MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

Pronóstico de la Variación de la Tasa de Referencia en México Mediante la Clasificación de los Texto a las Minutas del COPOM del BANXICO

Proyecto Aplicativo
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Gilberto Anzaldo San Vicente

Asesores:

Dr. Mario Graff Guerrero
Dr. Guillermo Benavides Perales

Ciudad de México, noviembre, 2023.

CONACYT

INFOTEC POSGRADOS

Autorización de impresión

GOBIERNO DE MÉXICO | CONAHCYT CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS | INFOTEC

BIBLIOTECA INFOTEC
VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencias de Datos e Información (MCDI)

Ciudad de México, 24 de noviembre de 2023

Unidad de Posgrados
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Pronóstico de la variación de la tasa de referencia en México mediante la clasificación de los texto a las minutas del COPOM del BANXICO"

Desarrollado por el alumno: **Gilberto Anzaldo San Vicente,** y bajo la asesoría del **Dr. Mario Graff Guerrero** y el **Dr. Guillermo Benavides Perales** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que el alumno no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Mtro. Carlos Josué Lavandeira Portillo
Director Adjunto de Innovación y Conocimiento

CJLP/jah

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento
Gilberto Anzaldo San Vicente.- Alumno de la Maestría en Ciencias de Datos e Información.- Para su conocimiento.

Las opiniones expresadas en este documento son sólo del autor y no necesariamente representan la opinión del Banco de México.

Agradecimientos

Yahvé Dios que, al temerte, uno principia en la sabiduría, ya que sólo los insensatos la desprecian así coma la doctrina. (Proverbios. 1:7).

A Jesús y Martha, quienes, al instruirme y enseñarme, me han dado una corona de gracia para mi cabeza y un collar para mi cuello, para examinar los pasos de mi pie para que sean rectos, no declinando ni a la derecha ni a la izquierda, apartándome del mal (Proverbios 1:8, 1:9; 4:26, 4:27).

Esta obra ha sido posible gracias al inconmensurable apoyo, amor y paciencia de mi amada esposa Jaqueline. Te amo por siempre.

Al CONAHCYT y al INFOTEC por darme la oportunidad de realizar mis estudios de posgrado en un tema que tanto me apasiona y que está marcando el destino de la humanidad para un mejor futuro.

Mi reconocimiento y gratitud a mis profesores por dedicarse a mejorar las condiciones de vida de muchas personas mediante la docencia, y a mis compañeros y compañeras del posgrado con quienes tuve el honor de compartir actividades académicas que no hubiesen sido resueltas y entregadas sin su valiosa colaboración[1].

Al Dr. Graff, por compartir su sabiduría en aprendizaje computacional y por sus consejos para superar con facilidad los problemas que enfrenté durante el desarrollo de este proyecto.

Al Dr. Benavides, por sus valiosos comentarios en las materias económico-financieras que se abordan en este proyecto. Sin ellos, este trabajo no hubiese sido posible.

[1] Véase el anexo IV para mayores detalles

Tabla de contenido

Índice de figuras

Índice de gráficos

Índice de cuadros

Siglas y abreviaturas

AC	Aprendizaje Computacional
Banxico	Banco de México, Instituto Central o Banco Central de México.
BoW	Bag of Words
COPOM	Comité de Política Monetaria
CRISP-DM	Cross Industry Standard Process for Data Mining
EvoDAG	Evolving Directed Acyclic Graph
EvoMSA	Evolutionary Multilingual Sentiment Analysis
Forex	Foreign Exchange
FT Fast Text	Fast Text
HA B4MSA	Human Annotated Base Multilingual Sentiment Analysis
INFOTEC	Centro de Investigación en Tecnologías de Información y Comunicaciones
IA	Inteligencia Artificial
Internet	International network
JSON	JavaScript Object Notation
MCDI	Maestría en Ciencia de Datos e Información
PLN	Procesamiento del Lenguaje Natural
SIE	Sistema de Información Económica
SVC	Support Vector Machine for Classificaction
SVM	Support Vector Machine
SVR	Suppert Vector Machine for Regression
TII	Tasa de Interés Interbancaria, tasa de referencia, o tasa objetivo
TIC	Tecnologías de Información y Comunicación
TF-IDF	Term Frequency – Inverse Document Frequency
TH Lexico	Thumbs Up-Down Lexico
TR M4MSA	TRainning set Base Multilingual Sentimen Analysis
UTF-8	8-bit Unicode Transformation Format

Coeficientes y constantes

Bit	Binary digit	Valor de 0 o 1
Byte	Combinación de un octeto	$2^8 = 256$ bits
KB	Kilobyte	$2^{10} = 1{,}024$ bytes
MB	Megabyte	$2^{20} = 1'048{,}576$ bytes
GB	Gigabyte	$2^{30} = 1{,}073'741{,}824$ bytes
TB	Terabyte	$2^{40} = 1''099{,}511'627{,}776$ bytes
PB	Petabyte	$2^{50} = 1{,}125''899{,}906'842{,}624$ bytes
EB	Exabyte	$2^{60} = 1'''152{,}921''504{,}606'846{,}976$ bytes-
ZB	Zetabyte	$2^{70} = 1{,}180'''591{,}620''717{,}411'303{,}424$ bytes
YB	Yotabyte	$2^{80} = 1''''208{,}925'''819{,}614''629{,}174'706{,}176$ bytes

Glosario

“A”

Activo financiero: Es un título o instrumento que otorga al comprador el derecho a recibir un ingreso futuro de parte del vendedor [31].

Agentes racionales: Término asignado a las personas que complementan sus análisis con información cuantitativa proveniente del análisis de los datos; o bien, considerando los resultados de modelos analíticos.

Agnóstico: Se refiere a los modelos analíticos o computacionales que procesan cualquier tipo de información de la misma forma.

Aprendizaje computacional: Conjunto de técnicas que se emplean para procesar la información para producir pronósticos o clasificaciones de los datos.

Arbitraje financiero: Es una estrategia que consiste en aprovechar la diferencia de precio de un mismo activo financiero en diferentes mercados para obtener ganancias [32].

“B”

Benchmark: Dentro del contexto financiero, se hace referencia a un valor que es utilizado para medir el rendimiento de diversos instrumentos de inversión.

Bag-of-word (BoW): Colección de palabras o términos que forman el léxico de varios documentos.

“C”

Ciberseguridad: Conjunto de políticas de negocio, criterios de riesgo y empleo de TIC que buscan reducir la posibilidad de que personas ajenas a una organización accedan su información electrónica o a los sistemas informáticos de ésta.

Corpus: Colección de términos empleados en la construcción de un vocabulario.

“D”

Dataset: Colección de datos que son empleados en el análisis o en la construcción de un modelo de aprendizaje computacional.

Deep learning: Arquitecturas de procesamiento de datos basadas en las redes neuronales artificiales. Este modelo se caracteriza por tener un número de capas ocultas superior a dos.

Deflactor. Es un coeficiente que se emplea en la actualización de los precios de los bienes y servicios que se comercializan en la economía real con el objetivo de estimar su valor presente y con ello hacerlo comparable con otros bienes deflactados.

Dummy: Se refiere a las variables de contenido dicotómico que son empleadas en los modelos econométricos.

“F”

Firefox: Nombre comercial de un navegador de Internet.

Framework: Nombre en inglés que hace referencia a un marco de trabajo en el procesamiento de la información dentro de un proyecto de ciencia de datos o de inteligencia de negocios.

Función impulso-respuesta: Mediante la representación de medias móviles asociadas con el modelo estimado, por ejemplo, un VAR o un modelo de AC, se estima el impacto de la variable dependiente ante variaciones abruptas en las variables independientes.

“I”

Information extraction: Es el proceso de obtener información de un conjunto de datos que es de interés procesar para construir un modelo analítico.

Information gain: Es una medida numérica que valora la cantidad de información que se obtiene al emplear un preprocesamiento en los datos que se emplearán en la construcción de un modelo computacional.

Information retrieval: Se refiere al acceso a la información que una persona o un sistema informático tiene para hacer su trabajo.

Internet: Es la red global de comunicación donde convergen millones de redes más pequeñas cuyas computadoras ofrecen datos e información de naturaleza diversa.

"L"

Lematizar: Reduce las palabras a términos a formas que sean lingüísticamente válidos mediante un análisis morfológico. Por ejemplo, "programaba" tendría el lema "programar".

"M"

Minuta: Se refiere al documento emitido por el Banxico donde se exponen los criterios que se emplearon para modificar la TII.

Modelos de ensamble: Colección de diversos modelos de aprendizaje computacional que se emplean secuencialmente para mejorar la calidad de los resultados de un pronóstico.

"O"

One Hot Encoder: Técnica empleada en la vectorización de textos donde en cada casilla se asigna un valor numérico que indica la presencia de éste con un valor 1, o en el caso de su ausencia con un cero.

Ontología: En las ciencias computacionales, se refiere al análisis de los componentes que conforman un campo de conocimiento que se desea modelar y ejecutar dentro de una computadora.

“P”

Phishing: Técnica empleada para robar la identidad digital de una persona, al proporcionarle una página Web similar a la que habitualmente utiliza.

Política monetaria: Colección de criterios económicos y financieros que se esgrimen para analizar el entorno económico de una economía y tomar la decisión de cuál será la variación de la TII.

Principio de Taylor/Regla de Taylor: En Macroeconomía, se emplea en las decisiones de política monetaria, como el COPOM del Banxico, para controlar la inflación en el corto plazo y promover el pleno empleo. Tiene los siguientes insumos: 1. La diferencia entre la inflación esperada con la inflación objetivo, 2. El PIB esperado y su tendencia en el largo plazo, 3. Un tipo de interés a corto plazo.

Programa maligno: Se refiere a un sistema informático poco complejo que busca alterar el buen funcionamiento de otros que son empleados en las organizaciones.

“R”

Reglas de asociación: Es una estructura matricial donde se relacionan valores que se desean analizar. Normalmente se emplean el valor uno para indicar su presencia y de cero su ausencia. La lectura es por renglón y se buscan estimar probabilidades a priori para hacer predicciones.

“S”

Semántica: Se refiere al análisis que se realiza en un documento para encontrar el significado de sus expresiones.

Spam: Transmisión de un mismo mensaje que es enviado a diversos destinatarios de correos electrónicos. Normalmente, esta es una técnica empleada en los ataques informáticos.

Spot: Se refiere al valor de mercado de un activo financiero.

Stopwords: Colección de palabras que no aportan valor en el PLN y por ende son descartadas.

Stemming: Es el proceso de obtener la raíz de una palabra dentro del PLN de un documento, mediante la extracción de los tallos que se encuentran presentes en las variaciones de una palabra. Por ejemplo, las palabras “programaba”, “programar”, “programó” tienen la raíz “program”.

“T”

Tamizar: Sinónimo del proceso de limpieza de los datos que se procesarán en el PLN.

Tasa de interés: Es el costo del dinero que es empleado para financiar el consumo de bienes, de servicios, o la materialización de un proyecto de inversión [35].

Token: Términos que se obtienen después del proceso de limpieza de un documento.

Introducción

Introducción

Motivación

El presente documento es un reflejo de mi interés personal por estudiar la teoría de los temas administrativos, económicos y financieros, su aplicación e impacto en el desarrollo de la sociedad moderna, y en específico, como promotores del desarrollo de las modernas TIC que componen la base del cuerpo de conocimiento de la ciencia de datos que han sido en la última década motores fundamentales en la creación de modelos analíticos basados en la IA, AC y el PLN que facilitan el procesamiento de los datos que coadyuvan a mejorar el rendimiento de los procesos de negocio.

En este tenor, la aplicación de la metodología CRISP-DM para abordar la construcción de un clasificador que compute el movimiento alcista, bajista o estable de la TII a partir del procesamiento de las minutas del COPOM del Banxico arrojó resultados interesantes que se valoraron con una precisión superior al 80%.

Hay que señalar que este proyecto puede considerarse innovador en la aplicación de la ciencia de datos en el ámbito financiero internacional, ya que aborda dos situaciones inéditas en la literatura estudiada.

La primera, se aborda la valoración del comportamiento de la tasa de interés de referencia que es un tema de política monetaria, que tiene un impacto directo en las decisiones empresariales y que no ha sido abordado con suficiencia. La segunda versa en la forma en que se pueden construir clasificadores útiles a partir del preprocesamiento de información escasa que apenas superaba los 80 MB, cuando en la literatura sólo se consideran volúmenes valorados en GB o superiores.

El contenido del documento se compone de cinco partes. El capítulo 1 describe la relevancia de la tasa de referencia en la economía y las finanzas en México y el porqué es importante su estudio dentro del contexto de la ciencia de datos. El capítulo 2 expone una revisión de la literatura desde 1984 hasta 2020 donde se documentan las aplicaciones del PLN en documentos financieros para

diferentes propósitos. A partir de su valoración, se opta por la construcción del clasificador del movimiento de la TII.

