three weeks for the entire database. So, it took us a lot of time process different data sets. For our research it was not a big deal but for SERNAC would be useless to wait that long. As a future task we will improve this implementation to reduce processing time. We could not find any other open sources of implementations of this model.

3.4.3 LSA model implementation

We proved with three implementations of LSA model. First implementation was programmed by Joseph Wilk in python[5]. We reproduced his implementation using a data set of 1,000 complaints. It takes about an hour and a half to process all complaints. When we use a data set with 7,000 complaints the program takes so much more hours that we don't let it finish. Despite the time, topics make no sense. It was hard interpret what the topics really mean. So, we dismiss this implementation.

In the blog of Joseph there were many comments about LSA and its implementations. One of them indicates that LSA doesn't scale well to large sizes because the matrix decomposition, even when it's calculated using sparse methods. It's O(docs^2+unique_terms^2). Thus, we decide to experiment with R, a statistical language that it has been highly used by data scientists[6].

From CRAN project[7] we downloaded a deprecated package for LSA. We proved with 1,000 complaints and the output was fast (order of milliseconds) and topics were interpretable. However, for

[4] https://github.com/vchahun/vpyp

[5] blog.josephwilk.net/projects/latent-semantic-analysis-in-python.html

[6] Discussions about it in datatau.com forum

[7] cran.at.r-project.org

data sets with more complaints than 10,000 we have a memory error. We tried to improve our code to avoid this problem obtaining some enhancements but anyway we were limited for the numbers of complaints with this implementation. Since we don't have more hardware, we looked for others implementations.

Scikit-learn is an open source package for machine learning in python. It has simple and efficient tools for data mining and data analysis[8]. It has a function that we used to apply LSA model. We were able to process a data set with 21,800 complaints and other with 38,500 complaints. Topics make sense. However we have to reduce the number of unique words - we don't consider words that appear less than 4 times in the data set - we were satisfied with this implementation.

This implementation receives the path of a folder in which each file contained is a document. Every document is represented by words separated by blank character. The parameters used are the number of topics (25) and the number of top words (30).

3.4.4 NMF model implementation

As we describe in Section 2.3.4, NMF involves an optimization problem. So there are many algorithms to implement NMF model. One of them was the "Projected Gradient Methods for Nonnegative Matrix Factorization" by Chih-Jen Lin[16]. Author demonstrated that his method converges faster than the multiplicative update approach and it is provided a simple MATLAB code for use it.

In author's web site[9] it is available a python code of this algorithm. We used in a dataset of 1,000 complaints. The algorithm converges in about 400 iterations. But topics were not interpretable. So we decide to look for other codes.

As we did with LSA, we looked for implementations in CRAN project. We found a NMF package[10]. Also, it has the same problems of LSA package in R. For an amount of 1,000 complaints it's works well. But processing data sets with more than 10,000 complaints produced a memory error. Therefore, we discard this implementation.

Finally, scikit-learn has a function for NMF model. Based on the tutorial[11] of DARIAH-DE[12] we applied this code in a data set of 38,500 complaints and another one of 21,800 complaints. We also considered just words with more than 3 appearances in the data set. We decided to keep going with this implementation.

As LSA model implementation, the code uses as input the path of a folder with all complaints. Each complaint is in one file in the folder. The parameters fixed are the number of topics - 25 - and the number of top words - 2,000.

[8] Source: scikit-learn.org/stable

[9] csie.ntu.edu.tw/c̃jlin/nmf

[10] http://cran.at.r-project.org/web/packages/NMF/index.html

[11] https://de.dariah.eu/tatom/topic_model_python.html

[12] https://de.dariah.eu/

3.5 Evaluation

To evaluate and create a benchmark of use of models, we calculates each metric described in Section 2.4 as follows.

3.5.1 Topic Coherence implementation

We programmed in python equation described in Equation 2.14. We were not really sure about how to interpret the numbers. Every model gives us a range of negative numbers but in different scales. Thus, we contact David Mimno and asked him if there is a cutoff to determine whether topics are coherent or not. His reply was "The best I can say is that topics with the lowest scores are likely to be obviously incoherent to users. I don't think there's a clear cutoff".

Therefore, this metric is not useful to compare topics among models, i.e. if a topic of LSA model has a topic coherence of -1,000 it doesn't mean that is more coherent than a topic of PYTM model with a topic coherence of -2,500. However, we used this metric for compare the semantic coherence of a topic with the other metrics proposed and then, characterize the model.

3.5.2 Word Intrusion Survey

To calculate this metric we modified the task. We adapted to our context and to avoid some biases.

First we created a pre-test. We constructed a survey consisting in ten topics of LDA model, with its top 5 words and a fixed intruder for each topic. We selected the intruder as in the original task but it will be the same in all the surveys. We polled 10 people. Then we calculated the word intrusion. Our appreciation of this pre-test is that it is necessary to poll a significant amount of people. We decided to poll 100 people. Also, as the original metric state, the intruder has to be random and vary among subjects.

Hence, we created a second pre-test consisting in 4 forms. Each form has 25 questions (one for each topic of LDA experiment). We mixed randomly the top 5 words with an intruder. We selected 4 intruders for each topic. So, the first form has 25 questions using the first intruder. The second form has 25 questions using the second intruder, and so on. These forms were made in *Google docs*[13]. Then, we decided to spread the 4 forms by Facebook and e-mail. The idea is obtain 25 answers from different people of each form. So, we will have 100 answers of each topic with 4 different intruders.

It took us 3 days to complete the 100 answers. The feedback was that the questionnaire was too long and people got confused because the words were repeated in every question. Additionally we observed answers differ in every questionnaire. For some topics every form has similar results but for other topics results highly differ. So, it's useful to use many intruders and the questionnaires have to be shorter. We decided not to survey all topics of every model.

[13]

Considering concerns described above, we designed the final experiment. We selected 10 topics randomly for NMF, LSA and PYTM model. We designed a form with 3 pages (one model in each page). Each page has 5 questions related to 5 topics. Each question has 5 top words of its related topic and one intruder. So, we created 2 phases: phase 1 has first 5 topics of each model and phase 2 the other topics.

Then we selected 4 intruders for each topic as before. Thus, for phase 1 we created 4 forms. Form 1 has one specific intruder. Form 2 has another intruder and so on. For phase 2 we did the same. In total, there were 8 forms. Each form has 15 questions.

The instruction was to choose just one form of each phase. Then, each person is answering 30 questions related to 10 topics per model. We spread the forms by Facebook and e-mail. The reason of using social media to spread the forms was to avoid greedy bias. If we pay someone for respond the questionnaire, the person can be willing to respond randomly just for finish quickly and obtain the payment.

Based on responds we calculate word intrusion.

