Trabajo de Maestría en la modalidad de profundización: Análisis de discurso de los máximos responsables de las empresas participantes en el COLCAP

Dairo Alberto Cuervo García

Director: Javier Orlando Pantoja Robayo
jpantoja@eafit.edu.co
Escuela de Finanzas, Economía y Gobierno,
Área de Macroeconomía y Sistemas Financieros,
Director del grupo de investigación de Sistemas Financieros y Macroeconomía,
Profesor Asociado, Universidad EAFIT.

Co-Director: Johan Armando Ceballos Cañon
jaceballosc@unal.edu.co
Doctor en Ingeniería Matemática,
Profesor Asociado, Universidad Nacional de Colombia.

Maestría en Ciencia de los Datos y Analítica
Universidad EAFIT.

Resumen

Este trabajo aborda el análisis y la extracción de información de las cartas de los máximos responsables incluidas en reportes integrados y públicos de las empresas colombianas pertenecientes al COLCAP[1]. Utilizando técnicas de *minería de texto*, *procesamiento de lenguaje natural* (NLP) y *aprendizaje automático* (ML), este ejercicio captura información del discurso de los líderes empresariales, buscando revelar aspectos del comportamiento de las empresas en su ecosistema, estableciendo relaciones entre el discurso ejecutivo y los indicadores de rendimiento influenciados por las decisiones de accionistas y *stakeholders*.

[1]La firma proveedora de índices MSCI desarrolló en alianza con la Bolsa de Valores de Colombia (BVC) el índice MSCI COLCAP como sucesor y reemplazo del índice COLCAP original, que fue creado en 2008 y estuvo vigente hasta el 27 de mayo de 2021 - https://www.bvc.com.co/msci-colcap.

Índice

Índice de cuadros

Índice de figuras

Cuadro 1: Listado de términos utilizados

Siglas	**Descripción**
GRI	Iniciativa de Reporte Global (Global Reporting Initiative)
TM	Minería de Texto (Text Mining)
NLP	Procesamiento de Lenguaje Natural (Natural Language Processing)
ML	Aprendizaje Automático (Machine Learning)
TF-IDF	Frecuencia de Término - Frecuencia Inversa de Documento (Term Frequency-Inverse Document Frequency)
LSA	Análisis Semántico Latente (Latent Semantic Analysis)
CAPM	Modelo de Valoración de Activos de Capital (Capital Asset Pricing Model)
LDA	Asignación Latente de Dirichlet (Latent Dirichlet Allocation)
VADER	Razonador de Diccionario Consciente de la Valencia y Análisis de Sentimiento (Valence Aware Dictionary and sEntiment Reasoner)
SA	Análisis de Sentimientos (Sentiment Analysis)
CRISP-DM	Proceso Estándar Inter-Industrias para Minería de Datoss (Cross-Industry Standard Process for Data Mining)
OCR	Reconocimiento Óptico de Caracteres (Optical Character Recognition)
BERT	Representación de Codificadores Bidireccionales para Transformers (Bidirectional Encoder Representations from Transformers)

1. Introducción

Las empresas se ven cada vez más impulsadas a revelar información relevante, un esfuerzo que anualmente se refleja en informes guiados por marcos, estándares e indicadores como la *Iniciativa de Reporte Global - GRI* - (GRI, 2023) , y documentos como el reporte integrado, el informe de gestión o reportes de sostenibilidad, en general, buscando comunicar nociones alrededor de las dificultades, logros, mejoras u objetivos de corto, mediano y largo plazo a sus accionistas y *stakeholders*. Un componente específico de estos documentos de comunicación son las cartas de los máximos responsables, estos son un elemento clave, donde el análisis de sus elementos textuales ofrece la oportunidad de explorar el discurso de los líderes empresariales, proporcionando una visión de la estrategia y dirección de las empresas, así como la comprensión de sus objetivos y prioridades estratégicas.

1.1. Descripción del problema

En un entorno empresarial dinámico y competitivo, comunicar de manera efectiva la estrategia, logros y metas de una empresa a sus accionistas e inversores se ha convertido en una necesidad. Los informes anuales, que incluyen cartas a los accionistas escritas por los máximos responsables de la empresa, son un medio valioso para lograr este propósito. Según Warren Buffett a lo largo de los años, han existido múltiples ocasiones donde leer la carta anual del máximo responsable "ha sido uno de los factores en su decisión de hacer algo o no hacer algo" respecto a la inversión en el mercado bursátil (Zweig, 2016).

Sin embargo, se convierte en un desafío inferir las intenciones y los resultados de la empresa en el discurso de los máximos responsables que pueda sugerir aspectos del comportamiento de las empresas, relacionando el contenido de estas cartas con las interacciones de las compañías con sus grupos de interés, esto último expresado en variables numéricas como el precio de las acciones o la variabilidad en los resultados a lo largo del tiempo en el mercado bursátil; lo anterior, debido a la dificultad de obtener características o métricas del contenido textual de las cartas que puedan ser analizadas respecto variables numéricas de la bolsa de valores.

La transformación de datos textuales en información numérica y el desarrollo de modelos que permitan cuantificar y comprender esta influencia es un reto; además de la variabilidad en la estructura y contenido de las cartas, así como la multicausalidad en los cambios de los indicadores bursátiles. Este problema plantea la necesidad de desarrollar enfoques que se centren en analizar el contenido de las cartas a través de técnicas de minería de texto, procesamiento de lenguaje natural y aprendizaje automático en busca de métricas que puedan indicar relaciones entre el contenido textual de las cartas a los accionistas y los comportamientos de variables bursátiles, en un esfuerzo por analizar patrones y tendencias.

1.2. Justificación

En el campo de la minería de texto y el procesamiento de lenguaje natural, el análisis de textos es esencial para desentrañar información valiosa y tomar decisiones informadas. La riqueza de datos contenidos en las cartas a los accionistas es un recurso valioso que, si se aborda con éxito, puede proporcionar *insights* fundamentales para la toma de decisiones empresariales.

Este proyecto de investigación busca aplicar técnicas de estos campos, con un enfoque en idioma español, para abordar los desafíos planteados, donde se espera tras el análisis de patrones y tendencias en el discurso de las cartas a los accionistas a lo largo del tiempo, que proporcione información sobre las estrategias y logros de las empresas. La solución de este desafío proporcionará una comprensión más profunda de cómo las comunicaciones empresariales impactan en el comportamiento de las empresas en el mercado, además de mostrar un panorama general de la toma de decisiones y la gestión de riesgos en un contexto empresarial en constante evolución.

2. Objetivos

2.1. Objetivo general

Explorar el uso de técnicas de minería de texto y procesamiento de lenguaje natural en el análisis del contenido de las cartas de los máximos responsables de las empresas colombianas a sus accionistas, identificando patrones y tendencias en el discurso empresarial que sugieran posibles comportamientos en los indicadores bursátiles.

2.2. Objetivos específicos

1. Procesar un conjunto representativo de cartas escritas por los máximos responsables de empresas colombianas listadas en el COLCAP a sus accionistas, utilizando técnicas de minería de texto y procesamiento de lenguaje natural.

2. Desarrollar métricas que extraigan información de las cartas a partir de la identificación de temas recurrentes, cambios de discurso en el tiempo, tópicos, cuantificar la polaridad (positivo, negativo o neutro) de las expresiones en cada uno de los documentos.

3. Analizar las métricas y patrones identificados en las cartas y su posible influencia en indicadores como el precio de las acciones y la rentabilidad.

4. Evaluar resultados alrededor de las técnicas planteadas, para la extracción de información de las cartas y su posible influencia en los comportamientos en indicadores bursátiles.

3. Estado del arte y marco teórico

3.1. Minería de texto y procesamiento de lenguaje natural

A pesar de los recientes avances, los primeros indicios de la disciplina del análisis de textos se encuentran a principios del siglo XIX, cuando eruditos se dedicaron a analizar la expresión de emociones en textos bíblicos mediante técnicas como la reorganización, clasificación y cálculo de frecuencias de palabras (McLaughlin et al., 2022). Cerca al año de 1940, los investigadores expandieron su enfoque e incluyeron el examen de diversos tipos de textos, como cartas, periódicos y libros.

Posteriormente, la adopción generalizada de Internet para los años posteriores a 1990, posibilitó el acceso a un extenso *corpus* de documentos basados en texto, donde se desarrollan diversas técnicas para extraer información de estos textos como la minería de datos, la minería de textos y la minería web. Este panorama evoluciona luego del año 2000 donde estas técnicas se ampliaron e incluyeron la identificación de temas cómo LDA, en ingles (*Latent Dirichlet Allocation*), y el modelado de tópicos, también la descripción de la polaridad, sentimiento o estado de ánimo - análisis de sentimientos o minería de opiniones - y análisis de redes sociales (McLaughlin et al., 2022).

Actualmente, el análisis de texto es una familia de métodos destinados a extraer información útil de colecciones de texto grandes, información no estructurada, esto en un proceso de recopilar, procesar e interpretar datos de texto (Eldén, 2007). En términos generales, la minería de texto transforma grandes volúmenes de texto en una estructura organizada que facilita un análisis más profundo para la identificación de relaciones y patrones, donde los datos estructurados resultantes pueden ser incorporados en bases de datos o almacenamientos específicos y se emplean para análisis descriptivos, inferenciales o predictivos (McLaughlin et al., 2022).

En general, el análisis de texto se lleva a cabo mediante el procesamiento del lenguaje natural, una disciplina que ha experimentado un crecimiento significativo en importancia, utilidad y sofisticación en los últimos años; el NLP no solo es capaz de gestionar grandes volúmenes de datos basados en texto de manera coherente, sino que también puede interpretar conceptos en contextos complejos y resolver algunas ambigüedades del lenguaje (McLaughlin et al., 2022).

El proceso base es el preprocesamiento del texto, cómo se muestra en la figura 1, en general se realizan los mismos procedimientos, en primer lugar, se hace una limpieza de los documentos, posteriormente se divide en palabras (tokenización), se eliminan *stopwords* para enfocarse en términos relevantes, se hace *stemming* y lematización para reducir palabras a formas base y por último se realiza una indexación para asignar índices numéricos. Una forma común en la literatura es la construcción de la Matriz Tf-Idf para transformar el documento y asignar pesos a las palabras según su importancia, Estos pasos optimizan el texto para análisis estadísticos y de aprendizaje automático (Eldén, 2007).

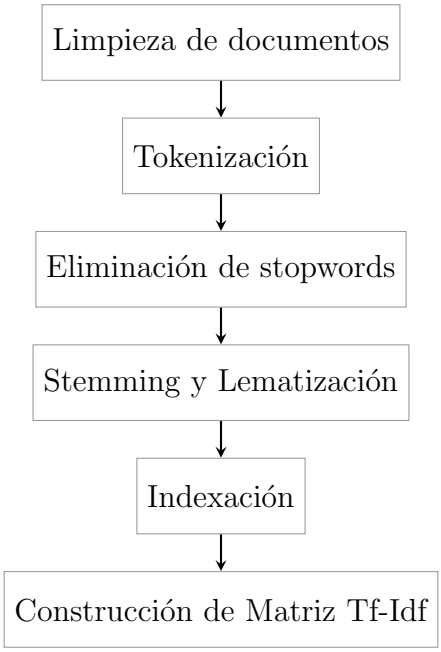

Figura 1: Diagrama del proceso de preprocesamiento de texto basado en la descripción de Eldén, 2007.

