ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Fundamentos Machine & Deep Learning

Diplomado Big Data y Analítica de Datos 2022

Profesor:
Rodrigo Sandoval U.

Contenidos en temas de NLP, NLU y NLG

Clase 6 NLP	• **Pre-procesamiento de Texto** • **Clasificación de texto** • **Extracción de entidades**
Clase 7 No-Supervisado	• Diferentes modelos Clustering • Deep Learning no-supervisado
Clase 8 NLU & NLG	• **Word Embedding** • **Contextualización y Secuencia** • **Transformers**

Y seguimos hablando de Redes Neuronales Profundas para lenguaje

Definir Uds, en sus propias palabras:

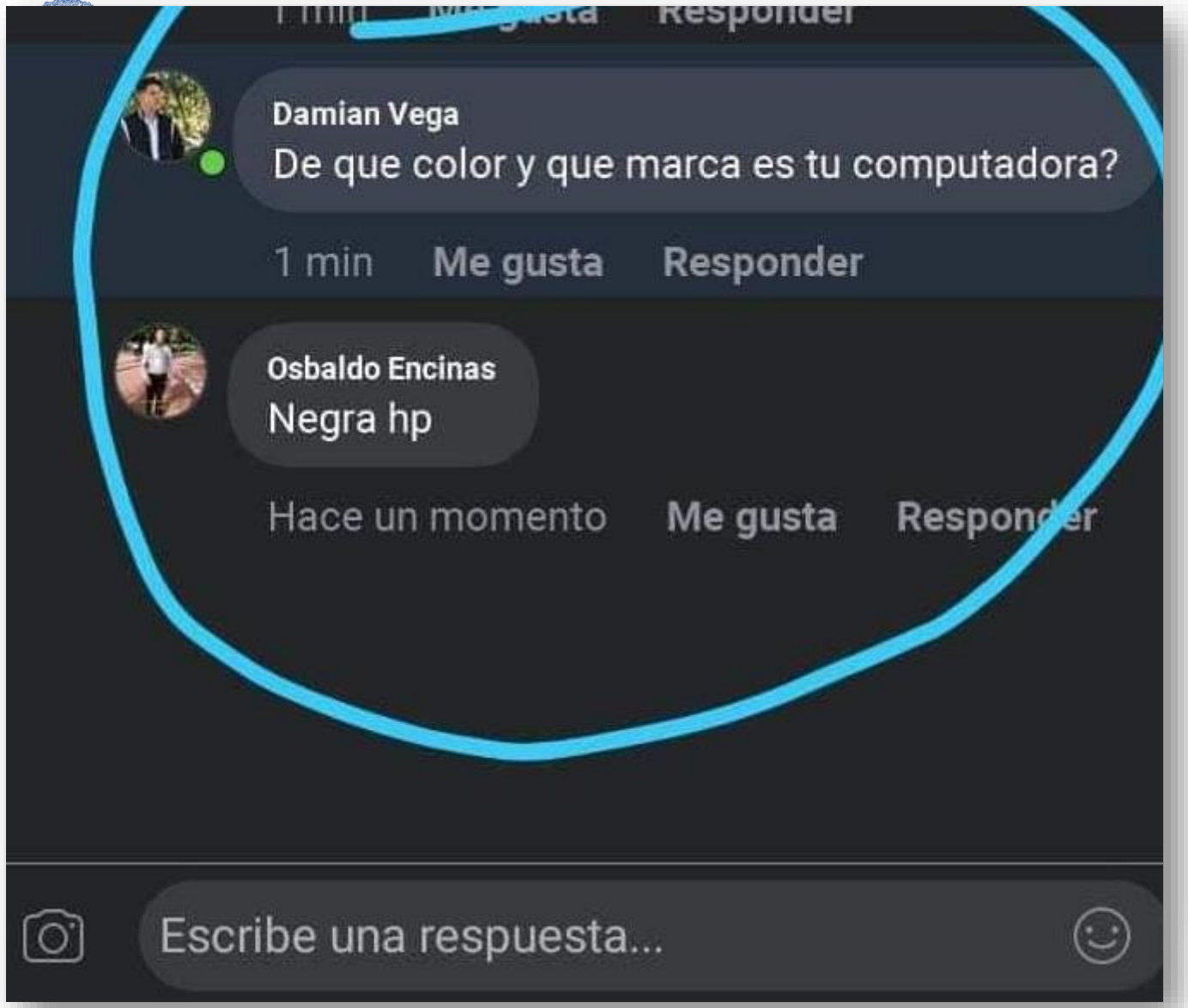
¿En qué consiste el procesamiento de lenguaje natural (NLP)?

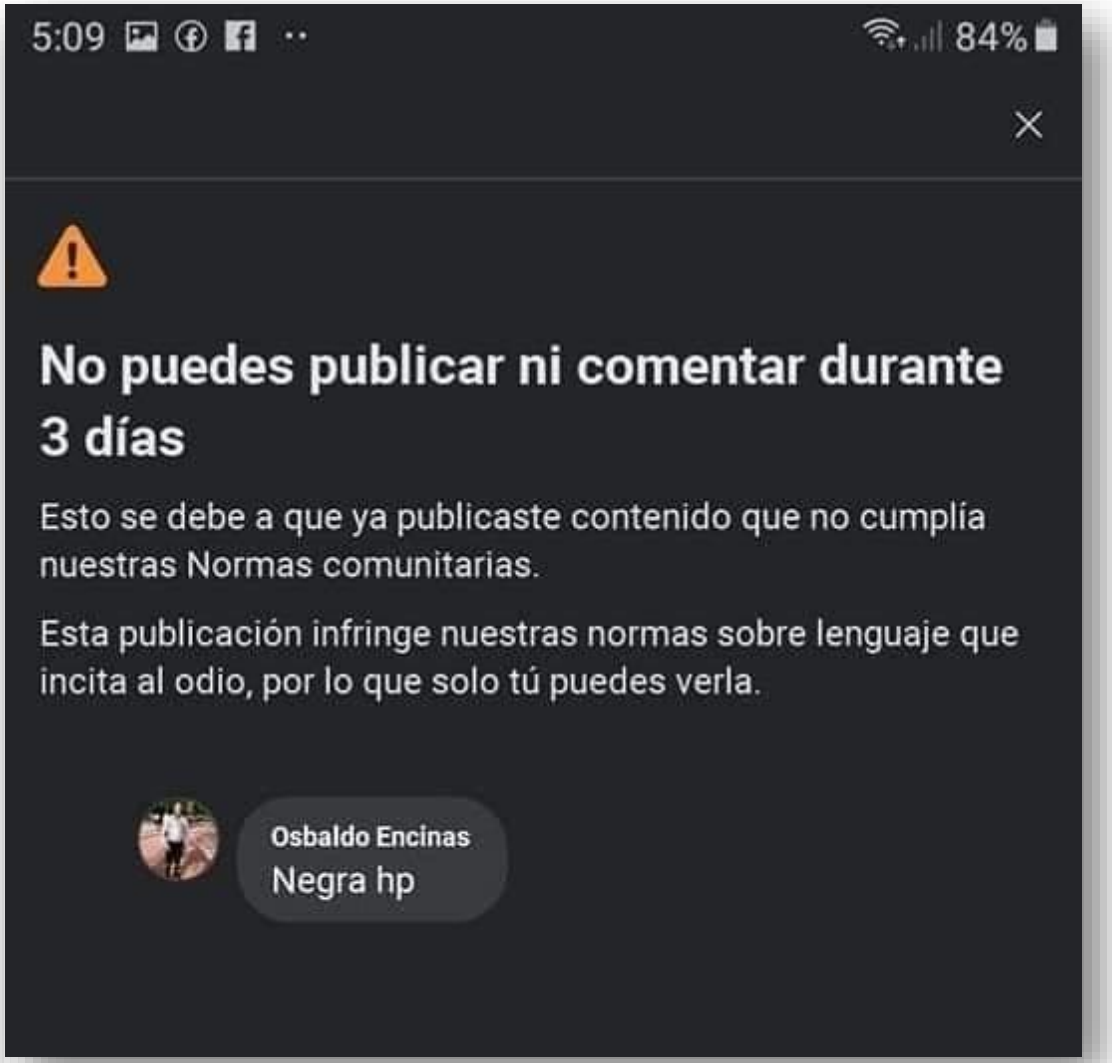
¿Qué ejemplos del día a día se les ocurre en NLP?

