41 Jornadas Nacionales de Administración Financiera
Septiembre 30 y Octubre 1, 2021

Encuestación y análisis predictivo en finanzas con apoyo de minería de textos

Gabriel R. Feldman
Universidad Nacional de Tucumán

SUMARIO

Para comentarios:
gabriel.feldman1@gmail.com

Resumen

Nos encontramos en un contexto en que los datos son los protagonistas. Hacer frente a la avalancha de datos que se procesan a través de redes de personas o dispositivos se ha vuelto cada vez más importante para la inteligencia empresarial. Junto con el análisis de big data, el campo de la minería de texto está evolucionando continuamente, entendiéndose como el proceso que intenta descubrir patrones en grandes volúmenes de datos textuales. Aunque haya una gran cantidad de información disponible, el uso de técnicas computacionales puede ayudar a procesar la información y analizar documentos completos, es así que la integración de los campos del conocimiento de las finanzas, datos y sistemas, resulta trascendente en pos del objetivo de maximización de valor. Este artículo comprende consideraciones relevantes para el proceso de encuestación y análisis de datos para investigaciones en el área económico-financiera, con énfasis en la minería de datos textuales.

1. Introducción

Como investigadores y como profesionales tenemos necesidades en relación a la gestión de encuestas, que involucra básicamente los procesos de captación de información, análisis de datos y elaboración de informes. Ya sea a fines académicos o de investigación de mercado, la encuestación comprende una compleja cadena de procesos, desde el diseño del cuestionario, las tareas de personalización, presentación divulgación, difusión, hasta su análisis de datos y presentación de resultados. Todas estas etapas generalmente se realizan con una combinación de herramientas informáticas, o integrándolas en un software, en la medida que éste cuente con todas las funcionalidades. Contar con conocimiento en programación contribuye a poder emplear distintas herramientas, caso contrario, es sugerible recurrir a softwares ad hoc, que ya cuentan con los desarrollos necesarios. El esquema en figura 1 sintetiza las etapas del proceso de encuestación, enfatizando las diversas formas de captación de datos.

Figura 1: Proceso de encuestación: detalle de las formas de captación de datos

Fuente: Software Le Sphinx

Queda evidenciada la diversidad de fuentes de datos que integra el proceso de captación, y por lo tanto la necesidad de contar con los mecanismos para su incorporación a la base de datos.

Una herramienta ampliamente difundida en la actualidad a efectos de la captación de datos es el formulario de Google, que, si bien es de uso gratuito, sus prestaciones son limitadas en comparación con aplicaciones especialmente diseñadas para este fin. Dichas limitaciones se materializan en la medida que las preguntas requieran condicionales, niveles de personalización, imagen institucional, entre otras funcionalidades habituales en el proceso.

En síntesis, Google Form, constituye una metodología "low cost" de hacer encuestas (al menos captar datos), pero requiere luego exportar e integrar con distintos softwares de análisis y presentación visual.

Es así que están disponibles también herramientas específicas como ser: Survey Monkey, Qualtrix, Typeform, Tableau, y otras, las cuales cubren una o más de las mencionadas etapas. Estos softwares permiten generar un cuestionario, y personalizarlo hasta cierto punto (unas en mayor o menor medida), así como la divulgación y recolección de datos en línea.

Como herramientas específicas para la segunda etapa, es decir el análisis de datos, pueden mencionarse Excel, Spss, Stata, Eview, mientras que, para la etapa de presentación, existen programas como Power BI, Infogram, Prezi, que posibilitan una visualización de datos con una calidad aun mayor, con infografías (presentación de datos a nivel de calidad un poco mayor), o la presentación de datos dinámicos (por ejemplo, a través de una URL). A su vez, algunas herramientas, como Le Sphinx permiten cubrir la totalidad del proceso en forma integrada, al conformar una suite de aplicaciones, posibilitando optimizar muchas tareas. Es sugerible contar con una herramienta que posibilite integrar las etapas de recopilación de información y su análisis, permitiendo diseñar el cuestionario, recoger datos, analizarlos, etc. El hecho de concebir todo el proyecto desde el principio al fin, posibilita comprender cómo tiene que pedir la información, en lo que comúnmente se conoce como trabajo de campo.

Desglosando las sucesivas etapas se obtiene el esquema de la figura 2, que menciona cada uno de los pasos involucrados.

El primer paso es clave, la planeación y estructuración, es decir trabajar esquemáticamente en forma previa a proceder a la redacción del cuestionario propiamente dicho. Ejemplo, una parte podría estar orientada a indagar sobre el uso de redes sociales por parte de la organización. La figura 3 ilustra el proceso.

La visualización, debe estar cada vez más pensada en el destinatario de la misma. Es decir, si se trabaja con información a nivel gerencial, ésta debe ser presentada en un modo amable y eficiente para el usuario, que no siempre es técnico en la materia.

La figura 4 sintetiza la importancia de la visualización de los datos, que se centra en tres ideas: conceptos que facilita, que permite y que requiere.

Un *dashboard* es un informe interactivo en que se puede simplificar una gran cantidad de datos en pocas páginas. Su elaboración implica que se mantiene un vínculo entre la herramienta de encuestación y la herramienta de análisis. Es importante su implementación dado que da a entender a los alumnos o investigadores, que lo que se realiza es algo mucho más parecido a la realidad que un ejercicio académico. La incorporación inmediata de la información a la base de datos, es un elemento que genera mucho atractivo a la metodología, dado que el usuario percibe que no se queda en el plano teórico, al visualizarse mucha similitud con