3.2. Análisis de sentimientos

El análisis de sentimientos (SA), inicialmente marcado por métodos léxicos y basados en reglas, es por lo general soportado en diccionarios con baja capacidad para captar orientaciones de sentimientos específicas del contexto; además, con inconvenientes en la creación de *corpus* extensos, una tarea poco práctica (Khan et al., 2016). En este enfoque, la puntuación global del sentimiento se estima construyendo una relación entre la frecuencia del sentimiento y la suma de los sentimientos positivos y negativos (Gupta et al., 2020).

Este enfoque léxico, posee varias limitaciones dado que se basa en la identificación de palabras o frases clave que están asociadas con sentimientos positivos, negativos o neutrales. Una limitación, es que los diccionarios de palabras clave pueden ser incompletos o inexactos, llevando a interpretaciones incorrectas del sentimiento del texto, en segundo lugar, la creación de *corpus* extensos, necesarios para entrenar estos modelos, puede ser una tarea costosa y laboriosa (Khan et al., 2016).

El enfoque léxico también presenta dificultades por la ambigüedad, la cual es manifestada en dos casos: homonimia y la polisemia. En general, la homonimia es la coincidencia entre palabras similares en su escritura o pronunciación, pero que pueden diferir en sus sentidos, estas se presentan a su vez como homógrafo (coincidencia gráfica en la escritura) y homófono (coincidencia en pronunciación, diferencia en la escritura), en cuanto a la polisemia esta refleja la coincidencia de palabras en su escritura y pronunciación, y un sentido relacionado con el mismo significado (Torres & de Alba, 2023).

Por lo anterior, han tomado fuerza un segundo enfoque con técnicas basadas en la clasificación de textos, la primera es el aprendizaje automático supervisado (SML), utilizando métodos como Naïve Bayes, máxima entropía o máquina de vectores de soporte (SVM), en segundo lugar se encuentra el aprendizaje automático no supervisado (UML), por último el aprendizaje automático semi-supervisado (Appel et al., 2016).

El aprendizaje automático supervisado, es un enfoque que se basa en el entrenamiento de un modelo a partir de un conjunto de datos de texto etiquetado con el sentimiento correspondiente(Tripathy et al., 2016), además, con ventajas sobre el enfoque léxico, siendo más preciso, ya que el modelo aprende a identificar patrones de sentimiento en el texto, en lugar de depender de un diccionario de palabras clave y no requiere la creación de un *corpus* extenso, ya que el modelo puede aprender a partir de un conjunto de datos relativamente pequeño. Sin embargo, el aprendizaje automático supervisado también tiene algunas limitaciones cómo la disponibilidad de un conjunto de datos etiquetado, lo que puede ser difícil o costoso de obtener, además de que la calidad de los datos puede afectar el rendimiento del modelo (Khan et al., 2016).

Los enfoques no supervisados y semi-supervisados pueden ser útiles para superar algunas de las limitaciones de los enfoques supervisados. Los enfoques no supervisados no requieren datos etiquetados, sino que agrupan los datos en función de sus similitudes sin tener una variable objetivo predefinida. Por su parte, los enfoques semi-supervisados combinan técnicas supervisadas y no supervisadas, utilizando una pequeña cantidad de datos etiquetados junto con una mayor cantidad de datos sin etiquetar (Khan et al., 2016).

Si bien estos enfoques no requieren datos etiquetados, poseen sus propias limitaciones, ya que tienden a ser menos precisos y pueden ser más difíciles de implementar y entrenar (Khan et al., 2016). Tanto el Aprendizaje Automático como los enfoques basados en léxico poseen características deseables, en este sentido, se han propuesto varios enfoques híbridos que potencien las fortalezas inherentes a las técnicas y logren mitigar sus limitaciones (Khan et al., 2016).

El análisis de sentimiento en textos tiene como objetivo principal la identificación de la polaridad en tres niveles distintos: documento, oración y aspecto. A nivel de documento, se busca determinar los sentimientos analizando el contenido completo del documento, mientras que a nivel de oración, se realiza un análisis más detallado con el propósito de identificar la polaridad de cada oración que compone el documento; por último, el análisis a nivel de aspecto se enfoca en la identificación de aspectos y atributos específicos expresados en las reseñas, clasificando las opiniones de los usuarios con respecto a elementos particulares (Kastrati et al., 2021).

3.3. Procesamiento de texto multilingüe

La diversidad lingüística es un desafío en el procesamiento de texto, ya que la mayoría de las técnicas y enfoques están inicialmente desarrollados para el inglés (Martínez-Cámara et al., 2011). Existen ejercicios con enfoque al idioma español, bajo la creación de *corpus* específico para el análisis de sentimientos y un enfoque semántico, se centran en la identificación de palabras y expresiones con significado emocional, revelando la importancia de adaptar las técnicas de análisis de sentimientos a las particularidades lingüísticas de cada idioma (Martínez-Cámara et al., 2011).

Para otros idiomas, cómo el alemán, se encuentran ejercicios donde un traductor automático convierte los textos al inglés y posteriormente se entrenan los modelos en este último con herramientas como *SentiWordNet* para asignar polaridad (Martínez-Cámara et al., 2011). Por último, se encuentran múltiples trabajos en diversos idiomas, desde análisis de sentimientos en chino con enfoques basados en reglas y aprendizaje automático (*SVM, Naïve Bayes y Decision Tree*) hasta los enfoques gramaticales en árabe, chino e inglés, utilizando noticias financieras, o el uso de *corpus*

cómo *EmotiBlog* de comentarios en español, inglés e italiano sobre diversos temas para comprender las sutilezas del análisis de sentimientos en un contexto determinado. (Martínez-Cámara et al., 2011).

En general los *corpus* y la generación de conjuntos de datos con emociones en español pueden ser un reto, mediante validación con evaluadores como kappa de Fleiss y la herramienta de *Sentic Computing BabelSenticNet* se han procesado comentarios y reacciones en este idioma en fuentes de gran diversidad como Facebook (Tessore et al., 2022).

3.4. Procesamiento de cartas de máximos responsables e informes de sostenibilidad

Las cartas de máximos responsables presentes en los informes de sostenibilidad desempeñan un papel crucial en la rendición de cuentas anuales de las compañías. Estas poseen elementos textuales que enriquecen la comunicación de las empresas a sus grupos de interés dado que son dirigidas directamente a sus lectores, contienen características que aumentan la credibilidad, como autoría explícita, apelación personal y opiniones del autor, destacando hechos y pruebas. Además, son consideradas la parte más leída y difundida del informe anual (Dontcheva-Navratilova et al., 2020).

El análisis de sentimiento en los informes de sostenibilidad, puede desempeñar un papel crucial en la toma de decisiones, especialmente para las partes interesadas de una empresa. Existen dos enfoques en el análisis de sentimientos, en primer lugar, utilizan la clasificación de documentos, donde se categoriza la polaridad de los textos (Harymawan et al., 2020). Un segundo enfoque es la clasificación de oraciones o cláusulas, que permite una evaluación más detallada de las opiniones expresadas en un texto dado que las procesa de forma individual, en categorías de sentimientos subjetivos u objetivos, donde dicha forma de análisis, permite encontrar más detalle e información en las opiniones expresadas en un texto (Liu, 2010).

Otros procesos, toman todo el informes de sostenibilidad para identificar temas comunes mediante enfoques que usan métodos cómo LDA; identificando temas y su distribución, LDA ayuda a encontrar una mezcla de temas en cada documento, utilizando una combinación de términos para describir cada tema, utilizando la implementación *Mallet* de LDA, que estima hiperparámetros y determina el número de temas (Szekely & Brocke, 2017).

Un esfuerzo reciente por capturar información en las cartas de los máximos responsables lo hace el *National Bureau of Economic Research* (NBER) que aplicó la técnica BERT. Este modelo de aprendizaje profundo versátil, pre-entrenado en un *corpus* de texto en inglés y luego adaptado al lenguaje empresarial de las cartas de los accionistas, fue utilizado para clasificar los párrafos de las cartas en términos de contener o no contener una meta. Además, se empleó un enfoque de clasificación multi-clase y multi-etiqueta para identificar diferentes tipos de metas en los párrafos clasificados (Rajan et al., 2023).

3.5. Medición clásica del riesgo en activos financieros

Esta última sesión aborda el modelo de valoración de activos de capital (CAPM, Capital Asset Pricing Model), cuyos indicadores son ampliamente usados en medición de riesgo. Este es un modelo simple que busca describir la relación entre los activos, los ingresos y el riesgo, en especial, uno de los conceptos relevantes para el trabajo es el riesgo sistemático no diversificable expresado en el coeficiente de riesgo beta (β), cuando el mercado de capitales se encuentra en un equilibrio competitivo. El coeficiente refleja la sensibilidad de los retornos de una empresa en un mercado de activos determinado ante los movimientos en el mercado principal y es considerado inherente e inevitable en el sistema financiero global. (Jianbao & Jingjie, 2009).

A continuación se definen las ecuaciones clásicas de Beta (β) y Alpha (α) en el modelo CAPM (eToro, 2023).

La ecuación CAPM básica es la siguiente:

$$R = R_f + \beta(R_m - R_f) \tag{1}$$

β puede calcularse así:

$$\beta = \frac{\text{Cov}(R, R_m)}{\text{Var}(R_m)} \tag{2}$$

α puede calcularse así:

$$\alpha = R - R_f - \beta(R_m - R_f) \tag{3}$$

Donde:

- R es el retorno esperado del activo.
- R_f es el retorno sin riesgo o asegurado de la acción.
- R_m es el retorno del mercado o índice de referencia.
- β es el riesgo de la acción en comparación con el mercado.
- $\text{Cov}(R, R_m)$ es la covarianza entre el retorno del activo R y el retorno del mercado R_m.
- $\text{Var}(R_m)$ es la varianza del retorno del mercado R_m.

Sin embargo, más allá del riesgo sistemático, que en teoría es inevitable para una empresa, existe también el riesgo idiosincrático o específico de la empresa, que juega un papel crucial en la dinámica de las inversiones; este riesgo puede ser atribuido a factores únicos de una empresa o industria, como la gestión empresarial, decisiones estratégicas o exposición a eventos sectoriales (Lee et al., 2023). A diferencia del riesgo sistemático, el riesgo idiosincrático puede mitigarse significativamente mediante la diversificación adecuada de la cartera. Por otro lado, este riesgo ha sido vinculado con retornos de acciones positivos, como se especifica en el estudio de (Lee et al., 2023), donde Merton encontró una relación positiva entre el riesgo idiosincrático y los rendimientos de las acciones, además de que investigaciones posteriores han demostrado que el exceso de rendimientos del mercado puede explicarse a través del riesgo idiosincrático (Lee et al., 2023).