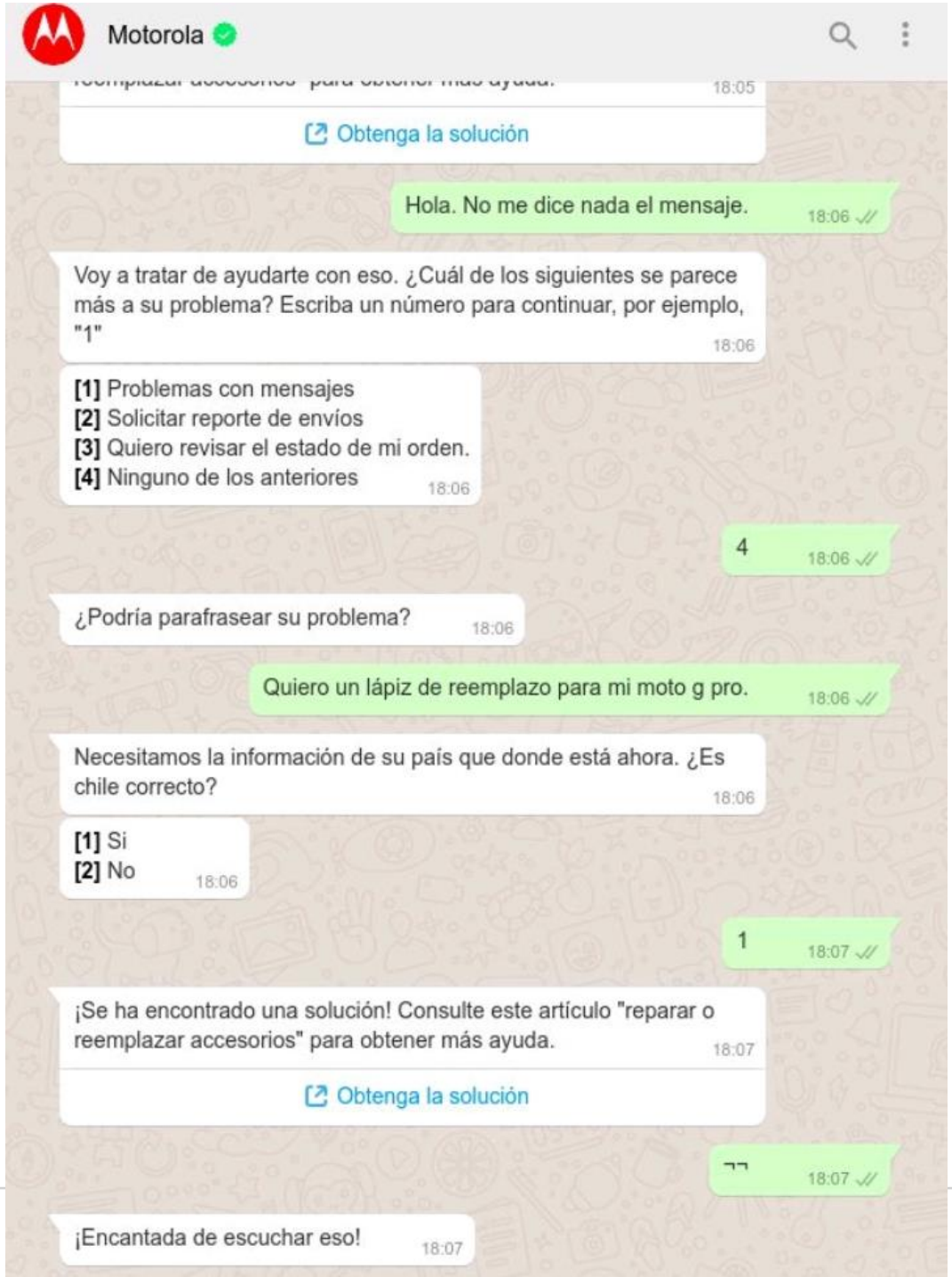
NLP – Procesamiento de Lenguaje Natural

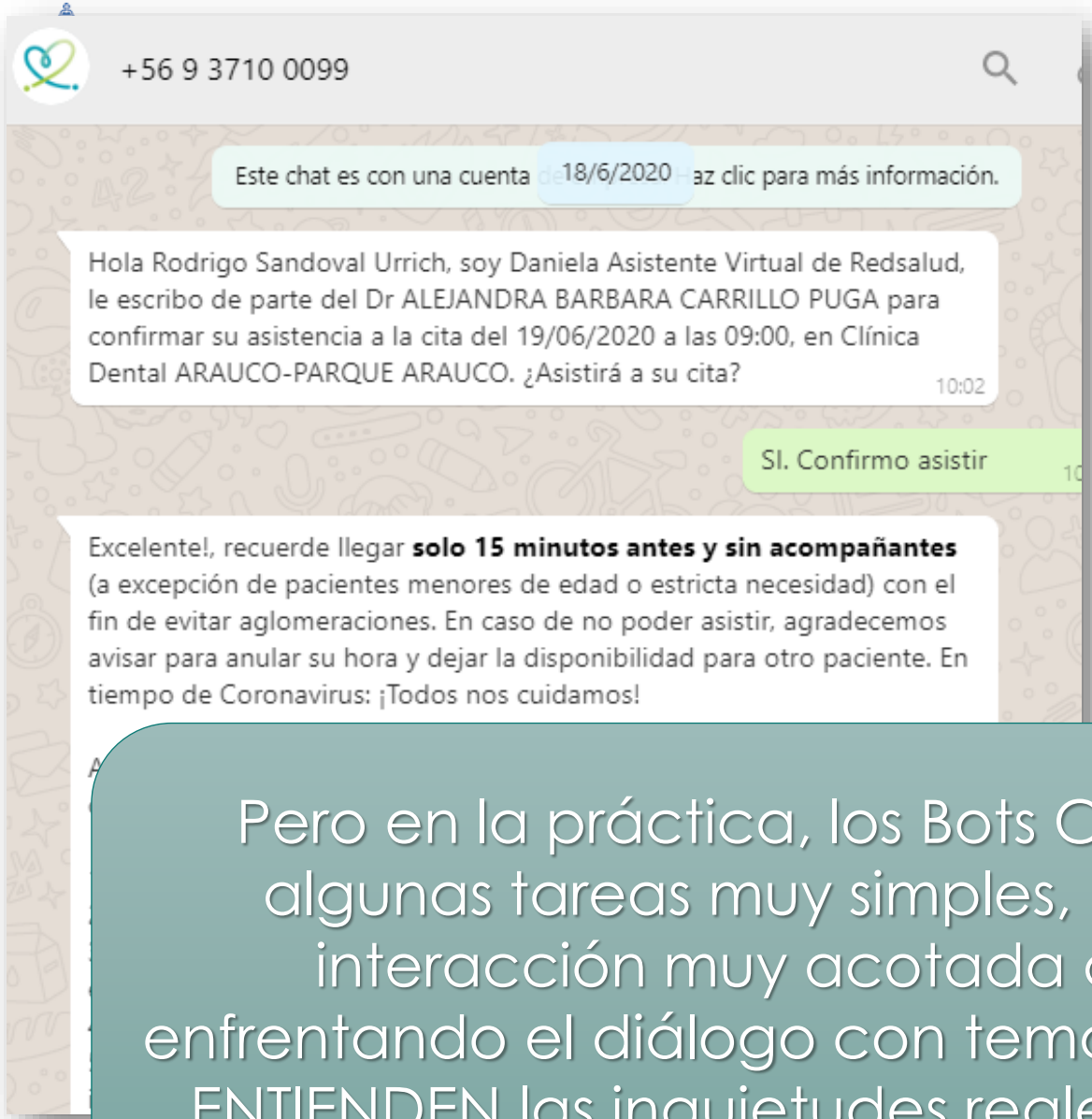
- *La manipulación automática del lenguaje natural, como texto o habla, por software.*

- *La extracción de información o datos útiles y significativos, desde texto natural, sin una estructura pre-definida.*


Damian Vega
De que color y que marca es tu computadora?
1 min
Me gusta
Responder
Osbaldo Encinas
Negra hp
Hace un momento
Me gusta
Escribe una respuesta...


5:09
84%
No puedes publicar ni comentar durante 3 días
Esto se debe a que ya publicaste contenido que no cumplía nuestras Normas comunitarias.
Esta publicación infringe nuestras normas sobre lenguaje que incita al odio, por lo que solo tú puedes verla.
Osbaldo Encinas
Negra hp


Motorola
18:05
Obtenga la solución
Hola. No me dice nada el mensaje.
18:06
Voy a tratar de ayudarte con eso. ¿Cuál de los siguientes se parece más a su problema? Escriba un número para continuar, por ejemplo, "1"
18:06
[1] Problemas con mensajes
[2] Solicitar reporte de envíos
[3] Quiero revisar el estado de mi orden.
[4] Ninguno de los anteriores
18:06
4
18:06
¿Podría parafrasear su problema?
18:06
Quiero un lápiz de reemplazo para mi moto g pro.
18:06
Necesitamos la información de su país que donde está ahora. ¿Es chile correcto?
18:06
[1] Si
[2] No
18:06
1
18:07
¡Se ha encontrado una solución! Consulte este artículo "reparar o reemplazar accesorios" para obtener más ayuda.
18:07
Obtenga la solución
18:07
¡Encantada de escuchar eso!
18:07



De: Daniela Asistente Virtual de Redsalud <daniela@.....cl>
Enviado el: lunes, 20 de julio de 2020 17:07
Para: rodrigo@.....com
Asunto: RE: Rodrigo, confirme su asistencia.

Queda confirmada su hora con DRA NNNN AAAAA, para el 27/07/2020 a las 09:45, en Clínica XXXXXXXXX.

Muchas gracias,
Daniela. Ejecutiva Clínica Dental ARAUCO-PARQUE ARAUCO

De: Rodrigo Sandoval <rodrigo@...com>
Si. Confirmo que asistiré el lunes 27

De: Daniela Asistente Virtual de Redsalud
<daniela@confirmaciones.redsalud.cl>
Enviado el: lunes, 20 de julio de 2020 17:05

...AA, para

...ail, de
...en
nuestro sistema su respuesta.

Pero en la práctica, los Bots Conversacionales sólo resuelven algunas tareas muy simples, basando su operación en una interacción muy acotada de lenguaje natural, siempre enfrentando el diálogo con temáticas específicas. Los Bots NUNCA ENTIENDEN las inquietudes reales de sus interlocutores humanos.

GranModa

Gran Moda SpA, RUT 55.666.444-3
Casa matriz: Los Alerces 4455, Puerto Montt
Sucursales: Las Tepas 6677 loc 12, Osorno; Arrayanes 1254 loc33, Castro; Pedro de Valdivia 9876, Valdivia.

Boleta Electrónica N° 9873865
Local: Internet e-commerce
Dirección: Casa Matriz, comuna Puerto Montt
Fecha: 12 enero 2022 Hora: 16:21
Vendedor: [00001 | e-commerce]

Código	Descripción	Cant	Precio
132556	Blusa cyan verano	01	$22.990
565464	Jeans Vermont rojo	01	$18.990
	SUBTOTAL		$41.980
	DESCTO		$4.200
	TOTAL		**$37.782**
	MCARD		$37.782

Cliente: Carola Santamaría Benavides 18.888.777-3
Puntos venta: 328 Puntos acumulados: 2.345

udarastyling INSPO -- LOOK TOTAL BLACK
Un look que nunca falla y que además puedes usar tanto en el día como en la noche y se verá bien!
★☆ -- Acá te dejamos 10 looks de inspiración para esta seman

¿Sabías que tenemos una página donde elegimos ropa SOLO PARA TI?
Para conocerla click al link in bio

#moda #styleinspo #style #outfit #ootd #moda #fashion #outf #estilo #looks #instafashion #outfitoftheday #streetstyle #tendencia #tipsdemoda #tenida #look

★☆
10 sem Ver traducción

elblogviajerodemarijane Que viva el negro, los abrigos largos y los pantalones pitillos (cuero o jeans) siempre pegan bien. 🙌
9 sem Responder Ver traducción

30 Me gusta
JUNIO 27

Reviews with images

See all customer images

Read reviews that mention

relaxed fit | another pair | true to size | business casual | different colors
work pants | good quality | great fit | wrinkle free | easy to care
second pair | even though | perfect fit | hard to find | day pant