Figura 2: Proceso de encuestación: etapas

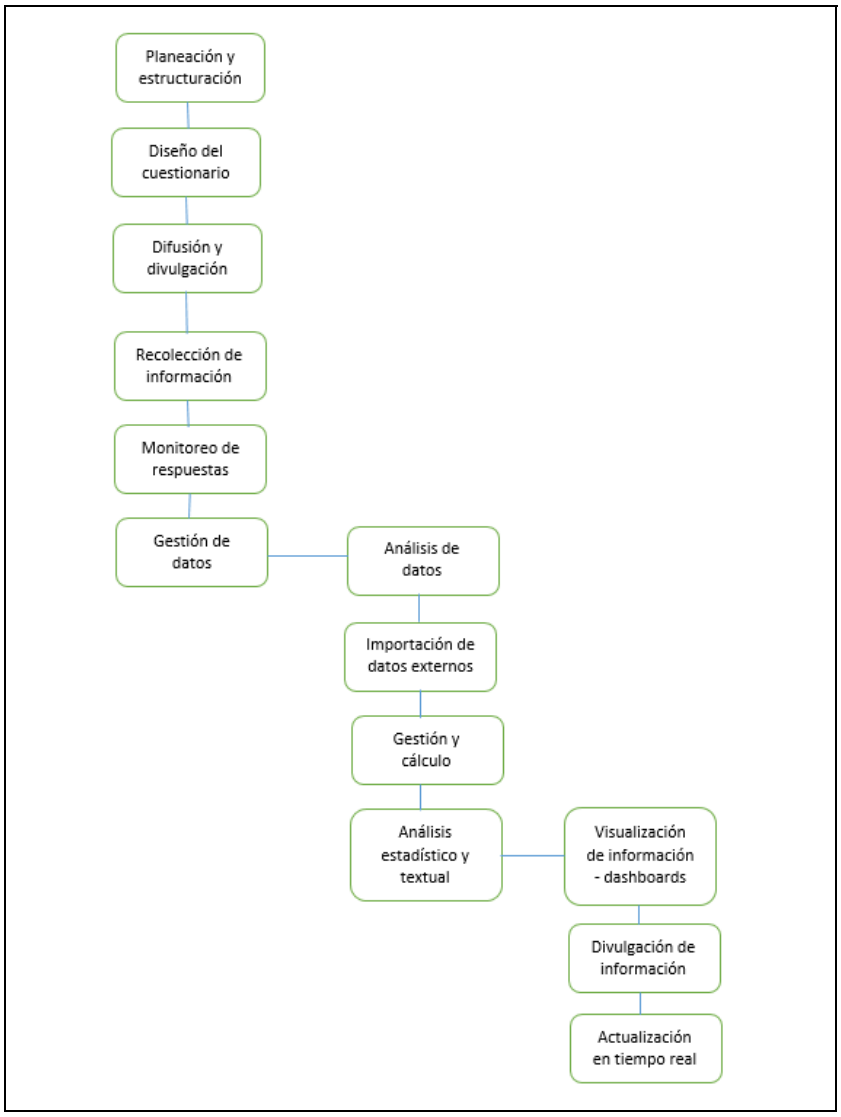


Figura 3: Planeación y estructuración de la encuesta

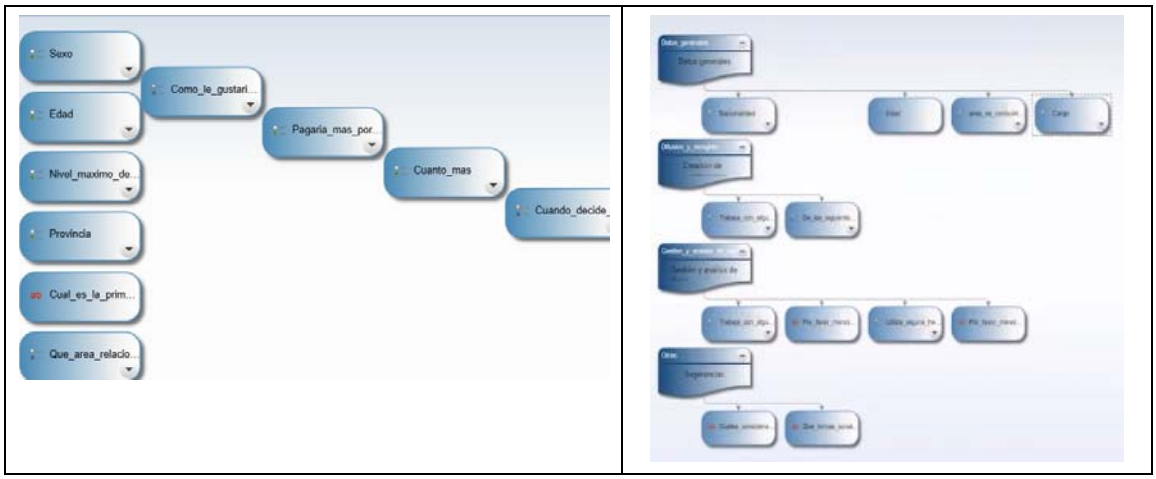


Figura 4: Visualización de la información

Importancia de la visualización de los datos	Facilita	Comunicación de ideas, mediante imágenes y gráficos
		Identificación de patrones y conexiones
		Relevancia y organización de la información
	Permite	Resumir y presentar gran cantidad de información
		Fácil revisión y análisis de los datos
		Toma de decisiones basadas en la información
	Requiere	Capacidad analítica
		Detección de lo relevante
		Selección de la información y datos claves

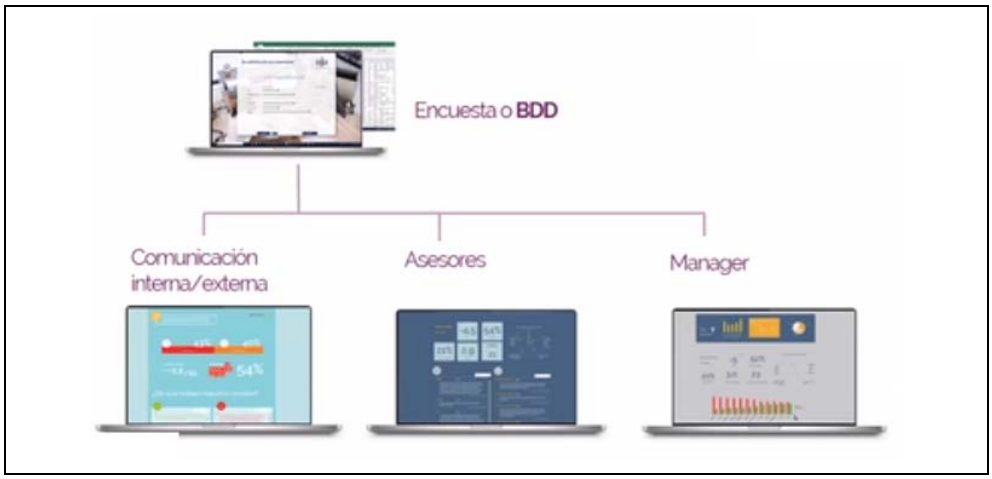
el estilo de reportes habituales en publicaciones de actualidad y periódicos. La satisfacción del usuario (especialmente alumnos), con el intercambio entre la teoría (que a veces parece abstracta), su aplicación práctica, y la comprensión de la capacidad del "para qué" y "cuando" pueden utilizarse las herramientas estadísticas.