En este sentido, si (β) representa la medida de riesgo inevitable (sistemático) al que una inversión está expuesta debido a los movimientos generales del mercado, el intercepto Alpha (α) puede determinar una proporción del riesgo que es idiosincrático; en esencia, α refleja el rendimiento adicional que no puede ser explicado por β y, por lo tanto, puede capturar parte de los fenómenos específicos de una empresa o industria que afectan su rendimiento (Lee et al., 2023).

En una aproximación al panorama colombiano, (A. & Rojas, 2022) indica que α debería no ser significativamente diferente de cero, lo que indicaría que el rendimiento adicional es atribuible a la gestión de riesgos idiosincráticos específicos de la empresa, por lo tanto, el α puede ser una medida indicativa de aquellos factores específicos que las empresas pueden mitigar, ya sea interna o externamente, a través de estrategias de gestión y decisiones operativas (A. & Rojas, 2022). Este, al igual que el β, según sea su signo, indica si una empresa ha logrado generar una rentabilidad extra, ha incurrido en pérdidas adicionales o ha tenido un rendimiento conforme al riesgo sistemático (McNulty, 2024).

4. Metodología

La Metodología CRISP-DM (IBM, 2023), vista en la figura 2, concebida por IBM, es un enfoque integral y estandarizado para la minería de datos. Su estructura, compuesta por seis fases interconectadas, para efectos de este trabajo solo se toman 5 fases, excluyendo la fase final de despliegue. En el contexto específico de este proyecto centrado en el análisis de cartas de máximos responsables en informes de sostenibilidad, la Metodología CRISP-DM ofrece una estructura sólida y reconocida para abordar cada etapa del proceso de extracción y análisis de información clave.

4.1. Entendimiento del negocio

En esta fase inicial, se busca comprender a fondo el contexto, centrando la necesidad de analizar las cartas de los máximos responsables de empresas colombianas, por lo tanto, reconocer la importancia estratégica de estas cartas en la comunicación empresarial y su posible influencia en indicadores bursátiles, especialmente en compañías que forman parte del COLCAP. Se destaca la diversidad de información presente en estas cartas, como logros, metas, y estrategias empresariales a lo largo del tiempo.

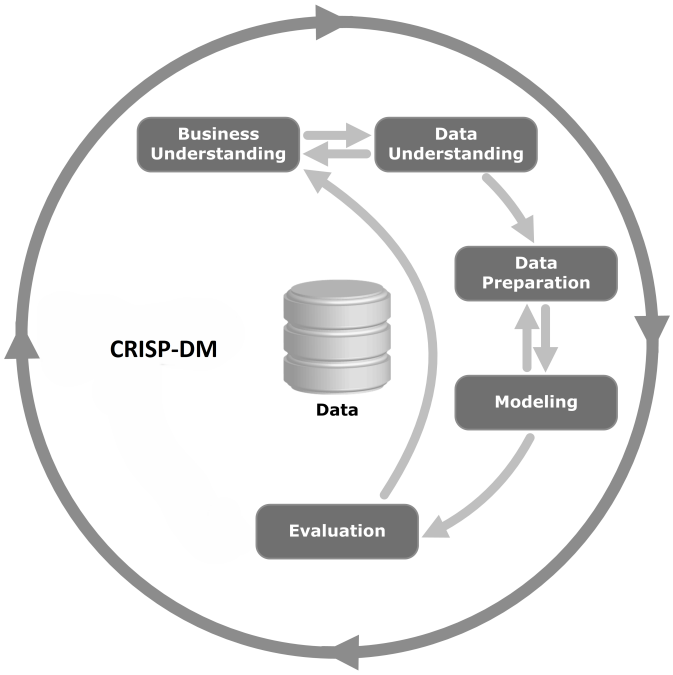

Figura 2: Adaptación del modelo CRISP-DM de IBM sin la fase de despliegue. (IBM, 2023).

El COLCAP, originalmente fue el principal índice bursátil de la Bolsa de Valores de Colombia desde su creación en 2008 hasta el 27 de mayo de 2021, por lo qué es ideal para el periodo de tiempo analizado en este trabajo. Este índice reflejaba el comportamiento de las 20 acciones más líquidas del mercado colombiano, ponderadas por su capitalización bursátil ajustada y es la referencia principal del mercado accionario colombiano durante ese período, proporcionando una medida del desempeño general de las acciones más representativas del país. Su nueva versión es calculada por MSCI en colaboración con la Bolsa de Valores de Colombia (Bolsa de Valores de Colombia, 2024).

4.2. Entendimiento de los datos

La segunda fase se enfoca en explorar y comprender los datos disponibles para el análisis, en este caso, las cartas de los máximos responsables, como se observa en el cuadro 2. Este ejercicio toma 21 empresas y explora 217 cartas entre los años 2012 a 2022 con algunas variaciones por la no disponibilidad de algunas empresas en algunos de los primeros años, todas estas son obtenidas desde los repositorios de cada una de las paginas web de las empresas, posteriormente se extrae de los informes las cartas en formato *.txt* donde cada carta representa un documento. Posteriormente al involucrarse el contraste con los resultados bursátiles, se retiran las cartas de aquellas empresas que no se encuentran datos de forma pública.

Año	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012
Cartas Iniciales	21	21	21	21	21	20	20	19	19	17	17
Cartas Finales	20	20	20	20	20	19	19	18	17	15	15

Cuadro 2: Documentos tomados por año (2012-2022)

También una clasificación relevante para los análisis que se abordan en la sección de resultados son los sectores en los que se agrupan las empresas, ver cuadro 3. Se realiza la selección de sectores que logra menor dispersión en el análisis de los resultados que un agrupamiento por industria, donde existen organizaciones que tienen industrias únicas para ellas, de esta forma se trabaja con 3 segmentos, los sectores energético y financiero contienen 14 de las 20 empresas totales y una categoría de otros sectores que agrupa los sectores restantes:

Sector	Número de empresas
Energético	7
Financiero	7
(Otros) Conglomerado	2
(Otros) Consumo masivo	1
(Otros) Materiales	2
(Otros) Telecomunicaciones	1

Cuadro 3: Conteo de empresas por sectores

4.3. Preparación de los datos

La preparación de los datos implica la transformación de las cartas a un formato procesable mediante técnicas de minería de texto y NLP. Se busca desarrollar métricas que extraigan información relevante, como identificación de temas recurrentes, cambios de discurso en el tiempo, y cuantificación de la polaridad de las expresiones en las cartas. La complejidad de la estructura y contenido de las cartas plantea desafíos en la transformación de datos textuales en información numérica.

A continuación, se nombran los principales pasos en el preprocesamiento del texto e insumo principal de las métricas:

4.3.1. Proceso de lectura y transformación inicial de datos

Carga de cartas/documentos: Cargar cada carta almacenadas en archivo *.txt*, abrir archivo con codificación específica *Latín 1*, convertir contenido a *UTF-8* y limpiar caracteres no válidos.

Crear registros individuales por documento: Para cada contenido procesado extraer el nombre de archivo y crear registro con (empresa, año, contenido de texto). Por último combinar todos los registros en un solo conjunto de datos y almacenar en *.csv*, una fila por documento de empresa por año.

4.3.2. Proceso de preparación de texto

Limpieza de texto: Reemplazar y eliminar caracteres o palabras no deseadas en el siguiente orden:

- Eliminar guiones.
- Convertir a minúsculas.

- Eliminar números, puntuación, caracteres especiales, espacios extra, saltos de línea, símbolos de moneda, números romanos y palabras de una sola letra.
- Eliminar comillas dobles.

4.3.3. Preprocesamiento de texto

El siguiente proceso es almacenado en una variable de salida con el texto preprocesado y lematizado.

Tokenización: Se convierte el texto en una lista de palabras o *tokens*, considerando solo palabras con más de tres letras y que sean alfabéticas, es decir, se asegura de que los *tokens* resultantes sean realmente palabras y no fragmentos de texto que podrían incluir otros caracteres no deseados.

Remoción de *stopwords* y palabras específicas: Se eliminan palabras comunes que no aportan significado al análisis, como artículos y preposiciones, así mismo, se crea una lista de palabras específicas para eliminar, que no aportan al análisis como nombres de empresas o términos financieros comunes.

Lematización: Reduce las palabras a su forma base, simplificando la representación del texto y agrupando variantes de una palabra en una forma común.

4.3.4. Indexación

Asigna índices a las palabras procesadas, creando una representación numérica del texto para facilitar el análisis. Para este caso, se emplea la Matriz Tf-Idf (Frecuencia de Término - Inversa de Frecuencia de Documento) para la transformación del documento a un formato procesable. Esta matriz, es dispersa y de alta dimensionalidad, dado que es aplicada a la variable de texto preprocesado y lematizado, asigna pesos a las palabras según su importancia relativa en el documento y en el conjunto de documentos(Eldén, 2007).

4.3.5. Similitud y evaluación

Una forma de reducir dimensionalidad es el uso de una matriz de similaridad por el método de coseno, una medida usada a menudo para estos ejercicios (Eldén, 2007). El calculo de la similitud de coseno se da entre los vectores TF-IDF de todos los pares de documentos, para obtener la matriz de similitud, posteriormente se calcula el número de condición de la matriz de similitud, para evaluar su estabilidad numérica y la dependencia lineal entre sus filas.

4.3.6. Recolección y procesamiento de datos accionarios del COLCAP

Para contrastar los resultados del análisis de las cartas de los máximos responsables, se utilizó el histórico de los rendimientos de las acciones en el mismo periodo de tiempo que los documentos analizados. Los datos fueron obtenidos de fuentes públicas y disponibles [5]. Se extrajo la serie temporal de los precios mensuales de las acciones y se calcularon sus rendimientos. Además, se obtuvieron los datos del indicador de referencia (COLCAP), que se emplearon para determinar el β y α , permitiendo así identificar tanto el riesgo sistémico como el posible riesgo idiosincrásico

de cada acción. Estos valores se obtienen mediante el proceso de contrastar cada resultado de las acciones con el índice COLCAP.

En este análisis, se han utilizado los valores deβ y α de las acciones, obtenidos de plataformas públicas como Investing.com (Investing.com, 2024) y Yahoo Finance.com (Yahoo Finance, 2024). Estos datos se han elegido por su accesibilidad pública y su actualización constante, lo que los convierte en una fuente confiable y práctica para obtener información financiera actualizada.

4.4. Modelado

En esta fase se implementan técnicas de aprendizaje automático y NLP para analizar patrones y tendencias en el contenido de las cartas. Esto incluye la aplicación de algoritmos no supervisados combinados con enfoques léxicos y de análisis de sentimiento, dándole atención al desarrollo de modelos que puedan cuantificar la influencia del contenido textual.