Top reviews

Top reviews from the United States

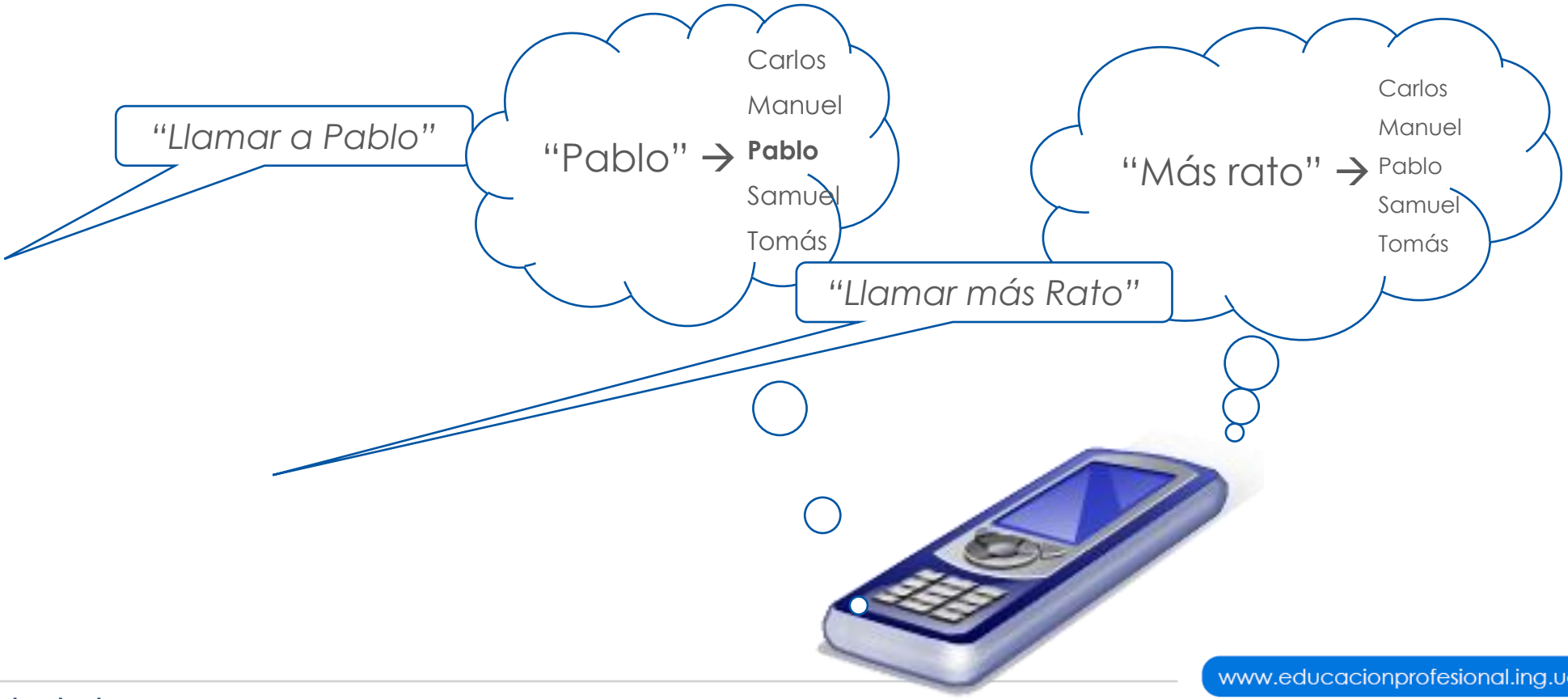
Elizabeth H
★★★★☆ **Very disappointed that this didn't fit me how i'd like, they're so comfortable!**
Reviewed in the United States on August 10, 2022
Size: 4 Short | Color: Flax | Verified Purchase

Okay, so I bought these pants because I needed them for a work/volunteer gig. I had high expectations because I have had other Lee pants and I really want them to fit properly. I'm 5'3 and 110-113lbs, not exactly the curviest person around. I ordered a 4-Short.
So...they do fit at the waist. Perfectly, I might add. Not tight, but definitely not loose. Very comfortably, exactly like how I think a pants should fit. However, they are both VERY baggy on my relatively skinny legs and a little bit long even though I bought the short version. When I put it on, I loved how comfortable they were, but disappointed how baggy and unkept they looked on me. I would get a 2, but I doubt the waist would fit. For my short queens (those sub 5'3), you might want to look elsewhere. I have a feeling they'd look rather long on y'all.

They are otherwise brilliant pants and great Lee quality. Such a bummer.

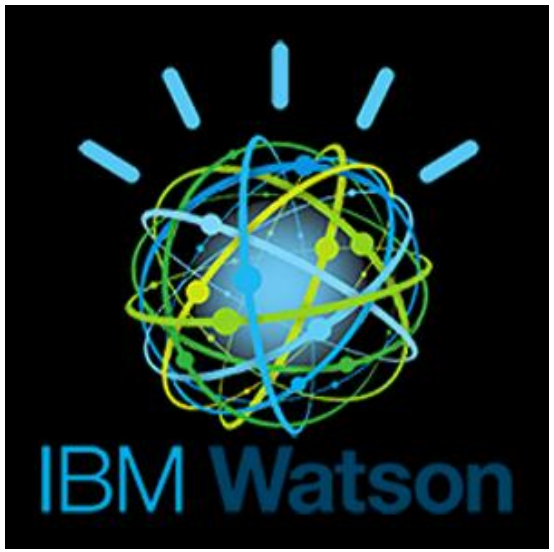
AI – Comprensión de Lenguaje

Primeras versiones …
¿comprensión de lenguaje?



Natural Language Processing with Watson

- IBM Watson: combina hardware y modelos altamente sofisticados, que analizan un texto no-estructurado, reconociendo elementos relevantes.
- Saltó a la fama cuando pudo ganarle a expertos en el juego de conocimientos Jeopardy



Entonces … ¿basta con procesar el lenguaje natural?

NLP
Natural Language Processing
(Qué se dijo)

NLU
Natural Language Understanding
(Qué significa)

NLG
Natural Language Generation
(Qué se responde)

Procesamiento Lenguaje Natural

Profesor: **RodrigoSandoval.net**

Procesamiento Lenguaje Natural

Objetivo: responder preguntas como éstas

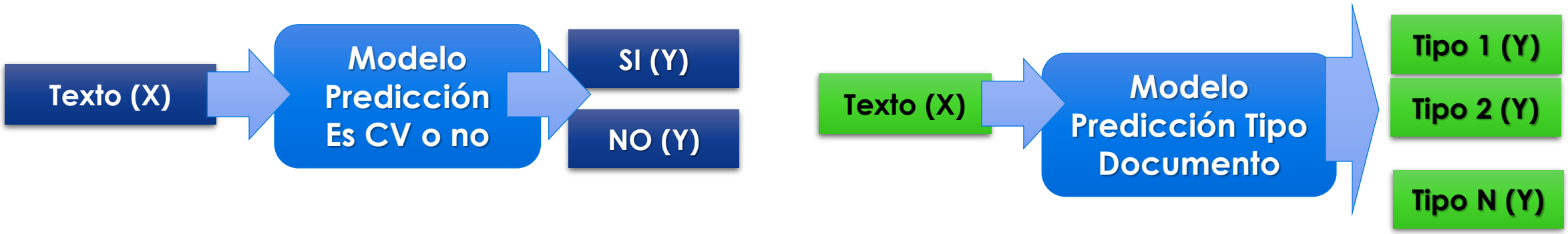
- ¿A qué se refiere un texto? ¿De qué clase es un texto en particular?
- ¿Qué cosas están mencionadas en un texto?
- ¿Cómo poder atender y reaccionar ante las necesidades de un usuario o cliente desde sus comentarios?
- ¿Cuál es la opinión o percepción de personas en sus comunicaciones?

Objetivos en NLP

Clasificación	• Procesar una porción de texto y asignarle una clase o categoría.
Extracción	• Desde una porción de texto, determinar (extraer) un dato de un tipo definido
Comparación	• Determinar la semejanza o similitud entre dos porciones de texto.

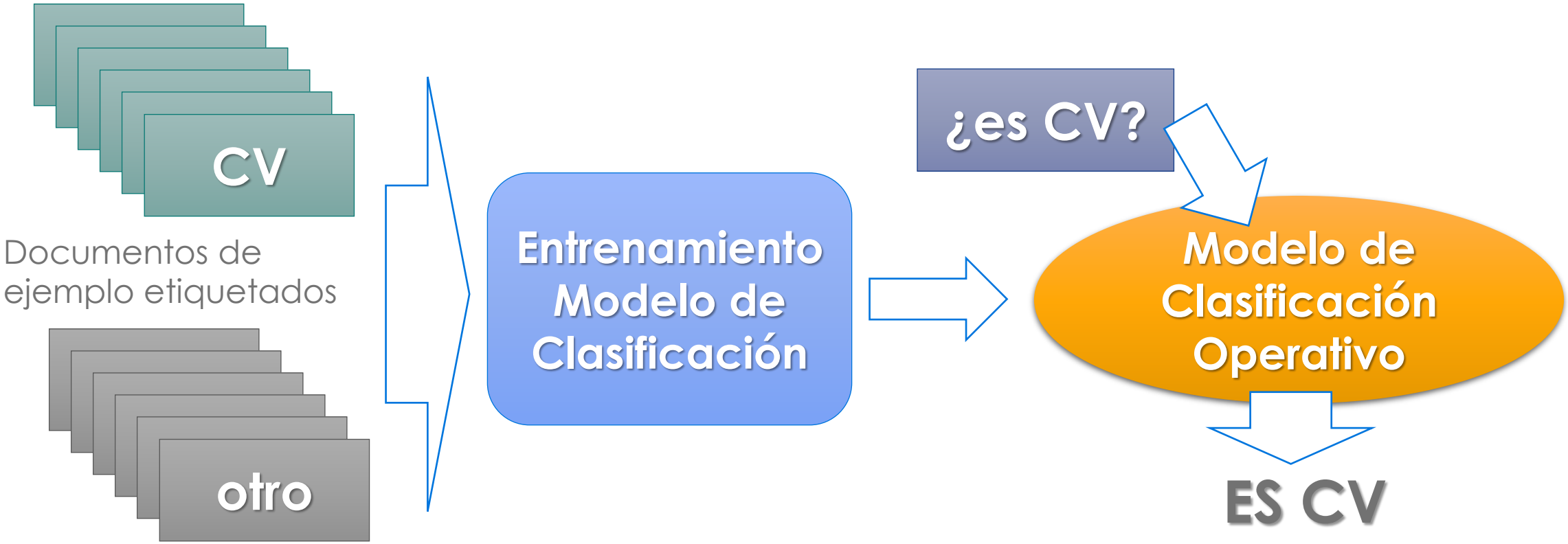
A. Clasificación de Texto No-Estructurado

Modelos de Clasificación Supervisada



- La predicción de que una porción de texto corresponda a una u otra categoría sigue la misma lógica y dinámica de cualquier modelo de clasificación supervisada.
- Se requieren ejemplos de cada clase, pero en el contexto de lenguaje natural, dada la gran diversidad de palabras, normalmente se requiere un mayor volumen de ejemplos para lograr un rendimiento aceptable.

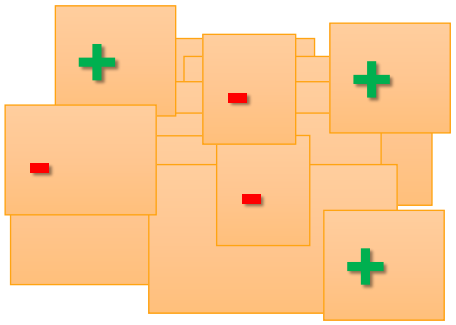
Modelos de Clasificación Supervisada



La determinación de que un documento entregado es un CV válido se basa en un modelo de clasificación, que al ser entrenado con varios ejemplos de CVs y de no-CV, es capaz de determinar que un nuevo documento es considerado un CV válido o no.