En el capítulo 3 se aplica el marco metodológico CRISP-DM, y se expone cómo se relaciona con las actividades para la construcción del clasificador del movimiento de la tasa de referencia. El capítulo 4 describe los pasos que se realizaron en el preprocesamiento de la información contenida en las minutas del COPOM, y la forma en que se amalgamaron con las series de tiempo de la TII proporcionadas por el SIE. Así mismo, se expone el uso de la biblioteca EvoMSA 2.0 y la aplicación de diversos análisis estadísticos para valorar la precisión de los resultados obtenidos del modelo.

Finalmente, en las conclusiones se comparten diversas ideas que se produjeron durante la elaboración del presente proyecto aplicativo y se esbozan algunas líneas futuras de investigación relativas al procesamiento de documentos financieros que pueden ser procesados con EvoMSA.

Capítulo 1

Definición del objetivo de investigación

Capítulo 1. Descripción del Problema

1.1 Banxico y la TII

El Banco de México, Banxico, es el Banco Central de México y tiene por objetivo el preservar el valor de la moneda nacional a lo largo del tiempo y con ello contribuir al bienestar económico de los mexicanos enfocando sus actividades en cuatro objetivos principales, 1. Estabilidad del sistema financiero, 2. Emisión y puesta en circulación del papel moneda, 3. Preservar el valor de la moneda y 4. Promover la estabilidad del sistema de pagos. Para materializar estos objetivos cuenta con un instrumento financiero, la Tasa de Interés Interbancaria, la TII, que también es conocida como la tasa de referencia [33, 34].

La publicación de esta tasa es relevante para el Instituto Central porque con ésta envía un mensaje acerca de cuál es su postura sobre el comportamiento de la inflación en el corto y mediano plazos en México. Un aspecto operativo relevante, mas no teórico, sobre la TII es que para que su valor sea considerado en los análisis y las decisiones de los agentes financieros, consiste en que ésta debe publicarse en el Diario Oficial de la Federación (DOF).

La relevancia social que tienen los anuncios del COPOM son muy relevantes para minimizar el estrés económico y financiero que se pudiera ocasionar si estos se contraponen con las expectativas de los agentes sobre el desempeño de la actividad económica, la cual fue construida con la información de las minutas del pasado.

Por su parte, los agentes económicos utilizan la TII para definir las tasas de interés que emplearán para estimar el costo del dinero en diversas actividades como son, 1. el otorgamiento de los créditos, 2. la valoración de los proyectos de inversión, 3. la composición de los portafolios de inversión que gestionan las instituciones financieras, 4. la determinación de los riesgos financieros que permiten tomar decisiones ejecutivas, entre otras muchas actividades públicas y privadas relacionadas con el dinero.

La tasa de interés interbancaria es publicada por el Banxico en fechas específicas. Para su determinación, el COPOM emplea diversas informaciones

cuantitativas y cualitativas de origen nacional e internacional que son analizadas por los miembros de la Junta de Gobierno del Banxico y consideran el comportamiento de corto, mediano y largo plazo de diversas variables macroeconómicas y financieras que son relevantes en México. El anuncio de la TII es presentado al público en general mediante documentos públicos en la página de este banco central [2].

1.2 Planteamiento del problema

En el mundo de las decisiones financieras y de la política monetaria en México, es de gran relevancia la estimación del valor de la tasa de interés interbancaria porque ésta tiene un doble uso. En el primer caso, incide en el valor futuro de los instrumentos de inversión donde los agentes tienen depositados sus recursos económicos. En el segundo caso, el Banxico emplea dicha tasa para mantener el valor de la inflación dentro de una banda, acotada entre el 2-4 porciento, que permita el buen desarrollo de las actividades económicas.

Es una práctica común entre los expertos del ramo el pronosticar la tasa de referencia empleando variadas técnicas econométricas que se emplean de forma independiente, combinadas; o bien, mediante la ecuación de Taylor[2]. Sin embargo, dichos modelos sólo emplean datos previamente tabulados que en el mejor de los casos representa el 20% de la información disponible. El 80% restante se encuentra en documentos de análisis e investigación que no forman parte del procesamiento analítico de los modelos [1]. Para mitigar esta deficiencia, los integrantes del COPOM valoran las informaciones cuantitativas, cualitativas y documentales, y posteriormente emiten cuál será el nuevo valor de la TII.

Dada la relevancia que el valor de la tasa de referencia tiene para los tomadores de decisión financiera en México, se propone como problema de investigación “El construir un modelo de pronóstico de la variación de la TII a partir de la categorización de las minutas del COPOM del Banxico”.

[2] Para mayor detalle véase el anexo V.

El ofrecer una tecnología que pueda procesar los documentos del COPOM, permitirá complementar los análisis financieros con un resultado analítico que pueda procesar el 80% de la información que se omite regularmente.

1.2 Objetivos

1.3.1 Objetivo general

Construir un clasificador que procese el contenido documental sobre temas económicos y financieros que proporcione un pronóstico del valor de la tasa de referencia señalando si ésta subirá, bajará o se mantendrá empleando técnicas de procesamiento de lenguaje natural y aprendizaje computacional que se estudiaron en el programa de la MCDI de INFOTEC en el período 2021-2023 y que se entregará en el segundo semestre del año 2023.

1.3.2 Objetivos específicos

Objetivo Específico 1. Obtener las 98 minutas que publica el Banxico con relación al COPOM del período 21/01/2011 al 28/02/2023 que se han publicado en el sitio de este instituto central [2] y que contienen información documental con los criterios para decidir los cambios en puntos base en la TII. Esta actividad se realizó durante los meses de mayo/2022 a junio/2023.

Objetivo Específico 2. Procesar la información de las 98 minutas del COPOM con programas realizados en Python que incorporen los algoritmos para procesar los textos relevantes durante el segundo semestre de mayo/2022 a junio/2023.

Objetivo Específico 3. Obtener la serie de tiempo de la tasa de referencia de la página web del Banxico del período especificado en el objetivo específico 1.

Objetivo Específico 4. Relacionar el contenido semiestructurado de las minutas (Objetivo Específico 2) con la serie de tiempo (Objetivo Específico 3) para construir el DATASET que se empleará en el modelo de aprendizaje computacional mediante las herramientas informáticas como Excel, Python y Orange para construir el archivo JSON, para entrenar un modelo computacional supervisado, que relacione el contenido documental de las minutas con valores numéricos se la serie de tiempo.

Objetivo Específico 5. Para construir el analizador de sentimientos que clasificará el contenido de las minutas del COPOM, se estudiará la biblioteca EvoMSA [21,22].

1.3.3 Resultados esperados

Entregar un modelo informático que pueda procesar la información documental sobre economía y finanzas para pronosticar el movimiento, al alza, baja o sin cambio, de la TII del Banxico.

Capítulo 2
Estado del arte

Capítulo 2. Proyectos relacionados

El objetivo de la presente revisión de la literatura ofrece un panorama de los avances que se han alcanzado en la aplicación de los algoritmos de aprendizaje computacional empleados en el procesamiento de documentos financieros comprendidos entre los años 2013 a 2023.

Se emplearon seis documentos de investigación que condensan el contenido de 236 artículos de investigación publicados y 31 que fueron presentados en conferencias, totalizando 267 documentos que abarcan las aplicaciones del procesamiento de textos financieros desde 1984 hasta 2020 [1, 3, 4, 5, 6, 7].

La principal fuente documental empleada fueron sitios públicos disponibles en la Internet. Estos pueden tipificarse en 1) información general sobre el mundo empresarial, 2) sitios de las organizaciones de gran tamaño económico donde presentan sus reportes financieros a sus inversionistas, y 3) sitios que presentan diversos análisis a los estados financieros [6].

De las aplicaciones encontradas, el 31% de éstas se subdividen en, 18% en las noticias con información empresarial. El 4% son índices accionarios, el 6% se relaciona con los reportes financieros que publican las empresas y en un 3% de los servidores de correos. En este rubro, los proveedores son las mismas empresas; o bien, agencias noticiosas especializadas.

El 28% de las fuentes de información empleadas para procesar los documentos, proveen un vocabulario económico financiero especializado, siendo la más utilizada la que ha publicado la Universidad de Harvard que representa el 14%. El 36% se ha especializado en el tema de la ciberseguridad. En este rubro hay diversas fuentes que están disponibles en la internet.

Las redes sociales aportan el 5% de la información financiera principalmente en los chats que emplean los usuarios de estas plataformas; siendo las principales, Twitter y Facebook, así como las secciones de noticias de estas.

El cuadro 1 presenta los porcentajes del tipo de informaciones empleadas donde se podrá apreciar que no se identificaron fuentes de ningún banco central.

Si el lector está interesado en ahondar en las fuentes de información [1, 3, 4, 7], puede revisarlas en el anexo I.

Categoría	Tipo de fuentes de información	Número de fuentes	Porcentaje	Porcentaje acumulado
Noticias empresariales[3]	Noticias e información financiera	40	18%	
	Índices accionarios	8	4%	
	Información corporativa	14	6%	
	Servidores de correos	6	3%	31%
Redes sociales[4]	Facebook	3	1%	
	Twitter	7	3%	
	Noticias en las redes sociales	1	<1%	5%
Corpus financieros[5]	DICTION	8	4%	
	GI/Harvard	31	14%	
	Términos financieros específicos	18	8%	
	Otros	4	2%	28%
Ciberseguridad	Phishing	16	7%	
	Spam	11	5%	
	Programa maligno	10	5%	
	Intrusión	7	3%	
	Detección de fraudes	36	16%	36%
		220	100%	100%

Cuadro 1 *Tipos de información y fuentes de datos empleados (1984-2020).*

Fuente: Elaboración propia.

[3] Para conocer las fuentes de información, véase el inciso 1 del anexo I.
[4] Para conocer las fuentes de información, véase el inciso 2 del anexo I.
[5] Para conocer las fuentes de información, véase el inciso 3 del anexo I.

2.1 Aplicaciones de la minería de textos y el PLN

Considerando las fuentes de los documentos y los datos del cuadro 1 se procedió a identificar cuáles han sido las principales aplicaciones que se han desarrollado con estas. Se estimó que el 96% se enfocaron en dos rubros principales, la ciberseguridad y el pronóstico de valores financieros [1, 5, 6].

La Ciberseguridad ocupa el 37% de la investigación, porque este tema tiene un impacto económico relevante. El 10% se ha enfocado en el phishing, el 8% en la identificación de correos spam, el 9% en proteger a los usuarios de los servicios financieros al identificar los riesgos asociados con los programas malignos, el 6% con intrusiones no autorizadas a los sistemas informáticos y el 4% a la prevención de fraudes bancarios.

En el procesamiento de los documentos financieros para tomar decisiones, representan el 59% de las investigaciones realizadas, donde el 17% se relaciona con la gestión de riesgos asociados a los portafolios de inversión. El 42% restante se ha abocado en construir modelos de pronóstico basados en modelos econométricos que estiman la volatilidad de diversos indicadores como el forex, índices bursátiles o los precios de acciones específicas, de manera individual o conjunta en proporciones de 5%, 26% y 11% respectivamente.

El 4% restante de las investigaciones se han enfocado en aplicaciones de CRM basadas en el minado de las opiniones de los clientes acerca de la calidad de los servicios adquiridos. En esta parte de la revisión de la literatura no se encontraron aplicaciones que se relacionen con las publicaciones de política monetaria. En el cuadro 2 se muestra la distribución de las aplicaciones aquí comentadas.

		Número de investigaciones	Porcentaje	Porcentaje acumulado
Ciberseguridad	Phishing	12	10%	
	Spam	9	8%	
	Programa maligno	10	9%	
	Intrusión	7	6%	
	Fraudes	5	4%	37%
CRM	Minar las opiniones de los clientes	5	4%	4%
Gestión de riesgos y portafolios	Minimizar la varianza de los portafolios	9	8%	
	Relación entre los comentarios negativos de las empresas vs su desempeño financiero	11	9%	17%
Pronósticos financieros	FOREX	6	5%	
	Índices bursátiles y el precio de acciones	30	26%	
	Combinación de indicadores	13	11%	42%
		117	100%	100%

Cuadro 2 *Distribución de las aplicaciones financieras de la minería de textos y PLN.*

Fuente: Elaboración propia.

2.2 Aplicaciones de la minería de textos en el pronóstico financiero

Desarrollando los rubros: fraudes, gestión de riesgos y portafolios y pronósticos financieros del cuadro 2 por estar relacionados con los objetivos de esta investigación, se encontró que representa el 38% de las aplicaciones de la minería de textos se han enfocado en 1) clasificar textos, 2) resumir el contenido de los documentos, 3) construir reglas de asociación y ontologías especializadas, 4)

tipificar el sentimiento de los reportes financieros de las empresas y, 5) pronósticos de diversas variables económicas y financieras.

