I Seminario sobre el impacto de las tecnologías de Big Data en las Finanzas

Análisis predictivo con datos textuales

Agosto de 2021

Autor:

Gabriel Feldman

.UBA200

.UBA económicas
FACULTAD DE CIENCIAS ECONÓMICAS

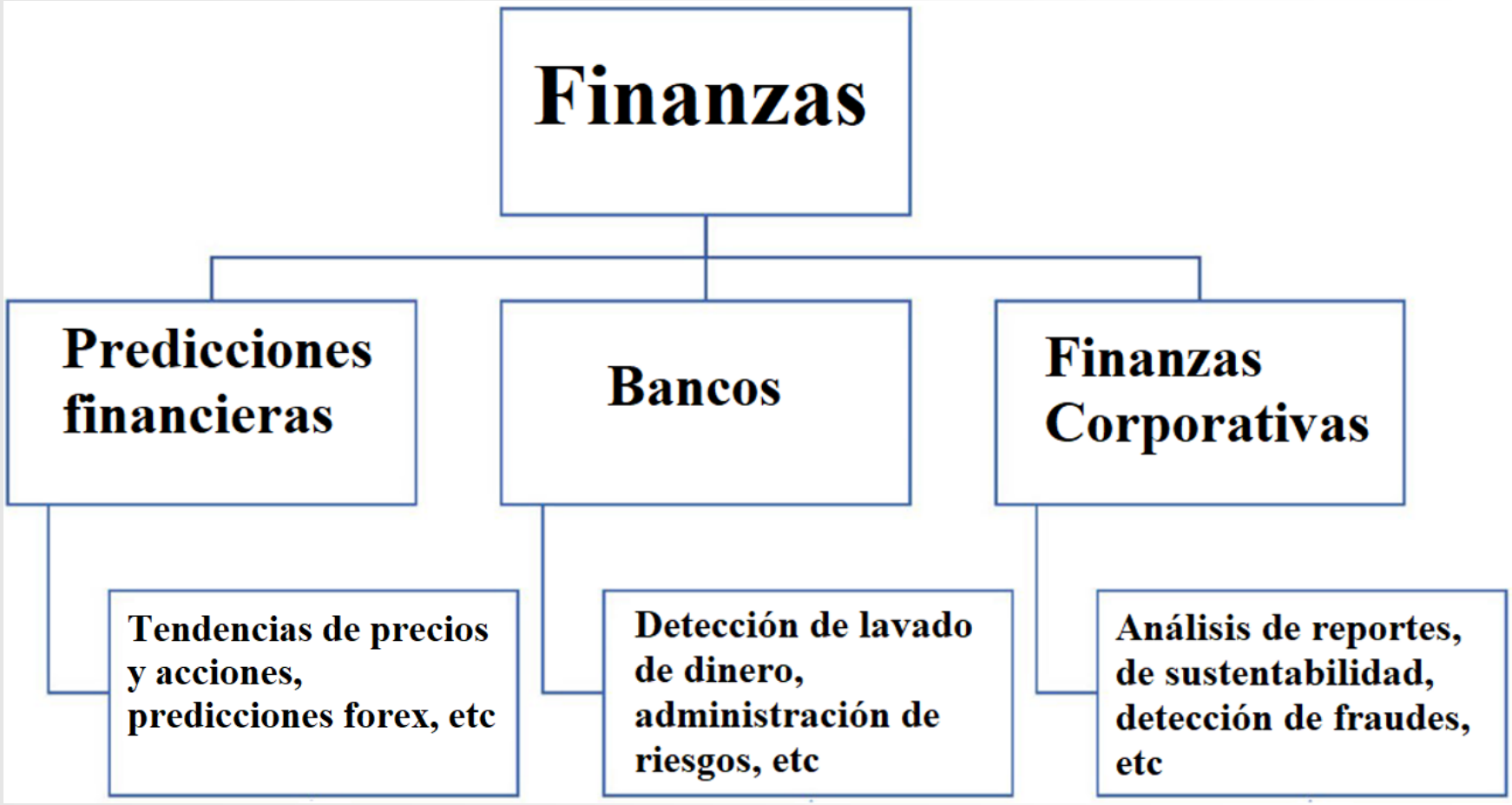
Contenido

- Los flujos de texto en las finanzas modernas
- Aplicaciones de minería de texto en finanzas
- Análisis temático: exploración y codificación
- Análisis de sentimiento: métodos disponibles
- Desafíos y futuros objetivos