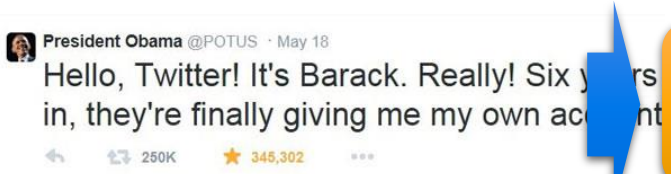
Otro Uso: Análisis de Sentimiento

Sentiment Analysis sobre texto natural o no-estructurado.



Ejemplos etiquetados

Entrenar



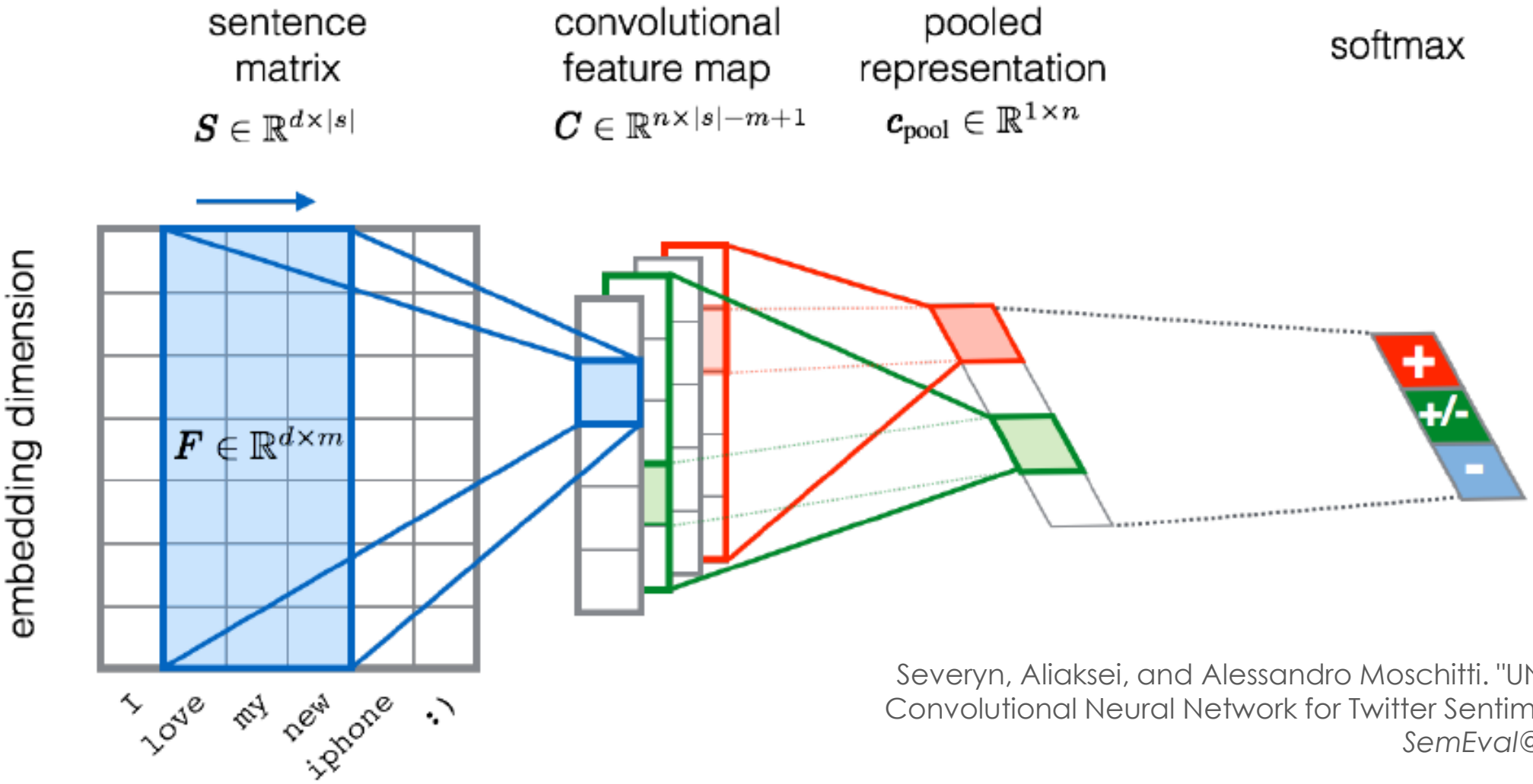
Clasificar

Modelo Clasificación Supervisada

Positivo

Negativo

Una alternativa: CNN para clasificación de texto



Severyn, Aliaksei, and Alessandro Moschitti. "UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification." *SemEval@ NAACL-HLT*. 2015.

Preguntas esenciales

¿Cuál es la dimensión del X de entrada si el texto es de largo variable?

¿Es factible/necesario diferenciar versiones de la misma palabra (p. ej: “Profesión”, “PROFESIÓN”, “Profesiones”, “Profesional”, …)?

¿Cómo se pueden tratar los errores ortográficos y gramáticos? (Ej: “Profesion”, “Profecion”)

Se necesita ...
Pre-Procesamiento de Texto No-Estructurado

Necesidad 1: reducción de diversidad lingüística

Se sabe que en lenguaje natural hay diferentes posibles verbalizaciones para decir lo mismo. El analizar texto en forma sintáctica estricta no permite reconocer la equivalencia entre estas posibles alternativas, por lo que se hace necesario encontrar formas "generales" del texto.

El pre-procesamiento de texto se centra en diferentes técnicas de "normalización". En el fondo, se busca generalizar las expresiones, de modo de poder compararlas con más facilidad.

TEXTO 1

Dentro de las características relevantes, está el "Cálculo de Indicador de Riesgo por Cliente", que se basa en una fórmula que combina una serie de datos de los antecedentes personales.

TEXTO 2

La característica más relevante es el "Cálculo de Indicador de Riesgo por Cliente", basado de acuerdo a una fórmula que combina antecedentes de las personas.

TEXTO COMÚN

caracteristic relevant calcul indic riesg client bas formul combin (serie) (dat) antecedent person

Normalización de texto (reducción diversidad)

Tokenización

- Grandes porciones de texto en segmentos más cortos (segmentación: párrafos, frases; tokenización: palabras), según la necesidad
- Más que una operación mecánica, requiere incorporar cierta lógica propia del lenguaje: qué palabras se podrían separar, etc.

Simplificación

- Todo a minúsculas.
- Eliminación de tildes y ~ (si es factible, dado el contexto).
- Eliminación signos puntuación.
- Eliminación de números o dígitos.
- Eliminación palabras innecesarias (stop words – dependiendo del contexto)

Semántica básica

- Stemming: base de la palabra. Ej: corriendo → corre
- Lemmatization: similar, unificando conceptos similares en una única palabra, conservando la idea de la palabra original. Ej. Mejor → Bueno

Referencia: https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html

Ejemplo de Normalización

"Este es un TEXTO de ejemplo, que sirve para ver qué términos quedan después de la limpieza de palabras, símbolos, y otros elementos."

[text ejempl sirv ve termin qued despu limpiez palabr simbol otr element]

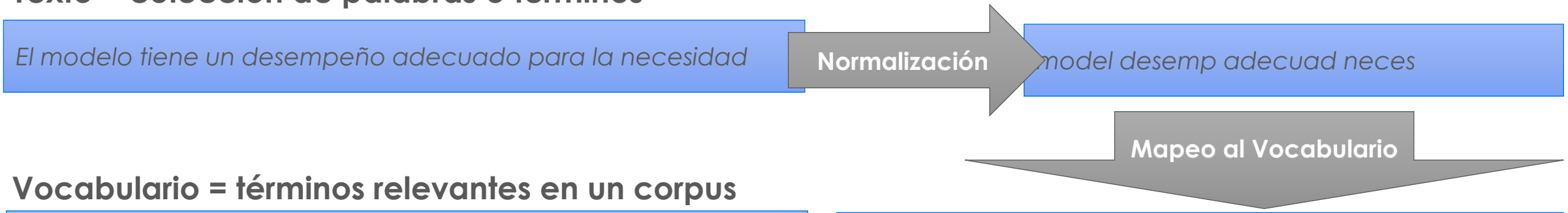
Necesidad 2: Uso de un Vocabulario para Clasificación

Todos los modelos de clasificación reciben un X de dimensión fija, pero los textos son de largo variable. Por ello, se vuelve necesario cambiar la representación de los vectores ("documentos" = porciones de texto) que reciben los modelos.

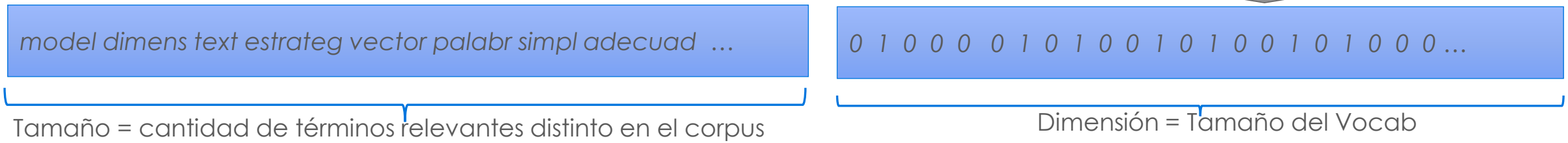
La mejor estrategia es convertir los textos en vectores que referencien palabras existentes en un Vocabulario.

Luego, cada nuevo texto a analizar se mapea a la ocurrencia de palabras en el vocabulario, simplificando su representación general.

Texto = colección de palabras o términos



Vocabulario = términos relevantes en un corpus



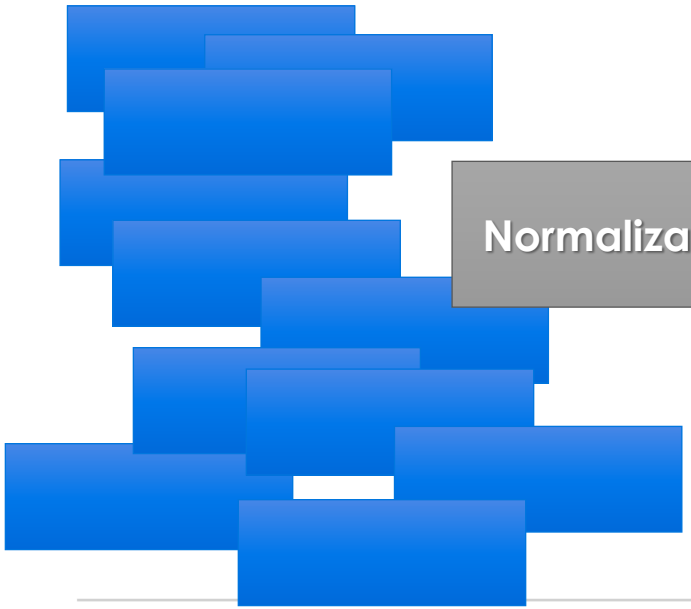
Tamaño = cantidad de términos relevantes distinto en el corpus

Dimensión = Tamaño del Vocab

Construcción del Vocabulario (parte 1)

Para construir un vocabulario se necesita un "corpus" o colección de "documentos" (ejemplos de texto del lenguaje de referencia). Luego de normalizar y armar la liste de términos diferentes, se determina la ocurrencia de éstos en los documentos de origen (Document Term Matrix).