La clasificación de los tópicos financieros que se han documentado utilizó bases de datos de noticias que para el año 2020 totalizaban 40 ZB, 50 veces mayor respecto al 2010 [3]. La principal aplicación del PLN se ha enfocado en clasificar el sentimiento de los comentarios en las redes sociales sobre eventos de coyuntura; o bien, la clasificación del sentimiento de los encabezados de las noticias. En todos los casos y dependiendo del contexto, se emplearon etiquetas binarias como {bueno, malo}, {alcista, bajista}, {comprar, vender}, entre otros más [1, 5, 7].

En este tenor las clasificaciones aplicadas a los reportes financieros fundamentales que publican las empresas buscan el otorgar una opinión sobre su contenido que pueda ser de utilidad a los auditores internos o externos a las organizaciones como {“evidencia de fraudes”, ”sin evidencia de fraudes”}, {confiable, ”no confiable”}, entre otras más. Sin embargo, estos documentos resultan ser un reto analítico elevado, ya que en su contenido hay notas a los estados financieros que son opiniones sobre los valores de las razones financieras expresadas. El reto analítico radica en que la interpretación de los valores de dichas razones no es siempre la misma para diferentes sectores industriales [1, 5, 6, 7].

Con relación al pronóstico de variables como los tipos de cambio (forex), los índices bursátiles, el precio de las acciones de manera aislada o combinada, se ha documentado que hay diversos métodos que se han empleado y que han buscado pronosticar la volatilidad o valores específicos.

Con relación al pronóstico de la volatilidad, el 37% de las investigaciones reportadas emplean una perspectiva econométrica basada en el procesamiento de series de tiempo utilizando principalmente variaciones del método GARCH [1,4,5].

Por otra parte, en las aplicaciones que buscan predecir valores específicos, ha predominado el uso de los modelos autorregresivos [1, 4]. En ambas perspectivas de modelado, se emplearon variables dummy para incorporar el contenido de las noticias buenas o malas de acuerdo con el criterio del investigador.

Sobre lo anterior, hay que informar que, durante el proceso de revisión del presente proyecto en los meses de abril y mayo de 2023, se descubrió la existencia

de un futuro documento de investigación que se encuentra próximo a publicarse y emplea el PLN a 199 minutas que ha publicado el Banco de Inglaterra entre los años 1997 a 2017 y se valora el impacto que estas han tenido en promover la estabilidad en tres rubros, que son de interés para los investigadores, 1) la estabilidad en la inflación, 2) la productividad y 3) el sector financiero en el Reino Unido de la Gran Bretaña.

Para ello, se utilizó un modelo VAR que relaciona el valor de la tasa de referencia británica con una matriz binaria con unigramas, construida con el método One Hot Encoder, que identifica los tokens presentes en cada minuta con el propósito de cuantificar las interacciones entre los tokens y el valor de la TII mediante funciones de impuso-respuesta [29]. Pese a esta limitación en el vocabulario, el PLN de los textos financieros es relevante debido a que se ha reportado que el 30% del valor pronosticado se vio afectado en el corto plazo por las malas noticias publicadas, experimentando variaciones entre el 22-56% en el precio. Por otra parte, cuando se valoraron activos estadounidenses, del 9-15% del precio diario se vio afectado por los anuncios del gobierno de los EE. UU. [1, 4, 6].

El resumir el contenido de los documentos, así como la construcción de reglas de asociación, se ha basado en el empleo de vocabularios económicos y financieros limitados y poco estandarizados ante la ausencia de una ontología que los unifique. Esto ha propiciado que no haya un consenso general acerca de la efectividad de los resultados obtenidos [3, 6].

En este mismo tenor, se ha documentado que en el caso de los anuncios de política monetaria del Sistema de la Reserva Federal de los EE. UU., cuando ha dado a conocer modificaciones abruptas a la TII, los mercados de capitales reaccionan abruptamente y liquidan sus posiciones afectadas, lo que incide negativamente en el comportamiento de la economía. Es por esta razón que este Instituto Central es muy cuidadoso en anunciar sus acciones para evitar movimientos abruptos en los sectores económicos [8].

En materia de la gestión financiera de los activos, matemáticamente se ha buscado la asignación eficiente de los recursos disponibles empleando la minimización de la volatilidad del portafolio de inversión. Dentro del contexto de la

minería de textos, se ha intentado identificar las tendencias, ganadoras o perdedoras, de los activos utilizados dentro de períodos de tiempo de 90 días [6].

Para amalgamar la optimización de los portafolios de inversión con la minería de textos, el 63% de las investigaciones han aplicado la categorización de textos combinando unigramas, bigramas o n-gramas para proporcionar un resultado binario, como {favorable, desfavorable}, {alcista, bajista} o {subir, bajar} [1, 3, 4, 5, 6]. A partir del resultado obtenido el gestor del portafolio tomará una de tres posibles acciones {comprar, vender, retener} los activos para optimizar sus inversiones.

Aunque existen diversos algoritmos de aprendizaje computacional para el procesamiento de la información financiera, se ha documentado que gran parte de las aplicaciones reportadas combinaron varios métodos para incrementar la precisión de los resultados, sugiriendo el empleo de métodos de ensamble.

En el contexto de la relación entre los clientes y las empresas, 41% de las aplicaciones que se exhiben en el cuadro 2, se relacionan con aplicaciones relativas a la ciberseguridad (37%) y el CRM (4%). En el rubro de la ciberseguridad, ésta construye agentes inteligentes que puedan identificar mediante la minería de texto, señales de riesgos potenciales en programas o en mensajes que son enviados a las personas o a las empresas que puedan incidir negativamente en su salud financiera. Estos riesgos se materializan cuando los atacantes explotan vulnerabilidades inherentes en los sistemas informáticos o en los procesos de comunicación al interior de las organizaciones [1].

En materia de la atención a los clientes de los servicios financieros, las aplicaciones de la minería de texto y el PLN en los sistemas CRM se minan las opiniones de las personas para identificar áreas de mejora en los procesos de negocio, incrementar la personalización de los servicios que se ofrecen para incentivar el consumo; o bien, mejorar los pronósticos de la demanda más probables. Lo anterior, permite generar indicadores estadísticos que coadyuvan en la toma de decisiones al interior de las empresas [1].

2.3 Supuestos teóricos empleados en las aplicaciones financieras

El principal supuesto teórico empleado es la Teoría de los Mercados Financieros Eficientes, elaborada por Eugene Fama en 1964 [1, 4, 5, 7]. Esta teoría esgrime que el precio de un activo refleja completamente la información disponible en el mercado al momento de que éste es asignado. Cuando una nueva información aparece, ésta es asimilada por el mercado y la refleja inmediatamente en el precio, como se asume que ocurre en las bolsas de valores [9, 10].

Este supuesto teórico está soportado en dos consideraciones. La primera señala que, los mercados son eficientes y las correcciones en los precios de los activos se realizan inmediatamente, propiciando que resulte imposible la predicción de los precios a partir de la información disponible y con ello cancelar la posibilidad de cualquier arbitraje. Por su parte, los participantes, en las transacciones de compra y de venta, son agentes racionales porque incorporan con inmediatez la nueva información disponible en sus decisiones.

El segundo supuesto es conocido como la teoría de la caminata aleatoria. Ésta fue propuesta inicialmente por Markop y posteriormente por Fama. Se asume que el predecir el precio de un activo a partir de la información del mercado lo hace indistinguible de una caminata aleatoria en el corto plazo [9, 10, 11].

Estos supuestos teóricos de las finanzas son relevantes porque con el empleo del PLN se puede procesar las noticias y valorar la inmediatez con la que estas se asimilan al precio de los activos subyacentes.

2.4 Las noticias financieras como fuente de Información en el PLN

2.4.1 ¿Son las noticias financieras relevantes para la toma de decisiones?

Se ha documentado que las informaciones emitidas por personas u organizaciones que son consideradas como fiables, inciden directamente en el desempeño financiero de las organizaciones. Esto se ha observado cuando se ha emitido una opinión negativa sobre el contenido de los estados financieros de empresas o sectores industriales completos y los mercados de capitales tienden a incrementar el volumen de sus operaciones, en comparación a las opiniones positivas [1, 4, 6].

Este comportamiento incide directamente en el precio de las cotizaciones de las acciones, por lo que se podría asumir una relación causal dinámica de reforzamiento como se ilustra en la figura 1.

En este tenor se construyó el sistema AZFinText para pronosticar el precio de una acción. Dado que la información empleada es pública, existe un rezago de 20 minutos entre el pronóstico y lo que publican los proveedores de información de noticias financieras. Los resultados estadísticos han alcanzado una precisión del 74% [4].

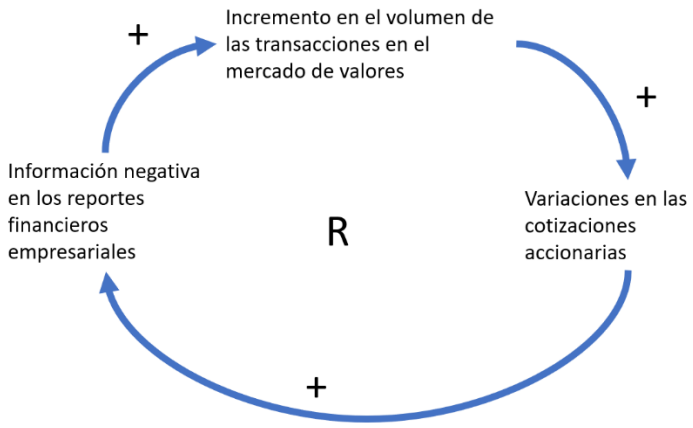


Figura 1 *Ciclo de reforzamiento de la opinión negativa de la información financiera.*

Fuente: Elaboración propia.

En el ámbito profesional, los reportes financieros son una fuente de información muy valiosa porque en ésta se describe la salud económica de las empresas. Se ha observado que al ser analizada correctamente se han podido identificar oportunidades de negocio; o bien, fraudes a los accionistas [4, 6]. Sin embargo, el modelar esta habilidad humana con modelos de inteligencia analítica ha resultado ser una tarea muy difícil.

El principal reto para la ciencia de datos consiste en poder plasmar en los modelos analíticos la flexibilidad de poder variar una clasificación de positiva a negativa bajo ciertas circunstancias, y la contraria bajo otras, como lo hacen los expertos humanos, cuando valoran las posiciones de las empresas dependiendo

de las condiciones por las que están transitando uno o varios sectores industriales en una economía nacional [5, 6].

2.5 Técnicas empleadas en el procesamiento de textos y documentos financieros

El procesar el volumen masivo de los datos existentes representa un reto científico y tecnológico y ha conllevado la creación y la aplicación de técnicas y metodologías diversas para extraer de ésta la mayor cantidad de información que pueda ser de utilidad para la toma de decisiones empresariales [3, 4, 5, 6].

En el caso del procesamiento de los textos, hay una colección de pasos agnósticos que se siguen sin que exista una metodología que las unifique. El paso inicial es cuando el científico de datos tiene una colección de documentos con los que trabajará. Posteriormente, los pulveriza en términos individuales, que son llamados tokens, para tamizarlos por diferentes filtros con el objetivo de obtener únicamente aquellos que aportan información.

El tamizado consiste en aplicar una colección de filtros donde se eliminan los tokens que se encuentran en listas de términos que no aportan valor (stopwords), o, por el contrario, recopilar los términos que deben ser considerados en los textos dentro de una bolsa de palabras (BoW). Entre estos pasos, se transforman los tokens a expresiones más simples mediante la extracción de sus raíces (stemming); o bien, transformarlos de acuerdo con el contexto económico y financiero que se esté abordando (lematización). El flujo de los pasos se presenta en la figura 2, siendo similar al presentado por Téllez et al para el procesamiento de textos en español [7, 12].

El producto de estos pasos permite generar una matriz de valores numéricos que serán el insumo de los modelos de aprendizaje computacional. La cabecera de esta matriz son los tokens depurados y sus valores pueden ser binarios, o mediante la transformación con TF-IDF [7].