El dashboard reporta sucesivas ventajas para el analista, que se listan a continuación:

- Compartir: se genera un link que se puede enviar por mail, redes sociales, etc.
- Adaptarse: puede hacerse varios dashboards por una encuesta, y puede adaptarse el mensaje acorde a quien se dirige, es decir, según quien se conecta verá una versión u otra.
- Enfocarse: puede tenerse una vista global de los resultados, y en la misma página una vista específica.
- Reaccionar: dado que los datos se incorporan en tiempo real, el dashboard va evolucionando y se pueden ver las tendencias y así reaccionar, sin tener que esperar al final del proceso. Esto se vincula con la gestión apoyada en inteligencia de negocios.
- Motivar: contar con información repercute en que se motive el personal involucrado en la toma de decisiones.
- Valorar: hace tomar conciencia del valor de los datos. Si bien el proceso de análisis puede hacerse rápido, pero el relevamiento es una labor estratégica y que toma un tiempo considerable. El dashboard hace que el estudio sea más visible, dándosele así más sentido al trabajo realizado. Ejemplo: todo el personal, está atento a que se está evaluando la satisfacción de los clientes, y a su vez los clientes saben que su opinión y es algo importante para la empresa.
- Ahorrar: Menor tiempo dedicado a la elaboración y procesamiento.

La "adaptación" a la que se hizo referencia en la enunciación previa, implica que puede generarse diversas formas de interactuar en una investigación en la actualidad. Con una base de datos pueden hacerse distintas reacciones (tanto en el ámbito empresarial como la investigación académica). Los análisis y visualizaciones son distintas, como se ilustra en la figura 5.

Figura 5: Adaptación de las visualizaciones mediante dashboard



Fuente: Le Sphinx

2. Prueba piloto

Como paso previo a llevar a la práctica el proceso descripto, se sugiere una prueba piloto (van Teijlingen y Hundley, 2001) seleccionando una muestra reducida, a fin de verificar la adecuación del instrumento de recolección de datos, determinar omisiones o inconsistencias, y hacer los ajustes necesarios para su aplicación en la investigación propiamente dicha.

Durante estas sesiones, se trabajará con una versión reducida de la encuesta, persiguiendo múltiples propósitos específicos, que se indican a continuación:

- Constatar la adecuación de la terminología empleada
- Indagar sobre la forma de contacto o distribución prevista
- Agregar/eliminar preguntas, analizando las sugerencias de los participantes
- Identificar problemas logísticos que pudieran afectar el proceso
- Confirmar la utilización de variables relevantes

El objetivo de la prueba piloto es lograr la mayor eficiencia en la encuesta, procurando anticipar los posibles puntos débiles y focalizar los aspectos deducidos del análisis teórico y práctico efectuado hasta ese momento. Se pretende así aportar validez y confiabilidad al proceso como herramienta metodológica (Greener, 2008), conectando los entramados teóricos con los datos. A su vez, aporta información preliminar que encausa la investigación hacia puntos considerados de interés, y detección de aspectos críticos. Para ello, contiene una sección abierta con el propósito que las personas se puedan expresar o relatar sus experiencias, que generalmente tienen que ver con referencias pasadas o presentes, así como lo que ellas esperan a futuro.

A su vez, el método en sí conforma una aproximación inicial al entorno, aportando en consecuencia una dinámica práctica para interactuar con las partes interesadas.

Con estas consideraciones, se pretende maximizar la tasa de respuesta, lo que es consecuencia de múltiples factores, y para lo cual el grupo de voluntarios es representativo de la población objetivo.

3. Minería de texto: concepto y aplicaciones en finanzas

El análisis con datos textuales representa una oportunidad para acercarse al día a día de las personas. El crecimiento de los textos financieros a raíz de big data ha desafiado a la mayoría de las organizaciones y ha generado una creciente demanda de herramientas de análisis.

En general, los flujos de texto son más difíciles de manejar que los flujos de datos numéricos. Ello es así porque los flujos de texto no están estructurados por naturaleza, pero representan expresiones colectivas que son valiosas en cualquier decisión financiera. Puede resultar abrumador, pero a la vez necesario dar sentido a los datos textuales no estructurados.

La minería de texto es un proceso mediante el cual el usuario obtiene información de alta calidad de un texto determinado, siendo una variante de la Minería de Datos, adoptando sus técnicas de aprendizaje automático para reconocer patrones y comprender nueva información.

Puede decirse que es un derivado combinado de técnicas como la minería de datos, el aprendizaje automático y la lingüística computacional. Tiene como objetivo extraer información y patrones de datos textuales. El enfoque trivial de la minería de texto es manual, en el que el humano lee el texto y busca información útil. Un enfoque más lógico es el automático, que extrae el texto de manera eficiente en términos de velocidad y costo. Ambos métodos serán descriptos en la sección aplicada del trabajo.

Presentada esta definición, resulta claro que revisar la literatura reciente sobre aplicaciones de minería de textos en finanzas puede ser útil para identificar áreas de investigación adicional.

En finanzas y contabilidad, en relación con los métodos cuantitativos utilizados tradicionalmente, el análisis textual se ha vuelto popular recientemente a pesar de su forma sustancialmente menos precisa. Uno de los aspectos notables de este boom de big data es la enorme amplitud de las interacciones entre los participantes que se pueden documentar. Como resultado de esta tendencia, la mayoría de los datos están cada vez menos estructurados, con una parte significativa del flujo de datos en formato textual. Las formas de dichos datos van desde comunicaciones por correo electrónico y redes sociales hasta informes corporativos y anuncios de noticias en diarios especializados.