Primero, se obtiene una métrica que evalúa la similitud entre los documentos de las mismas empresas y en orden cronológico, utilizando la métrica de similitud del coseno. Posteriormente, se recurre a modelos específicos de análisis latente semántico para descubrir temas subyacentes en los documentos y análisis de sentimiento utilizando *TextBlob* y *VADER* que evalúan la polaridad y subjetividad del contenido.

4.4.1. Evaluación de la similitud entre documentos consecutivos

Si bien el cálculo de similaridad se hace para todos los documentos, dado que esto es el insumo de varios modelos, para el ejercicio, también es importante extraer la métrica de los documento respecto a su año anterior, esto para cada empresa y por año ordenado cronológicamente, usando empresa y año como índice para esta matriz.

1. Se inicializa una lista de similaridades, comenzando con un valor de 0 para el primer documento (ya que no tiene predecesor).
2. Para cada documento, desde el segundo en adelante, se compara con el documento anterior:
    a. Si ambos documentos son de la misma empresa, se calcula la similaridad del coseno entre estos dos documentos usando la matriz de similitud y se agrega a la lista.
    b. Si son de diferentes empresas, se añade un 0 a la lista (indicando ninguna similaridad con un documento de otra empresa).

4.4.2. Análisis latente semántico

El análisis latente semántico (LSA), se emplea para descubrir relaciones subyacentes en el contenido, este modelo permite reducir la dimensionalidad de los datos y extraer temas ocultos que pueden no ser evidentes a través del análisis superficial, por lo que se utiliza LSA para identificar patrones temáticos en las cartas.

Esto se mide en un coeficiente que está entre -1 y 1 donde los valores positivos indican que un tema es relevante o significativo para el documento, los valores negativos pueden sugerir que el tema es menos relevante o tiene una relación inversa con el contenido del documento, por último Valores cercanos a cero implican que el tema tiene baja o nula relevancia con el documento.

[5]https://www.investing.com/ https://finance.yahoo.com/.

4.4.3. Análisis de sentimiento con TextBlob

El primer modelo de analisis de sentimiento que se usa es *TextBlob*, que utiliza *corpus* y diccionarios preentrenados para realizar análisis de sentimiento en el contenido de las cartas, esta es una herramienta de procesamiento de lenguaje natural que permite analizar la polaridad (positiva, negativa, neutro) y la subjetividad del texto. El objetivo es cuantificar cómo el sentimiento expresado en las cartas con un *score* específico, que se traduce a una polaridad (positivo, negativo, neutro), puede proporcionar una métrica adicional para evaluar la comunicación corporativa.

4.4.4. Análisis de sentimiento con VADER

El segundo modelo de Análisis de Sentimiento que se usa es *VADER*, que utiliza diccionarios pre-entrenados. VADER es una herramienta de procesamiento de lenguaje natural optimizada para texto corto, que permite analizar la polaridad y la intensidad del sentimiento, a través de un *score* entre -1 y 1.

4.5. Evaluación

Durante la fase de evaluación, se analizaron los resultados obtenidos a partir de las métricas desarrolladas y los modelos implementados, que se almacenan por cada documento de empresa por cada año en un formato tabular, para completar los resultados se utiliza la medición del riesgo del activo mediante el cálculo de β, que representa la sensibilidad de la rentabilidad del activo con respecto al mercado, y α, que mide el rendimiento adicional del activo sobre el rendimiento esperado, estos cálculos se realizan bajo una regresión lineal de cada uno de los rendimientos mensuales de cada empresa en el periodo de un año respecto a los rendimientos del COLCAP arrojando ambas métricas (Ossa González & Rojas Dominguez, 2023).

5. Resultados

Esta sección se divide en tres partes principales: en primer lugar, se exploran los efectos de la mejora en el procesamiento de documentos en español desde su formato inicial hasta obtener los insumos finales para el modelado. Posteriormente, se discuten las métricas y los modelos utilizados para evaluar estos documentos, finalmente, se analizan los resultados obtenidos de combinar los modelos de NLP con enfoques centrados en el riesgo sistemático e idiosincrático, calculado a partir de los resultados de las acciones en el COLCAP.

5.1. Lectura, preprocesamiento y procesamiento de texto

Una de las métricas más indicativas de la efectividad del proceso fue el número de condición en la matriz de similitud. Este número ayuda a detectar la sensibilidad de la matriz frente a errores de redondeo, se observa que a medida que se mejora la limpieza de los textos, el número de condición disminuye, lo que indica una mejora en la calidad de los datos para el análisis.

5.1.1. Evolución del número de condición

A continuación, se presenta un resumen de cómo el número de condición evolucionó a través de las etapas del proceso de aplicado a los textos:

1. El número de condición inicial era mayor a 500, debido a que los textos se copiaban directamente de los archivos *.pdf* a archivos *.txt* sin considerar el orden o la integridad del contenido, incluyendo a menudo símbolos inusuales dado que algunos archivos *.pdf* poseen formatos de diseño que se traducen en símbolo y caracteres ilegibles.

2. Se logró una reducción del número de condición a aproximadamente 291 cuando los textos fueron cuidadosamente copiados por segmentos para minimizar problemas de transcripción y se realizaron conversiones OCR necesarias cuando los documentos eran imágenes.

3. El número de condición se redujo a 205 al eliminar el proceso de *stemming*, el cual a menudo fragmentaba los *tokens*, reduciendo su relevancia y comprensibilidad.

4. Mediante la implementación de una serie de pasos de limpieza descritos detalladamente en la metodología, se logró disminuir el número de condición a menos de 100.

5. Finalmente, se ajustó la lista de palabras a remover en cada iteración del proceso. Estas palabras, aunque por sí solas no impactaban significativamente el número de condición, sí afectaban la calidad de los modelos de NLP al repetir *tokens* que no eran relevantes para el análisis.

5.2. Revisión de resultados usando métricas del COLCAP

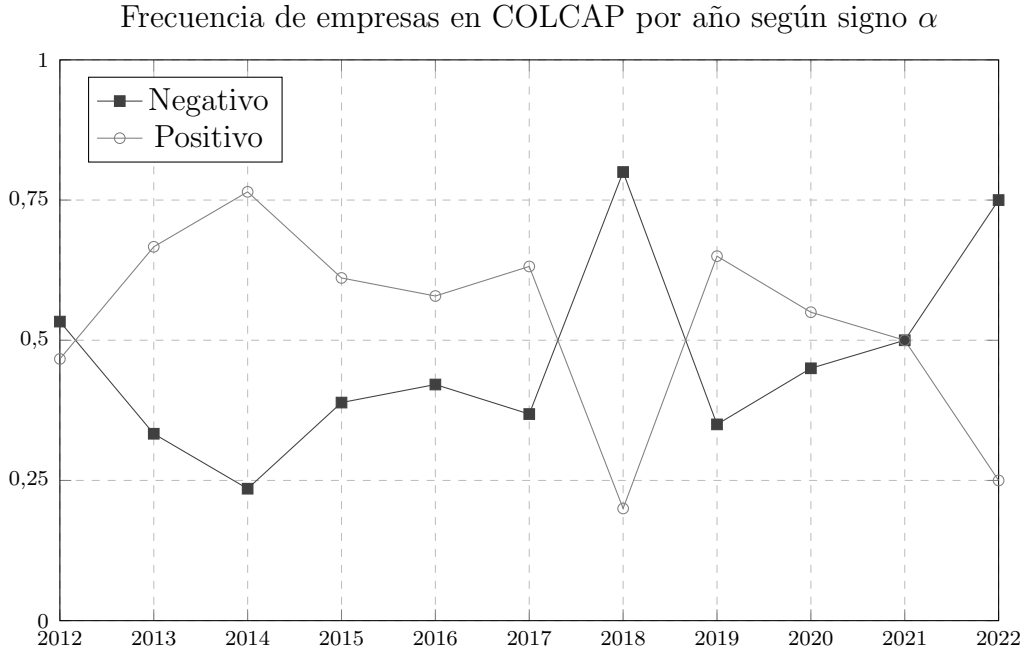

Figura 3: Evolución de la proporción de signo α desde 2012 hasta 2022.

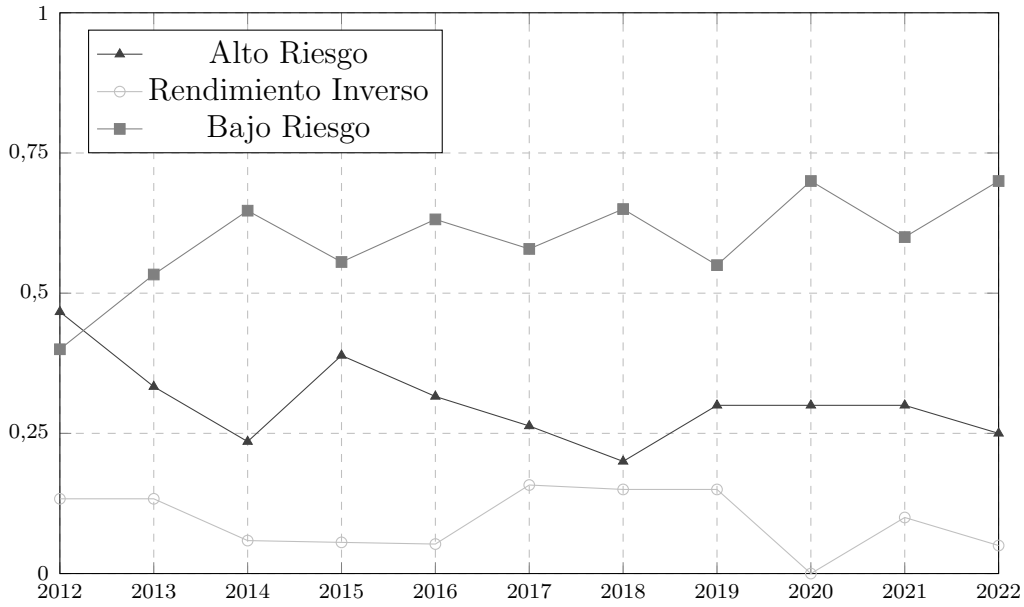

Figura 4: Frecuencia de empresas en COLCAP por año según riesgo de activo β.

En esta subsección, se examinan las fluctuaciones de los indicadores financieros de las empresas en el índice COLCAP. Este análisis es fundamental para evaluar las posibles conexiones entre los discursos de los máximos responsables de las empresas y los movimientos en el mercado de valores. Específicamente, se discuten los resultados anuales de la polaridad de α, en la figura 3, y el coeficiente de riesgo sistémico β, en la figura 4, presentando ambos en términos de frecuencias anuales (porcentaje de empresas anuales). Estos indicadores sirven como insumos para los análisis subsiguientes en esta sección, permitiendo una comparación cuantitativa de cómo las percepciones y estrategias expresadas en los discursos corporativos podrían influir o reflejar el comportamiento del mercado.