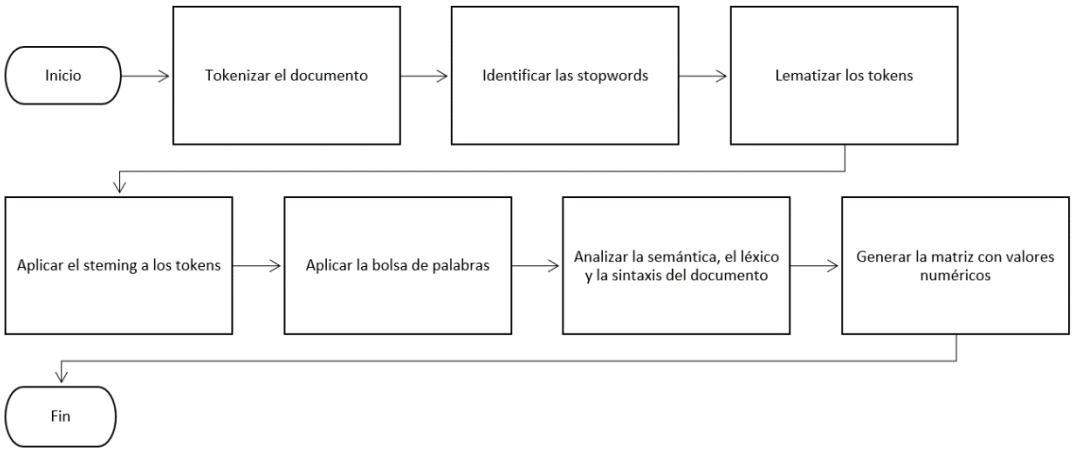


Figura 2 *Diagrama de flujo lineal del tamizado de textos.*

Fuente: Elaboración propia.

Junto con el procesamiento de los documentos que se utilizarán, diversas investigaciones la complementan con mediciones sobre la relevancia que los tokens tienen. De esta forma se depuran aún más los insumos que se emplearán en la construcción del modelo [1,2], como se muestra en la figura 3.

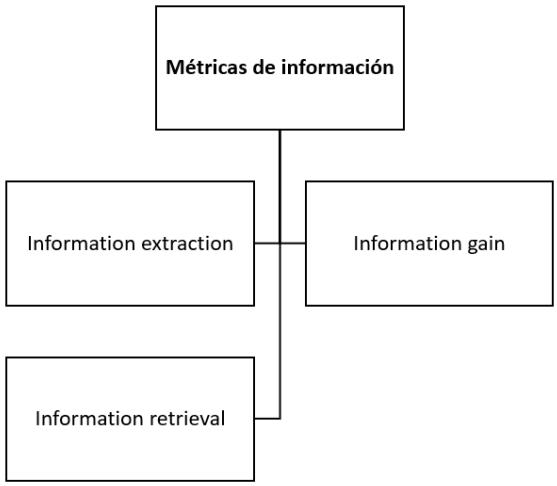


Figura 3 *Métricas de la información de los datos.*

Fuente: Elaboración propia.

Cuando los textos han sido transformados a tokens que aportan información que se considera valiosa, son procesados para construir los modelos de aprendizaje computacional. En el rubro de las aplicaciones financieras se agrupan en cuatro categorías. 1) el análisis de series de tiempo, 2) el análisis de sentimiento, 3) los sistemas de preguntas y respuestas del tipo (IF/THEN) y, 4) la optimización de los portafolios de inversión [1, 3, 4, 7], como se presentan en la figura 4.

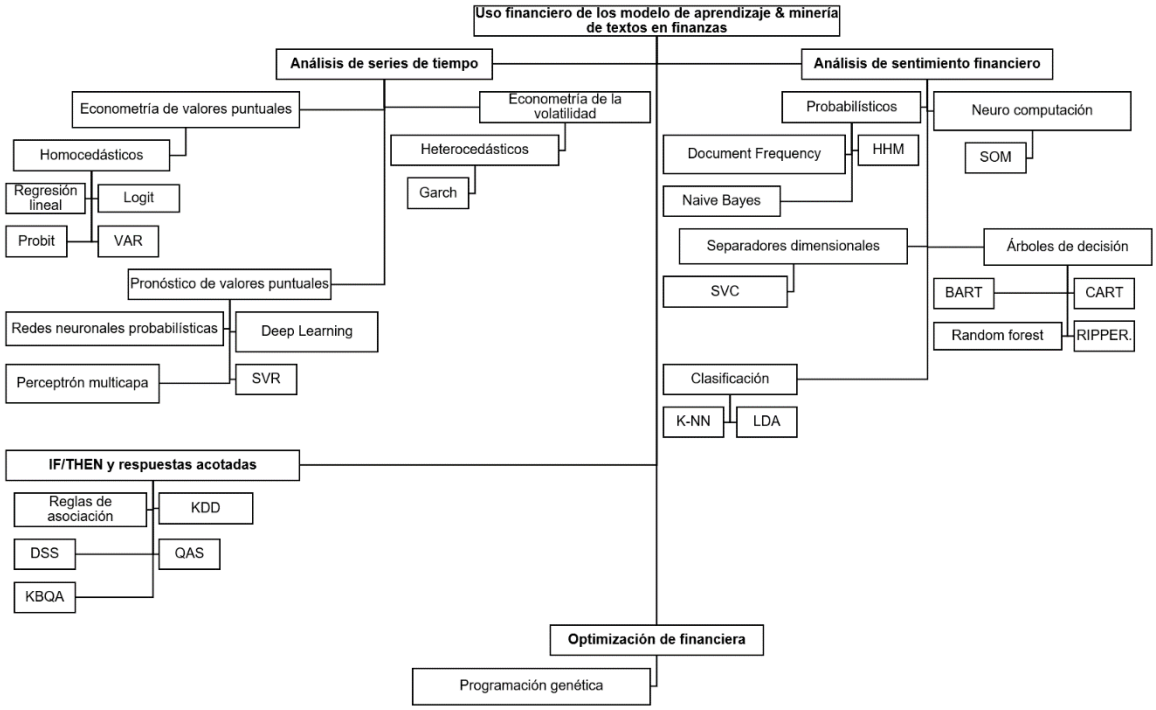


Figura 4 *Clasificación de los modelos de aprendizaje computacional.*

Fuente: Elaboración propia.

En materia del análisis de series de tiempo, la principal perspectiva que se ha documentado es la econométrica, que es el campo de estudio que se ocupa de procesar la información numérica que se encuentra indexada al tiempo. Lo que se ha buscado es pronosticar la volatilidad mediante las variaciones del modelo GARCH.

Para el pronóstico de valores financieros basado en los modelos de regresión, se han basado en la minimización de la varianza, ya sea ésta homocedástica o heterocedástica. Los modelos más empleados son los regresores lineales, logit, probit y los vectores autorregresivos (VAR). Para la incorporación de informaciones no cuantitativas, es común el empleo de variables dummy.

Por otra parte, se emplean diversos modelos de redes neuronales o máquinas de soporte vectorial (SVR), ambos empleados para hacer regresiones. En estos modelos, es posible incorporar la matriz de valores numéricos como producto del preprocesamiento de los textos como el exhibido en la figura 2.

El análisis de sentimiento busca clasificar el contenido de los documentos financieros para tipificarlos en valores binarios. Para obtener la clasificación, se emplean diversos modelos como los mapas autoorganizados (SOM), la máquina de soporte vectorial para clasificar (SVC), árboles de regresión bayesianos (BART), árboles de regresión y de clasificación (CART), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), K vecinos cercanos (K-NN), Latent Dirichlet Allocation (LDS), variaciones de métodos autorregresivos en los modelos basados en transformers conocidos como Bidirectional Encoder Representations from Transformers (BERT), entre otros más.

Las investigaciones relacionadas con la construcción de sistemas del tipo IF/THEN, han creado sistemas de consulta del conocimiento (KDD), sistemas que apoyan la toma de decisiones (DSS), sistemas de preguntas y respuestas sobre temas específicos (QAS/KBQA) [7, 13, 14, 15]; o bien, construir reglas de asociación a partir de consultas al contenido de los documentos ya preprocesados en la figura 2.

Es una práctica común que la optimización financiera de los portafolios de inversión se utilice la optimización de portafolios propuestos por Markowitz que minimiza la varianza de los retornos esperados de los instrumentos que lo componen [16]. Sin embargo, dichos modelos emplean varianzas homocedásticas

que tienen limitaciones para capturar el comportamiento no lineal de los activos empleados.

Desde la perspectiva del aprendizaje computacional se ha investigado en la construcción de portafolios que buscan modelar las no linealidades mediante la implementación de algoritmos evolutivos en modelos de optimización múltiple, o empleando optimización dinámica estocástica complementada con funciones difusas [17,28].

Cuando el modelo de aprendizaje computacional ya fue construido y entrenado, éste es evaluado estadísticamente. Dependiendo del tipo de resultado se emplean las técnicas para los modelos de regresión como el error porcentual absoluto medio (MAPE), el error cuadrático medio (MSE). Otra forma es mediante la construcción de una clasificación como las curvas de Receiver Operating Characteristic (ROC), el área bajo la curva de ROC (AUC), entre otras, como se presenta en la figura 5.

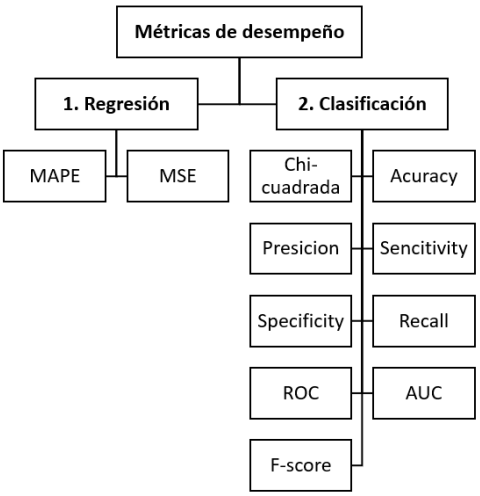


Figura 5 *Métricas del rendimiento de los modelos de aprendizaje computacional.*
Fuente: Elaboración propia.

2.6 Áreas de oportunidad y de manejo de las aplicaciones del aprendizaje computacional en finanzas

Con base en la revisión de la literatura estudiada, se indica que las principales aplicaciones se enfocan en la ciberseguridad, gestión de inversiones y la identificación de tendencias o pronósticos de los activos específicos mediante el procesamiento de documentos. Sin embargo, se observó que aún hay limitantes o problemas que esperan ser resueltos. Para los efectos de esta investigación, se desarrollaron cuatro grupos de oportunidades en este campo.

1. En materia *metodológica* y con foco en las finanzas, se ha observado la ausencia de una metodología universalmente aceptada para predecir el valor futuro de un activo financiero, por lo que se considera relevante el contar con un marco de trabajo que facilite el modelado de soluciones. Hasta el momento las ontologías desarrolladas se han limitado a la información de los instrumentos que se comercializan en los mercados de capitales, omitiendo por completo los anuncios relacionados con la política monetaria [1].
2. En materia de *procesamiento de los textos financieros*, desde la perspectiva de la ciencia de datos, no existe un sólo modelo de aprendizaje computacional que sea capaz de ofrecer los mejores resultados. Para mejorar la precisión de los pronósticos o de las clasificaciones, es necesario construir modelos de ensamble. Se ha observado que no se cuenta con técnicas especializadas por lo que se aplican técnicas agnósticas [1, 5].

Amalgamando el PLN y las finanzas, el análisis de la subjetividad contenida en los documentos aún es un problema por resolver, que se ha motivado por la ausencia de una ontología relacionada con la información económica, financiera y de aseguradoras [1].

No se ha investigado lo suficiente los motivos por los cuales el incremento de la imprecisión de la predicción o la clasificación de los valores de los instrumentos se incrementa cuando se incorporan noticias desfavorables, con independencia del modelo empleado [1, 6].

La mayoría de las investigaciones en el procesamiento de textos, han empleado textos cortos, o los encabezados de las noticias. Estos contenidos no

proporcionan la suficiente información a los modelos, por lo que se ha vuelto necesario el sofisticarlos para incrementar su precisión [1, 5, 6, 7].

3. En materia de las *fuentes de información financiera*, la oferta de un vocabulario financiero para el procesamiento de los documentos, ofrecen en lo individual un corpus limitado, poco estandarizado que, adolece de calidad en su contenido. Se considera pertinente un estudio de benchmark para ponderar su valor de acuerdo con el tipo de aplicación que se desea abordar [1, 3, 6]. Hay poca investigación sobre los análisis automatizados de los estados financieros de las empresas [1].

Pese a existir documentación en materia de política monetaria, la cual incide directamente en el comportamiento económico de un país, no se encontró evidencia que dichos documentos hayan sido investigados en el campo de las ciencias de datos en el período de 1984-2020. Sin embargo, se encontró uno para su próxima publicación en el año 2023 que emplea el PLN en un modelo VAR que es valorado con funciones de impulso-respuesta [29].

4. En materia de la ciberseguridad, se cuenta con pocas fuentes de datos, por lo que se recomienda el incrementar la oferta para que los modelos computacionales de aprendizaje incrementen su eficiencia al momento de identificar las amenazas informáticas que valoran, evolucionando de modelos estáticos a dinámicos [1].

A nivel de sistemas informáticos, es necesario construirlos para que sean tolerantes a fallas, lo que sugiere la implementación de metodologías de ingeniería de software correctamente aplicadas [1].