1 Corpus de referencia



Normalización

2 Normalizar texto → Palabras únicas

P1	P2	P3	P4	P5
P3	P5	P8	P9	P3
P4	P7	P9	P12	P13
…				

Calcular Frecuencia de c/palabra en c/texto

3 Generar una DocTermMatrix

	P1	P2	P3
Txt1	1	0	1
Txt2	0	0	2
Txt3	1	2	0
Txt4	0	0	1
Txt5	0	1	0
Txt…	0	0	0

Construcción del Vocabulario (parte 2)

Contando con la DocTermMatrix, que típicamente tiene una alta dimensionalidad (cantidad de términos diferentes en el corpus), se puede **reducir su dimensionalidad** al eliminar términos según su nivel de sparse ("escaso"). Se eliminan aquellos que tengan al menos un porcentaje "sparse" de elementos vacíos (muchos 0s en la columna).

DocTermMatrix

	P1	P2	P3	P4	P5	P6	P7	…
Txt1	1	0	1	0	0	1	0	
Txt2	0	0	2	0	1	0	0	
Txt3	1	2	0	0	0	0	0	
Txt4	0	0	1	0	0	0	0	
Txt5	0	1	0	0	0	0	2	
Txt..	0	0	0	0	0	0	0	

Dimensión = Tamaño original del Vocab

4

Se eliminan términos con un factor "sparse" menor que un umbral.

DocTermMatrix
Dimensión reducida

	P1	P5	P6	…
Txt1	1	0	1	
Txt2	0	1	0	
Txt3	1	0	0	
Txt4	0	0	0	
Txt5	0	0	0	
Txt..	0	0	0	

Dimensión = Tamaño reducido del Vocab

5

Arma Vocab

Vocabulario Reducido

P1
P5
P6
P10
P17
P22
…

Algunas consideraciones

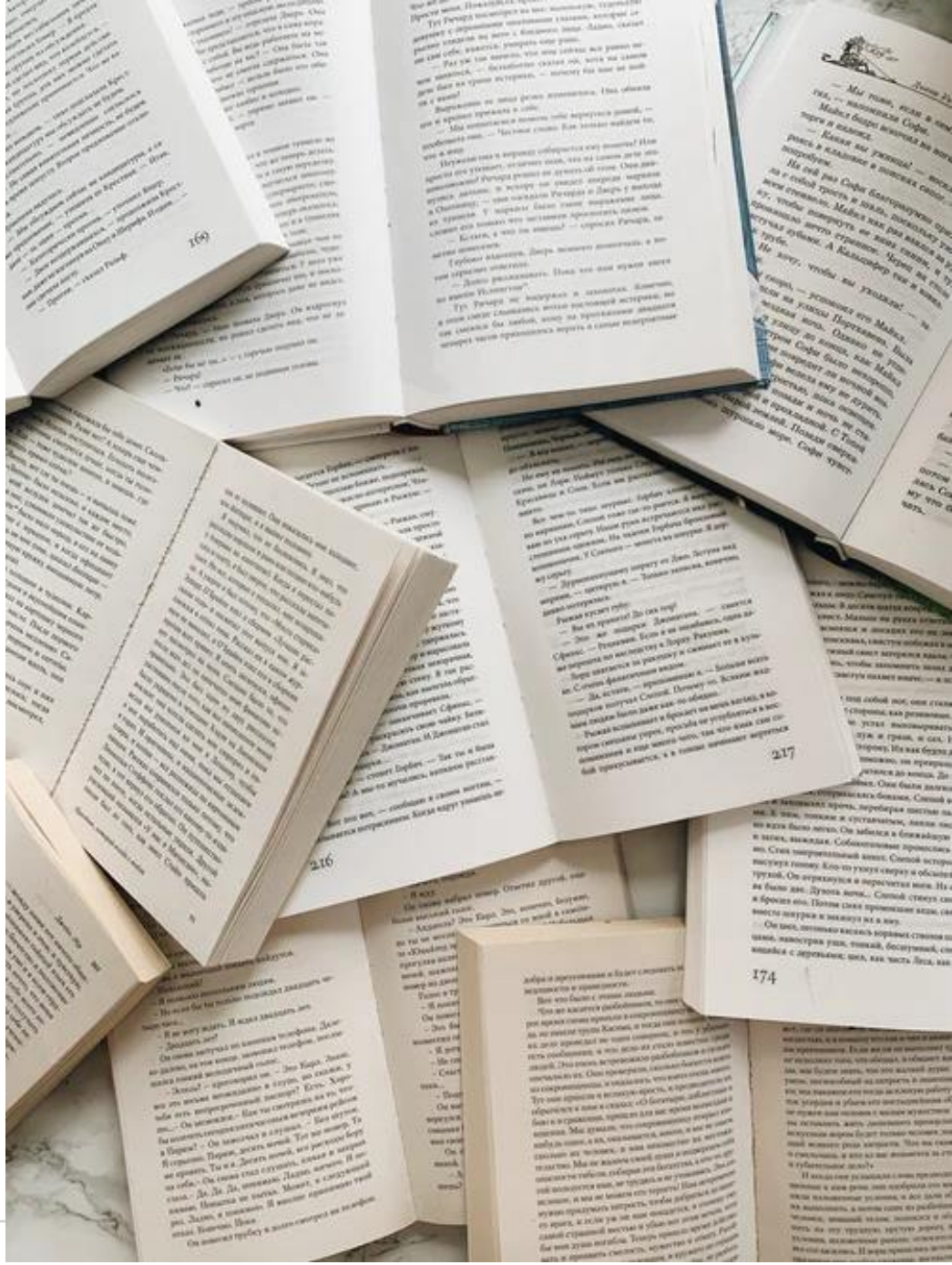
Diccionario ≠ Vocabulario

- Diccionario: lista de palabras de un lenguaje (definición de diseño)
- **Vocabulario**: lista de palabras encontradas en un texto (empírica)
- Por tanto, **Vocabulario** se usa en la práctica (que puede contener errores de tipeo y ortográficos, pero que son empíricos)

Representación vectorial mapeando el Vocabulario no considera:

- Orden de las palabras en el texto
- Cantidad ocurrencias de la palabra (aunque depende de la implementación)

Aplicando la Clasificación de Texto con Pre-procesamiento



Noticias y Comentarios

EDUCACIÓN PROFESIONAL

Noticias | Economía | Deportes | Espectáculos | Tendencias | Autos | Servi

Chile | Mundo | Tecnología | Educación | Documentos | Multimedia | Juegos | Especial Coronavirus | Pasatiempo

Buscar Comentarista | Buscar

PANDEMIA DE CORONAVIRUS

PORTADA | NOTICIAS | CASOS EN CHILE | CUARENTENAS | CASOS EN EL MUNDO | MIRADA GLOBAL

Presidente de la Cámara dice que hay "voluntad" para discutir este jueves proyecto de 10%

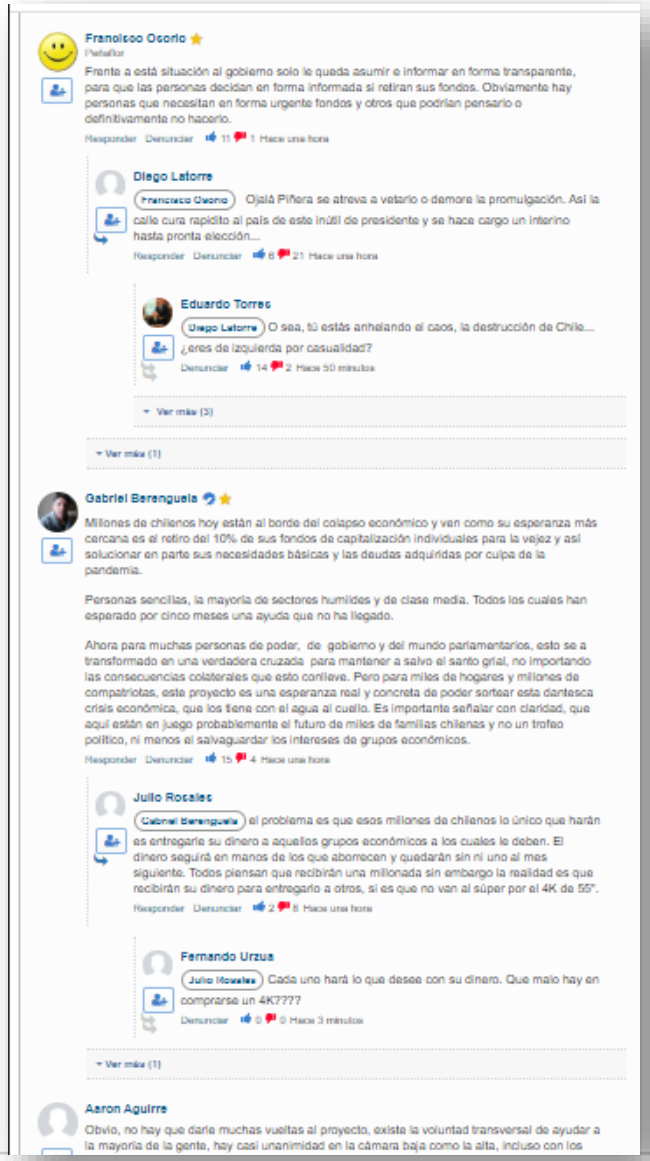
Asimismo el Senado adelantó la sesión de este miércoles y programó que a partir de las 13:00 horas se votará en general y en particular la iniciativa.

21 de Julio de 2020 | 17:54 | Por Verónica Marín, Emol 190

El presidente de la Cámara anunció que hoy hay reunión de comité a las 19:00 horas donde se discutirá la tabla de los próximos días.

Aton

El presidente de la Cámara de Diputados, **Diego Paulsen (RN)**, aseguró que existe voluntad **para discutir este jueves el proyecto que permite el retiro del 10%** de los fondos de pensiones en el contexto de la pandemia, esto considerando que mañana será votado en el Senado y pasaría a un tercer trámite.



Francisco Osorio

Frente a está situación al gobierno solo le queda asumir e informar en forma transparente, para que las personas decidan en forma informada si retiran sus fondos. Obviamente hay personas que necesitan en forma urgente fondos y otros que podrían pensarlo o definitivamente no hacerlo.

Responder Denunciar 11 1 Hace una hora

Diego Latorre

Francisco Osorio Ojalá Piñera se atreva a vetarlo o demore la promulgación. Así la calle cura rapidito al país de este inútil de presidente y se hace cargo un interino hasta pronta elección....