A medida que este flujo de datos continúa expandiéndose rápidamente, se vuelve cada vez más importante desarrollar técnicas para repasar innumerables páginas de textos digitalizados y seleccionar la información útil que está oculta a simple vista. Para los profesionales de los mercados financieros, el auge de esta información es especialmente desafiante, porque que hay muchas organizaciones de noticias y comentaristas que ofrecen información y opinión sobre los mercados a través de sitios web tradicionales, publicaciones en Twitter u otros medios de comunicación social.

En relación a la hipótesis del mercado eficiente, la eficiencia de los mercados se basa en la difusión de información de mercado a los inversionistas de manera oportuna y correcta. Pero a medida que la información producida por los medios financieros y los datos del mercado se expanden a un ritmo rápido, las decisiones racionales y perfectamente informadas suelen ser

inalcanzables, debido a las limitaciones cognitivas de la mente de los inversores y la cantidad finita de tiempo que tienen para tomar decisiones. Existe disyuntiva entre los distintos modelos teóricos que intentan describir los mercados desde puntos de vista opuestos como lo son, la hipótesis del mercado eficiente y la de la economía conductual. En tal sentido, la teoría de las finanzas conductuales, plantea que existe una relación entre los sentimientos que la gente expresa (en redes sociales) con los ciclos que experimentan los precios de los mercados.

La teoría financiera clásica asume que los inversores son entidades racionales, no emocionales y los precios representan el equilibrio de los rendimientos esperados y los posibles riesgos. Esto no deja margen para que el sentimiento de los inversores desempeñe un papel en las determinaciones de precios. Sin embargo, las finanzas conductuales modernas reconocen tanto a los inversores sentimentales como a los inversores racionales. Es así que en este contexto la pregunta ya no es, como lo era hace unas décadas, si el sentimiento de los inversores afecta los precios de las acciones, sino más bien cómo medir el sentimiento de los inversores y cuantificar su efecto

Estas técnicas que incorporaran la información textual no solo pueden revelar las últimas tendencias en el estado de ánimo del público tal como se reflejan en los medios de comunicación, sino que también proporcionan pistas para analizar posibles ramificaciones y reducir los riesgos de realizar transacciones en mercados financieros caóticos.

Los datos textuales no estructurados han planteado muchos desafíos a los analistas. Obviamente, estos datos no son secuencias aleatorias de caracteres y, en cualquier caso, no pueden normalizarse y prepararse fácilmente para un análisis literal. Es difícil captar el orden de las palabras y su significado.

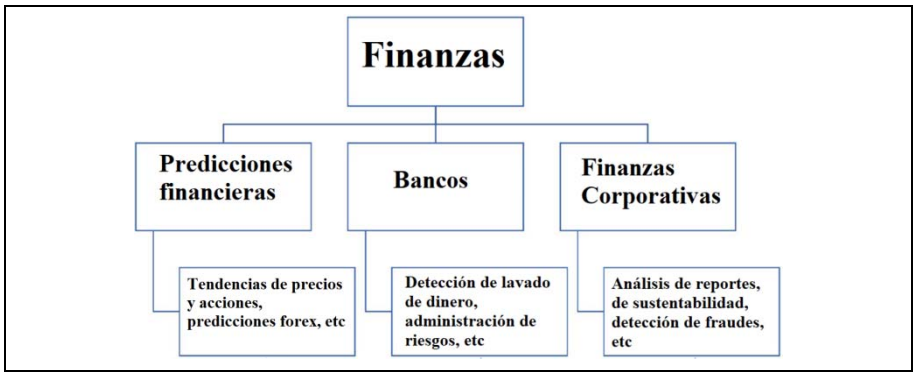
Las tecnologías de minería de textos han afectado sustancialmente a las industrias financieras. Dado que los datos en todos los sectores de las finanzas han crecido enormemente, la minería de textos se ha convertido en un importante campo de investigación en el ámbito de las finanzas. FinTech es un desarrollo en la industria financiera, que ha se ha definido como una unión de las finanzas y la tecnología de la información, y ha popularizado el uso de datos en la industria financiera. Estos datos están sustancialmente en forma de texto estructurado o no estructurado. Es decir, tradicional y técnicamente, se puede considerar que los datos textuales han sido siempre un elemento primordial y esencial en el sector financiero. Se han incrementado con las Fintech y las nuevas tecnologías.

La doctrina financiera trata el uso de minería de texto en tres áreas principales de aplicación: predicciones financieras, banca y finanzas corporativas, tal como se visualiza en el siguiente esquema de la figura 6.

En primer lugar, en el ámbito financiero, la predicción del mercado de valores es una de las aplicaciones en las que la minería de texto se ha utilizado para predecir las tendencias y los precios futuros del mercado de valores a partir del análisis de artículos de noticias financieras. Entre las muchas ideas cubiertas en la previsión financiera, se mencionan la predicción de la tasa de inflación, y una gran proporción de enfoque está en el mercado de valores y predicción de Forex.

Segundo, la banca es una de las industrias más grandes y de más rápido crecimiento en esta era de globalización, y lógicamente se encamina a adoptar las prácticas más eficientes para cada uno de sus departamentos. En medio de una revolución de Tecnologías de la información (TI), razones competitivas han llevado a la creciente importancia y adopción de la automatización bancaria. TI permite la implementación de diversas técnicas, incluyendo la minería de

Figura 6: Aplicación de minería de texto en finanzas



Fuente: Adaptado de Gupta et al.

textos, para el control de riesgos y el flujo fluido de transacciones a través de medios electrónicos. La evaluación de riesgos, la evaluación de la calidad, la detección del blanqueo de capitales y la gestión de las relaciones con los clientes son algunos ejemplos del amplio grupo de posibles aplicaciones de minería de textos en la banca.

Por último, las finanzas corporativas son un aspecto importante del dominio financiero porque integra el funcionamiento de una empresa con su estructura financiera. Varios documentos corporativos, como los informes anuales de una empresa, tienen mucho contexto financiero oculto. Se pueden emplear técnicas de minería de textos para extraer esta información oculta y también para predecir la futura sostenibilidad financiera de la empresa.