3.5.3 Expert Judgment Interviews

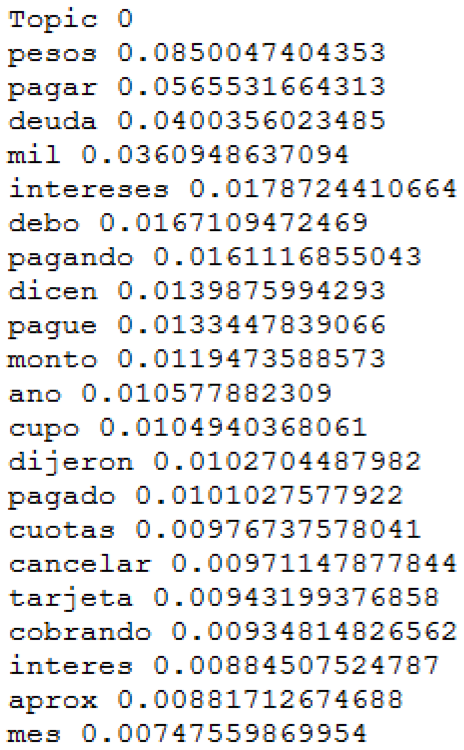
We talked several times with executives of SERNAC. We showed them the output of the model. This output is shown in Figure3.3 Then, they told us if words makes sense or not. Also, they provided us extra information in order to understand better SERNAC context, consumers and legal terms. Later, we compared documents with its related topics by reading them. In general terms, most documents were about what executive thought. Hence, executive's information was vital to complete their vision with our vision.

We talked with 4 people. 2 executives helped us explaining consumer's common complaints. The other 2 executives helped us labeling complaints and telling us about their experience dealing with consumers and companies. Finally, for each topic we obtained a description about it.

3.5.4 Value's Interpretation

We presented to SERNAC Managers (2) the topics with its explanation and number of complaints related (topic's size). We marked which topics belong to which model using a different color for each model. In Table 3.2 it is shown an example of this.

Then, each manager marked the topics that are relevant for him. After that, we identified which model has more valuable topics and compare this information with the metrics.



```
Topic 0
pesos 0.0850047404353
pagar 0.0565531664313
deuda 0.0400356023485
mil 0.0360948637094
intereses 0.0178724410664
debo 0.0167109472469
pagando 0.0161116855043
dicen 0.0139875994293
pague 0.0133447839066
monto 0.0119473588573
ano 0.010577882309
cupo 0.0104940368061
dijeron 0.0102704487982
pagado 0.0101027577922
cuotas 0.00976737578041
cancelar 0.00971147877844
tarjeta 0.00943199376858
cobrando 0.00934814826562
interes 0.00884507524787
aprox 0.00881712674688
mes 0.00747559869954
```

Figure 3.3: Example of an output of a model.

Topic	Size	Description
Topic 16	1,646	Cobros realizados con tarjetas de falabella o a clientes de falabella totalmente desconocidos, de días que no han ido o adicionales que no tienen.
Topic 18	416	Reclamos por cobros de gastos de cobranza, intereses moratorios e intereses usureros en los estados de cuenta

Table 3.2: Presentation of topics to SERNAC Managers

3.5.5 Cosine Similarity comparison

We selected the 30 top words with its probabilities in each topic for every model. First, we had to normalize the values because all the models have their own scale. So, we normalized to 0-1 scale. Then, we compared two topics mixing the words of both vectors (and keeping the unique words) and calculating cosine similarity between them. We implemented a code and we used the *cosine function* in scipy library.

3.6 Deployment

To show our results to SERNAC we created an infographic with the methodology applied over the complaints. It is easy to visualize the impact of our research in the institution with the graphics and descriptions shown in the infographic. This infographic is shown in Apendix A.

Chapter 4

Results

We divided our work in 2 phases. The first one is "Topic Extraction". We used all the experiments done with LDA model to validate our research and to create a methodology for the analysis of complaints. In the second phase, we focus in one case (one data set) applying the 4 models to do the benchmark.

4.1 Topic Extraction

As we explained in section 3.4, we used LDA model for first experiments. We present some results for 2 data sets: banks (38,528 claims) and department stores credit cards (21,863 claims). The final report to SERNAC it is present in Appendix A.

The distribution of complaints in each topic is shown in Figure 4.1 (every data set has quite same distribution) and about a third of the topics correspond to easily identifiable legal reasons. In Table 4.1, as an example, we show documents classification in 3 topics and how SERNAC labeled each topic. Those topics confirm some categories of SERNAC classification.

Moreover, there are many topics that reflect common consumer's problems, like persons scared for being drowned in debts or people worried about their credit report.

We plotted the monthly numbers of complaints for each topic. This visualization gave us 2 different insights: companies/consumer behavior during year and events. For example, in Figure 4.2 is observable that months with less complaints are February and September. This consumer behavior can be explained by summer holidays and national holidays. Most people in Chile go on vacations

Topic Label	Sernac Classification
Improper Charges	55% of complaints classified as "Improper Charges"
Calls from Collection Agencies	61% of complaints classified as "Extrajudicial Collection"
Unilateral Renegotiation	50% of complaints classified as "Unilateral Renegotiation"

Table 4.1: Topics related to legal motives in Sernac Classification.

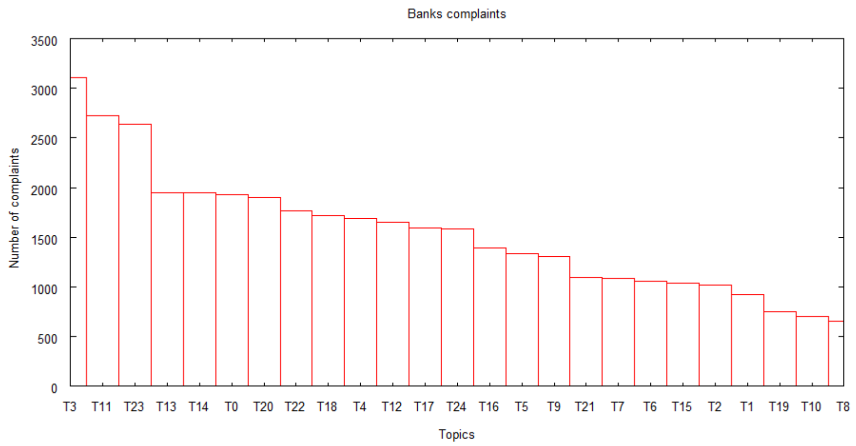


Figure 4.1: Claims for each Topic in Banks data set in descending order

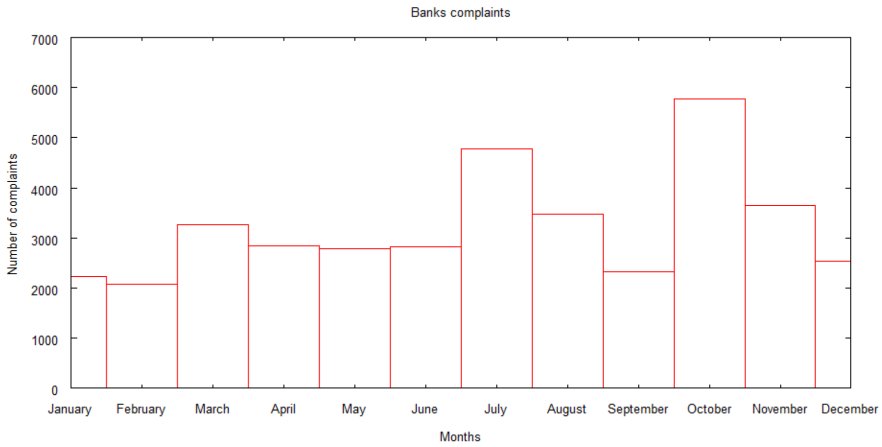


Figure 4.2: Claims per Month in Banks data set

on February and in September we celebrate Chile's national holiday usually for a week.