Los flujos de texto en las finanzas modernas

- Big data e inteligencia empresarial
- Impacto en la eficiencia del mercado
- Minería de texto
- Fintech

Aplicaciones de minería de texto en finanzas



Aplicaciones de minería de texto en finanzas

Predicciones financieras	Bancos	Finanzas corporativas
Yadav, Sharan, y Vaish (2020): Demostraron la correlación entre los sentimientos de las noticias financieras y la variación del mercado de valores. Las finanzas conductuales modernas reconocen tanto a los inversores sentimentales como a los inversores racionales	**Gao y Ye (2007)** Propusieron un marco para prevenir el lavado de dinero con la ayuda de los historiales de transacciones de los clientes. Lo hicieron identificando datos sospechosos de varios informes textuales de las agencias de historial de datos.	**Guo et al. (2017)** Implementaron algoritmos de minería de texto. Fusionaron la base de datos de Thomson Reuters News y bases de datos de Noticias. La primera proporciona noticias originales y la segunda puntuaciones de sentimiento con puntuaciones positivas, negativas y neutrales
Carreño Giscafré (2020) Investigó si es que existe una relación entre las emociones plasmadas en comentarios de twitter, y las variaciones en las tendencias de precios de los valores. La extracción de textos se realizó mediante un método de web-scraping. La NLP utilizada en este estudio, es la librería del lenguaje de programación python llamada "Textblob". Propuso un índice de sentimientos de mercado en base a la clasificación de los comentarios.	**Bach et. al (2019)** Realizaron un análisis de sentimiento para analizar las opiniones de los clientes, lo cual es crucial para el funcionamiento de un banco. El análisis de redes sociales proporcionó una perspectiva sobre cómo los clientes están conectados en ellos y qué tan impactantes fueron al compartir información. Este análisis de redes sociales podría combinarse con la minería de texto para identificar las palabras clave que corresponden a el interés común de los clientes.	**Holton (2009)** Implementó un modelo de prevención del fraude financiero empresarial. Consideró la insatisfacción de los empleados como un indicador oculto responsable del fraude. Utilizó un conjunto de datos de mensajes de comunicación dentro de la empresa y correos electrónicos en grupos de discusión en línea, para proponer un modelo para la evaluación del riesgo de fraude en organizaciones.

Proceso de análisis de datos textuales

- Importar y estructurar los datos

- Sintetizar/explorar: identificar los temas
Sintetizar la información (nube de palabras, clasificación)
Explorar (verbatim, contexto)

- Codificar: captar la frecuencia de los temas
Manualmente (codificación)
Automáticamente (por diccionarios)

- Captar la orientación: para qué sirve, métodos.
Manual
Automático (cognitivo, machine learning)

- Indicadores y análisis de la información

De qué datos hablamos

Ejemplo de aplicación

“Realmente es muy difícil tener sustentabilidad en un país como el nuestro”
Fabricante – Dueño – Menos de 10

“Es necesaria la promoción de lineamientos básicos de sustentabilidad en las pymes que no cuentan con acceso a programas específicos en este sentido.”
Comerciante – Dueño – Menos de 10

“En fábricas chicas es muy difícil invertir en estos momentos.”
Fabricante – Gerente – Menos de 10

“Considero importante que todo estudio tenga la posibilidad de ser viable para implementarlo en la mayoría de los lugares e independiente del tamaño de empresa o comercio.”
Comerciante – Dueño – Menos de 10

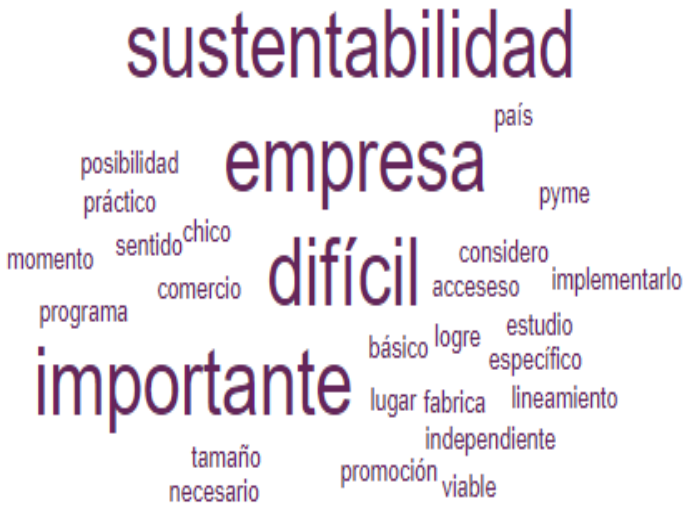
“Considero muy importante se logre implementar”
Fabricante – Dueño – 10 a 20