Descripción general de la evolución en las frecuencias de empresas en el índice COLCAP por año según el rendimiento de una inversión en relación con el mercado o índice de referencia α :

- Signo negativo: Se observa una notable fluctuación en la proporción de empresas con α negativo, con un pico significativo en 2018 del 80 %, seguido de un descenso a valores anteriores en 2019 (35 %) y volviendo a subir para cerrar en 75 % en el año 2022.
- Signo positivo: La frecuencia de empresas con α positivo muestra una disminución correspondiente, especialmente marcada en 2018 y 2022 con 20 % y 25 % respectivamente. Esto podría interpretarse como un aumento en la cautela entre los inversores o una reevaluación de las expectativas de crecimiento.

Descripción general de la evolución en las frecuencias de empresas en el índice COLCAP por año en relación con los niveles de riesgo de activo β:

- Alto Riesgo: La proporción de empresas clasificadas como de alto riesgo ha mostrado una tendencia general a la disminución entre 2012 y 2018 (47 % a 20 %) con una leve variación en el año 2015. posteriormente se estabiliza en cerca del 30 % hasta cerrar en el año 2022.

- Bajo Riesgo: Este grupo ha visto un aumento en su proporción hasta el año 2018 (40 % al año 2012 y cerrando en 65 %), en el año 2019 cae a 55 % y se estabiliza con sus mayores proporción de empresas en los años 2020 y 2022 (70 %).
- Rendimiento Inverso: Las empresas con rendimientos opuestos al mercado general mantienen una frecuencia baja y constante, lo que puede indicar estabilidad en el número de firmas que contrarrestan la tendencia general del mercado. sin embargo, en el año 2020 ninguna de las empresas perteneció a este segmento de riesgo.

5.3. Similaridad semántica entre documentos anuales

La consistencia en la comunicación corporativa puede reflejar estabilidad y coherencia en las estrategias, operaciones y toma de decisiones de una empresa. Para esto, se analiza la similaridad semántica utilizando la similitud del coseno, que como métrica, muestra cómo los temas tratados en los documentos se mantienen o varían año tras año. Esta medida no solo refleja la coherencia temática, sino que también puede indicar cambios estratégicos o adaptaciones a nuevas normativas o condiciones de mercado.

Sector	**Menor 10 %**	**10 %-20 %**	**20 %-30 %**	**30 %-40 %**	**40 %-50 %**	**50 %-60 %**	**Mayor 60 %**	**Total**
Energético	2.19 %	7.10 %	12.02 %	12.02 %	2.19 %	2.19 %	0.55 %	35.52 %
Financiero	0.55 %	3.83 %	6.56 %	12.02 %	7.10 %	0.55 %	0.55 %	32.24 %
Otros	1.09 %	3.83 %	12.57 %	7.10 %	2.19 %	3.83 %	1.64 %	32.24 %
Total	3.83 %	14.75 %	31.15 %	31.15 %	11.48 %	4.37 %	3.28 %	100.00 %

Cuadro 4: Distribución de la similaridad semántica entre documentos consecutivos por sector

Los resultados se segmentan en distintos rangos de similaridad, proporcionando una visión detallada de la consistencia en las comunicaciones entre años. Esta distribución de similaridades sugiere variaciones significativas en la manera en que las empresas en diferentes sectores responden a sus entornos operativos y de mercado. Un mayor porcentaje podría interpretarse como una estabilidad en la visión y estrategia corporativa, donde los cambios significativos son mínimos. Por otro lado, aquellos con similaridades bajas pueden estar asociados a cambios dinámicos en sus operaciones o en el entorno regulatorio, lo que podría indicar una adaptación estratégica a condiciones cambiantes o una reorientación en las prioridades corporativas.

Para efectos del análisis, dada la distribución de frecuencia general de las similidades entre documentos, se clasifica según su valor específico en los siguientes rangos:

1. Similaridad baja: Se presenta en la cola izquierda de la distribución por debajo del 20 % de similaridad donde se encuentra el 18.6 % de los documentos.
2. Similaridad moderada: Entre el 20 % y el 40 % de similaridad se encuentra la mayor frecuencia de documentos (62.3 %), en consecuencia, presentan cambios de discurso moderados.
3. Similaridad alta: Se presenta en la cola derecha de la distribución con porcentajes de similaridades mayores al 40 % y con una frecuencia documental del 19.1 % (muy similar a la frecuencia documental de la cola izquierda).

- Durante años con predominancia de similaridad moderada, como 2014 a 2017 (68.75 % a 73.68 %), se puede observar una posible relación con la estabilidad de α donde las proporciones de empresas con resultados negativos y positivos es estables. Estos años podrían indicar

Rango de similaridad	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Alta	21.43 %	12.50 %	11.67 %	11.11 %	15.79 %	15.79 %	20 %	25 %	30 %	25 %
Baja	35.71 %	18.75 %	17.65 %	16.67 %	10.53 %	21.05 %	20 %	20 %	15 %	15 %
Moderada	42.86 %	68.75 %	70.59 %	72.22 %	73.68 %	63.16 %	60 %	55 %	55 %	60 %

Cuadro 5: Frecuencia de documentos por rango de similaridad semántica y por año

que, aunque hubo ajustes y discusiones estratégicas, las empresas mantuvieron un enfoque relativamente consistente.

- En los años 2019 a 2022, la similaridad alta muestra un aumento constante de 20 % a 30 %, mientras que en los mismos años, se observa un aumento en la proporción de empresas con β de bajo riesgo, alcanzando hasta el 70 % en 2020 y 2022. Esto sugiere una coincidencia numérica entre mayor consistencia temática y disminución del riesgo.
- En el año 2018, el porcentaje de documentos con similaridad baja fue de 21.05 %, un año caracterizado también por un pico en la frecuencia de signo negativo de α 80 %. Este es un dato numérico que muestra una concurrencia entre baja similaridad semántica y un alto porcentaje de α negativo.

5.4. Encontrando relaciones con LSA

El LSA es una técnica utilizada para descubrir las relaciones subyacentes entre términos y documentos, esto basándose en la premisa de que los términos que aparecen en los mismos contextos tienden a tener significados relacionados, en este caso, se han identificado y agrupado los temas predominantes en las cartas anuales de las empresas, proporcionando una visión detallada de los enfoques temáticos y las tendencias discursivas en los diferentes años.

El número óptimo de temas fueron 5, donde cada uno refleja una dimensión particular de las preocupaciones y enfoques estratégicos por sectores. A continuación, se presentan algunas observaciones clave sobre estos temas, según la distribución de los documentos por sectores y su coeficiente temático que son clasificados en rangos para sintetizar el análisis.

5.4.1. Tema 1: gestión y desarrollo empresarial

Las palabras claves de este tema son: empresa, resultado, gestión, desarrollo, crecimiento, valor, negocio, sostenibilidad, financiero, mercado.

Sector	0.48/0.6	0.6/0.7	0.7/0.8	0.8/0.9	0.9/0.98	Total
Energético	0 %	4.42 %	17.49 %	43.58 %	34.50 %	100.00 %
Financiero	8.36 %	15.17 %	11.22 %	26.51 %	38.74 %	100.00 %
Otros	5.10 %	2.46 %	18.45 %	39.94 %	34.04 %	100.00 %
Total	4.26 %	7.16 %	15.83 %	37.06 %	35.68 %	100.00 %

Cuadro 6: LSA Tema 1 - frecuencia documental de los coeficientes que indican la cercanía temática con el Tema 1.

- Este tema es predominante en la mayoría de los documentos, como se muestra en cuadro 6, los coeficientes son positivos y todos sobre el 0.48, indicando que el tema de gestión y desarrollo empresarial es consistentemente dominante y relevante en las cartas anuales de las empresas

- El 72 % de los documentos tienen un coeficiente mayor al 0.8, esto podría reflejar una fuerte concentración en la gestión efectiva, desarrollo de negocio, y crecimiento sostenible a lo largo de los años.

5.4.2. Tema 2: finanzas y economía

Las palabras claves de este tema son: tasa, inflación, cartera, peso, fiscal, precio, semestre, dólar, gasto, desempleo.

Sector	**-0.55/-0.3**	**-0.3/-0.1**	**-0.1/0**	**0/0.1**	**0.1 / 0.3**	**0.3/0.77**	**Total**
Energético	8.33 %	56.94 %	11.11 %	15.28 %	6.94 %	1.39 %	100.00 %
Financiero	24.24 %	12.12 %	9.09 %	9.09 %	21.21 %	24.24 %	100.00 %
Otros	26.15 %	32.31 %	10.77 %	6.15 %	18.46 %	6,15 %	100.00 %
Total	19.21 %	34.48 %	10.34 %	10.34 %	15.27 %	10,34 %	100.00 %

Cuadro 7: LSA Tema 2 - frecuencia documental de los coeficientes que indican la cercanía temática con el Tema 2.

- Los coeficientes muestran una amplia gama de valores negativos y positivos desde moderados hasta significativos, por lo que los resultados varían ampliamente entre documentos y sectores.
- El sector energético acumula un 56 % de sus documentos con negatividad moderada entre -0.3 a -0.1, también cuenta con un porcentaje muy bajo de cartas con coeficiente positivo que se pueda considerar moderado, en este sentido y dado su sesgo hacia valores negativos podría reflejar preocupaciones o retos en el contexto económico durante dichos años, como los desafíos o perspectivas respecto a la volatilidad de los precios y otras dinámicas del mercado energético.
- El sector financiero cuenta con una distribución mucho más uniforme, sin embargo, presenta comportamientos donde muestra mayor porcentaje de documentos acumulándose en sus extremos de negatividad y positividad de moderada a alta, causado por coeficientes consecutivamente altos o bajos en empresas del sector que se comportan diferente.
- En otros sectores podemos ver una acumulación del 58 % de los documentos entre negatividad moderada a alta, este fenómeno esta jalonado por dos sectores específicos que contienen cuatro empresas, el primer sector es el de materiales que contiene el 100 % de sus documentos con un coeficiente entre (-0.55 a -0.1) y el segundo el sector de conglomerados que tiene el 70 % de sus documentos con coeficiente negativo.

5.4.3. Tema 3: Servicios financieros y atención al cliente

Las palabras claves de este tema son: vivienda, digital, cartera, cliente, crédito, producto, canal, enriquecer, persona, mensaje.

- El sector financiero muestra una alta proporción de documentos con coeficientes positivos en el rango de 0.3 a 0.6, indicando un enfoque fuerte y positivo en la temática de servicios financieros y atención al cliente, donde el 53.03 % de los documentos en este sector reflejan una alineación alta con la temática.