For topics that identified common behaviors and problems with financial markets and companies, we discovered that some of them have been declining since the government opened the "SERNAC Financiero" in March. Others had declined for a few months and then raise again. Examples of those cases are shown in Figure 4.3.

We also observed that some monthly graphics of topics shown a peak in one month or claims are mostly in a specific date. Those topics are events. For example, one topic in banks data set has all documents in July, as is shown in Figure 4.4. The main terms (words with highest probabilities) in that topic are "enviar (send)", "recibir (receive)", "cartola (account statement)" and "personal (personal)", also words like "error(mistake)" and "Julio (July)". Justly in July 2012, there was a controversy with a Chilean bank because account statements were sent to wrong persons. Thus, debts and personal information were exposed to others, violating personal privacy [1]. The name of this bank also appeared as a main term.

[1] http://www.biobiochile.cl/2012/07/03/banco-de-chile-reconoce-error-en-envio-de-datos-personales-a-traves-de-correo-electronico.shtml

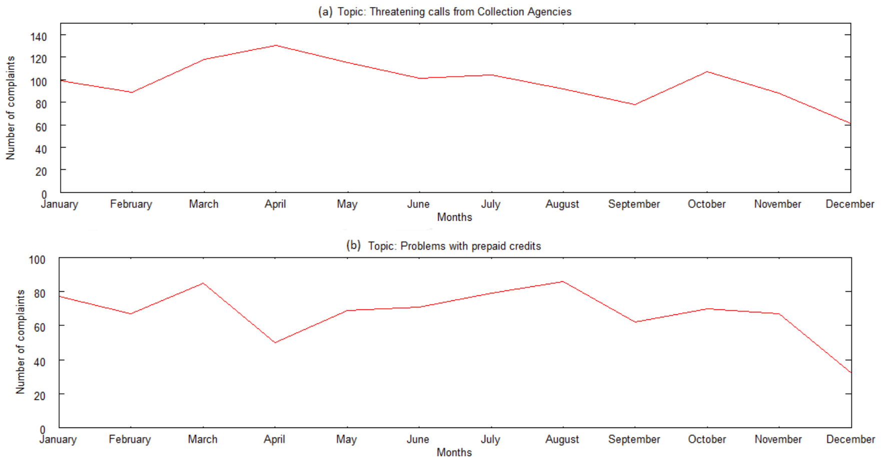


Figure 4.3: Claims per Month in 2 Topics of Department Stores Credit Cards data set. In (a) after March complaints increased in April and later decreased every month. In (b) complaints decreased in April but later started to increase.

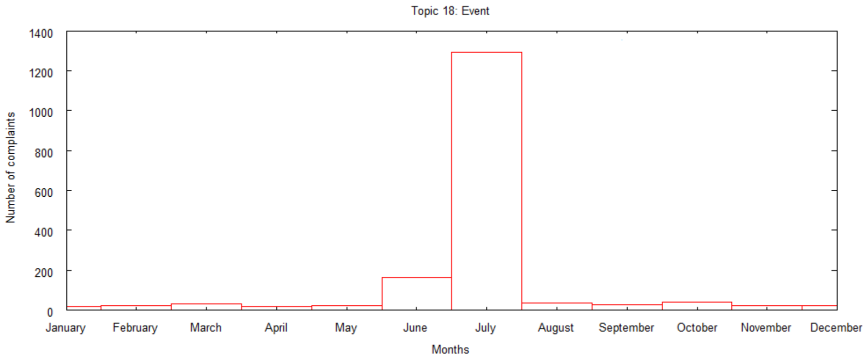


Figure 4.4: Claims per Month in Topic 18 of Banks data set.

Like that event, we found 2 more in banks data set.

Finally, there are some topics that did not change in number of complaints during year. One of them corresponds to a lawsuit that SERNAC won in January 2014. With the new law that supports SERNAC Financiero, the institution demand a copy of contracts of department stores credits cards. Analyzing those contracts, SERNAC found illegal charges and took legal action against companies involved. SERNAC won the dispute and companies have to return 2,000 million Chilean pesos (almost 4 million USD) to consumers [2]. We have a topic that discovered this case with 874 complaints related. Could be used as a evidence or SERNAC could identify this case earlier. Similar to this case (either past lawsuit or actual lawsuit) there are 2 topics more, besides all the topics with no changes.

[2] http://www.sernac.cl/empresas-del-retail-restituiran-cerca-de-dos-mil-millones-de-pesos-a-consumidores/

4.1.1 Topic Models Validation

As you can observe, further the extraction of topics, we developed a structured form for analyze them in order to get valuable information to improve management in SERNAC. Many of these topics are related to legal issues that should be reformulated. Between topics we found: common problems for SERNAC, consumer's behavior, organizational behavior, special events and specific problems with products/services.

We presented those results to SERNAC and the institution was satisfied. For them, topics models are a suitable technique to do complaint's analysis and to extract valuable information. Looking all the information extracted they chose those topics that are valuable to SERNAC. Based on its election, we could define our metric "Value", explained in section 2.4.4.

4.1.2 Methodology for Complaint's Analysis

Considering the results presented, the analysis done and the structure of SERNAC's information, we propose a methodology for the analysis of complaints:

1. **Characterize Topics:** The first step is interpret each topic. Using the top words in each topic it is possible to interpret the underlying context of the topic. Also, we labeled each topic using top words. So, the label summarizes the topic described in it. Once the topic is labeled, we can figure which kind of topic is it (a common problem, an event, among others).
2. **Topic Size:** After identifying the topic, we classified the complaints using topics. Therefore, a complaint belongs to a topic if the highest probability of that complaint is related to that topic. After this, we counted how many complaints belong to each topic. The size of the topic gives us an approach to dimension the problem related.
3. **Trends:** Every complaint has its date and location. So, we counted how many complaints are each month and in each location of the country. This allows us to find trends in months or places. Maybe a specific problem occurs in a specific location or during a specific month.
4. **News:** We compare some important events in the country with the evolution of complaints in some topics. New policies or events in consumer field can affect the trend of some topics.
5. **Legal Aspects:** We related each topic with legal categories. To do this, we counted the complaints classified in each legal category in each topic. This crossing of information shows how important or bad can be the topic found and leads to take actions against it.
6. **Entities:** Some complaints talk about the company involved in the problem. Thanks to this information (that we identified as Key Entities, see section 3.3.2), we connected companies with topics identifying which companies has which kind of problems.