En las siguientes secciones se van a abordar preguntas claves relacionadas con la explosión de interés en cómo extraer información de datos no estructurados y cómo determinar si esta información proporciona alguna pista sobre las tendencias, ya sea de los mercados financieros o de la variable bajo análisis.

4. Análisis predictivo con datos textuales: Análisis temático

Está relacionado a la visualización y análisis dinámico de información aplicado a la comunicación eficiente de resultados, estudios y datos estadísticos. El análisis de datos textuales al que se referirá en lo sucesivo, corresponde a las preguntas abiertas de las encuestas. Las respuestas en las encuestas pueden ser cerradas (en este aspecto las consideraremos cuantitativas), y abiertas (en este aspecto las consideraremos cualitativas).

Figura 7: Publicaciones destacadas sobre minería de texto en administración financiera

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 al 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.

Figura 8: Proceso de análisis de datos textuales

1) Captar o Importar y estructurar los datos

2) Sintetizar/explorar: identificar los temas

Sintetizar la información (nube de palabras, clasificación)

Explorar (verbatim, contexto)

3) Codificar: captar la frecuencia de los temas

Manualmente (codificación)

Automáticamente (por diccionarios)

4) Captar la orientación: para qué sirve, métodos.

Manual

Automático (cognitivo, machine learning)

5) Indicadores y análisis de la información

Fuente: Software Le Sphinx

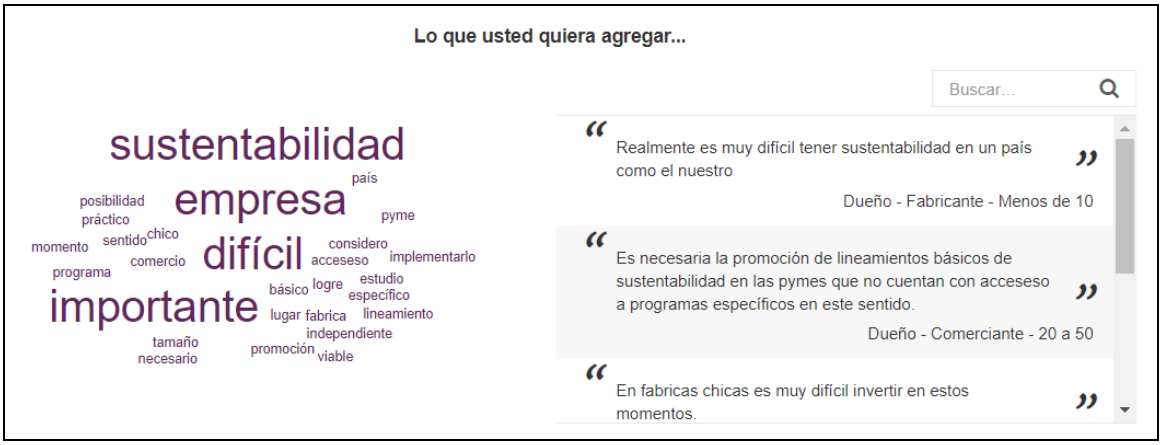
Desarrollo

1) Captación de los datos. El primer paso se refiere al proceso de captación de datos presentado al inicio, y que involucra diversas fuentes posibles de datos:

- Comentarios de la web que se importan (sistema de web-scraping).
- Encuestas: puede ser cualquier herramienta de encuestación. Normalmente proporciona un esquema de datos muy útil, ya que brinda en la forma de una línea por persona que contesta, y una columna por pregunta.
- Entrevistas

2) Exploración. La exploración tiene dos objetivos: i- Explorar para la gente que evalúa o hace su análisis temático, ii- Para que un tercero pueda consultar la información.
A este propósito, la nube de palabra proporciona un primer contacto con los datos, como la que se expone a continuación[1].

Figura 9: Nube de palabras



Puede apreciarse que aparecen las palabras y a la derecha una lista de los comentarios. Es un primer paso necesario, ya que brinda un resumen visual (rápido) de qué habla la gente, y al posicionarse sobre cada palabra accedo al comentario completo que contiene dicha palabra. Este método de explorar los datos tiene dos propósitos: i – acceder fácilmente a los datos, ii- tener una síntesis de la información de la base de datos. Sus ventajas son:

- Brindar una imagen visual rápida de las respuestas. Ello posibilita intuir algunos temas, lo cual va a resultar útil en la etapa siguiente. Esta nube permite una síntesis de los motivos expresados por los encuestados, aún sin leer los comentarios.
- El software lo que hace es lematizar los comentarios, lo que significa un análisis morfológico que permite simplificar el texto. Es un proceso lingüístico que simplifica el corpus, reduce las palabras a su "lema", o sea a su forma lo más sencilla posible. La lematización implica: verbos en infinitivo, adjetivos y sustantivos en masculino singular.

[1] Elaborada con software Le Sphinx

- Distinguir lo que se conoce como “canal” es decir, por ejemplo, en este caso los encuestados destacan la DIFICULTAD. Y el canal o factor que lo implica: tamaño, momento.
- Al verse palabras como “empresas”, “fábricas”, que serían los sujetos afectados

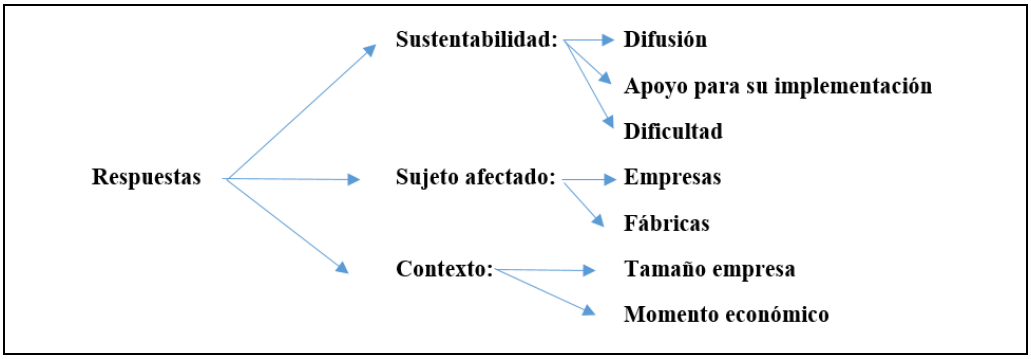
3) Codificación. El próximo paso, Codificar, tiene por objetivo analizar la frecuencia de los temas que se han identificado, es decir, una vez que se cuenta con una base de datos, lo que se intentará es identificar cuáles son los temas, para luego intentar ver cuántas veces cada uno de esos temas ocurre en la base de datos. Una forma de apreciar claramente este aspecto, es exponerlo en la forma de tabla, como se muestra a continuación, en que la columna de la derecha es la que se generará a partir del proceso de codificación.