Responder Denunciar 6 21 Hace una hora

Eduardo Torres

Diego Latorre O sea, tú estás anhelando el caos, la destrucción de Chile.... ¿eres de izquierda por casualidad?

Denunciar 14 2 Hace 50 minutos

Ver más (3)

Ver más (1)

Gabriel Berenguela

Millones de chilenos hoy están al borde del colapso económico y ven como su esperanza más cercana es el retiro del 10% de sus fondos de capitalización individuales para la vejez y así solucionar en parte sus necesidades básicas y las deudas adquiridas por culpa de la pandemia.

Personas sencillas, la mayoría de sectores humildes y de clase media. Todos los cuales han esperado por cinco meses una ayuda que no ha llegado.

Ahora para muchas personas de poder, de gobierno y del mundo parlamentarios, esto se a transformado en una verdadera cruzada para mantener a salvo el santo grial, no importando las consecuencias colaterales que esto conlleva. Pero para miles de hogares y millones de compatriotas, este proyecto es una esperanza real y concreta de poder sortear esta dantesca crisis económica, que los tiene con el agua al cuello. Es importante señalar con claridad, que aquí están en juego probablemente el futuro de miles de familias chilenas y no un trofeo político, ni menos el salvaguardar los intereses de grupos económicos.

Responder Denunciar 15 4 Hace una hora

Julio Rosales

Gabriel Berenguela el problema es que esos millones de chilenos lo único que harán es entregarle su dinero a aquellos grupos económicos a los cuales le deben. El dinero seguirá en manos de los que aborrecen y quedarán sin ni uno al mes siguiente. Todos piensan que recibirán una millonada sin embargo la realidad es que recibirán su dinero para entregarlo a otros, si es que no van al súper por el 4K de 55".

Responder Denunciar 2 8 Hace una hora

Fernando Urzua

Julio Rosales Cada uno hará lo que desee con su dinero. Que malo hay en comprarse un 4K????

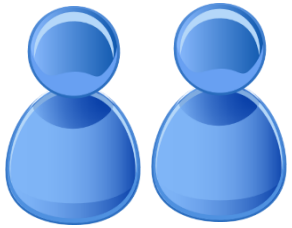
Denunciar 0 0 Hace 3 minutos

Ver más (1)

Aaron Aguirre

Obvio, no hay que darle muchas vueltas al proyecto, existe la voluntad transversal de ayudar a la mayoría de la gente, hay casi unanimidad en la cámara baja como la alta, incluso con los

Líderes Políticos



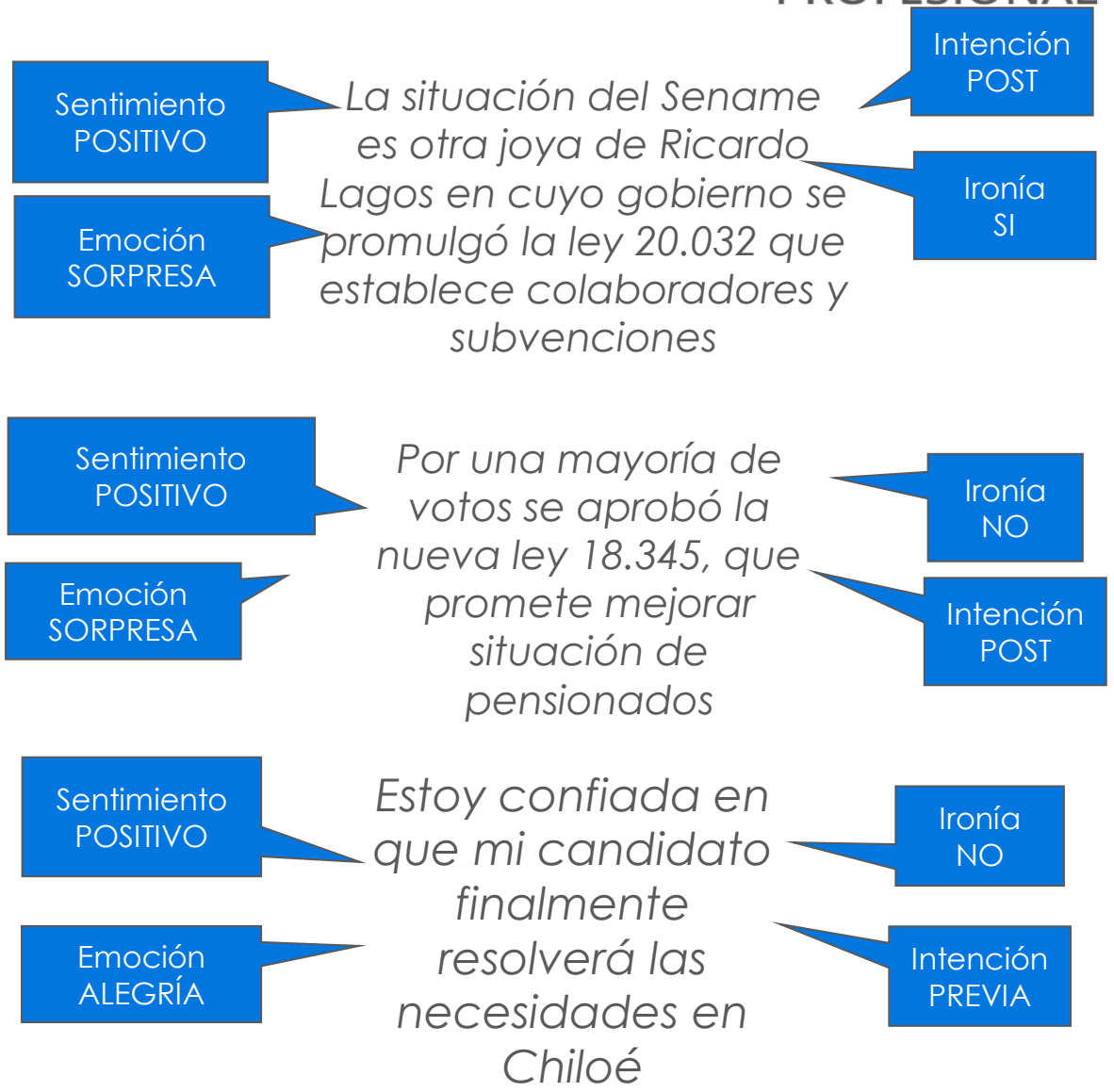
Asesores comunicacionales

www.educacionprofesional.ing.uc.cl

Clasificación multi-dimensión

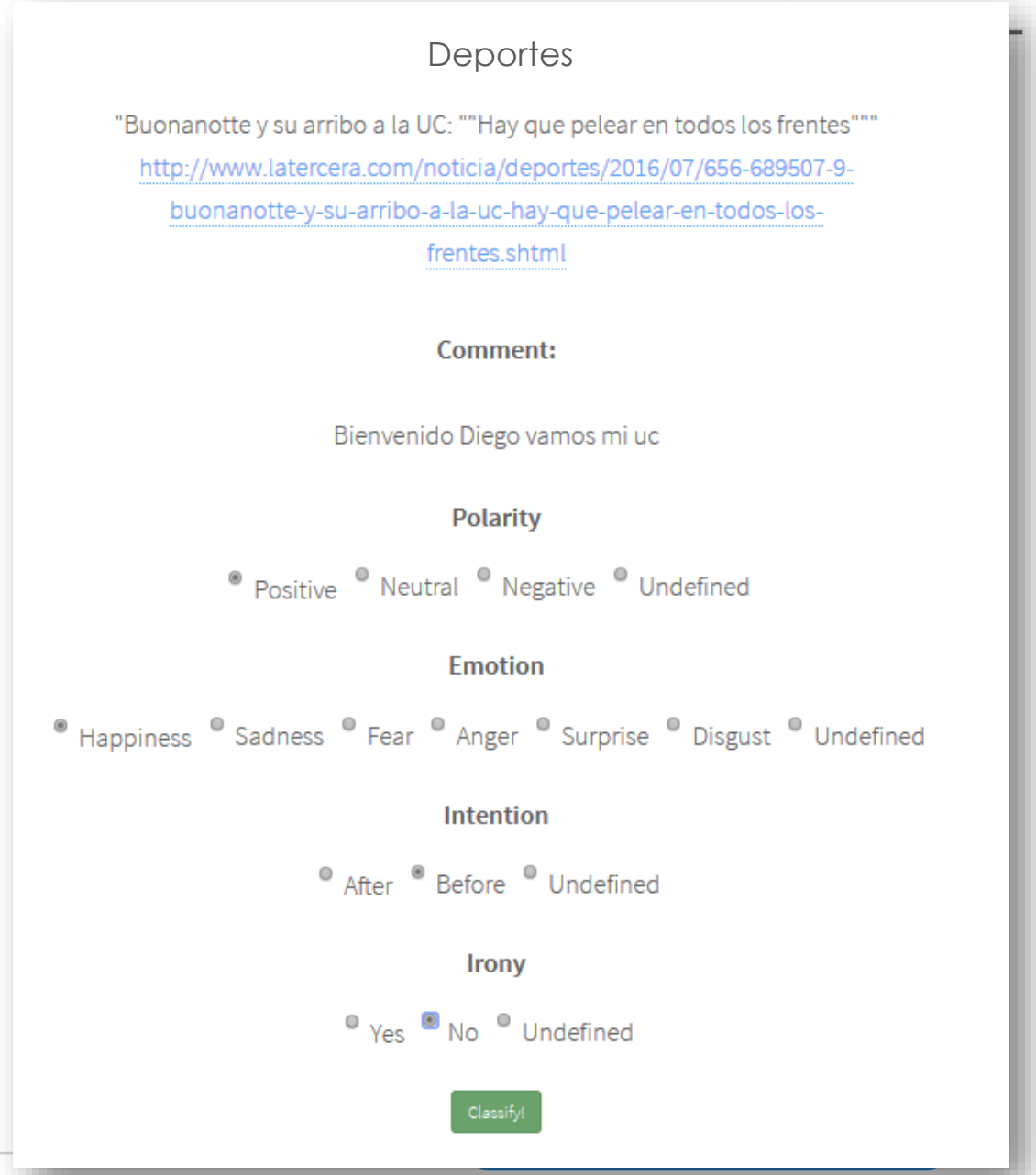
Dimensiones:

- **Sentimiento**: positivo, negativo, neutro. Se considera una opción de no mostrar neutros → en política.
- **Emoción**: enojo, sorpresa, alegría, etc.
- **Intención**: previo, posterior (por ej, diferenciar una intención de compra de una evaluación de compra).
- **Ironía**: es o no es.
  En este caso, se pretende usar la ironía de un comentario como "negador" del sentimiento y emoción.



Entrenamiento Contextual

- Los humanos que clasifican manualmente comentarios de referencia, tienen disponible un contexto y potencialmente links a las noticias de origen.
- Esta clasificación contextual permite mejorar sustancialmente la calidad de clasificación.
- Áreas contextuales: Política-Economía, Artista, Deportes, Empresa Servicios, Empresa Producto.