Other options to add to the methodology are (not explored for us):

- Connecting people with companies. We don't have the information about which person claim against which company. But it is possible to identify which topics matters more between companies and people by creating this relation.
- Identifying Communities. Having the relationships described above, it is interesting to analyze if there are communities of consumers complaining about the same or against the same

company. This information could lead to apply legal actions.

- Better Categories. SERNAC has its own classification of complaints. We propose to contrast this classification with the classification made with topics. Some topics can show hidden categories for SERNAC.
- Educating people. Crossing the topics with FAQ's, it would be helpful to update them. Also, compare if people are following the correct process or not.

4.2 Benchmark of Models

For this phase we work with department stores credit cards data set (21,863 claims). In this section we show the results of applying the 3 other models: Pitman-Yor Topic Model, Latent Semantic Analysis and Non-Negative Matrix Factorization. First we discuss about the general results of each model (overview). Then, we specified their behavior in terms of coherence, interpretability and similarity. Finally, we summarize the comparison to conclude considering all the research.

We want to remind that we extracted 25 topics in each model. All the topics were labeled and tested with the metrics and measures (except the word intrusion test, see section 3.5.2 for more explanation).

4.2.1 Overview and Value

In table 4.2 we present some general aspects about the results of each model. Topics Size refers to the range between the topic with most complaints related and the topic with fewer complaints related. The positive aspects remark some good appreciations about the results and the negative aspects refer to bad observations about the results. LDA has topics with equal distributions of complaints (the difference between topics is not really big). About a third part of them are not useful o weird (not a specific topic). 7 topics refer to known problems that are potentially valuable if we look deeper (doing more analysis). This model has 4 topics considered as valuable. One it is for its specificity (against a company) and the other three for being serious problems (illegal). But in general, topics are easy to interpret reading the top words and the model extracted known themes for SERNAC. Anyway, it was necessary to read some complaints strongly related to the topic to make it clearer.

Aspect	LDA	PYTM	LSA	NMF
Topics Size	5.91% - 2.37%	7.55% - 1.74%	21.23% - 0.6%	13.2% - 0.96%
Positive Aspects	Easy to interpret	Easy to interpret	Polysemy and synonymy	Valuable topics are easy to interpret
Negative Aspects	Very standard and general	Slow process	Topic's size and difficult to understand	Not easy to interpret

Table 4.2: Models Overview

Now, comparing the 2 Bayesian models, PYTM has a bigger difference in topics size than LDA. It was also necessary to read some complaints to define better the labels of the topics. But, the valuable topics found in PYTM (4 topics) are better defined for its words than LDA. PYTM has 9 potential topics to add value. As we explained in section 3.4.2, the implementation used is very slow.

Frequency models are quite different from Bayesians in results, especially LSA. First, if we choose the closer topic to relate the complaints, we obtain a big topic with the 66% of complaints related. It is completely ambiguous, without a single theme in it. For this reason, if the difference between the first topic and the second closer topic is small (less than 1), we choose the second topic. Re-organizing complaints and topics with this new criterion, the biggest topic has 21% of the complaints related. Anyway, the size difference between topics in LSA is the biggest of the 4 models. In NFM was not necessary to change the criterion to relate complaints with topics. But as you can observe, the topics size difference is bigger than Bayesian models too.

Analyzing the top words in frequency models, we checked that LSA has polysemy and synonymy properties. It has topics in which consider the semantic use of a word and also, in top words can appear a word just in singular (or plural) and not both. Besides those good attributes, the topics are not easy to understand. The top words in each topic seem not to be strongly related making difficult the task of interpret the topic. It was completely necessary to read complaints related. After that, we could identify 4 valuable topics and 5 potential topics.

NMF has quite the same problem. The difference was that the valuable topics were easy to interpret but not the rest. Moreover, NMF presented 5 valuable topics and many topics seems to be closer to Bayesians topics (we will verify this in section 4.2.4) because they are specific.

Finally, in table 4.3 we show how many topics are valuable and potentially valuable in each model.

Number of topics	LDA	PYTM	LSA	NMF
Valuable	4	4	4	5
Potentially valuable	7	9	5	6

Table 4.3: Number of topics valuable and potentially valuable in every model

4.2.2 Coherence

The authors of Topic Coherence worked with experts that classified topics in good or bad. A topic is good if the words of the topic belong to a single concept. Then, they calculated their metric and also applied the word intrusion task with the experts. They realized that good topics have high accuracy identifying the intruder and bad topics have uniform accuracy. But topic coherence was successful identifying good topics (good topics presented higher topic coherence than bad topics).

So, their contribution was a metric capable of identify a class of topics with low quality that can't be detected with word intrusion tests. Based on that, topics with higher coherence should be "good" topics. Therefore, they should have a high accuracy in word intrusion task, because their words represent a single concept.

In table 4.4 we present topics from each model with high topic coherence and their value in word intrusion task. Position in coherence is the position of the topic ordered by coherence, being the first position the topic with highest coherence in the model. As it is observed, topics with high coherence have not high value in word intrusion. All of them, except one, have a value below 0.5, which means that less of the half people polled, could identify the intruder.

Model	Topic	Position in Coherence	Word Intrusion Value
PYTM	Topic 23	1st	0.25
PYTM	Topic 3	4th	0.45
LDA	Topic 18	4th	0.38
LSA	Topic 24	2nd	0.34
LSA	Topic 5	4th	0.73
NMF	Topic 22	2nd	0.48
NMF	Topic 16	3rd	0.50

Table 4.4: Topics from the 4 models with high topic coherence value and their word intrusion value.

Then we checked the bad topics (topics with lowest coherence in their models) comparing them with our value scale (if they are valuable to SERNAC or not). Value is 0 if the topic is not valuable, 0.5 if it is potentially valuable and 1 if it is valuable. Results are in table 4.5. It is seems that bad topics are not really bad topics for some models (LDA and LSA), because some bad topics are valuable. So, in our opinion, this metric is not helpful identifying good topics.

Model	Topic	Position in Coherence	Value
PYTM	Topic 14	25th	0.5
PYTM	Topic 0	24th	0.5
LDA	Topic 4	25th	0.5
LDA	Topic 1	24th	1
LSA	Topic 0	25th	0
LSA	Topic 8	24th	1
NMF	Topic 23	25th	0.5
NMF	Topic 3	24th	0

Table 4.5: Topics with low topic coherence value and their value.