Figura 10: Análisis de datos textuales

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”	¿?

El mapa cognitivo es muy importante hacerlo para estructurar la forma de analizar y crear lo que se conoce como *codebook.*

Figura 11: Mapa cognitivo



La codificación puede realizarse con un procedimiento manual o automático, y la elección del mismo depende de muchos factores, que se listan a continuación:

- 1er aspecto: frases cortas vs largas (y por lo tanto, ideas más complejas): puede impactar en mucho el método de análisis temático.
- 2do aspecto: pocos datos (y por lo tanto apto para análisis manual), o muchos datos (y por lo tanto sugerible análisis automático).

- 3er aspecto: los datos son definitivos (encuesta terminada) o provisorios (encuesta abierta). Incluso puede que vengan muchos datos más en el futuro. Por lo tanto, hacerse un análisis manual en primer lugar y luego automatizar.
- 4to aspecto: la forma de compartir los resultados. Si bien depende de los datos, cuando es un análisis puntual, que inicia y termina, puede convenir un análisis puntual. Si hay que compartir, es más adaptable un análisis automático, ya que preciso elaborar un dashboard para compartir. Es decir, esto se refiere al objetivo del estudio, ya que el objetivo puede ser un informe puntual, o un dashboard interactivo, que la gente consulte en tiempo real y por lo tanto se tienen que actualizar los temas de los comentarios.
- 5to aspecto: tiempo disponible para analizarlos. Analizar manualmente 10.000 respuesta implica un mes de trabajo.
- 6to aspecto: grado de conocimiento del analista respecto de lo que habla la gente.

Figura 12: Consideraciones para el método de codificación

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

En la codificación manual, el analista procede a leer cada frase e indicar indicando que tema/subtema corresponde. Cuando aparece un nuevo tema, entonces hay que agregarlo para poder imputar el comentario al mismo. Es un método muy bueno y traduce la inteligencia del que va codificando. ¿Cuándo es sugerible esta codificación manual?

- Cuando se dispone mucho tiempo. Esta codificación manual, conlleva mucha “calidad” en su resultado, por eso sigue siendo utilizada cuando se la puede hacer. Es decir, aplicar el método es “proporcional” al tiempo y la cantidad de datos.
- Se cuenta con pocos datos. En caso de más de 200 respuestas, se sugiere directamente ir a un método automático.
- Son ideas complejas o hay ironía. Difícilmente un soft pueda resolver estos casos. Incluso al hacerse manualmente, puede haber diferencias de criterio si participa más de una persona. Paralelamente, en la codificación manual, puede no solo ir codificando, sino también indicarse si es “positiva” o “negativa”, captando incluso la orientación en forma perfecta, aunque haya ironía.
- Es decir, una ventaja de la codificación manual es que, al mismo tiempo de hacer la codificación, podemos captarse la orientación. En tal caso, cabe añadir un tema que titulado “positivo” o “negativo” y trabajarlo en paralelo a la codificación.
- Cuando es preciso preparar un informe, ya que permite extraer secciones de comentarios, para incorporarlos al mismo, brindando mayor calidad.

El método automático no es incompatible con el visto antes, sino que incluso puede resultar del mismo. Si bien es automático, cuando se cuenta con muchos datos, conviven una parte manual para obtener el codebook para proceder con la parte automática. Es por eso que también se puede llamarle “semi-automático”.

Síntesis de las consideraciones de usar este método automático:

- Cuando la base de datos es grande o enorme (como el caso de bancos o compañías telefónicas)
- Cuando se quiere seguir en tiempo real, actualización frecuente de datos.
- Cuando se quiere compartir los resultados en un dashboard.
- Cuando se dispone de tiempo al principio, pero no se quiere repetir el análisis a posteriori.
- Consultoras que trabajan con empresas clientes que son de la misma rama de actividad. Entonces ya tienen pre-configurados los temas y se apoyan en dicho historial para “duplicar los diccionarios” y así configurar los análisis de encuestas de nuevos clientes del mismo/similar sector.

En concreto, lo que se hace es a las palabras identificadas por la herramienta, las asigna a los temas (a esta asignación de palabras a los temas es lo que le llama “diccionarios”). Entonces el soft asigna todos los comentarios a dicho tema, así como los que llegarán en el futuro, y contengan dicha palabra. Con ello, el software capta la cantidad de ocurrencias y la cantidad de observaciones, y debería llegarse a un resultado similar al del método anterior, habiendo seguido un camino diferente.

5. Análisis de sentimiento

Este análisis complementa al análisis temático. Luego de haber analizado el proceso de codificación, a continuación, se describe los siguientes aspectos del análisis de sentimiento:

I) Para qué sirve
II) Métodos disponibles

Desarrollo

I) Para qué sirve. Previamente vimos de qué habla la gente, y ahora veremos cómo hablan. O sea, si lo hacen de forma positiva, negativa, mixta, etc.

La gente expresa su opinión ya sea como clientes, ciudadanos, estudiantes, inversores, usuarios (de productos/servicios). El término “análisis de sentimiento” refiere a “determinar la polaridad de la subjetividad” (positiva o negativa) y la “fuerza de la polaridad” (fuertemente positiva, levemente positiva, débilmente positiva, etc.) de un texto dado; en otras palabras, determinar la opinión del escritor. El esquema en la figura 13 ilustra la terminología.