“Apoyar a las empresas para llevar esto a la práctica”
Comerciante – Dueño – 10 a 20

CONTEXTO			PREGUNTA ABIERTA	Temas
Actividad	**Rol**	**Planta**	**Comentario**	**¿?**
Fabricante	Dueño	Menos de 10	Realmente es muy difícil tener sustentabilidad en un país como el nuestro	**¿?**
Fabricante	Dueño	10 a 20	Considero muy importante se logre implementar	**¿?**
Comerciante	Dueño	10 a 20	Apoyar a las empresas para llevar esto a la práctica	**¿?**

Sintetizar y explorar

Lo que usted quiera agregar...



Buscar...

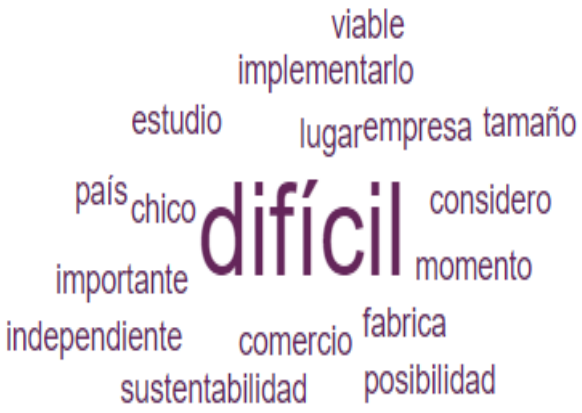
> Realmente es muy difícil tener sustentabilidad en un país como el nuestro
>
> Dueño - Fabricante - Menos de 10

> Es necesaria la promoción de lineamientos básicos de sustentabilidad en las pymes que no cuentan con acceseso a programas específicos en este sentido.
>
> Dueño - Comerciante - 20 a 50

> En fabricas chicas es muy difícil invertir en estos momentos.

Filtrado para empresas con planta menor de 10

Lo que usted quiera agregar...



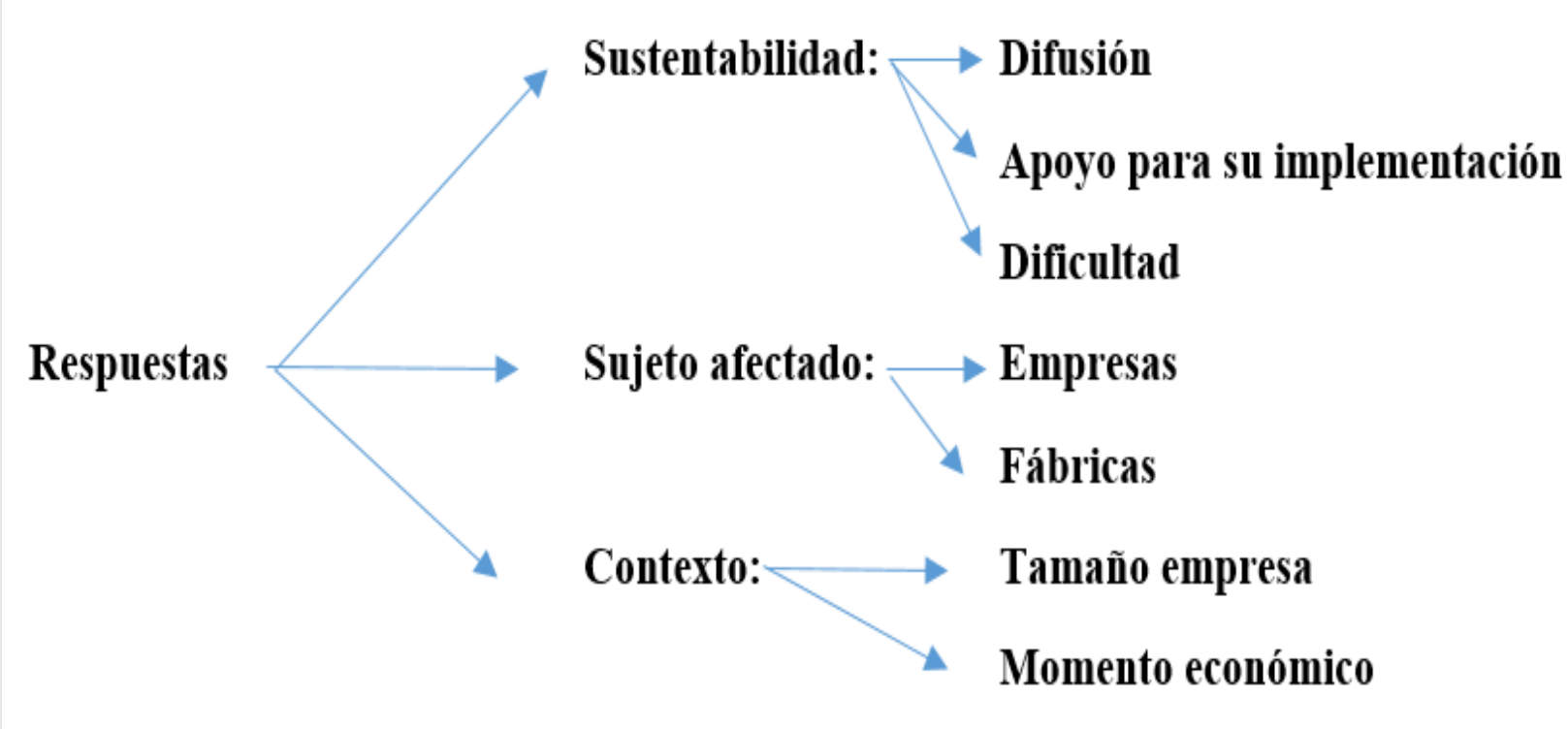
Buscar...