Resultados Procesamiento Mensajería Clientes

ANTES (MANUAL)	DESPUÉS (AUTOMÁTICO)
CAPACIDAD Un ejecutivo de centro de atención puede revisar hasta 30 mensajes por hora	**CAPACIDAD** El robot revisa 60 mensajes por minuto
PRECISIÓN **85% en ejecutivos humanos** Tasa de error de 15%	**PRECISIÓN** **92% en robot** Tasa de error de 8% en el robot
ESTANDARIZACIÓN Ciertos criterios dependen de la persona	**ESTANDARIZACIÓN** Criterios dependen sólo de los ejemplos de entrenamiento

Recibí la chaqueta y me pareció de una excelente calidad y terminaciones

Área:	Post-venta
Tópico:	Felicitaciones
Producto:	Chaqueta
Sentimiento:	Positivo
Emoción:	Alegría
Intención:	Post-evento

Muchos de estas dimensiones de interpretación siguen siendo modelos de Machine Learning (clasificación supervisada) y otros son NER.

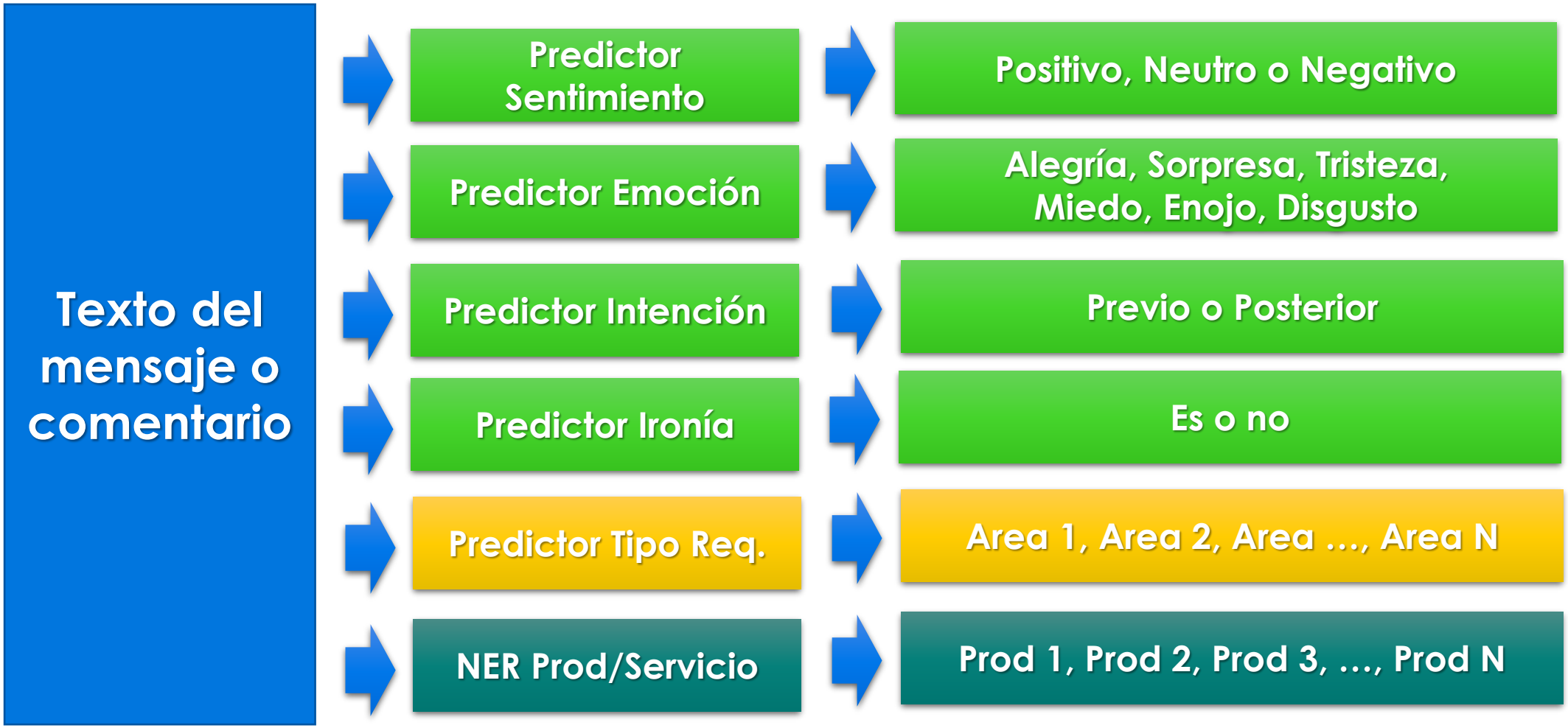
Tengo serios problemas con el servicio de Internet. Se corta repetidamente, a veces por varios minutos. He llamado y no contestan.

Área:	Soporte técni
Tópico:	Reclamo
Producto:	Internet
Sentimiento:	Negativo
Emoción:	Enojo
	Post-evento

URGENCIA

Lo interesante es incorporar aspectos "humanos", como es el caso de emociones e intenciones, dentro de los elementos a interpretar.

Clasificadores de texto



Caso atención servicios telefónicos

Análisis de Mensaje de Cliente

> *Estimados. Solicito bloquear el envìo de SMS para las lìneas adjuntadas en el excel. Saludos Cordiales*

La petición del cliente es de u

Análisis de Mensaje de Cliente

> *Estimados buenas tardes. Junto con Saludar, agradecería gestionar Eliminación de Plan de Datos 9 Gb (Plan w89) a Línea N°989899998. Muchas Gracias.*

de un **plan de datos**

Análisis de Mensaje de Cliente

> *por favor revisar las lineas 99999909999 y 99999989909, tienen problemas con la recepcion de mensajes. se llamó a servicio tecnico y lo que indican es lo siguiente No tiene recepción de mensajería entrante solo voz*

Este es claramente un **reclamo** respecto a una **línea**. Nótese que se siente disgusto en el mensaje.

Caso: clientes de banco

Análisis de Mensaje de Cliente

> *Hola Adrián, no puedo hacer transferencias, se queda pegado en la llamada de la clave dinámica.*

La petición del cliente es de una **consulta sobre pro**
productos y servicios, subcódigo **6**, que pide que oc

Solicitud: **Consulta sobre Productos y Servicios / ot**
Subcódigo: **6**
Emoción: **sorpresa**
Intención: **que pide que ocurra**

Análisis de Mensaje de Cliente

> *El 1 de marzo compre el SOAP con mi tarjeta de credito mastercard y a la fecha aun no recibo los creditos cabify ofrecidos por la compra.*

Ojo, que el cliente está presentando un **reclamo (problema con un producto)**
/ tarjeta de crédito mastercard respecto a un código **7**, que ya ocurrió.

Análisis de Mensaje de Cliente

> *Al revisar el día de hoy mi cuenta me percato que me han cobrado de forma automática el pago mínimo de la linea de crédito cargándomelo en la corriente, siendo que este fue pagado el 28 de febrero. Poseo el comprobante de pago realizado el 28 de febrero, y también el comprobante del cobro que me han realizado el día de hoy.*

Ojo, que el cliente está presentando un **reclamo (problema con un producto)**
/ cuenta corriente y servicios respecto a un código **1**, que pide que ocurra.
Nótese que se siente disgusto en el mensaje.

Solicitud: **Reclamo (problema con un producto) / Cuenta Corriente y Servicios**
Subcódigo: **1**
Emoción: **disgusto**
Intención: **que pide que ocurra**

B. Extracción de Información
desde Texto No-Estructurado

NER – Named-Entity Recognition

Técnica para la extracción de entidades desde un texto en lenguaje natural.

No se basa en reglas predefinidas (por ej, entregar una lista de nombres de personas para encontrarlos), sino que se "entrena" con ejemplos para reconocer Nombres, Lugares, Fechas, Montos, etc.

Poder Especial

Por el presente documento, en Santiago a 14 de abril del año 2018, declara don **Juan Pérez Soto**, en representación de la empresa ACME Ltda., para declarar que se autoriza a don **Miguel José Soto González** para realizar, en nombre de la empresa, todo tipo de trámites ante el servicio de impuestos internos, incluyendo, pero no limitándose a …

Se busca encontrar, en el texto extenso, los nombres de personas mencionadas.
NER permite reconocer estas porciones de texto, sin tener una lista de nombres previamente definida, sino que en base a ejemplos

Desafíos de la ambigüedad

2010: ¿año o cantidad?

Washington: ¿persona o lugar?

Julio: ¿mes o persona?

Etc.

En base a ejemplos, no a lista de alternativas

- Ejemplos de dónde aparece un nombre (NNN AAA):
  - … se confiere el poder a don NNN AAA a partir de…
  - … cuyo representante es NNN AAA, quien acepta …
  - … comparece NNN AAA, declarando …
- Entonces, no se necesita una lista de nombres; basta con ver las palabras alrededor.
- Esto permite enfrentar desafíos como el siguiente:
  - … confiere poder a don ARMANDO CABEZAS DEL RIO a …

NER - Named-Entity Recognition

Nombre de Persona

Nombre de Organización

Elon Musk se convirtió en el mayor accionista de **Twitter** al comprar una participación del 9,2% el pasado **4 de abril**. No obstante, **Parag Agrawal**, el CEO de la aplicación, ha anunciado que el empresario ha decidido no formar parte de la junta directiva, aseguró desde la oficina central, en **San Francisco, California**.

Fecha

Nombre de Ubicación

- Por medio de ejemplos, es capaz de reconocer palabras que representan diferentes tipos: fechas, números o montos, nombres de personas, nombres de instituciones, lugares, …
- Se entrena en forma similar a modelos de ML.
  - Internamente hay, primero, un clasificador, que toma un término o conjunto de palabras y determina si es de un tipo, de otro, o ninguno (clasificador supervisado).

NER - ¿Cómo funciona?