Figura 13: Análisis de sentimiento

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

II) Métodos disponibles. Básicamente pueden describirse dos métodos:

- Manual: aplicable cuando hay pocas observaciones. Tiene algunas ventajas, pero también tiene límites.
- Automatizado: aplicable cuando hay muchos datos.

El método manual queda ilustrado en la figura 14. Como puede apreciarse, es complementando el análisis visto en la primera parte de la presentación. Es decir, se agregan "temas" (que en realidad no son temas sino "orientaciones"), y luego el analista va imputando a cada uno de ellos. Tendría que trabajar "comentario por comentario" realizando esta tarea de asignación a temas, y sus respectivas orientaciones.

Figura 14: Análisis de sentimiento 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"	☐	☐	☐	☐	☐

Las características de este método se listan a continuación:

Ventajas

- Permite conocer mejor los datos
- Capta bien la ironía (conceptos complejos, doble sentido, etc)
- Casi no requiere herramienta informática

- Se puede realizar en paralelo a la codificación temática

Límites

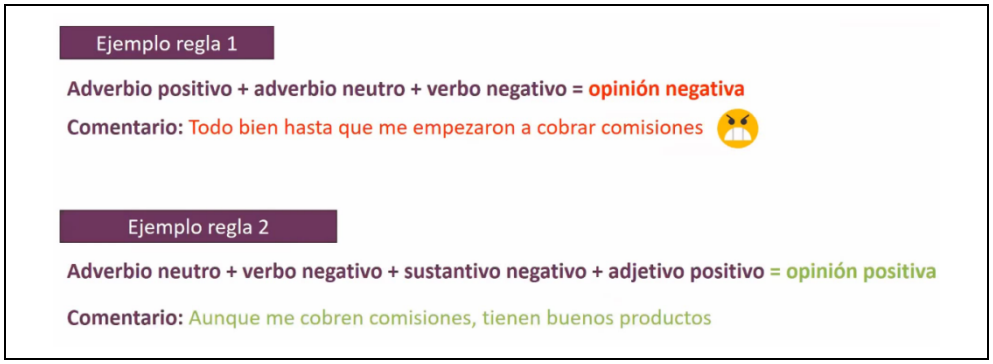
- No es adaptado para un volumen grande de comentarios
- No se puede realizar en tiempo real
- Objetividad
- Diferencias entre evaluadores

El método automatizado requiere tecnología, para que el operador no tenga tanto trabajo que hacer luego, sino preparar la rutina. Hay dos métodos: i) Cognitivo (o tradicional, es el que la gente ha utilizado desde hace muchos años; y ii) *Machine Learning* (más reciente).

Desarrollo de estos métodos

i) Cognitivo (o tradicional). Lo emplean los que son "lingüistas" o expertos en lenguaje. Consiste en la creación de bases de datos terminológicas, series ordenadas de palabras a partir del léxico, creando reglas usando análisis sintáctico y semántico. Estas reglas permiten clasificar entre positivo y negativo. Es decir, lo que se describe es la construcción de la frase, y en base a ello se le da una orientación.

Figura 15: Método automatizado cognitivo de análisis de sentimiento



Fuente: Le Sphinx

El límite o desventaja de este método (en comparación con el que veremos luego) es que cada vez que añadimos reglas, éstas pueden alterar la calidad del análisis, ya que pueden ir "contra" antiguas reglas. Ello hace que el mantenimiento de este método sea dificultoso a través del tiempo, y por ende los resultados no sean tan buenos.

ii) Machine learning. Ha empezado en el año 2018. *Machine learning* es una forma de la Inteligencia Artificial que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. De lo que se trata es de asignarle tareas a un algoritmo capaz de entender el lenguaje humano. Ejemplo, en los dispositivos que comprenden cuando se les indica: "apaga la luz" lo hace. En nuestro caso también le asignaremos tareas: la de clasificación de opiniones. A este algoritmo, para asignarle tareas hay que entrenarlo, esa es la diferencia con el método previo, que no requería entrenamiento. Entrenarlo significa indicarle lo que queremos que haga.

Una organización no necesita de *big data* para utilizar las técnicas del *machine learning*; sin embargo, big data puede ayudar a mejorar la precisión de los modelos de *machine learning*. Con *big data*, ahora es posible virtualizar los datos para que se puedan almacenar de la forma más eficiente y eficaz en función de los costos, ya sea en el entorno local o en la nube.

Este modelo en un principio está vacío, y hay que darle ejemplos para entrenarlo. Lo que se hace es importarle miles y miles de comentarios, ya codificados por un operador.

Lo bueno de este método, es que cuando la máquina se equivoca, se "auto-mejora". Decir que el sistema "auto-mejora" es a medias, ya que es el operador que debe tomar los datos equivocados, corregirlos, o añadírseles nuevos comentarios. Se corrige a partir de errores detectados, se incorporan nuevos hábitos, o nuevos idiomas. Se dice que auto mejora, porque las frases que el algoritmo no logra interpretar, es lo que el programador le vuelve a configurar para que en el futuro ya pueda hacerlo, ya que va estar entrenado para ello. El nombre *machine Learning* proviene justamente de dicha capacidad para "aprender" y mejorar.

Figura 16: Método automatizado Machine Learning de análisis de sentimiento

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

Es así que, si utilizamos un algoritmo diseñado para un sector de actividad y queremos emplearlo en otro, entonces buscaríamos los términos específicos de esta nueva actividad, y lo importaríamos en el modelo, para que éste progrese y luego sea mejor para futuros trabajos. Esa es una ventaja del *machine Learning*.

Resumen de las características de automatización 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 o ideas complejas.