Las investigaciones se han enfocado en el acceso a las llamadas a las funciones de las bibliotecas de los sistemas operativos, lo que hace necesario incrementar la investigación en otras formas de intrusión en los sistemas informáticos [1].

La mayoría de los trabajos realizan clasificaciones del tipo de ataque, de estos el método más empleado es K-NN, lo que denota un uso limitado en los modelos de aprendizaje computacional [1].

Capítulo 3
Marco metodológico

Capítulo 3. Marco metodológico

3.1 Metodología

Para este proyecto se empleará la metodología CRISP-DM. Ésta considera seis pasos básicos para abordar cualquier proyecto de ciencia de datos, iniciando con 1) entendimiento del negocio, 2) entendimiento de los datos, 3) preparar los datos, 4) construcción del modelo, 5) evaluación del modelo y, 6) puesta en producción. Cada paso contiene diversas actividades [18] como se presenta en la figura 6.

El método CRISP-DM fue elegido porque es un método iterativo explícito entre sus diversas fases de ejecución. Así mismo, proporciona una guía de los pasos a seguir, lo que resultó de gran ayuda en el desarrollo del presente proyecto aplicativo.

Para adecuar el uso de CRISP-DM a los requerimientos académicos de INFOTEC, se optó por responder los siguientes puntos que el programa de maestría está solicitando y que se detalla a continuación.

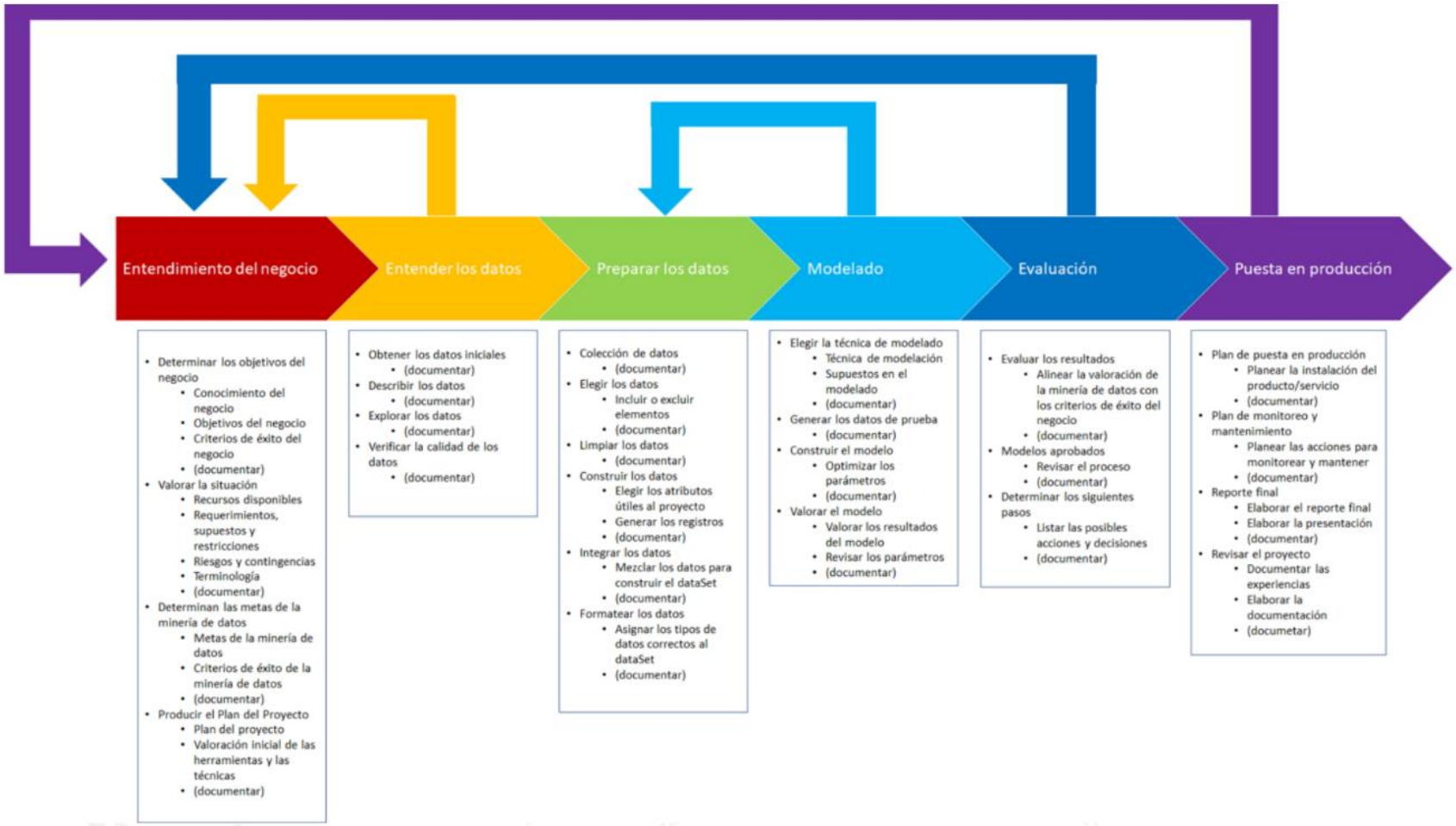


Figura 6 *Los seis pasos de la metodología CRISP-DM.*

Fuente: [18], 2006.

3.2 Propuesta de investigación

Considerando los pasos de CRISP-DM, se emplearán: 1) Entender el negocio, el cual se satisface con la introducción, los objetivos y la revisión de la literatura que se ha presentado.

2) La factibilidad de la investigación, considerando criterios éticos y de INFOTEC. La información que se empleará es del dominio público y su uso no está limitado para ser utilizado en el presente proyecto. Así mismo, no se han identificado hasta el momento, que las actividades que se desarrollarán contravengan principios morales o éticos en los campos académicos, profesionales o religiosos.

3) El análisis y la confiabilidad de los datos a utilizar. Se realizarán las actividades de los pasos 2. Entender los datos y 3. Preparar los datos, donde se otorgará un significado de negocio a la información que se empleará. Así mismo, se determinará si es necesario aplicarles procesos de limpieza y preprocesamiento para incrementar su calidad.

Al trabajar con documentos y con series de tiempo, las técnicas que se emplearán para cada conjunto de datos serán diferentes como se presentan a continuación.

3.1) Para analizar los documentos, inicialmente se utilizarían los pasos descritos en la figura 2 que identifican los tokens más relevantes en los documentos.

3.2) Para las series de tiempo, se verificará manualmente si todos los datos están presentes, y en el caso de encontrar ausencias, se procederá a generarlos mediante el criterio de calcular el promedio entre los valores anterior y posterior al que se está analizando[6]. Una vez identificado que los datos están completos, se procederá a separarlos en, 70% datos de entrenamiento y 30% datos de pronóstico.

3.3) Considerando que la estimación de la tasa de referencia es parte de las actividades de política monetaria del Banxico, se consultará al Dr. Guillermo Benavides Perales las dudas que en este tema se presenten.

4) Aplicar los algoritmos matemáticos y probar la precisión de sus resultados. Se aplicarán los pasos 4. Modelado y 5. Evaluación de CRISP-DM, considerando la clasificación de la variación en puntos base de la TII que puede ser visto como un problema de clasificación de textos. Se utilizará la biblioteca EvoMSA y se consultará al Dr. Mario Graff Guerrero para su uso.

5) Predicciones y análisis de resultados. Se empleará el modelo entrenado y se probará con los datos de prueba que se eligió en el inciso 3.2. La precisión estadística se computará con las métricas precisión, recall y F1-score, figura 5. Con los resultados obtenidos se elaborarán las conclusiones.

3.3 Cronograma de actividades

Las actividades involucradas en el desarrollo del proyecto consideran 13 puntos que se encuentran distribuidos desde el segundo semestre hasta el cuarto del programa de maestría. Las primeras actividades iniciaron el 8/enero/2022 y se espera terminarlas el 28/octubre/2023. El avance obtenido hasta el 12/febrero/2023 ronda el 94% de aquellas que se identificaron en el segundo semestre como se presenta en la figura 7.

Semestre 2/4	Porcentaje de avance
1. Definir el marco teórico	
• Estudiar el estado del arte	81%
• Delimitar los capítulos	81%
• Valorar la pertinencia de los subtemas	81%
2. Redactar el marco teórico	
• Definir cuál será la forma de citar textos, cuadros, gráficas y tablas	100%
• Definir el estilo para citar las referencias bibliográficas, hemerográficas, sitios Web, y cualquier otra fuente de información	100%
• Considerar las normas y los criterios que Infotec tenga sobre este apartado	100%
3. Identificar el problema de investigación	
• Delimitar el problema a investigar	100%
• Elaborar los objetivos de la investigación	100%
• Elaborar las hipótesis de investigación	100%
4. Seleccionar los datos	
• Definir los documentos de textos a utilizar	100%
• Establecer los criterios para determinar si la muestra de documentos ofrece información confiable y estable	100%
5. Procedimientos analíticos	
• Elegir los criterios y las pruebas estadísticas que se emplearán en la recolección de los datos	100%
• Elegir las pruebas estadísticas más adecuadas para el análisis de los datos	83%

Semestre 3/4	Porcentaje de avance
6. Análisis inicial de los resultados	0%
• Determinar cuán vinculado se encuentran las hipótesis con los resultados obtenidos	0%
• Considerar las normas y los criterios de Infotec	0%
7. Descripción de los resultados	
• Determinar cuán vinculado se encuentra el marco teórico con los resultados obtenidos	0%
• Considerar las normas y los criterios que Infotec tenga sobre este apartado	
8. Presentación de los resultados	
• Organizar los resultados de acuerdo a la estructura del documento	0%
• Estandarizar la presentación de los resultados, cuadros, gráficos y tablas	0%
• Considerar las normas y los criterios que Infotec tenga sobre este apartado	0%
9. Alcance y limitaciones de los resultados	0%
• Identificar las posibles líneas de investigación futuras a partir de los huecos identificados en el proyecto de investigación	
• Determinar el desarrollo de nuevos conocimientos a partir de los resultados	0%
• Considerar las normas y los criterios que Infotec tenga sobre este apartado	0%
10. Bosquejar las conclusiones	
• Organizar los apartados	0%
• Considerar las normas y los criterios que Infotec tenga sobre este apartado	0%

Semestre 4/4	Porcentaje de avance
11. Redacción del documento final	0%
• Vincular los resultados con el planteamiento del problema	0%
• Vincular los resultados con el objetivo	0%
• Vincular los resultados con el marco teórico	0%
12. Elaboración de los índices	
• Índices de tablas, cuadros y gráficos	0%
• Índices para los anexos	0%
13. Elaborar los alcances y las limitaciones del estudio	
• Redactar las limitaciones metodológicas identificadas a lo largo del proyecto	0%
Elaborar las sugerencias para reducir las limitaciones metodológicas	0%

Figura 7 *Avance de las actividades del proyecto.*

Fuente: Elaboración propia.

3.4 Recursos

Para el desarrollo del proyecto se cuenta con diversos recursos de información y de procesamiento de datos, que se encuentran disponibles, ya que varios de ellos son propiedad del investigador, o son de libre acceso en la Internet o en la biblioteca del Banxico. Estos se presentan a continuación.

- Fuentes de información para la construcción del modelo de aprendizaje publicados por el Banxico:
    - Las minutas de las reuniones del COPOM
    - Las series de tiempo disponibles en el SIE
- Recursos informáticos
    - Cómputo
        - Laptop de 6 núcleos físicos y 12 virtuales.
        - 16 GB de RAM.
        - Sistema operativo Windows 10.
        - Lenguajes de programación.
            - Python versión 3.11.
        - Herramientas para el procesamiento de datos.
            - Orange versión 3.31.
            - LabSOM versión 1.0.0.0. Laboratorio de Dinámica no lineal. Facultad de Ciencias. UNAM.
            - EvoMSA. Paquete informático desarrollado en INFOTEC especializado en el análisis de sentimiento multilenguaje.
    - Microsoft Office 365.
- Materiales bibliográficos, hemerográficos y documentos de investigación relacionados con los temas económicos y financieros.
    - Libros físicos
        - Baca G., Marcelino M., Ingeniería Financiera, Grupo Editorial Patria, 2018.