Sector	-0.6/-0.3	-0.3/0.1	-0.1/0	0/0.1	0.1/0.3	0.3/0.6	Total
Energético	43,06 %	36.11 %	6.94 %	2.78 %	11.11 %	0 %	100.00 %
Financiero	0 %	3.03 %	9.09 %	25.76 %	9.09 %	53.03 %	100.00 %
Otros	0 %	33.85 %	24.62 %	12.31 %	29.23 %	0 %	100.00 %
Total	15.27 %	24.63 %	13.30 %	13.30 %	16.26 %	17.24 %	100.00 %

Cuadro 8: LSA Tema 3 - frecuencia documental de los coeficientes que indican la cercanía temática con el Tema 3.

- En el caso del sector energético, la mayoría de los documentos, cerca del 43 %, muestran coeficientes altamente negativos (-0.6 a -0.3), que muestra un interés contrario o bajo en estos temas, dado que este puede ser distante por el tipo de industria presente en este sector.
- Los otros sectores muestran una variedad de enfoques con cerca del 29.23 % de los documentos alineados positivamente con el tema, indicando una incorporación de prácticas relacionadas con la atención al cliente y los servicios.

5.4.4. Tema 4: información y reportes

Las palabras claves de este tema son: información, informe, aspecto, versión, materialidad, documento, marco, principio, global, interés.

Sector	-0.55/-0.3	-0.3/-0.1	-0.1/0	/0.1	0.1/0.3	0.3/0.59	0.59/0.82	Total
Energético	18.06 %	38.89 %	20.83 %	11.11 %	9.72 %	0 %	1.39 %	100.00 %
Financiero	0 %	31.82 %	27.27 %	10.61 %	25.76 %	4.55 %	0 %	100.00 %
Otros	10.77 %	6.15 %	13.85 %	15.38 %	21.54 %	21.54 %	10.77 %	100.00 %
Total	9.85 %	26.11 %	20.69 %	12.32 %	18.72 %	8.37 %	3.94 %	100.00 %

Cuadro 9: LSA Tema 4 - frecuencia documental de los coeficientes que indican la cercanía temática con el Tema 4.

- El sector energético muestra que un 77 % de los documentos cuentan con un coeficiente de símbolo negativo, este sesgo de negatividad puede sugerir que la información y reportes en este sector pueden tener menor foco.
- Para el sector financiero predomina un enfoque más balanceado, con un 59 % de los documentos con una negatividad moderada y el 41 % de los documentos con coeficiente positivo.
- Para otros sectores, se tienen una variada distribución en los coeficientes, sin embargo, un sector (telecomunicaciones) contiene el 100 % de sus documentos con coeficiente positivo, mientras el resto de sectores (conglomerado, consumo masivo, y materiales) tienen su mayoría de documentos con coeficientes positivos.

5.4.5. Tema 5: responsabilidad social y ambiental

Las palabras claves de este tema son: petróleo, barril, comunidad, crudo, vivienda, reserva, ambiental, producción, minero, comprometido.

- El sector Energético muestra una considerable variabilidad en sus coeficientes, con un coeficiente en temas relacionados con la responsabilidad ambiental entre 0.1 y 0.53 para aproximadamente el 25 % de sus documentos, sin embargo, la negatividad moderada en un rango

Sector	-0.44/0.3	-0.3/-0.1	-0.1/0	0/0.1	0.1/0.3	0.3/0.53	Total
Energético	6.94 %	33.33 %	15.28 %	6.94 %	23.61 %	13.89 %	100.00 %
Financiero	1.52 %	12.12 %	19.70 %	18.18 %	27.27 %	21.21 %	100.00 %
Otros	24.62 %	40.00 %	13.85 %	6.15 %	4.62 %	10.77 %	100.00 %
Total	10.84 %	28.57 %	16.26 %	10.34 %	18.72 %	15.27 %	100.00 %

Cuadro 10: LSA Tema 5 - frecuencia documental de los coeficientes que indican la cercanía temática con el Tema 5.

similar (-0.44 a - 0.1) llega al 40 % de sus documentos, donde podría reflejar debates o críticas sobre la sostenibilidad en las operaciones o retos constantes en este tema para el sector.

- El sector financiero tiende a coeficientes positivos en la discusión sobre responsabilidad social y ambiental con más del 48 % de los documentos en el rango de 0.1 a 0.53, lo que sugiere un enfoque en sostenibilidad y prácticas de responsabilidad social.

- Para los otros sectores se observa una alta incidencia de valores negativos, especialmente en el rango de -0.44 a -0.1, con el 64 % de los documentos, lo que podría reflejar críticas o retos significativos en la implementación de prácticas sostenibles o responsables. En el único sector que se denota coeficientes positivos moderados, es en el de materiales con un 44 % de los documentos en el rango de 0.1 a 0.53 y ninguno de los otros 3 sectores que lo conforman en este rango.

5.4.6. Análisis de temas

Para simplificar el análisis en el periodo de tiempo, se mostrara el porcentaje de empresas por año con coeficientes negativos, para este caso se excluye el tema 1 cuyos resultados son altamente positivos para todas las compañías y existe 0 % en frecuencia de coeficientes negativos.

Coeficiente Negativo	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Tema 2	53 %	80 %	59 %	61 %	68 %	58 %	60 %	70 %	65 %	65 %	65 %
Tema 3	67 %	60 %	65 %	61 %	63 %	53 %	50 %	45 %	45 %	45 %	40 %
Tema 4	33 %	53 %	53 %	33 %	37 %	53 %	65 %	65 %	70 %	70 %	80 %
Tema 5	60 %	53 %	53 %	50 %	53 %	58 %	65 %	60 %	50 %	55 %	60 %

Cuadro 11: Frecuencia documentos con coeficiente negativo por tema y año

- Tema 2 - Finanzas y Economía: Es consistente a lo largo de la serie una alta frecuencia de compañías por año con coeficientes negativos, con un pico en 2013 del 80 % de las compañías, al año siguiente se encuentran los picos más significativos en la frecuencia de coeficientes positivos de α con cerca de 75 % de las empresas con este signo, también es este año 2014 donde se encuentra un resultado alto en la frecuencia de bajo riesgo para β con un 64.7 %. El segundo año con alta frecuencia de empresas (70 %) con signo negativo para el tema 2 es el año 2019, un año donde el coeficiente de α vuelve a ser positivo en la mayor parte de las empresas.

- Tema 3 - Servicios financieros y atención al cliente: La tendencia decreciente en la frecuencia de coeficientes negativos desde un 67 % en 2012 hasta un 40 % en 2022 podría reflejar mejoras en la percepción de los servicios financieros y la atención al cliente.

- Tema 4 - Información y Reportes: Observamos un incremento en la frecuencia de coeficientes negativos, particularmente notorio desde el año 2017 donde los coeficientes positivos pasan a tener menor frecuencia, coincidiendo en este mismo periodo con el pico de frecuencia de valores negativos para α, su tendencia creciente se mantiene hasta 2022, alcanzando el 80 % de frecuencia de empresas con retos significativos en este tema.

- Tema 5 - Responsabilidad Social y Ambiental: Aunque el coeficiente negativo es bastante alto a lo largo del período analizado, fluctuando alrededor del 50 % a 60 %, esta tendencia podría sugerir desafíos constantes, para el periodo de tiempo los valores positivos no alcanzaron en ningún año una frecuencia mayoritaria.

5.5. Extracción de polaridad con análisis de sentimientos

Para el análisis de sentimiento se usa *TextBlob* y *VADER*, dos modelos basados en métodos de aprendizaje automático y análisis lingüístico que determinan la polaridad del sentimiento (positivo o negativo), sin embargo, ambos tienen enfoques ligeramente diferente. Por un lado, *TextBlob* mide la subjetividad del texto, es decir, si el contenido se comporta como una opinión personal que un hecho objetivo y tiende a estar entrenado para texto formal. Por otro lado, *VADER* está especializado en el análisis de textos cortos y con algún grado de informalidad, y tiende a captar la polaridad basado en la intensidad del sentimiento expresado.

Para el presente trabajo se analizan complementariamente los modelos de SA *TextBlob* y *VADER*. La construcción de un modelo híbrido que combine ambos puede ofrecer una evaluación robusta de los sentimientos, este enfoque permite no perder expresiones negativas, que a menudo tienen un impacto significativo. Esta combinación es simple y se basa en dar prioridad a los resultados negativos de cualquiera de los dos modelos, el sistema puede captar de manera efectiva tanto la subjetividad detrás de opiniones más elaboradas como la intensidad emocional en expresiones breves y un poco más coloquiales.

5.5.1. Sistema híbrido usando los modelos de SA *TextBlob* y *VADER*

Este análisis de polaridad a través del modelo híbrido que combina *TextBlob* y *VADER*, y siendo en principio una combinación simple que prioriza los documentos con resultados negativos, en general los resultados enfatizan la prevalencia de la polaridad positiva en los documentos (55 %), lo cual es indicativo de un tono generalmente optimista en las comunicaciones de la empresa. Sin embargo, la presencia de una cantidad significativa de documentos con polaridad negativa (25 %), sugiere que también se abordan temas sensibles o desafiantes, por último un 20 % de los documentos presentan neutralidad.

Si bien la serie de tiempo muestra una predominancia de documentos con polaridad positiva se presentan variaciones que podrían enriquecer el análisis del presente trabajo, un primer momento podría agruparse entre los años 2012 y 2018.

- El porcentaje de documentos positivos disminuyó del 53 % para el año 2012 al 33 % en el año 2015 (estando por debajo del total de documentos neutrales), para luego incrementar al 80 % para el 2018.

- Cómo se evidencia en las figura 6 de*TextBlob* y la figura 7 de *VADER*, se refuerza la observación de una reducción en el tono positivo en las comunicaciones corporativas entre 2012 y

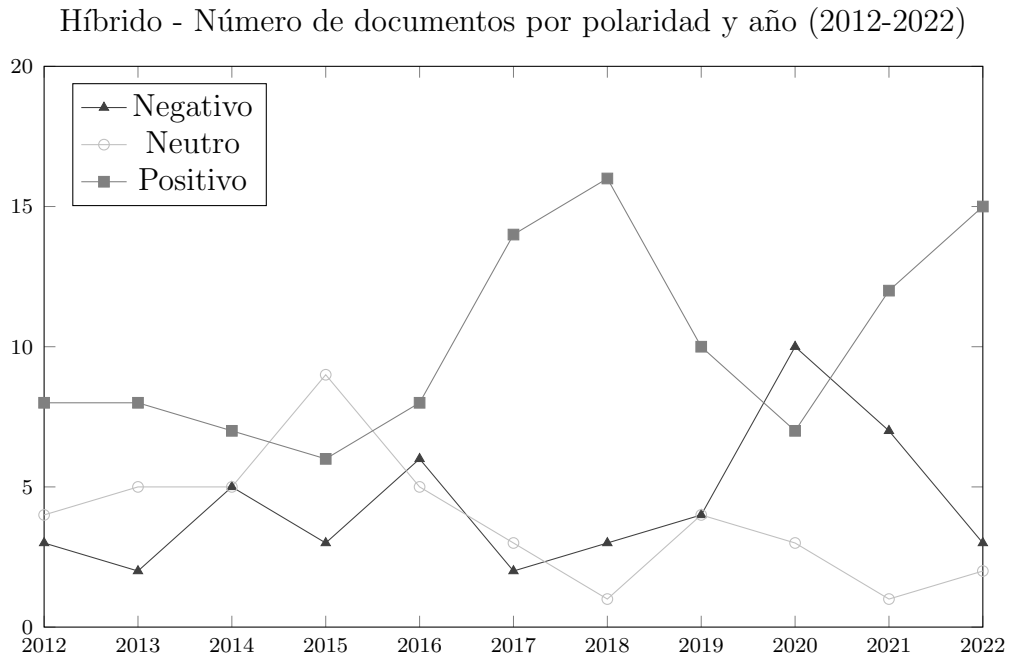

Figura 5: Sistema Híbrido - Número de documentos por polaridad desde 2012 hasta 2022.