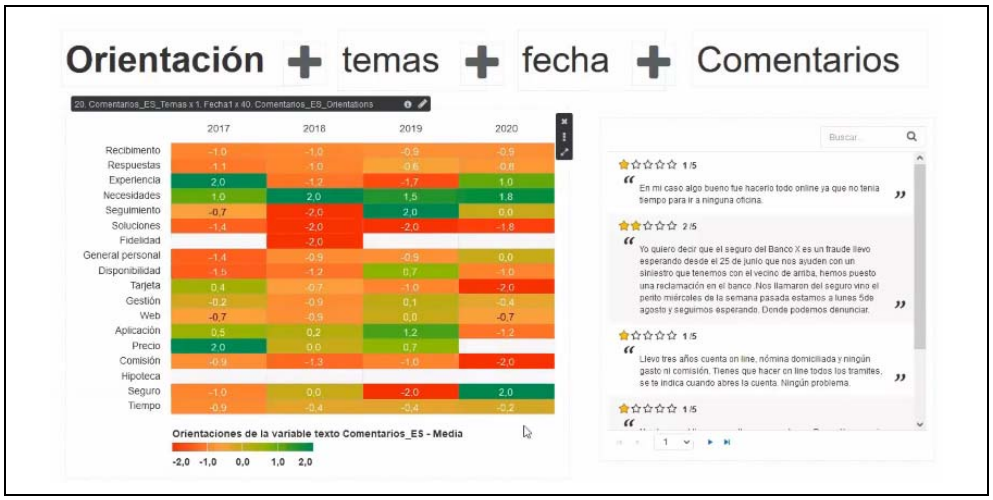
Indicadores. Ya han sido presentados los métodos de análisis de sentimiento, que son los "ingredientes" para desarrollar los indicadores. Un primer análisis que podemos hacer una vez que hemos encontrado la orientación es asignar porcentajes a cada una de las categorías, es decir parametrizar un baremo interno para cada una de las categorías, como se ilustra a continuación:

Figura 17: Baremo para análisis de sentimiento

De ese modo puede obtenerse la media de todos los datos, que representa el "tono" de las respuestas. El signo de este indicador (positivo o negativo), ya constituye una información de interés. A su vez, puede analizarse si se aprecia una evolución del tono a través del tiempo, lo cual es particularmente relevante para aspectos de decisiones financieras. Con ello, puede evaluarse, por ejemplo, qué hizo la empresa en cierto mes en que la media de las opiniones fue mejor, peor, etc., constituyendo una señal de resiliencia.

También es relevante cruzar este análisis con los temas, ya que al agregársele el análisis de sentimiento puede apreciarse la orientación de esos temas. De ese modo saber, por ejemplo: cuantas veces han hablado bien de cierto tema, cuantas veces han hablado mal de cierto tema, o ver las proporciones correspondientes. Lo interesante es que se estará cruzando información que previamente no existía en la base de datos, ya que son las 2 variables que se han creado. De esta forma, el análisis resultará de utilidad para reaccionar rápido y saber qué hacer para mejorar. En tal sentido se considera apropiado el esquema de mapa de calor para crear un tablero de comando, que represente un "semáforo" para la toma de decisiones.

Figura 18: Mapa de calor



Fuente: Le Sphinx

6. Desafíos y futuros objetivos

El sector financiero es un impulsor significativo de la industria en general, y la creciente cantidad de datos en este campo ha dado lugar a una serie de aplicaciones que pueden utilizarse para mejorar el campo y lograr objetivos comerciales.

Los datos deben someterse a un procesamiento previo adecuado antes de que puedan utilizarse para el análisis. Aunque las listas de léxicos están disponibles para varios dominios, el sector financiero debe tener un diccionario específico para tales enfoques, a fin de asignar pesos adecuados a los aspectos correspondientes en el documento que se analice.

Cabe destacar que aún existe un acceso restringido a la información clasificada, lo que es un obstáculo importante.

Por otra parte, las técnicas actuales se centran en obtener resultados de forma estática que son verdaderos durante un período de tiempo determinado. Existe la necesidad de un sistema que realice técnicas de minería de texto en datos obtenidos dinámicamente para generar resultados en tiempo real para permitir una mejor comprensión.

La combinación de técnicas de minería de texto y análisis de datos financieros puede producir un modelo que potencialmente pueda ser el modelo más eficiente para este tipo de problemas. Los resultados obtenidos de la minería de datos textuales se pueden integrar con los del análisis financiero, proporcionando modelos que se centran en datos históricos, así como opiniones de diversas fuentes.

La extracción manual de datos es cara, laboriosa y propensa a errores. Como resultado, existen enormes oportunidades para la extracción de datos automatizada a gran escala para transformar esta faceta de las finanzas en un campo más cuantitativo y rico en datos. Más aun teniendo en cuenta que la metodología permite integrar datos de origen no técnico, en el análisis financiero. Otro de los retos importantes es optimizar el uso de grandes volúmenes de datos para extraer patrones de ellos. Es necesario adecuar el almacenamiento de esos datos, indexarlos, y que el acceso sea lo suficientemente rápido para que pueda escalar.

REFERENCIAS

Bach, M.P., Krstic, Z., Seljan, S. & Turulja, L. (2019). *Text mining for big data analysis in financial sector: A literature review*. Sustainability, 2019, 11 (5): 1277.

Carreño Giscafré, F. (2020). *Construcción de un índice de sentimientos con Twitter para el mercado argentino*. Tesis de posgrado. FACE, UNT.

Chang, S. & Chong, M. (2017). *Sentiment analysis in financial texts*. Decision Support Systems, 94: 53-64.

Gao, Z. & Ye, M. (2007). *A framework for data mining-based anti-money laundering research*. Journal of Money Laundering Control, 10 (2): 170-179.

Guo, L, Shi, F. & Tu, J. (2017) *Textual analysis and machine learning: Crack unstructured data in finance and accounting*, The Journal of Finance and Data Science, 2 (3): 153-170.

Gupta, A., Dengre, V., Kheruwala, H.A. & Shah, M. (2020). *Comprehensive review of text-mining applications in finance*. Financial Innovation, 6: 39.

Holton, C. (2009). *Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem*. Decision Support Systems, 46 (4): 853-864.

Le Sphinx, videos instructivos. https://www.lesphinx.es/curso-analisis-textual

Nasukawa, T. & Yi, J. (2003). *Sentiment analysis: Capturing favorability using natural language processing*. Proceedings of the 2nd International Conference on Knowledge Capture, Florida, 23-25 October 2003, 70-77.

Yadav, A., Jha, C., Sharan, A. & Vaish, V. (2020). *Sentiment analysis of financial news using unsupervised approach*. Procedia Computer Science, 167: 589-598.