“Realmente es muy difícil tener sustentabilidad en un país como el nuestro”

“En fabricas chicas es muy difícil invertir en estos momentos.”

“Considero importante que todo estudio tenga la posibilidad de ser viable para implementarlo en la mayoria de los lugares e independiente del tamaño de empresa o comercio.”

Codebook

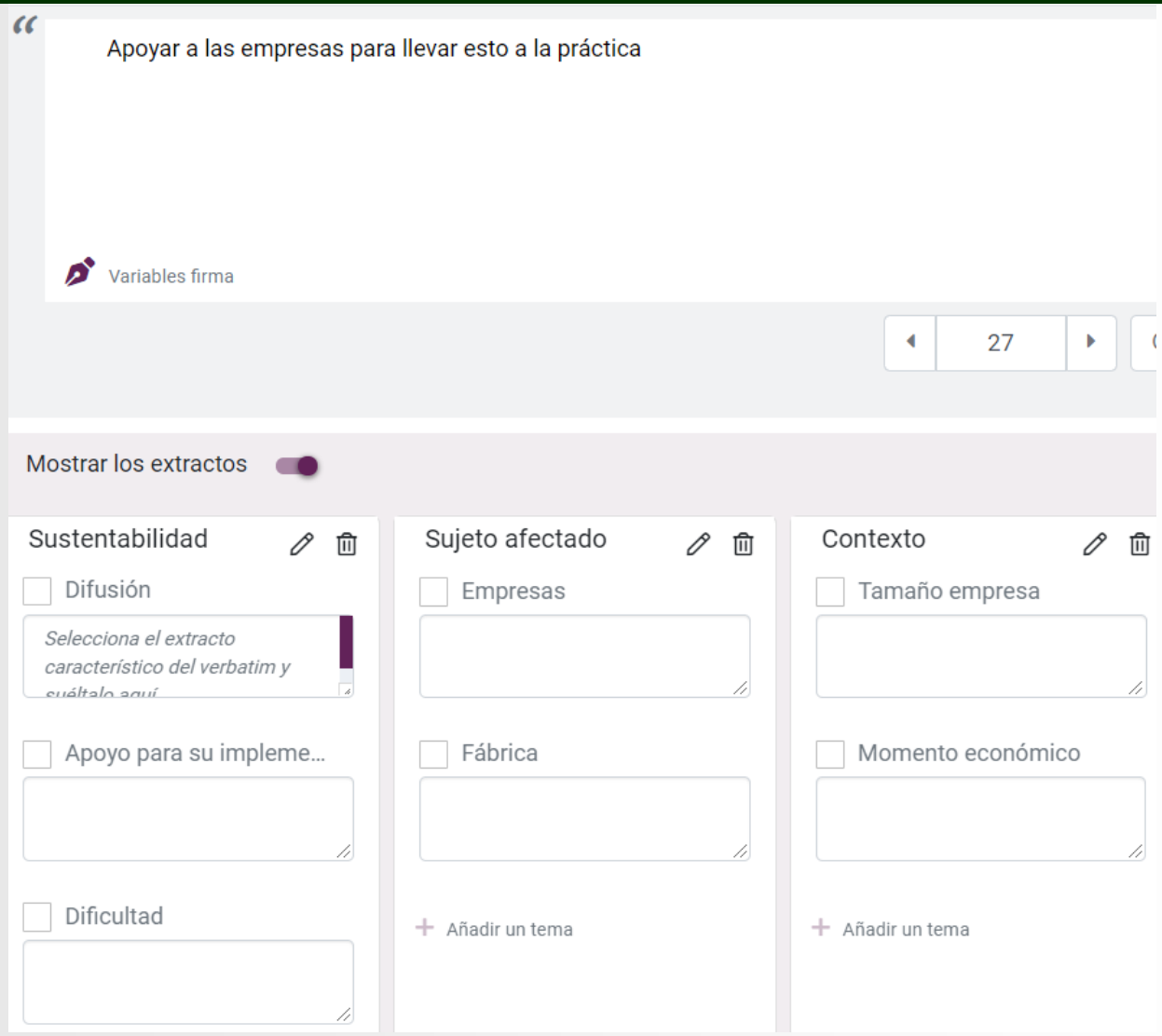

Respuestas