We decided to do a new experiment based on those results. In the paper the word intrusion task was done with experts. We polled common people. So, we decided to poll our experts and compare. We repeated the task with 2 SERNAC employees. Results are in table 4.6 for good topics. SERNAC word intrusion value has the same trend than our previous value. Even the single "good topic" with high word intrusion has a 0 for SERNAC. So, can we compare the word intrusion metric with topic coherence metric? And what can we say about the value of the topics with topic coherence metric?.

Since we couldn't have the same results obtained in the original work of topic coherence, we decided to take a different approach. We observed if valuable topics in each model are coherent

Model	Topic	Position in Coherence	Word Intrusion	SERNAC WI
PYTM	Topic 23	1st	0.25	0
PYTM	Topic 3	4th	0.45	0.5
LDA	Topic 18	4th	0.38	0
LSA	Topic 24	2nd	0.34	0
LSA	Topic 5	4th	0.73	0
NMF	Topic 22	2nd	0.48	0
NMF	Topic 16	3rd	0.50	0

Table 4.6: Topics with high topic coherence, their word intrusion value and SERNAC word intrusion value (SERNAC WI).

or not. If they are, for SERNAC could be easier to identify valuable topics. In other words, when SERNAC is analyzing topics, it should be easier to classify between valuable or not if the model is capable of generate coherent-valuable topics.

We calculated the average of coherence in each model and we counted how many valuable topics are over the average. We counted how many valuable topics are top 4 in coherence too. Table 4.7 show the results. The only model that has all its valuable topics in the most coherent list of topics is PYTM. But none of its topics is top 4 in coherence. NMF instead has 2 of 3 valuable topics in top 4.

Number of topics	LDA	PYTM	LSA	NMF
Valuable	4	4	4	5
Over Average	2/13 (15%)	4/15 (26%)	1/14 (7%)	3/10 (30%)
Top 4	1/4 (25%)	0/4 (0%)	1/4 (25%)	2/4 (50%)

Table 4.7: Number of valuable topics in each model over the average of coherence and top 4 in coherence

4.2.3 Interpretability

Chang et al. consider Perplexity (described in section 2.4.1) and held-out likelihood as not useful metrics to explore common goals in topic models. Those metrics are useful just to evaluate how predictive is a model. Therefore, the goal of authors of word intrusion is to measure the successfulness of interpret topic models over number of topics and model assumptions.

They found that traditional metrics are negative correlated with topic quality. Also, extracting high number of topics produces topics more fine-grained but less useful to humans.

Moreover, it is known that the use of words fits a Zipf-law distribution. According to Ferrer and Solé[13], people make the least effort when they have to think in words to express what they want. Hence, people tend to use the most ambiguous words in their language. It is a receptor task to interpret words and extract the semantic context.

Considering this, it is possible to obtain low values in word intrusion task, because the test expose people to the top words, probably then the most ambiguous words in the topic. Therefore, we believe that models with high values in word intrusion are capable to define semantic context mixing ambiguous words with specific words (co-occurence). So, can we say that highly interpretable topics are good topics?.

As we explained in section 3.5.2, we didn't create surveys for all the topics. So, we select the valuable topics in each model that were polled. Then, we counted how many were over 0.7 in word intrusion task and over 0.5. Table 4.8 show the results. The percentage was calculated counting the number of polled topics over 0.7 (or 0.5 respectively). Because bias (we polled 10 of 25 topics from each model, randomly chosen), we can't say that a model find more interpretable topics than other (counting topics). But we can analyze the relation between the valuable topics and their interpretability in each model. It is observable that the only model that has all its valuable-polled topics over 0.5 is PYTM. Over 0.7 LSA and LDA have one valuable and highly interpretable topic.

Number of topics	LDA	PYTM	LSA	NMF
Valuable	4	4	4	5
Valuable polled	3	3	4	3
Over 0.7	1/1 (100%)	0/1 (0%)	1/1 (100%)	0/2 (0%)
Over 0.5	2/3 (66%)	3/8 (38%)	1/3 (33%)	1/5 (20%)

Table 4.8: Number of valuable topics over 0.7 and 0.5 in word intrusion task.

4.2.4 Similarity

We want to observe if the topics extracted in one model are present in other models in similar topics or in more than one topic. In other words, we want to find common topics along models, topics represented in many topics in other model and topics that just were extracted in one model. Moreover, we want to analyze if based on topics we can find complementary models, instead of choose just one for SERNAC.

To do this analysis we applied cosine similarity and we counted how many complaints related to one topic in one model are in other topic in other model. Also we compared the label and description of "similar topics" to define better the differences and similarities.

LDA and PYTM have more similar topics between them. The same it is observed between LSA and NMF. Since our research is focus on value, we decided to analyze deeper the valuable topics. The figure 4.5 shows the intersection of valuable topics between models.

Single Topics

LDA has its 4 valuable topics present in the other models. In fact, 3 of them are present in all the other models. The topic 1 of PYTM refers to a specific problem in a specific company. It has some similarity with topic 14 in LDA but the problem is completely different (the company involved is the same).

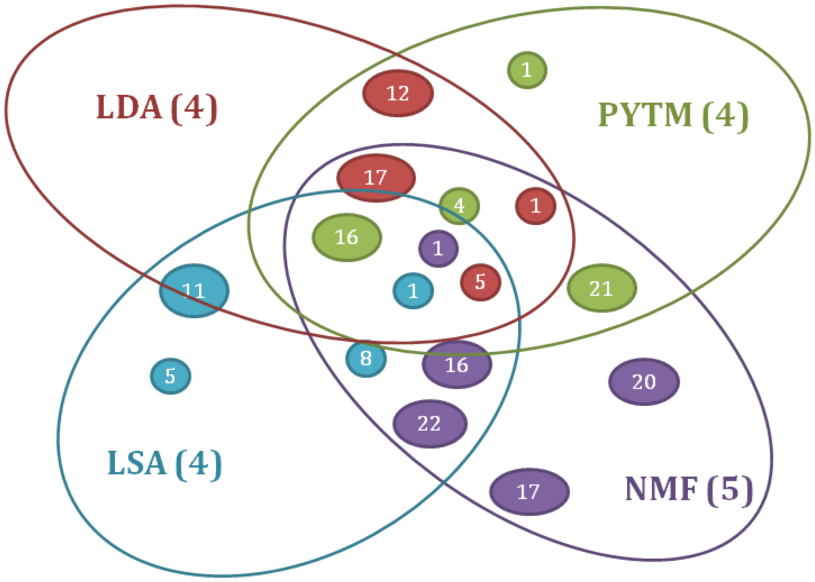


Figure 4.5: Intersection of valuable topics between models

The topic 5 in LSA model refers to many known and serious problems against retailers. So, the problems related are important but it is not a specific topic. We found the same in topic 17 in NMF. While topic 20 in NMF is the opposite. Topic 20 refers to many serious problems in a supermarket chain. None of them repeated in other models as clear as in their models. Table 4.9 present all the topics described in this section.