2015, pasando del 73 % al 61 % en *TextBlob* y del 66 % al 38 % en *VADER*. También, ambos modelos reflejan un aumento hacia mayor pico de positividad para el 2018.

- El porcentaje de documentos neutrales incrementó del 26 % para el año 2012 al 50 % en el año 2015 (estando por encima de la polaridad positiva), para luego descender al 5 % para el 2018.

- El pico del año 2015 en neutralidad se explica por la influencia de *VADER*, cómo se ve en la figura 7, para este año la mitad de los documentos son neutrales.

- El porcentaje de documentos negativos varió constantemente en el periodo, iniciando en 20 % en el año 2012 y cerrando con 15 % al año 2018, sin embargo, presentó un pico con el 31 % en el año 2016 (único año del periodo donde se encuentra por encima de los documentos neutrales).

- Para el año 2014, la figura 6 muestra que *Textblob* incide en la negatividad del primer pico gráfico en un 29 %, para el pico de negatividad del año 2016 ambos modelos arrojan resultados parecido.

- Para el año 2014 cerca del 75 % de las empresas mostraron un α positivo y un 64.7 % presentaron bajo riesgo en β, lo que sugiere una relación entre un clima comunicacional optimista y un rendimiento superior al del mercado.

- Para el año 2018, a pesar de un pico de positividad en el análisis de sentimientos, reflejando un clima comunicacional optimista, no se ve reflejado en el indicador α donde se muestra un pico de negatividad. un contraste que podría sugerir que las empresas intentaron proyectar una imagen positiva a través de sus comunicaciones, enfrentaban simultáneamente rendimientos desfavorables en comparación con el mercado.

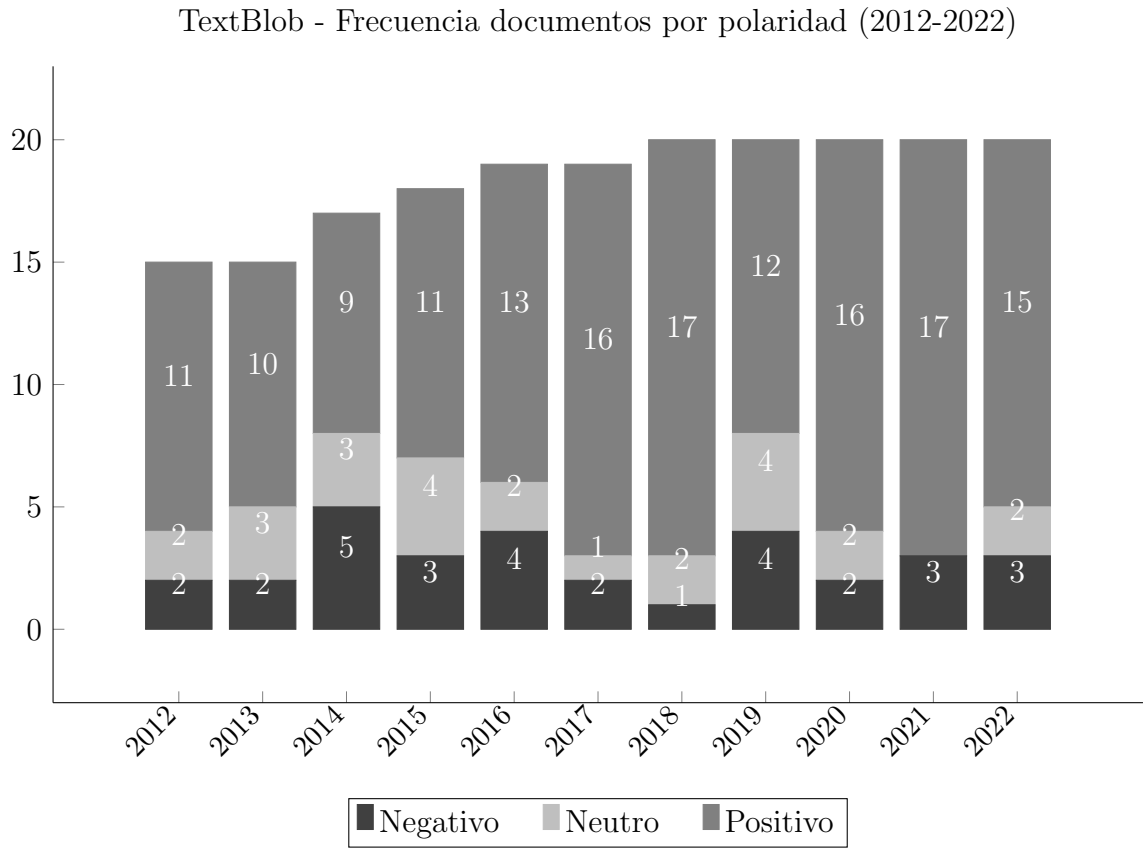

Figura 6: SA TextBlob - Número de documentos por polaridad desde 2012 hasta 2022.

El segundo momento de análisis se presenta entre los años 2018 y 2022, que puede estar directamente relacionado con los desafíos globales impuestos por la pandemia de COVID-19, que afectó a numerosos sectores, mostrando en los discursos los desafíos en las operaciones y preocupaciones en sostenibilidad o ajustes estratégicos para superar la crisis.

- El mayor grado de optimismos de las empresas en la serie de tiempo fue en el año 2018, desde aquí, inicia un descenso rápido a 50 % en 2019 y 35 % en 2020, siendo superada en proporción por los documentos con polaridad negativa para este año, los dos últimos años sube, cerrando en 75 % para el año 2022.

- Este es el periodo donde la negatividad toma mayor relevancia, estando en porcentaje anual por encima de los resultados neutrales, con frecuencias del 30 % en el año 2019, y un pico del 50 % de los documentos en el 2020 (superando la frecuencia de polaridades positivas), posteriormente un 35 % de participación en 2021, cerrando en 15 %.

- La figura 7 de *VADER* destaca un aumento en los documentos con polaridad negativa en 2020, alcanzando un 40 % de los documentos, con los que tiene una alta influencia en el resultado de este año. *VADER* también baja considerablemente los dos últimos años, primero bajando al 25 % en 2021 y cerrando en 0 % de los documentos, lo que puede explicar el repunte en la positividad en 2022.

- En este periodo la frecuencia de documentos neutrales fue baja,se podría destacar que la frecuencia en documentos en sus dos años más significativos fue de 20 % en 2019 y 15 % en

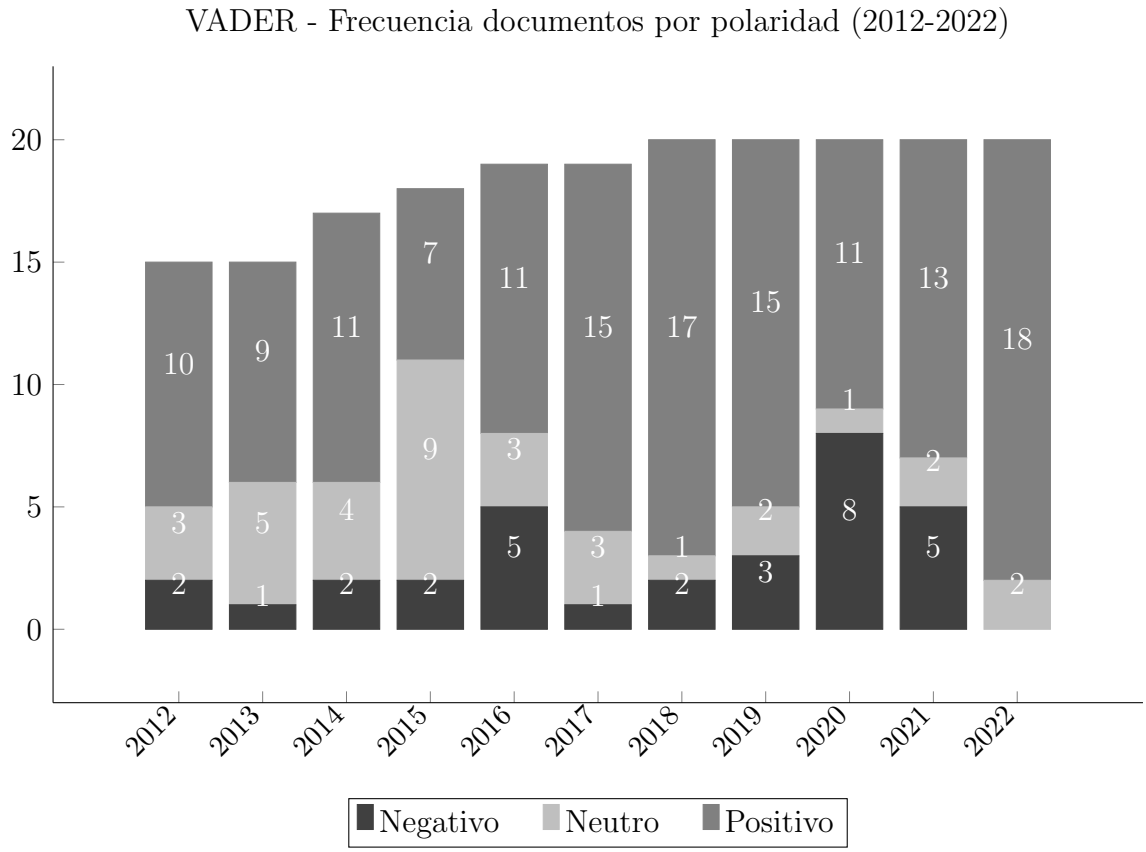

Figura 7: SA VADER - Número de documentos por polaridad desde 2012 hasta 2022.

2020.

- El notable aumento en la negatividad durante 2020, refleja eventos globales críticos que se manifiestan en la comunicación corporativa, este año se caracteriza por un β de riesgo bajo de mayor frecuencia, indicando menor volatilidad comparado con el mercado en general. Adicionalmente, no se registra ningún caso de empresas con riesgo inverso, y se observa un aumento en la frecuencia de retornos positivos de α lo que sugiere que, aunque el contexto era desafiante, algunas empresas lograron rendimientos por encima del mercado.
- La recuperación en la positividad en 2022, podría indicar una recuperación de la confianza en el mercado, sin embargo, se evidencia un incremento en la frecuencia de coeficientes negativos de α llegando al 75 % de las empresas. Este contraste destaca una posible disonancia entre la percepción de mejora transmitida a través de las comunicaciones corporativas y la realidad de un rendimiento aún desafiante frente al mercado.