Sustentabilidad:
Difusión
Apoyo para su implementación
Dificultad
Sujeto afectado:
Empresas
Fábricas
Contexto:
Tamaño empresa
Momento económico

Codificar – Captar la frecuencia

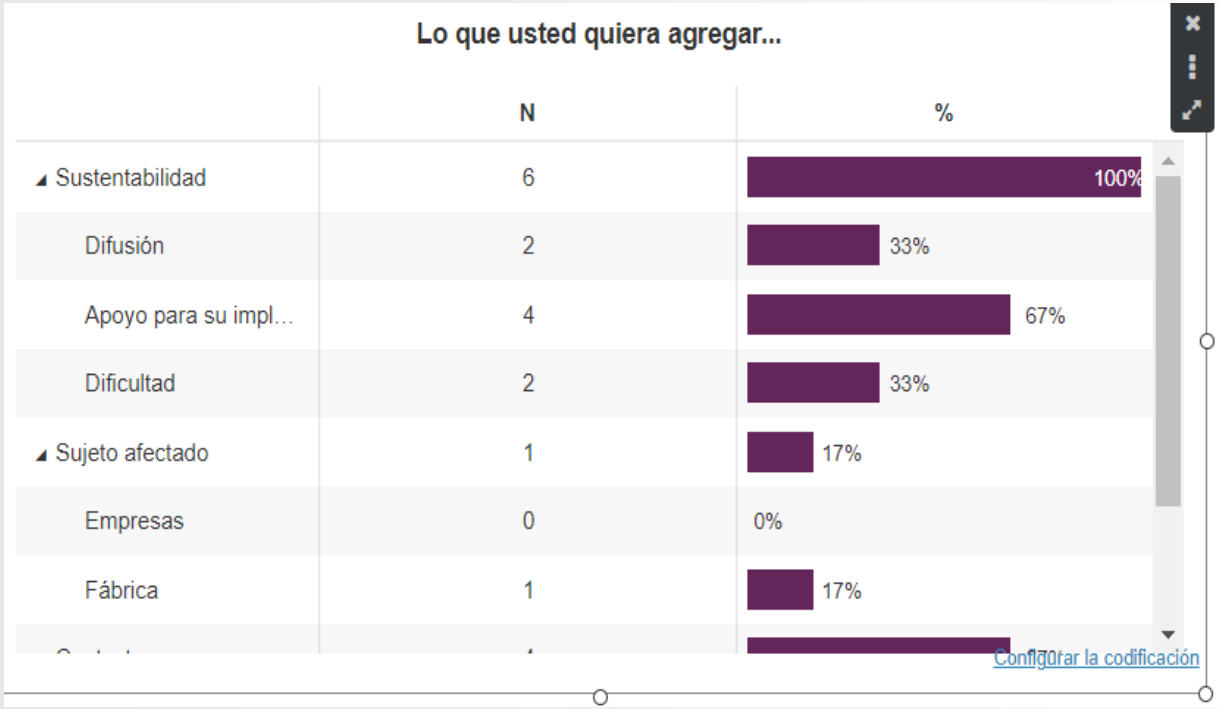
MANUAL	Vs.	AUTOMÁTICO
Pocos datos		Muchos datos
Datos definitivos		Datos provisorios
Informe puntual		Dashboard interactivo
Mucho tiempo disponible		Poco tiempo disponible