Common Topics

Most valuable topics are present in all the models. The differences are in how they were described and interpreted. For example, topic 12 in LDA is present in PYTM. But in LDA this topic has value and in PYTM was classified as potentially valuable. The reason is in PYTM the topic is quite specific and too common. So, SERNAC has no interest in it. But in LDA is one level up (more general). Then, it can lead to serious problems.

Topic 5 in LDA, topic 16 in PYTM, topic 1 in LSA and topic 1 in NMF are quite the same. The difference is in how they were interpreted. Table 4.10 shows the description of those 4 topics. In particular, topic 5 (valuable), 14 (not valuable) and 17 (valuable) from LDA are present in topic 16 of PYTM. Topic 1 in NMF, topic 17 in LDA and topic 4 in PYTM refer to the same theme too. Because how they are described and interpreted make them different from Topic 1 in LSA. This can be appreciate it in table 4.11. Topic 1 in LDA is present in PYTM and NMF but the topics in those models are not valuable. In PYTM the topic (potentially valuable) refers to 3 sub problems in the theme and in NMF there was not a clear problem. Instead, in LDA the topic is against a specific credit card.

There are many topics similar between LSA and NMF. Specifically, topic 1, 16 and 22 in NMF are better described than in LSA. Topics 1 in LSA is also better described in NMF and the topic related

Topic	Model	Description	Translation
Topic 1	PYTM	Pagos con tarjetas CMR visa o a Falabella y se pide reembolso o reversa por devolución de productos o porque la tarjeta no funcionó y figura el cargo	Payments made with CMR visa or to Falabella. Consumer ask the refund because the product was returned or the credit card didn't work but the charge appears anyway
Topic 5	LSA	Problemas de personas endeudadas. Les repactaron o los obligaron a repactar, sus deudas estan muy altas y no tienen como pagar. Les exigen pagos minimos que no estan claros. También algunos estan informados en dicom y ya pagaron	People drown in debts. High debts and they can't afford them. Minimum payments are not clear. Some of them are informed in dicom and they already pay
Topic 17	NMF	Repactaciones unilaterales o aumentos excesivos de las deudas	Unilateral Renegotiation or excessive increases in debts
Topic 20	NMF	Reclamos contra Presto por mantener cargos que ya fueron cancelados o borrados. También por anulación de ventas y devuelución del monto original del producto pero no de los intereses de las cuotas pactadas. También cargos por mantención de tarjetas que no son utilizadas y gente sin deuda con tarjeta bloqueada o reportada en dicom	Complaints against Presto for keeping improper charges or charges already pay. Also it is because the company doesn't refund the interests of the purchase. They charge commissions in credit cards that have not been used and inform people to Commercial Bulletin for debts in blocked credit cards

Table 4.9: Topics extracted in their respective models that are not present in other models

in NMF have value too. Topic 8 in LSA is against a company referring to many problems. The problems are present in topics in other models but not strongly related to this company, they are focus more in the problem itself.

Topic 11 in LSA is close to topic 1 in LDA and other topics related to unilateral renegotiations. But has no specificity about the process or company related. It needs more analysis.

Finally, topic 21 in PYTM was extracted in NMF too. But in NMF has no specificity in the interpretation. This topic refers to an event happened in 2011 in the country (unilateral renegotiations done by La Polar).

4.2.5 Summary

Table 4.12 summarizes the analysis made in sections above. Appendix B shows the values of every metric in each model. Those results confirm our first appreciations about the models. Bayesian

Topic	Model	Description	Translation
Topic 1	LSA	Problemas con tarjetas de créditos. Fraudes, cobros indebidos, bloqueos porque si, entre otros	Credit cards problems. Frauds, improper charges, cards unreasonably blocked, among others
Topic 1	NMF	Tarjetas bloqueadas, cerradas o nunca utilizadas que siguen teniendo cobros de seguros, cobranzas o fraudes	Credit cards blocked, closed or never used that have insurance charges, collection commissions or frauds
Topic 5	LDA	Cobros improcedentes por fraude/clonación/robo de tarjetas. Varias bloqueadas y aun así fueron utilizadas	Improper charges for frauds/clonation/stolen credit cards. Many of them were blocked and used anyway
Topic 16	PYTM	Cobros realizados con tarjetas de Falabella o a clientes de Falabella totalmente desconocidos, de días que no han ido o adicionales que no tienen	Charges in Falabella credit cards or to Falabella clients completely unknown

Table 4.10: Common topics over models

Topic	Model	Description	Translation
Topic 17	LDA	Tarjeta “muy vieja” (no ocupa hace tiempo) y la quiso ocupar y no pudo	Consumer couldn’t use an old credit card that it was not use for long time
Topic 1	NMF	Tarjetas bloqueadas, cerradas o nunca utilizadas que siguen teniendo cobros de seguros, cobranzas o fraudes	Credit cards blocked, closed or never used that have insurance charges, collection commissions or frauds
Topic 4	PYTM	Cobros en tarjetas cerradas, bloqueadas o no usadas	Charges in closed, blocked or not used credit cards

Table 4.11: Common topics over models NMF, LDA and PYTM

models are easier to understand and interpret than LSA and NMF (word intrusion values). But LSA and NMF focus more in some specific words that can add value to the topic (in our cases, words referring to companies). NMF present some improvements over LSA that make it more understandable (coherence) and close to LDA and PYTM.

Analyzing the similarities between valuable topics, we should apply the 4 models to obtain all those results. We believe that intersect results from a Bayesian model with a frequency model should be reasonable to obtain valuable topics with specificity. Besides that, our observations and analysis lead us to choose PYTM for SERNAC. NMF is still difficult to interpret for a person who doesn’t work with topic models. Therefore, considering the coherence and interpretability of valuable topics and our appreciations about similarities with other models, we would implement PYTM in SERNAC to do analysis of complaints. For this purpose, it should be necessary program a new code, because the implementation used in this research is too slow for SERNAC requirements.

Number of topics	LDA	PYTM	LSA	NMF
Valuable	4	4	4	5
Valuable polled	3	3	4	3
Over Average (Coherence)	2	4	1	3
Over 0.5 (Word Intrusion)	2	3	1	1

Table 4.12: Summary of metrics: value, topic coherence and word intrusion

Conclusion

We can say that all models extract valuable topics. The differences are in how easy is to interpret the topic and if it is necessary extra information (deeper analysis) to understand it better. We can't declare a model as the best absolutely. Every model has its good and bad attributes depend on the dimension analyzed. But for the reason explained in section 4.2.5, we choose PYTM for SERNAC.