        - De Lara A., Medición y Control de Riesgos Financieros, Limusa, 2012, 2001.
        - De Lara A., Productos Derivados Financieros. Instrumentos, Valuación y Cobertura de riesgos, Limusa, 2012.
        - Hull J., Risk Management and Financial Institutions, Fourth Edition, Wiley Finance Series, 2009.
        - Lyuu Y., Financial Engineering and Computation. Principles, Mathematics, Algorithms, Cambridge University Press, 2002.
        - Marín J., Rubio G., Economía Financiera, Antoni Bosch editores, 2011.
        - Pickford J., Máster en Inversiones. La Guía Completa sobre Inversiones, Ediciones Deusto, 2004.
        - Tarkin L., Tarquin A., Ingeniería Económica, McGraw Hill, 8ª edición, 2020.
        - Kozikowski Z., Finanzas Internacionales, McGraw Hill, 2007.
        - Van Deventer D., Imai K., Mesler M., Advanced Financial Risk Management, Second Edition, Wiley Finance Series, 2013.
        - Venegas F., Riesgos Financieros y Económicos, Segunda Edición, Cengage, 2008.
    - Documentos electrónicos en la Internet.
        - Banxico, Minutas del Comité de Política Monetaria del Banco de México.
        - Heath, J., Lo que Indican los Indicadores. Cómo utilizar la información estadística para entender la realidad económica de México, INEGI, 2012.
        - Heath, et al, Lecturas en Lo que Indican Los Indicadores. Cómo utilizar la información estadística para entender la realidad económica de México, volumen I, II y III, INEGI-MIDE-Banxico.
- Bases de datos de documentos de investigación.
    - Elsevier.
- Aprendizaje Computacional.

- Alpaydin E., Introduction to Machine Learning, Second Edition, The MIT Press, 2010.
- Bently J.L., et al., An Almost Optimal Algorithm For Unbounded Searching, November 1975.
- Bengfort B., Bilbro R., Ojeda, T., Applied Text Analysis with Python. Enabling Language-Aware Data Products with Machine Learning, O'reilly, 2018.
- Hernández J., Ramírez M.J., Ferri C., Introducción a la Minería de Datos, Pearson Prentice Hall, 2007.
- Manning C.D., Raghavan P., Scütze H., An Introduction to Information Retrieval, Cambridge University Press, 2009.
- Rusell S., Norving P. Inteligencia Artificial. Un Enfoque Moderno, 2da. Edición, Pearson Prentice Hall, 2004.
- Téllez E., et al, A case study of Spanish text transformation for twitter sentiment analysis, Elsevier, 2017

- Metodología de la Investigación
  - Chapman, P., Khabaza T., Shearer C., CRISP-DM 1.0. Step-by-step data mining guide, SPSS Inc., 2000.
  - Han, J., Kamber M., Data Mining: Concepts and Techniques, Second edition, Elsevier Morgan Kaufmann Publishers, 2006.
  - Ramírez F., Notas del curso Seminario de Proyectos I, INFOTEC, 2022.
  - C. Pérez López, Minería de datos: técnicas y herramientas, Thomson, 2007.
  - L. Joyanes Aguilar, Big Data: análisis de grandes volúmenes de datos en organizaciones, México: Alfaguara, 2013.
  - V. Mayer-Schönberg, C. Kenneth, Big data: la revolución de los datos abiertos, Madrid: Turner publicaciones, 2013.
  - M. García-Alsina, Big data: gestión y explotación de grandes volúmenes de datos, Barcelona: Editorial UOC, 2017.

Capítulo 4

Modelando el clasificador de las variaciones de la tasa de referencia

Capítulo 4. Construcción del modelo de clasificación

El proceso de construcción del modelo de clasificación se basará en los pasos de la metodología CRISP-DM que se presentó en la unidad 3 y que se desarrolla a continuación.

4.1 Entendimiento del negocio

Como se documentó en el capítulo 1, la TII proporciona información relevante a los tomadores de decisión en el corto plazo. Su valor es dado a conocer por el Banxico a través de la publicación de las minutas que se publican en su página de Internet [2].

Dichas minutas son documentos que exponen la información económica y financiera relevante con la que se basan los miembros del COPOM para decidir si la tasa de referencia se reducirá, aumentará o mantendrá su valor futuro.

4.2 Entendimiento de los datos

Considerando que el objetivo de este proyecto es el procesar la información del contenido de las 98 minutas del COPOM para tipificar las variaciones en puntos base de la TII. Se han recolectado todos los documentos disponibles, cuya portada se presenta en la figura 8 y la colección de archivos PDF se ilustran en la figura 9.

Figura 8 *Portada de la sesión 93 del COPOM.*

Fuente: Banxico, Minuta número 93 del COPOM del Banxico, 2022

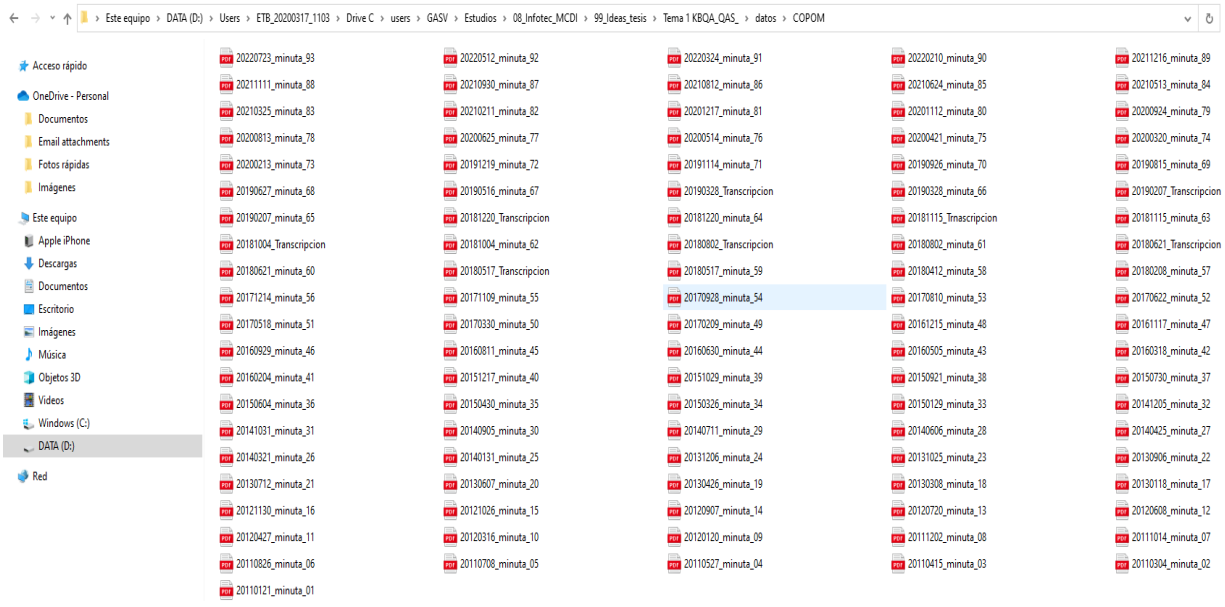


Figura 9 *Archivos PDF de las minutas del COPOM.*

Fuente: Elaboración propia.

Tomando en cuenta que este proyecto se puede tipificar como un modelo de clasificación que amalgama como variables de entrada textos y como salida la

descripción de la variación de la tasa de referencia, se recolectaron en el archivo de Excel “Datos numéricos para JSON.xlsx” las series de tiempo de las variables TII, TIIE, FIX, INPC y se calcularon sus variaciones en puntos base. Esta información se obtuvo del SIE del Banxico y se ilustran sus valores en la figura 10.

In [19]:

```
# Obtener los datos numéricos de las series de tiempo
# El archivo se encuentra en la ETB en: *C:/Users/J10240/Desktop/MCDI-SIE/dat*

import pandas as pd
#ruta="../dat/Datos originales.xlsx"
ruta="../dat/xls/Datos numericos para JSON.xlsx"
pestaña="P04_Salida_JSON"
archivo_excel=pd.read_excel(ruta, sheet_name=pestaña,index_col=0)
archivo_excel

```

Out[19]:

Minuta	TII	TIIE_promedio	FIX	INPC	Cambio_TII	Cambio_TIIE_promedio	Cambio_FIX	Cambio_INPC	Cambio_IPC-BMV
1	4.5	4.556667	12.0482	3.78	0.00	0.000000	-0.0421	0.0	-263.570313
2	4.5	4.480000	12.0064	3.04	0.00	0.000000	-0.0304	0.0	-232.140625
3	4.5	4.503333	11.7090	3.36	0.00	-0.006667	-0.0589	0.0	-81.582032
4	4.5	4.353333	11.6256	3.25	0.00	0.013333	-0.0700	0.0	76.988281
5	4.5	4.453333	11.6337	3.55	0.00	0.000000	0.0599	0.0	-83.437500
...	...	...	...	...	...	...	...	...	...
89	5.0	5.003333	20.9238	7.36	0.50	0.466667	-0.5056	0.0	1215.710938
90	5.5	5.496667	20.4148	7.28	0.50	0.553333	-0.0820	0.0	467.351563
91	6.0	5.953333	20.1313	7.45	0.50	0.556667	-0.2223	0.0	281.371093
92	6.5	6.546667	20.3200	7.65	0.50	0.513333	-0.1149	0.0	303.667969
93	7.0	6.980000	20.0365	7.99	0.75	0.700000	-0.1758	0.0	0.000000

93 rows × 9 columns

Figura 10 *Series de tiempo de la TII.*

Fuente: Elaboración propia.

Las fuentes de datos que se ilustran en las figuras 9 y 10 se deberán amalgamar en un solo archivo para poderlos preprocesar. Para este efecto, se construyó el archivo COPOM.JSON como se presenta en la figura 11.

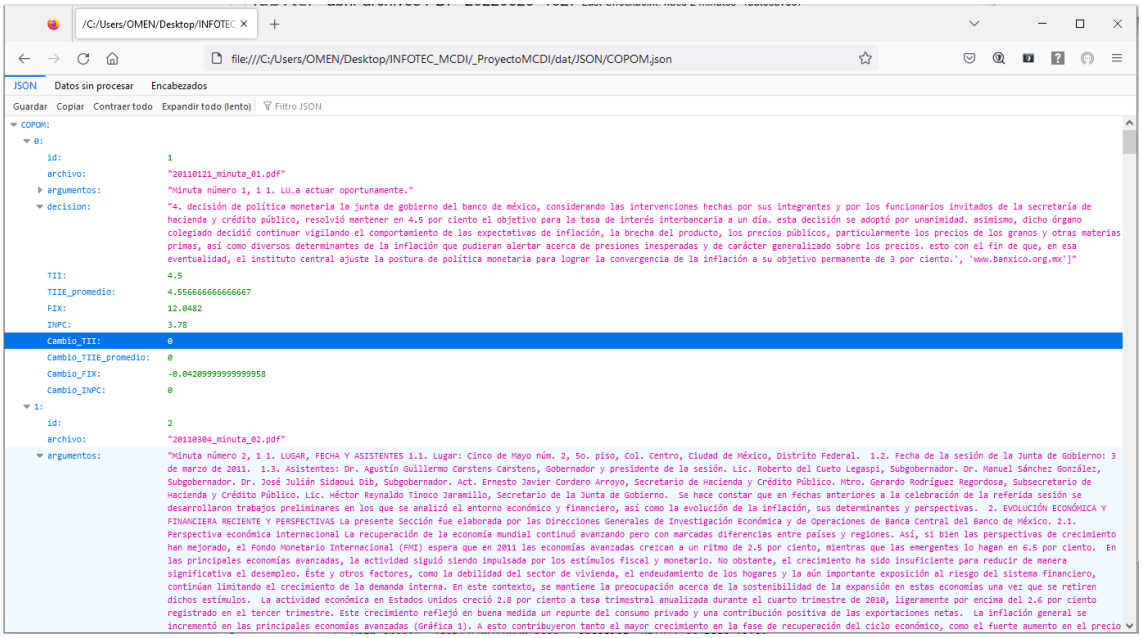


Figura 11 *Contenido del archivo COPOM.JSON.*

Fuente: Elaboración propia.

El contenido de la figura 11, se compone de 7 campos que contienen entre sus dos principales valores, cada minuta y las variaciones en puntos base de la TII, en los campos “argumentos” y “Cambio_TII”. El resto de los campos se conservaron para fines informativos como, el número de la sesión del COPOM “id”, el nombre del archivo de cada minuta “archivo”, los valores observados de las tasas como la “TII“ y “CAMBIO_TII“. En el cuadro 3 se presentan los metadatos de los citados campos.

A partir de la información contenida en el archivo COPOM.JSON, se procedió a limpiar el campo “argumentos” considerando los pasos descritos en la figura 2. Estos fueron codificados en Python (3.11) en la función *obtenerTokens(.)* que incorpora una serie de parámetros para configurar la limpieza y al terminar este proceso proporciona una lista con los tokens depurados. Los parámetros empleados se presentan en el cuadro 4 y un ejemplo de procesar todos los documentos del COPOM en la figura 12.