- Sé o no de qué habla la gente
- Cortas vs largas…ideas complejas o ironía

Codificación manual



Captada la frecuencia



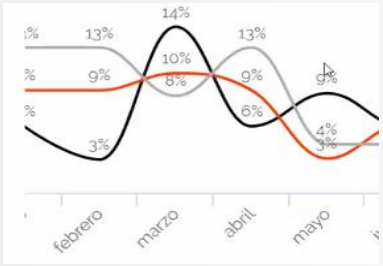
CODIFICACIÓN MANUAL

Codificación por diccionario (automática)

Ejemplos codificación manual vs automática

TIPO DE DATOS	SINTESIS	CODIFICACION MANUAL
Encuesta con fines académicos	200 respuestas en total	- Cuantificar los temas - Asignar una orientación a cada respuesta - Extraer respuestas ilustrativas de cada tema

TIPO DE DATOS	SINTESIS	CODIFICACION AUTOMATICA
Encuesta perfil del inversor	200 respuestas por día	- Cuantificar los temas - Actualizar los temas en tiempo real para un dashboard



• TIPO DE DATOS	SINTESIS	CODIFICACIÓN AUTOMÁTICA
Análisis de 30.000 tweets (u otro registro de comentarios financieros)	Aplicar una combinación de codificación manual (para crear un codebook a partir de 200 comentarios aprox)	- Cuantificar los temas - Actualizar los temas en tiempo real para un dashboard - Filtrar según un período del tema

Síntesis de exploración y cuantificación

"Realmente es muy difícil tener sustentabilidad en un país como el nuestro"
Fabricante – Dueño – Menos de 10

"Es necesaria la promoción de lineamientos básicos de sustentabilidad en las pymes que no cuentan con acceso a programas específicos en este sentido."
Comerciante – Dueño – Menos de 10

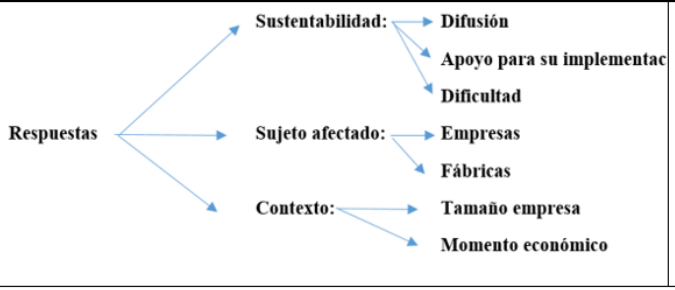
"En fábricas chicas es muy difícil invertir en estos momentos."
Fabricante – Gerente – Menos de 10

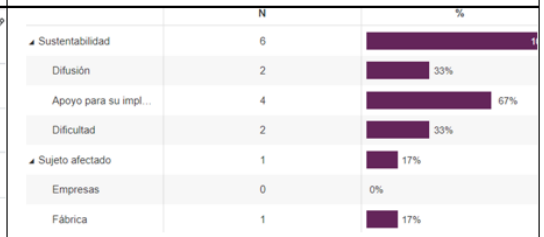
"Considero importante que todo estudio tenga la posibilidad de ser viable para implementarlo en la mayoría de los lugares e independiente del tamaño de empresa o comercio."
Comerciante – Dueño – Menos de 10

"Considero muy importante se logre implementar"
Fabricante – Dueño – 10 a 20

"Apoyar a las empresas para llevar esto a la práctica"
Comerciante – Dueño – 10 a 20

CONTEXTO			
Actividad	**Rol**	**Planta**	**Comentari**
Fabricante	Dueño	Menos de 10	Realmente e como el nue
Fabricante	Dueño	10 a 20	Considero m
Comerciante	Dueño	10 a 20	Apoyar a las