NER Involucra la identificación de nombres y palabras propias en texto y su clasificación en una categoría, de un set predeterminado de alternativas

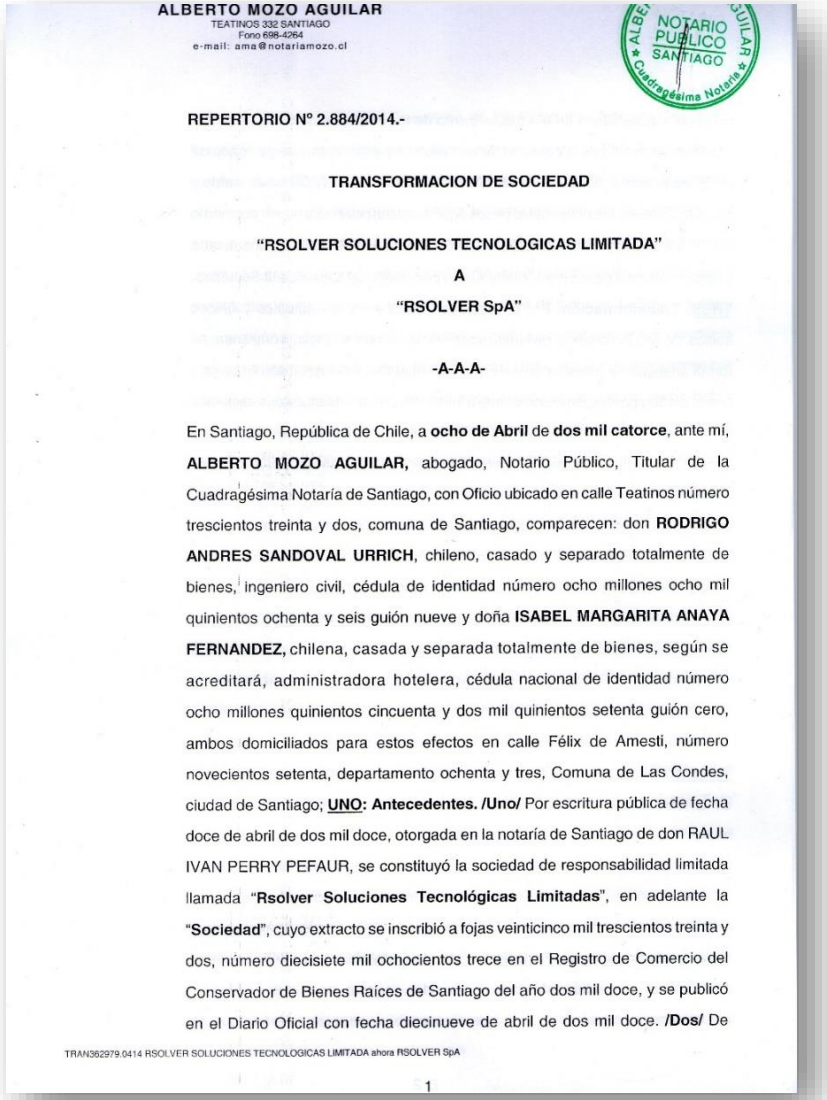
El caso particular de nombres ("Named-Entity") considera:

- Nombre de Persona
- Nombre de Organización
- Nombre de Lugar
- Nombre de Marca

Otros (Entity Recognition o Entity Identification) considera otros elementos:

- Fecha, monto, cantidad, …

Extendiendo y Complementando el Modelo



ALBERTO MOZO AGUILAR
TEATINOS 332 SANTIAGO
Fono 698-4264
e-mail: ama@notariamozo.cl

NOTARIO PÚBLICO SANTIAGO
Cuadragésima Notaría

REPERTORIO Nº 2.884/2014.-

TRANSFORMACION DE SOCIEDAD

"RSOLVER SOLUCIONES TECNOLOGICAS LIMITADA"
A
"RSOLVER SpA"

-A-A-A-

En Santiago, República de Chile, a **ocho de Abril** de **dos mil catorce**, ante mí, **ALBERTO MOZO AGUILAR,** abogado, Notario Público, Titular de la Cuadragésima Notaría de Santiago, con Oficio ubicado en calle Teatinos número trescientos treinta y dos, comuna de Santiago, comparecen: don **RODRIGO ANDRES SANDOVAL URRICH**, chileno, casado y separado totalmente de bienes, ingeniero civil, cédula de identidad número ocho millones ocho mil quinientos ochenta y seis guión nueve y doña **ISABEL MARGARITA ANAYA FERNANDEZ,** chilena, casada y separada totalmente de bienes, según se acreditará, administradora hotelera, cédula nacional de identidad número ocho millones quinientos cincuenta y dos mil quinientos setenta guión cero, ambos domiciliados para estos efectos en calle Félix de Amesti, número novecientos setenta, departamento ochenta y tres, Comuna de Las Condes, ciudad de Santiago; **UNO: Antecedentes. /Uno/** Por escritura pública de fecha doce de abril de dos mil doce, otorgada en la notaría de Santiago de don RAUL IVAN PERRY PEFAUR, se constituyó la sociedad de responsabilidad limitada llamada "**Rsolver Soluciones Tecnológicas Limitadas**", en adelante la "**Sociedad**", cuyo extracto se inscribió a fojas veinticinco mil trescientos treinta y dos, número diecisiete mil ochocientos trece en el Registro de Comercio del Conservador de Bienes Raíces de Santiago del año dos mil doce, y se publicó en el Diario Oficial con fecha diecinueve de abril de dos mil doce. **/Dos/** De

TRAN362979.0414 RSOLVER SOLUCIONES TECNOLOGICAS LIMITADA ahora RSOLVER SpA

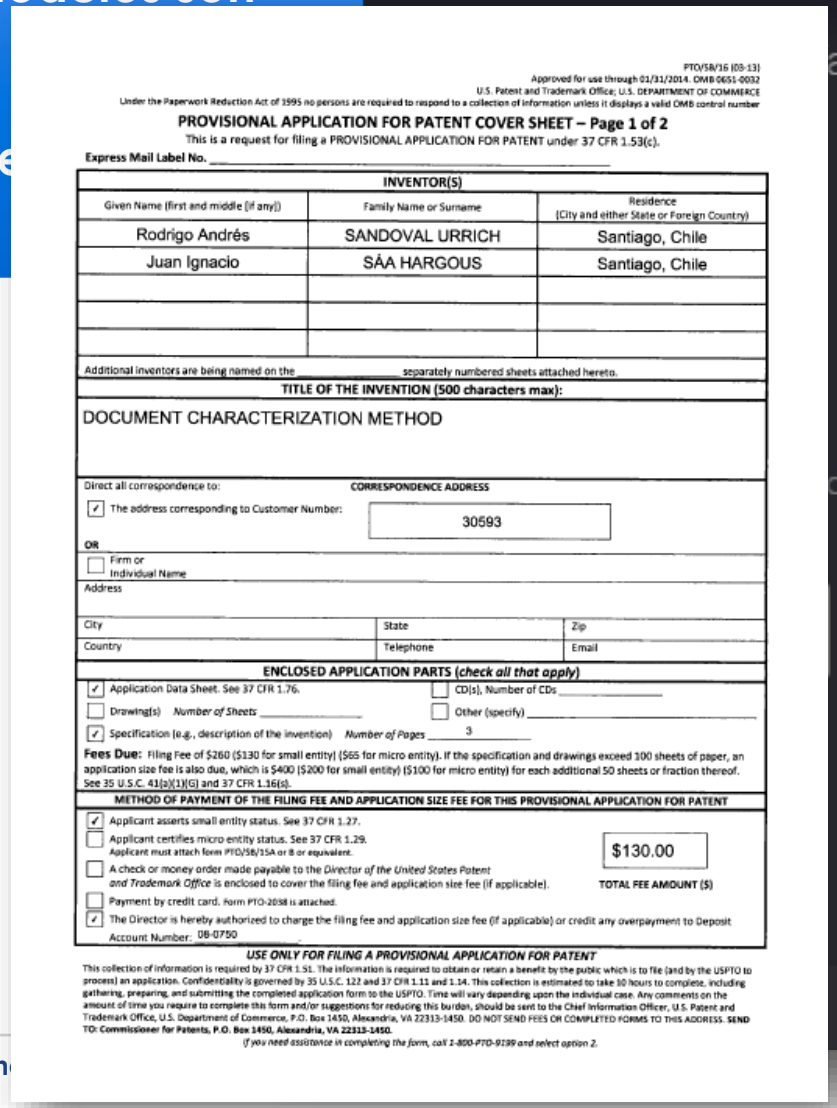
1

- Datos o Caract… …mplementarias
- ¿Cuál es la fech… …documento?
- ¿Cuáles son las partes nom…
- Si es un contrato de arriendo, ¿cuán… …cuál es el valor de arriendo definido?
- Etc.

Caracterización Documental

Patente

Métodos y modelos son patentables, software o el fuente en que escritos.



PTO/SB/16 (03-13)
Approved for use through 01/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

PROVISIONAL APPLICATION FOR PATENT COVER SHEET – Page 1 of 2

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. _____

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Rodrigo Andrés	SANDOVAL URRICH	Santiago, Chile
Juan Ignacio	SÁA HARGOUS	Santiago, Chile

Additional inventors are being named on the _____ separately numbered sheets attached hereto.

TITLE OF THE INVENTION (500 characters max):

DOCUMENT CHARACTERIZATION METHOD

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☑ The address corresponding to Customer Number: 30593

OR

☐ Firm or Individual Name

Address

City	State	Zip
Country	Telephone	Email

ENCLOSED APPLICATION PARTS (*check all that apply*)

☑ Application Data Sheet. See 37 CFR 1.76. ☐ CD(s), Number of CDs _____

☐ Drawing(s) *Number of Sheets* _____ ☐ Other (specify) _____

☑ Specification (e.g., description of the invention) *Number of Pages* 3

Fees Due: Filing Fee of $260 ($130 for small entity) ($65 for micro entity). If the specification and drawings exceed 100 sheets of paper, an application size fee is also due, which is $400 ($200 for small entity) ($100 for micro entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).

METHOD OF PAYMENT OF THE FILING FEE AND APPLICATION SIZE FEE FOR THIS PROVISIONAL APPLICATION FOR PATENT

☑ Applicant asserts small entity status. See 37 CFR 1.27.

☐ Applicant certifies micro entity status. See 37 CFR 1.29. Applicant must attach form PTO/SB/15A or B or equivalent.

☐ A check or money order made payable to the *Director of the United States Patent and Trademark Office* is enclosed to cover the filing fee and application size fee (if applicable).

$130.00

TOTAL FEE AMOUNT ($)

☐ Payment by credit card. Form PTO-2038 is attached.

☑ The Director is hereby authorized to charge the filing fee and application size fee (if applicable) or credit any overpayment to Deposit Account Number: 08-0750

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

This collection of information is required by 37 CFR 1.51. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 10 hours to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

C. Comparación (simple)
de Texto No-Estructurado

Comparación (simple) según mapeo a vocabulario

TEXTO 1
Dentro de las características relevantes, está el "Cálculo de Indicador de Riesgo por Cliente", que se basa en una fórmula que combina una serie de datos de los antecedentes personales.

Mapeo al Vocabulario

0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 0 1 0 1 0 0 0 ...

TEXTO 2
La característica más relevante es el "Cálculo de Indicador de Riesgo por Cliente", basado de acuerdo a una fórmula que combina antecedentes de las personas.

Mapeo al Vocabulario

0 1 0 0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 1 0 ...

Comparación de conjuntos indexados

Coincidencia del 85%
→ Suficientemente iguales

Gracias

rsandova@ing.puc.cl
rodrigo@RSolver.com
@RSandovalSolver
/in/RodrigoSandoval
www.RodrigoSandoval.net
www.RSolver.com