Nombre del campo	Tipo de dato (python)	Descripción
id	int	Número de la sesión del COPOM de la que se obtuvieron los datos del registro.
archivo	str	Nombre del archivo PDF que contiene la información de la reunión del COPOM.
argumentos	str	Contenido de texto que contiene los razonamientos que se esgrimieron para tomar la decisión.
decisión	str	Contenido de texto que contiene la resolución sobre la variación que se le aplicará a la tasa de referencia vigente.
TII	float	Valor numérico de la "tasa de interés interbancaria", también nombrada como "tasa de referencia".
Cambio_TII	float	Variación en puntos base de la tasa objetivo.
Descripcion_Cambio_TII	str	Contiene uno de los valores disponibles: {Sin_cambio, Bajar, Subir}

Cuadro 3 *Metadatos del archivo COPOM.JSON.*

Fuente: Elaboración propia.

Parámetro	Tipo de dato (Python)	Propósito
strTexto	str	Cadena de texto que contiene los tokens.
eliminarPalabrasLargas	Booleano	Eliminar las palabras que tengan una longitud mayor a 30 caracteres[7].
eliminarRepeticiones	Booleano	Eliminar del resultado los tokens repetidos.
usarBowCopom	Booleano	Utilizar la bolsa de 10,714 términos que se consideran válidos dentro de los documentos del COPOM.
usarStopwords	Booleano	Eliminar las palabras que no aportan información al texto.
lematizar	Booleano	Aplicar la función *stem(.)* para extraer la raíz de cada token.
ordenarSalida	Booleano	Entregar los tokens ordenados ascendentemente.

Cuadro 4 *Parámetros empleados en la función obtenerTokens(.).*

Fuente: Elaboración propia.

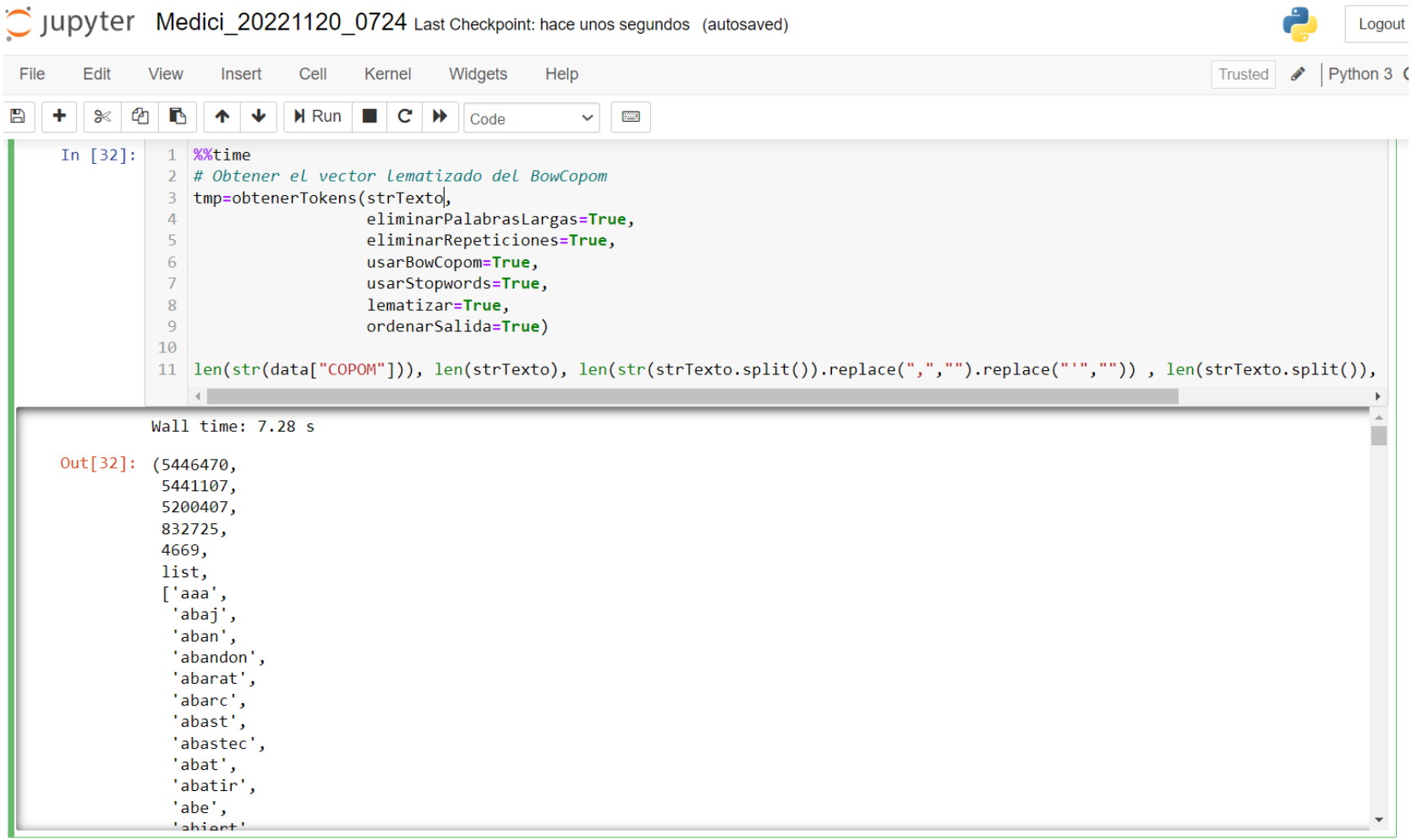


Figura 12 *Análisis de los tokens de las minutas del COPOM.*

Fuente: Elaboración propia.

El procesamiento de los 98 archivos del COPOM arrojaron que conjuntamente contienen, 5.4 millones de caracteres originales que al ser depurados se redujeron en 4.52% produciendo 5.2 millones. De estos últimos caracteres, se estimó que existen alrededor de 833 mil tokens, que al aplicar el proceso de limpieza se redujeron a 4.7 mil tokens, representando una reducción del 99.4%. En el Cuadro 5 se presentan los resultados obtenidos.

Concepto del cómputo	Resultado obtenido
Número de documentos PDF	98
Número de caracteres UTF-8 (texto original)	5’446,470
Número de caracteres UTF-8 (texto limpio	5’200,407
Número de tokens sin depurar	831,961
Número de tokens depurados	4,666

Cuadro 5 *Estadísticas del preprocesamiento de las minutas del COPOM.*

Fuente: Elaboración propia.

Recapitulando. Se puede decir que el modelo de clasificación procesará 98 documentos que contienen más de 5 millones de letras, de las cuales cerca de 5 mil son tokens relevantes para alimentar al modelo de aprendizaje computacional. Esto indica que con menos del 1% de los datos sería posible representar todas las minutas del COPOM dentro de una matriz de $98 \times 5{,}000$ que contendrá la presencia de $490{,}000$ tokens relevantes.

4.3 Preparación de los datos

Para procesar los más de 5 millones de palabras de las 98 minutas, se procedió a programar en Python la construcción del archivo COPOM.JSON que se comentó en la figura 11. Para ello, se aplicará el paso 3 de CRISP-DM considerando el diagrama de flujo de la figura 13 que se presenta a continuación.

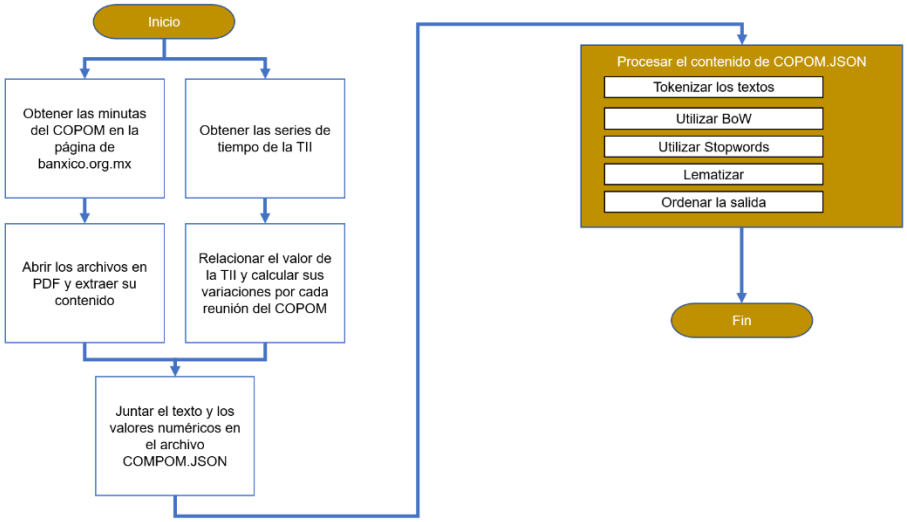


Figura 13 *Diagrama del procesamiento de los textos del archivo COPOM.JSON. Paso 3 de CRISP-DM.*

Fuente: Elaboración propia.

4.3.1 Obtener las minutas del COPOM de la página web del Banxico

Las minutas son archivos PDF que contienen los argumentos y las informaciones para tomar la decisión de cuál será la próxima variación de la tasa objetivo. Para extraer los textos, se colocaron las minutas en una carpeta y, el archivo rutaArchivosCOPOM.gil se almacenaron sus rutas absolutas.

Posteriormente, se iteró sobre los archivos para abrirlos y extraer los textos con la función *obtenerTextoPDF(.)* para almacenarlos en el diccionario dic_documentos como se ilustra en la figura 14.

```
%%time
#
# abrir una colección de documentos PDF, obtener sus textos y contar sus tokens
# la fuente de los PDF de COPOM es: https://www.banxico.org.mx/publicaciones-y-prensa/minutas-de-las-decisiones-de-politica-
conjuntoPalabras=set() #almacena las palabras únicas
palabrasDocumento=0 #cuenta las palabras de todo el documento

print("> Inicia el proceso de contar todas la palabras de todos los documentos...")

rutasArchivos="../dat/gil/rutaArchivosCOPOM.gil"

dic_documentos=dict() #{} # contenido de los documentos
cnt=0 #número de archivos procesados

with open(rutasArchivos) as rutas:
    for linea in rutas:
        linea=linea.replace("\n","")

        cnt=cnt+1
        if (cnt % 10)==0:
            print(" > Número de archivos procesados:",cnt)

        #
        # Obtener las palabras del documento en turno
        #
        rutaPdf=linea
        nombreArchivo, tmp=f.obtenerTextoPdf(rutaPdf) # obtener el contenido de un archivo PDF

        dic_documentos[nombreArchivo]=tmp # almacenar el contenido del archivo

        for pagina in tmp:
            palabrasDocumento=palabrasDocumento+len(pagina.split(" "))
            conjuntoPalabras=conjuntoPalabras.union(pagina.lower().split(" "))
rutas.close()
print("> Fin del proceso")

print("-"*100)
print("Número de archivos procesados..",cnt)
print("Número de elementos en el diccionario..",len(dic_documentos))
print("Palabras contadas...",palabrasDocumento)
print("Palabras diferentes..",len(conjuntoPalabras))
```

Figura 14 *Apertura de los archivos PDF de las minutas del COPOM.*

Fuente: Elaboración propia.

Con los textos almacenados en el diccionario dic_documentos, se extrajeron los argumentos y la decisión del valor de la TII de cada PDF. Cada argumento se almacenó en las listas lstArgumento y lstDesicion respectivamente, como se ilustra en la figura 15.

```
%%time
#
# iterar sobre los resultados para obtener los textos a procesar
# * Nombre del archivo
# * Argumentos
# * Desición de la política monetaria
#

lstArgumento=[]
lstDesicion=[]
lstArchivo=[]

for k,v in zip(dic_documentos.keys(),dic_documentos.values()):
    # valorar el contenido de la subcadena
    # extrar a partir de la decisión de política monetaria
    if str(v).lower().find("3. decisión de política monetaria")>0:
        tmp=str(v)[str(v).lower().find("3. decisión de política monetaria"):].lower()
    elif str(v).lower().find("4. decisión de política monetaria")>0:
        tmp=str(v)[str(v).lower().find("4. decisión de política monetaria"):].lower()
    elif str(v).lower().find("5. decisión de política monetaria")>0:
        tmp=str(v)[str(v).lower().find("5. decisión de política monetaria"):].lower()
    elif str(v).lower().find("6. decisión de política monetaria")>0:
        tmp=str(v)[str(v).lower().find("6. decisión de política monetaria"):].lower()
    else:
        strDecisionPoliticaMonetaria=":(" #No se encontró la decisión

    if tmp.find("votación")==-1 and tmp.find("anexo")==-1:
        #almacenar el contenido tal y como se extrajo
        strDecisionPoliticaMonetaria=tmp.strip()

    elif tmp.find("votación")<tmp.find("anexo"):
        #almacenar el contenido desde el inicio y antes de votación
        strDecisionPoliticaMonetaria=tmp[:tmp.find("votación")-4].strip()


    strArgumento = str(v)[:str(v).lower().find(strDecisionPoliticaMonetaria)].strip()
    #
    # eliminar caracteres especiales
    #
    strArgumento=strArgumento.replace("'" ,"" ,1_000_000)
    strArgumento=strArgumento.replace("[" ," " ,1_000_000)
    strArgumento=strArgumento.replace("]" ," " ,1_000_000)
    strArgumento=strArgumento.replace("\'"," " ,1_000_000)
    strArgumento=strArgumento.replace('"' ," " ,1_000_000)
    strArgumento=strArgumento.replace('\\x' ," " ,1_000_000) # control de escapa inválido
    strArgumento=strArgumento.replace('\\s' ," " ,1_000_000) # control de escapa inválido

    lstArchivo.append(k)
    lstArgumento.append(strArgumento)
    lstDesicion.append(strDecisionPoliticaMonetaria)

    #print(k,"\t",strDecisionPoliticaMonetaria[:33],"\t",strArgumento[-50:])
#
# almacenar el resultado en una lista
#

```

Figura 15 *Separación de los argumentos y las decisiones de cada minuta del COPOM.*

Fuente: Elaboración propia.