Análisis de sentimiento

A- Para qué sirve

B- Métodos disponibles: manual vs automatizado

C- Indicadores

CONTEXTO			SATISFACCIÓN	PREGUNTA ABIERTA	ANÁLISIS TEMÁTICO	ANÁLISIS DE SENTIMIENTO
Actividad	**Rol**	**Planta**	**¿Implementaría sustentabilidad?**	**Comentario**	**Temas**	**Orientación**
Fabricante	Dueño	Menos de 10	Sí	Realmente es muy difícil tener sustentabilidad en un país como el nuestro	Dificultad	Negativa
Fabricante	Dueño	10 a 20	Sí	Considero muy importante se logre implementar	Difusión	Positiva
Comerciante	Dueño	10 a 20	Sí	Apoyar a las empresas para llevar esto a la práctica	Apoyo	Positiva

Método manual

Codificación manual					
	Temas			Orientación	
	Difusión	Apoyo	Dificultad	Positivo	Negativo
"Realmente es muy difícil tener sustentabilidad en un país como el nuestro"	☐	☐	☑	☐	☑
"Considero muy importante se logre implementar"	☐	☐	☐	☐	☐
"Apoyar a las empresas para llevar esto a la práctica"	☐	☐	☐	☐	☐

Lo que usted quiera agregar...

	N	%
Contexto	4	67%
Tamaño empresa	3	50%
Momento económico	1	17%
Orientación	6	100%
Positiva	3	50%
Negativa	2	33%
Neutra	1	17%

Características del análisis manual

- Ventajas
  - Permite conocer mejor los datos
  - Capta bien la ironía (conceptos complejos, doble sentido, etc)
  - Casi no requiere herramienta
  - Se puede realizar en paralelo a la codificación temática
- Límites
  - No es adaptado para un volumen grande de comentarios
  - No se puede realizar en tiempo real
  - Objetividad
  - Diferencias entre evaluadores

Automatizado: cognitivo vs machine learning

Método cognitivo, ejemplo:

Ejemplo regla 1

Adverbio positivo + adverbio neutro + verbo negativo = opinión negativa

Comentario: Todo bien hasta que me empezaron a cobrar comisiones 

Ejemplo regla 2

Adverbio neutro + verbo negativo + sustantivo negativo + adjetivo positivo = opinión positiva

Comentario: Aunque me cobren comisiones, tienen buenos productos

Machine Learning

Paso 1) Entrenar el modelo (pre-training) Importar ejemplos bien formados para reproducir un comportamiento humano	
Esta acción es una bomba	Muy positivo
Vaya ladrones	Muy negativo
La mejor inversión de mi vida	Muy positivo

Paso 2) Datos para analizar 	
Esta acción es una bomba	Muy positivo
Vaya ladrones	Muy negativo
La mejor inversión de mi vida	Muy positivo

Paso 3) Optimización del modelo (fine-tuning)
• Entrenamiento con errores detectados
• Entrenar el modelo sobre nuevos ámbitos (inversiones, academia, empresas)
• Anadir nuevos idiomas

Características del método ML

Automatizado por Machine Learning

Ventajas

- Permite analizar muchos datos
- Análisis en tiempo real
- Objetividad
- Acuerdo entre evaluadores
- Transferencia entre ámbitos
- Transferencia entre idiomas
- Fácil de mejorar