Frequency models are difficult to interpret and understand. Hence, it is not possible to observe if the semantic space extracted in frequency models is better or not than Bayesian models. We just can say that the visibility of semantic space is higher in Bayesian models. Thus make us to prefer them for complaints analysis.

Anyway, we believe that NMF has a potential based on the topics similarities analyzed. Also it is because it extracted more number of valuable topics than the rest. We would like to see in future more investigations with this model and improvements in its interpretability. For LSA, we can't conclude it is an obsolete model. It has its properties that leave us curious about how really is the semantic space extracted. However, NMF has more potential to continue researching as Bayesian models as well.

As far as we concern, actual metrics are not enough to analyze the quality of models or their use for humans. We use a completely subjective metric called value and we related this metric with previous ones (word intrusion and metric). Thanks to that we could characterize each model. But nor traditional metrics or metrics used in this work leave us satisfied.

We propose to research in developing metrics which allow researchers to increase the number of valuable topics extracted. One idea is increase the number of specific words in the description of topics (top words) and mixes them in some proportion with ambiguous words. Also, it should be possible to create an indicator to predict how many valuable topics are possible to extract considering the model used and the number of topics.

Another investigation in our interests is to mix sentiment analysis with topics extraction. At least for SERNAC would be helpful to have topics with important problems and consumers feelings related too.

Bibliography

[1] David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[2] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - going beyond svd. In *FOCS*, pages 1–10. IEEE Computer Society, 2012.

[3] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[4] David M. Blei and John D. Lafferty. Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press, 2006.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] István Bíró, Jácint Szabó, and András A. Benczúr. Latent dirichlet allocation in web spam filtering. In Carlos Castillo, Kumar Chellapilla, and Dennis Fetterly, editors, *AIRWeb*, ACM International Conference Proceeding Series, pages 29–32, 2008.

[7] Jonathan Chang, Jordan L Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, volume 22, pages 288–296, 2009.

[8] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark) DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands), 2000.

[9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[10] Ali Faisal, Jussi Gillberg, Gayle Leen, and Jaakko Peltonen. Transfer learning using a non-parametric sparse topic model. *Neurocomput.*, 112:124–137, July 2013.

[11] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to

knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[12] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[13] Ramon Ferrer i Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):pp. 788–791, 2003.

[14] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63, November 1977. Supplement 1.

[15] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[16] C.J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[17] Camilo Lopez and Sebastian Rios. Diseno y construccion de una plataforma de clasificacion de texto basada en textmining aplicada sobre una red de blogs para betazeta networks s.a. In *Memoria para optar al titulo de Ingeniero Civil en Computacion e Ingeniero Civil Industrial*. Universidad de Chile, 2012.

[18] David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272. ACL, 2011.

[19] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[20] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 91–100, New York, NY, USA, 2008. ACM.

[21] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1):4:1–4:38, January 2010.

[22] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 673–682, New York, NY, USA, 2010. ACM.

[23] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[24] Akira Utsumi. Evaluating the performance of nonnegative matrix factorization for construct-

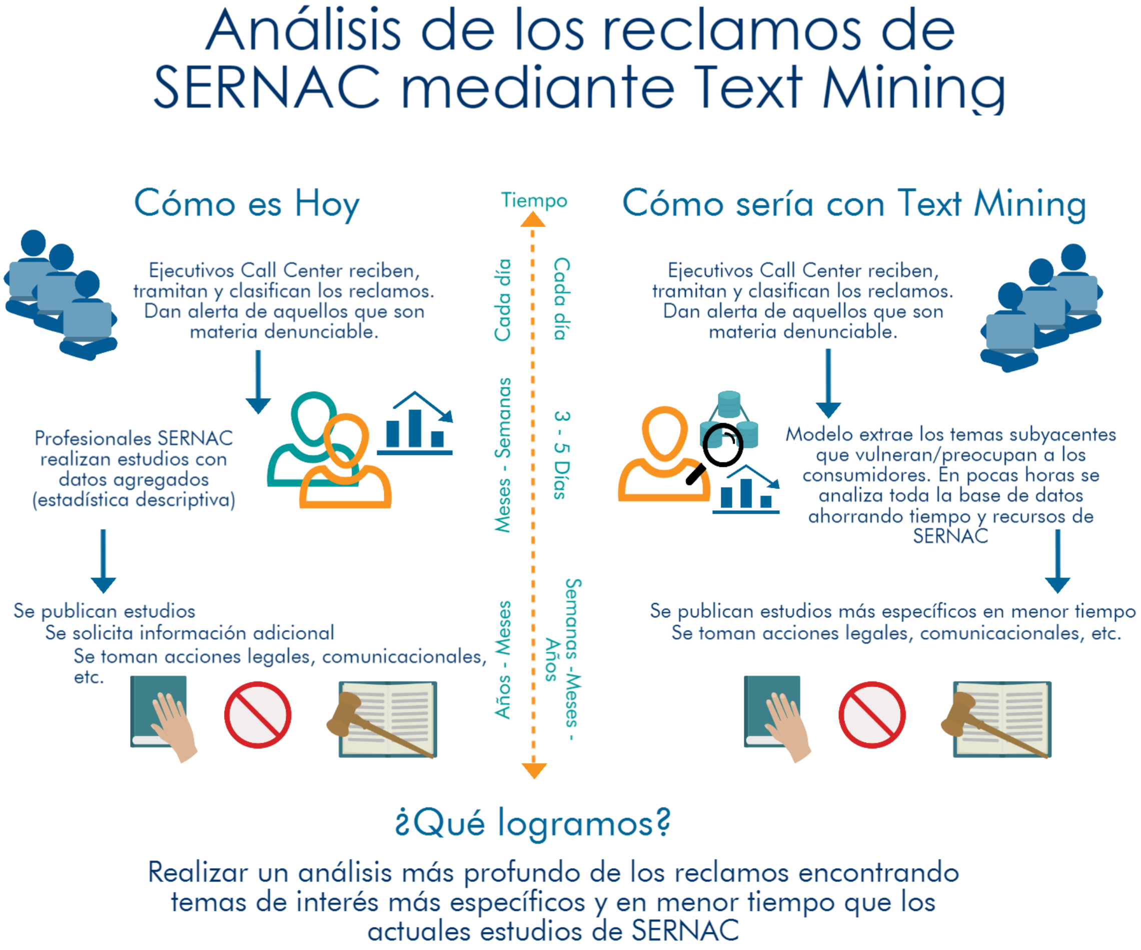
ing semantic spaces: Comparison to latent semantic analysis. In *SMC*, pages 2893–2900. IEEE, 2010.

[25] A. N. K. Zaman and Charles Grant Brown. Latent semantic indexing and large dataset: Study of term-weighting schemes. In *ICDIM*, pages 1–4, 2010.