6. Conclusiones

Este documento explora una variedad de técnicas y métodos en el procesamiento de lenguaje natural y minería de texto. Entre las primeras conclusiones resaltan las técnicas que ayudan a mejorar los procesamientos de texto en español.

- Es importante la captura correcta y precisa de textos en formatos legibles y manejables, como archivos de extensión .txt, lo cual es fundamental para garantizar la integridad y la calidad de los datos iniciales.
- Una preparación meticulosa del texto es crucial para mejorar la efectividad de las métricas y los modelos aplicados en el análisis. El número de condición es un indicador valioso en este proceso, pues una disminución en este número refleja una mejora en la calidad de los datos preparados, facilitando un análisis más estable y preciso.
- Es esencial adaptar, tras cada iteración, las técnicas de procesamiento de texto estándar al idioma específico del conjunto de datos. En este caso del español, eliminar el proceso de stemming ha probado ser beneficioso para mantener la integridad de palabras clave relevantes para los análisis, dado que simplifica las palabras a su raíz, rompiendo estructuras léxicas importantes en español.

En términos generales, los modelos y métricas utilizados facilitaron la extracción de información relevante a partir de textos, permitiendo su comparación con variables cuantitativas de contexto, lo que enriquece significativamente las posibilidades de análisis.

- La medida de similaridad para evaluar la consistencia del discurso a lo largo del tiempo es clave para identificar la estabilidad en la comunicación corporativa y los cambios estratégicos. La similaridad del coseno entre documentos de años consecutivos proporciona una visión de cómo las empresas mantienen o ajustan su enfoque comunicativo, con altos niveles de similaridad indicando continuidad y bajos niveles señalando posibles cambios o adaptaciones a nuevas circunstancias.
- El análisis semántico latente (LSA) revela los temas subyacentes en los discursos corporativos y su evolución a lo largo del tiempo, con este se identifica los focos de interés recurrentes y permite ver cómo ciertos temas ganan o pierden relevancia en diferentes períodos, ofreciendo insights sobre las prioridades estratégicas y operativas de las empresas.
- El análisis de sentimientos, utilizando un enfoque híbrido que combina TextBlob y VADER, proporciona una evaluación del tono emocional de los textos corporativos, el enfoque permite capturar tanto la subjetividad como la intensidad de los sentimientos expresados, siendo fundamental para entender cómo las emociones transmitidas en los discursos pueden influir en la percepción de los *stakeholders* y potencialmente en el comportamiento del mercado.

Es importante contar con variables de contexto que permita identificar que los datos capturados de los documentos pueden realmente relacionarse con aspectos puntuales del mercado, en este caso se toman las variables del modelo CAPM más simplificado, en un esfuerzo académico por realizar un análisis de los resultados y así identificar tendencias y patrones que aporten en la comprensión del discurso empresarial.

- Se observa una posible relación entre la similaridad discursiva y la estabilidad de mercado, donde a altos niveles de similaridad en los discursos corporativos tiende a relacionarse con menores valores de riesgo de mercado, indicando que una comunicación coherente y estable puede asociarse con menor volatilidad en las acciones de la empresa.

- Los cambios temáticos detectados por el análisis LSA, se asocian con las fluctuaciones en el indicador de rendimiento. Específicamente, temas que ganaron relevancia durante ciertos períodos a menudo coincidieron con aumentos en este indicador, lo que sugiere que la adopción de ciertas estrategias o enfoques comunicados efectivamente a los accionistas y el mercado podría estar relacionada con rendimientos superiores al promedio del mercado.

- Los resultados del análisis de sentimientos mostraron que los periodos con tonos más negativos en la comunicación corporativa se alineaban frecuentemente con un incremento en el riesgo del mercado y una disminución en sus rendimientos. Esto indica que las expresiones de preocupación o cautela por parte de los líderes empresariales pueden ser interpretadas por el mercado como señales de un aumento en el riesgo percibido o de un rendimiento potencialmente más bajo en comparación con el mercado.

Por último, al escribirse este documento (primer semestre del año 2024) el análisis de textos se ha diversificado enormemente, incorporando técnicas avanzadas como los modelos generativos adversariales (GAN) y los modelos basados en transformers como GPT (Generative Pre-trained Transformers). Sin embargo, en este trabajo se ha optado por emplear métodos clásicos, los cuales permiten una diferenciación más clara de los inputs entregados al modelo y una interpretación detallada de las categorías analizadas. Estos enfoques tradicionales, que preceden el auge de las tecnologías de procesamiento del lenguaje natural más recientes, ofrecen ventajas significativas en términos de claridad interpretativa y manejo de datos no estructurados. A pesar de los avances representados por modelos como GPT y los chatbots, los métodos clásicos siguen siendo valiosos para el análisis de contenido, proporcionando una mayor transparencia en los procesos analíticos y facilitando la identificación de efectos sobre variables específicas. Esta aproximación permite una comprensión más profunda de los datos y apoya la toma de decisiones informadas al mantener la integridad y la interpretabilidad de los resultados.

Referencias

A., O. G., & Rojas, M. (2022). Modelo CAPM para la valoración de acciones de las empresas en el mercado de la construcción durante el periodo 2015 - 2020. *REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMA Y LA EMPRESA*, 389-403. https://doi.org/10.46661/revmetodoscuanteconempresa.7350

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, *108*, 110-124. https://doi.org/10.1016/j.knosys.2016.05.040

Bolsa de Valores de Colombia. (2024). MSCI COLCAP [Accessed: 2024-02-03]. https://www.bvc.com.co/msci-colcap

Dontcheva-Navratilova, O., Adam, M., Povolná, R., & Vogel, R. (2020). *Persuasion in Specialised Discourses*. http://www.palgrave.com/gp/series/14534

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Society for Industrial; Applied Mathematics.

eToro. (2023). Alpha y Beta: Riesgo de Inversión [Accessed: 2024-02-03]. https://www.etoro.com/es/investing/alpha-and-beta-investment-risk/

GRI. (2023). *Iniciativa de Reporte Global (GRI)*. https://www.globalreporting.org/

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*(1), 39. https://doi.org/10.1186/s40854-020-00205-1

Harymawan, I., Nasih, M., Ratri, M. C., Soeprajitno, R. R. W. N., & Shafie, R. (2020). Sentiment analysis trend on sustainability reporting in Indonesia: Evidence from construction industry. *Journal of Security and Sustainability Issues*, 1017-1024. https://doi.org/10.9770/jssi.2020.9.3(25)

IBM. (2023). CRISP-DM Help Overview [Accessed: 2024-03-02]. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Investing.com. (2024). Historical Data on Beta and Alpha for COLCAP [Accessed: 2024-02-02]. https://www.investing.com/

Jianbao, C., & Jingjie, W. (2009). An Empirical Study on the Stability and Time Variation of Betas in Shenzhen Stock Market. *2009 International Forum on Computer Science-Technology and Applications*, 351-354. https://doi.org/10.1109/IFCSTA.2009.208

Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. *Applied Sciences*, *11*(9). https://doi.org/10.3390/app11093986

Khan, F. H., Qamar, U., & Bashir, S. (2016). eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. *Information Sciences*, *367-368*, 862-873. https://doi.org/10.1016/j.ins.2016.07.028

Lee, M.-H., Hooy, C.-W., & Brooks, R. (2023). A New Measure for Idiosyncratic Risk Based on Decomposition Method. *Journal of Risk and Financial Management*, *16*(1), 43. https://doi.org/10.3390/jrfm16010043

Liu, B. (2010). Sentiment Analysis and Subjectivity. En *Handbook of Natural Language Processing* (Second, pp. 627-666). CRC Press.

Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2011). Opinion Classification Techniques Applied to a Spanish Corpus. https://doi.org/10.1007/978-3-642-22327-3_17

McLaughlin, J. E., Lyons, K., Lupton-Smith, C., & Fuller, K. (2022). An introduction to text analytics for educators. *Currents in Pharmacy Teaching and Learning*, *14*(10), 1319-1325. https://doi.org/10.1016/j.cptl.2022.09.005

McNulty, D. (2024). Bettering Your Portfolio With Alpha and Beta. *Investopedia*. https://www.investopedia.com/articles/07/alphabeta.asp

Ossa González, G. A., & Rojas Dominguez, M. (2023). Modelo CAPM para la valoración de acciones de las empresas en el mercado de la construcción durante el periodo 2015 - 2020. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *35*, 389-403. https://doi.org/10.46661/revmetodoscuanteconempresa.7350

Rajan, R., Ramella, P., & Zingales, L. (2023). *What Purpose Do Corporations Purport? Evidence from Letters to Shareholders*. https://jasonzweig.com/its-time-for-investors-to-re-learn-the-lost-art-of-reading/

Szekely, N., & Brocke, J. V. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE*, *12*(4). https://doi.org/10.1371/journal.pone.0174807

Tessore, J. P., Esnaola, L. M., Lanzarini, L., & Baldassarri, S. (2022). Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. *Cognitive Computation*, *14*(1), 407-424. https://doi.org/10.1007/s12559-020-09800-x

Torres, F. N., & de Alba, M. B. P. C. (2023). Desarrollo de un sistema de aprendizaje automático supervisado para la desambiguación léxica automática utilizando DAMIEN. *Revista Electrónica de Lingüística Aplicada*, *21*(1), 150-178. https://doi.org/10.58859/rael.v21i1.504

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126. https://doi.org/10.1016/j.eswa.2016.03.028

Yahoo Finance. (2024). Historical Data on Beta and Alpha for COLCAP [Accessed: 2024-02-02]. https://finance.yahoo.com/

Zweig, J. (2016). *It's Time for Investors to Re-Learn the Lost Art of Reading*. https://jasonzweig.com/its-time-for-investors-to-re-learn-the-lost-art-of-reading/

7. Anexos

7.1. Repositorio

El repositorio de GitHub contiene todos los códigos fuente utilizados para el análisis, así como los datasets completos y los notebooks de procesamiento. Este recurso es crucial para replicar los estudios realizados o para extender el análisis a otros conjuntos de datos.

- Visita el repositorio para acceder a estos materiales en:

  Repositorio del Proyecto

7.2. Visualización de datos

Explora los resultados de este estudio a través de un dashboard interactivo que permite visualizar y manipular las métricas analizadas. Este dashboard proporciona una herramienta dinámica para explorar tendencias, comparaciones y patrones en los datos analizados de las cartas de los máximos responsables de las empresas.

- Accede al dashboard a través del siguiente enlace:

  Dashboard análisis de discurso de cartas de máximos responsables.