Límites

- Puede ser complicado entender la ironía

Indicadores

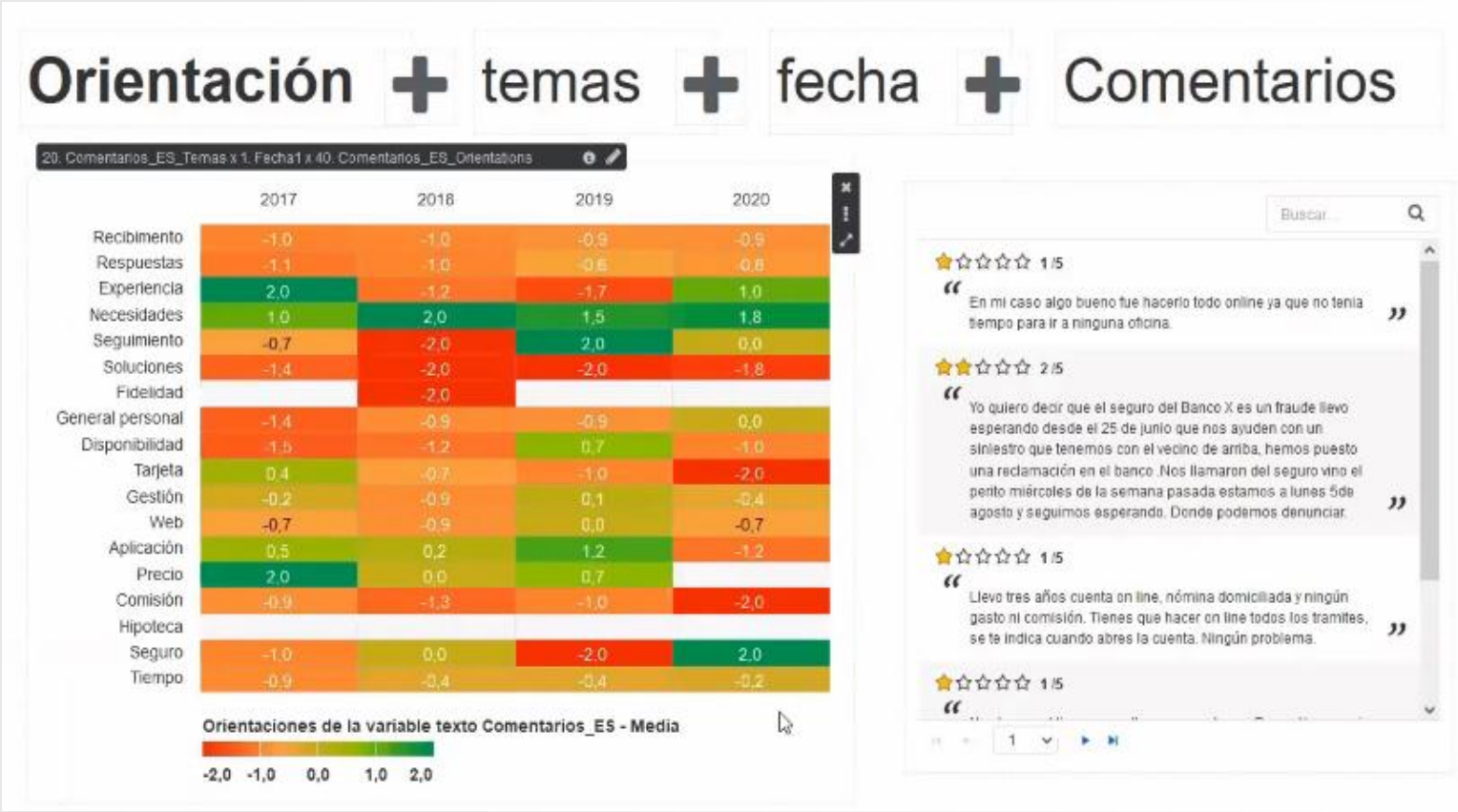
Definir la ponderación

Transforma la variable nominal en numérica, aplicando un peso a cada opción de respuesta.

Aplicar pesos...

OPCIONES DE RESPUESTA	PONDERACIÓN
Claramente negativo	-2
Bastante negativo	-1
Compartir	0
Bastante positivo	1
Claramente positivo	2
Sin opinión	

Indicadores



Desafíos y futuros objetivos

- Diccionario específico, a fin de configurar el baremo.
- Acceso restringido a la información clasificada.
- Necesidad de un sistema que realice técnicas de minería de texto en datos obtenidos dinámicamente.
- Combinación de técnicas de minería de texto y análisis de datos financieros.
- Oportunidades para la extracción de datos automatizada a gran escala.

Bibliografía

- Bach et. al (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review
- Carreño Giscafré (2020). Construcción de un índice de sentimientos con Twitter para el mercado argentino.
- Chang, Chong (2016). Sentiment Analysis in Financial Texts.
- Gao y Ye (2007). A framework for data mining-based anti-money laundering research
- Guo, Shi, Tu (2017). Textual Analysis and Machine Leaning: Crack Unstructured Data in Finance and Accounting.
- Gupta et. al (2020). Comprehensive review of text-mining applications in finance.
- Holton (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem
- Nasukawa, Yi (2003). Sentiment analysis: Capturing favorability using natural language processing
- Yadav, Sharan, Vaish (2020). Sentiment analysis of financial news using unsupervised approach,

Consultas y comentarios

Gabriel Feldman

- Profesor titular de Finanzas de Empresas I (UNT)
- Magister en Disciplinas Bancarias (UNLP)
- Doctorando (UBA)

gfeldman@face.unt.edu.ar

MUCHAS GRACIAS