Appendices

Appendix A

Infographic



Análisis de los reclamos de SERNAC mediante Text Mining

Cómo es Hoy

Ejecutivos Call Center reciben, tramitan y clasifican los reclamos. Dan alerta de aquellos que son materia denunciable.

Profesionales SERNAC realizan estudios con datos agregados (estadística descriptiva)

Se publican estudios
Se solicita información adicional
Se toman acciones legales, comunicacionales, etc.

Cómo sería con Text Mining

Ejecutivos Call Center reciben, tramitan y clasifican los reclamos. Dan alerta de aquellos que son materia denunciable.

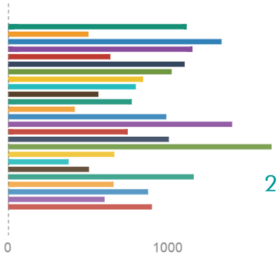
Modelo extrae los temas subyacentes que vulneran/preocupan a los consumidores. En pocas horas se analiza toda la base de datos ahorrando tiempo y recursos de SERNAC

Se publican estudios más específicos en menor tiempo
Se toman acciones legales, comunicacionales, etc.

¿Qué logramos?

Realizar un análisis más profundo de los reclamos encontrando temas de interés más específicos y en menor tiempo que los actuales estudios de SERNAC

Ejemplo



21.863 Reclamos

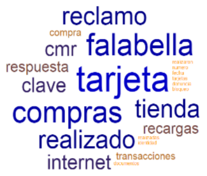
Tarjetas de Multitiendas Año 2012

25 temas encontrados

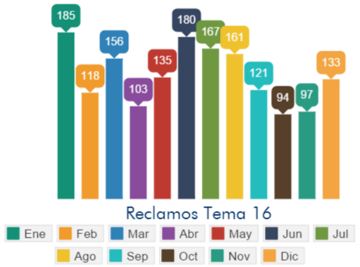
Tema 16

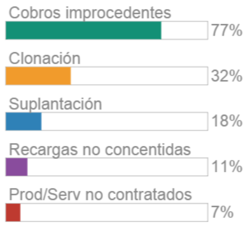
Cobros realizados a clientes de Falabella completamente desconocidos en tarjetas adicionales que no poseen o en días que no han realizado compras.

1.650 reclamos



Peak en Enero y Junio

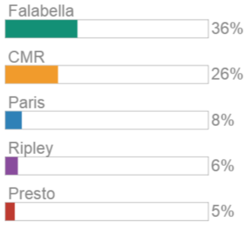




77% Cobros Improcedentes

Categoria Legal

30% Materia Denuniciable



36% menciona a Falabella

Otros temas relevantes

Tema 1 (505 reclamos)

Pagos realizados con CMR visa o con cheques a Falabella por los cuales se pide reembolso o reversa de cargos (por devolución de productos o porque la tarjeta no funcionó y figura el cargo)

Tema 4 (641 reclamos)

Cobros en tarjetas cerradas, bloqueadas o no usadas.

¿Cómo lo hicimos?

Analizamos los reclamos utilizando algoritmos de Semántica Latente

Y establecimos una metodología de análisis semi-automático de reclamos mediante extracción de tópicos

¿Qué podríamos lograr?

- Vincular a las empresas con los temas encontrados
- Identificar comunidades de consumidores en torno a los temas más importantes

Appendix B

Metrics for each model

We summarized the metrics for each model in tables B.1, B.3, B.2 and B.4 for LDA, PYTM, LSA and NMF model respectively.

B.1 LDA topics

Topic	Value	Size(5.91%, 2.37%)	Word Intrusion	Topic Coherence(-4,002.25, -3,537.98)
Topic 0	0.5	3.27%	0.14	-3,894.95
Topic 1	1	5.66%	0.64	-3,939.88
Topic 5	1	5.18%	0.37	-3,639.32
Topic 12	1	4.72%	0.88	-3,845.34
Topic 14	0	4.66%	0.49	-3,709.64
Topic 18	0	2.37%	0.38	-3,601.52
Topic 19	0	5.74%	0.46	-3,718.81
Topic 20	0.5	4.31%	0.24	-3,672.23
Topic 22	0	5.63%	0.42	-3,830.65
Topic 23	0.5	5.06%	0.62	-3,653.32

Table B.1: Metrics calculated in topics obtained with LDA model

B.2 LSA topics

Topic	Value	Size(21.23%, 0.6%)	Word Intrusion	Topic Coherence(-3,417.22, -2,964.58)
Topic 0	0	5.16%	0.5	-3,417.22
Topic 1	1	5.28%	0.47	-3,130.70
Topic 4	0	2.93%	0.64	-3,167.26
Topic 5	1	3.54%	0.73	-3,028.71
Topic 8	1	6.42%	0.31	-3,202.59
Topic 9	0.5	7.55%	0.44	-3,038.67
Topic 11	1	3.05%	0.24	-3,140.23
Topic 12	0	3.02%	0.24	-3,162.53
Topic 22	0	4.01%	0.13	-3,041.20
Topic 24	0.5	2.77%	0.34	-2,983.12

Table B.2: Metrics calculated in topics obtained with LSA model

B.3 PYTM topics

Topic	Value	Size(7.55%, 1.74%)	Word Intrusion	Topic Coherence(-4,383.29, -3,620.94)
Topic 0	0.5	5.16%	0.68	-4,332.97
Topic 3	0	5.28%	0.45	-3,909.74
Topic 4	1	2.93%	0.69	-3,994.97
Topic 10	0.5	3.54%	0.55	-4,312.99
Topic 13	0.5	6.42%	0.72	-4,188.13
Topic 16	1	7.55%	0.53	-4,008.03
Topic 17	0.5	3.05%	0.62	-4,199.07
Topic 21	1	3.02%	0.66	-3,984.36
Topic 22	0.5	4.01%	0.60	-4,298.51
Topic 23	0	2.77%	0.25	-3,620.94

Table B.3: Metrics calculated in topics obtained with PYTM model

B.4 NMF topics

Topic	Value	Size(13.20%, 0.96%)	Word Intrusion	Topic Coherence(-3,002.69, -2,641.01)
Topic 0	0	10.23%	0.49	-2,951.15
Topic 2	0	5.96%	0.72	-2,811.39
Topic 6	0	7.21%	0.68	-2,911.85
Topic 13	0	2.86%	0.18	-2,887.87
Topic 16	1	2.55%	0.50	-2,707.46
Topic 17	1	2.43%	0.30	-2,861.43
Topic 18	0.5	1.91%	0.82	-2,878.71
Topic 21	0.5	1.49%	0.28	-2,857.11
Topic 22	1	1.56%	0.48	-2,692.79
Topic 23	0.5	1.83%	0.57	-3,002.69

Table B.4: Metrics calculated in topics obtained with NMF model