4.3.2 Obtener las series de tiempo de la TII

Para obtener los datos de la TII, se consultó el SIE del Banxico. De éste se extrajeron diversos indicadores económicos y financieros que se exportaron al archivo “Excel Datos numéricos para JSON.xlsx”. En éste se calcularon las variaciones de cada variable. El código que construye el archivo se muestra en la figura 16.

```python
# Obtener los datos numéricos de las series de tiempo
# El archivo se encuentra en la ETB en: *C:/Users/J10240/Desktop/MCDI-SIE/dat*

import pandas as pd
#ruta="../dat/Datos originales.xlsx"
ruta="../dat/xls/Datos numericos para JSON.xlsx"
pestaña="P04_Salida_JSON"
archivo_excel=pd.read_excel(ruta, sheet_name=pestaña,index_col=0)
archivo_excel
```

Minuta	TII	TIIE_promedio	FIX	INPC	Cambio_TII	Cambio_TIIE_promedio	Cambio_FIX	Cambio_INPC	Cambio_IPC-BMV	Unnamed: 10
1	4.5	4.556667	12.0482	3.78	0.00	0.000000	-0.0421	0.0	-263.570313	NaN
2	4.5	4.480000	12.0064	3.04	0.00	0.000000	-0.0304	0.0	-232.140625	NaN
3	4.5	4.503333	11.7090	3.36	0.00	-0.006667	-0.0589	0.0	-81.582032	NaN
4	4.5	4.353333	11.6256	3.25	0.00	0.013333	-0.0700	0.0	76.988281	NaN
5	4.5	4.453333	11.6337	3.55	0.00	0.000000	0.0599	0.0	-83.437500	NaN
...	...	...	...	...	...	...	...	...	...	...
93	7.0	6.980000	20.0365	7.99	0.75	0.700000	-0.1758	0.0	0.000000	0.5
94	8.5	NaN	NaN	NaN	1.50	NaN	NaN	NaN	NaN	1.5
95	9.2	NaN	NaN	NaN	0.70	NaN	NaN	NaN	NaN	0.7
96	10.0	NaN	NaN	NaN	0.80	NaN	NaN	NaN	NaN	0.8
97	10.5	NaN	NaN	NaN	0.50	NaN	NaN	NaN	NaN	0.5

97 rows × 10 columns

Figura 16 *Series de tiempo obtenidas del SIE del Banxico.*

Fuente: Elaboración propia.

4.3.3 Abrir los archivos PDF y extraer su contenido

Para abrir cada archivo PDF se utilizó la biblioteca PdfReader que obtiene una instancia de cada hoja del documento que es leída con la función *pages(.)* y almacenada en una variable de texto dentro de un ciclo. El resultado final es el nombre del archivo procesado y una lista con el contenido de todas sus páginas como se muestra en la figura 17.

```python
def obtenerTextoPdf(rutaPdf):
    """
    Objetivo... Obtener el texto de un archivo PDF
    Entrada.... rutaPdf.............. str. Ruta completa del archivo que se procesarán
    Salida..... strNombreArchivo..... str. Nombre del archivo PDF que se procesó
                lstPagina............ list. Texto que contiene el documento PDF
    Fuente..... https://urldefense.proofpoint.com/v2/url?u=https-3A__blog.facialix.com_extraer-2Dtexto-2Dy-2Ddatos-2Dde-2Dun-2Darchivo-2Dpdf-2Dcon-2Dpython_-255Cn&d=DwIGaQ&c=AKs6EwELrBZKOG9H-C2eL9nCFyT6KLG5z2zMuwOnNTA&r=N6WJs1Q98ajQ7lXHjYvXWE7uSZkSmL_zRXrKXQYqKro&m=oHBWwUAZwZpJI-zKdVcY1Nu80JE4jVZ3iA5WHiF0BpDjelqzgIhG4Qa_sZk3UPjb&s=vVthllCfQ5F9uUf75EaU-BjNInK-cURNoGRllsyMi3Y&e=     Programó... GASV_20220603;1225
                https://pypdf2.readthedocs.io/en/latest/user/extract-text.html
    """
    from PyPDF2 import PdfReader

    reader=PdfReader(rutaPdf) # abrir el archivo
    #
    # Obtener el nombre del archivo
    #
    strNombreArchivo=""
    for i in range(len(rutaPdf)-1,0,-1):
        if rutaPdf[i]=="\\" or rutaPdf[i]=="/":
            strNombreArchivo=rutaPdf[i+1:]
            break
    #
    # abrir cada una de las hojas del archivo pdf
    #
    lstPagina=[]
    for x in range(len(reader.pages)):
        page=reader.pages[x]
        try:
            texto=page.extract_text() #Obtener el texto de la página
            #
            # Limpiar el texto
            #
            texto=texto.replace(" \n"," ") #eliminar los saltos de línea antes de una palabra
            texto=texto.replace("\n "," ") #eliminar los saltos de línea después de una palabra
            texto=texto.replace("\n" ," ") #eliminar los saltos de línea entre las palabras o cifras

            texto=texto.replace("\\x" ," ") #eliminar caracter de escape inválido
            texto=texto.replace("\\s" ," ") #eliminar caracter de escape inválido

            for _ in range(0,10):
                for i in[2,3,5,7,11,13,17]:
                    texto=texto.replace(" "*i, " ")

            texto=texto.strip() #eliminar los espacios en los extremos de la cadena
            lstPagina.append(texto)
        except:
            texto="" # no hay texto que procesar

    return strNombreArchivo,lstPagina

```

Figura 17 *Código para abrir y extraer el contenido de un archivo PDF.*

Fuente: Elaboración propia.

4.3.4 Juntar el texto y los valores numéricos en el archivo COPOM.JSON

Para combinar los datos, se procedió a iterar sobre las diferentes listas que contienen la información. Para ello, se almacenó el contenido utilizando un formato JSON que almacenará en la variable tmp el contenido del archivo COPOM.JSON como se muestra en las figuras 18 y 19.

Se puede apreciar en las líneas 40 y 46, figura 18, que se analiza la variación de la TII. Si ésta tiene un valor de cero, se asignó una clasificación Sin_cambio. Si hubo un incremento se otorgó la descripción Subir, en otro caso se asignó Bajar. Con estas etiquetas se entrenará el modelo de aprendizaje computacional para clasificar las minutas.

```
%%time
#
# convertir los datos en un archivo JSON
#
import numpy as np
tmp=""
cnt=0

for k,a,d,l1,l2,l3,l4,l5,l6,l7,l8 in zip(lstArchivo,lstArgumento,lstDesicion,
                                         list(archivo_excel.TII),
                                         list(archivo_excel.TIIE_promedio),
                                         list(archivo_excel.FIX),
                                         list(archivo_excel.INPC),
                                         list(archivo_excel.Cambio_TII),
                                         list(archivo_excel.Cambio_TIIE_promedio),
                                         list(archivo_excel.Cambio_FIX),
                                         list(archivo_excel.Cambio_INPC)):
    if (cnt % 10)==0:
        print("> Avance:{}, {}%".format(cnt, np.round(100*cnt/len(lstArchivo),0)))

    cnt=cnt+1
    tmp = tmp + "\n"
    tmp = tmp + '{'
    tmp = tmp + "\n"
    tmp = tmp +         '"id":'  + str(cnt) + ', '
    tmp = tmp + "\n"
    tmp = tmp +    '"archivo":"' + k        + '", '
    tmp = tmp + "\n"
    tmp = tmp + '"argumentos":"' + a        + '", '
    tmp = tmp + "\n"
    tmp = tmp +   '"decision":"' + d        + '", '
    tmp = tmp + "\n"

    tmp = tmp +   '"TII":'                  + str(l1) + ', '
    tmp = tmp + "\n"

    tmp = tmp +   '"Cambio_TII":'           + str(l5) + ', '
    tmp = tmp + "\n"

    tmp = tmp +   '"Descripcion_Cambio_TII":'
    if l5==0:
        tmp = tmp +   '"Sin_cambio"'
    elif l5>0:
        tmp = tmp +   '"Subir"'
    else:
        tmp = tmp +   '"Bajar"'

    tmp = tmp + "\n"

    tmp = tmp + '},'

    #if cnt>2:
    #    break

tmp=tmp.strip()
tmp=tmp[: len(tmp)-1]
tmp='{"COPOM":[' + tmp + ']}'

print("Tamaño de la cadena JSON=",len(tmp))

```

Figura 18 *Código para construir el contenido del archivo COPOM.JSON.*

Fuente: Elaboración propia.

```
%%time
#
# grabar el contenido en u archivo JSON
#
with open('../dat/JSON/COPOM.json', 'w', encoding='utf-8') as f_out:
    f_out.write(tmp)
f_out.close()

```

```
Wall time: 24.1 ms
```

Figura 19 *Almacenamiento de los datos en COPOM.JSON.*

Fuente: Elaboración propia.

4.3.5 Limpiar los textos de aquellos elementos que no aportan valor

En este punto ya se cuenta con la información para obtener los tokens que son relevantes para la construcción del modelo de aprendizaje computacional. Inicialmente se procedió con la apertura del archivo COPOM.JSON y posteriormente, se aplicaron los diversos tamizados, figura 20.

```
# importar las funciones para el procesamiento de los textos del COPOM

import funcionesCOPOM_20230115_0835 as f
```

```
%%time
#
# abrir el archivo JSON
# Paso 100. Obtener los datos
#
import json
with open('../dat/JSON/COPOM.json', encoding='utf8') as json_file:
    data=json.load(json_file)
json_file.close()
```

```
Wall time: 38.9 ms
```

```
#data["COPOM"][62]
```

```
#type(data["COPOM"][62])
del data["COPOM"][62] # Eliminarlo por contener caracteres no legibles desde el origen
```

```

```

Figura 20 *Apertura del archivo COPOM.JSON.*

Fuente: Elaboración propia.

Se aplicó un diccionario de palabras válidas que se almacenó en el archivo BoW_Copom.gil que consta de más de 10 mil términos que fue elaborado por el tesista.

Con la apertura del archivo COPOM.JSON, figura 20, se calculó la distribución de las clasificaciones de las variaciones de la TII, donde el 17% fueron decisiones a la baja, el 55% a mantener su valor y el 28% al alza como se muestra en el gráfico 1.

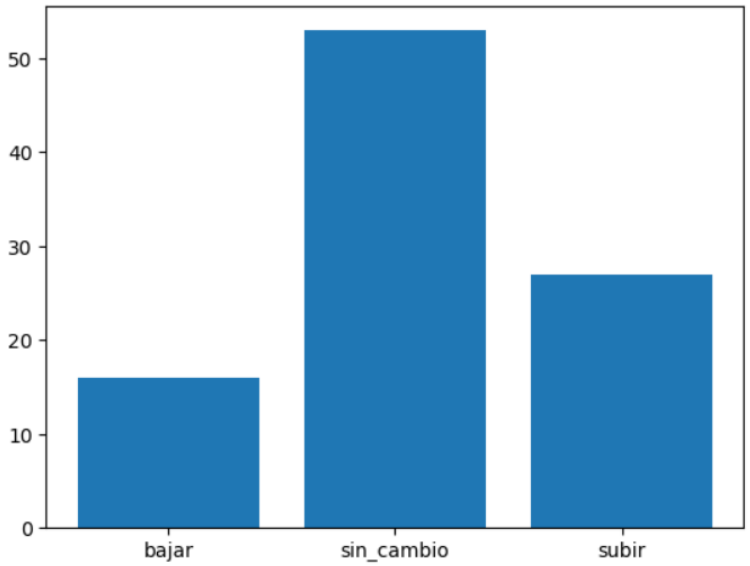


Gráfico 1 *Distribución de las clasificaciones de los argumentos para modificar el valor de la TII.*

Fuente: Elaboración propia.

4.3.6 Separar los datos de entrenamiento y de prueba

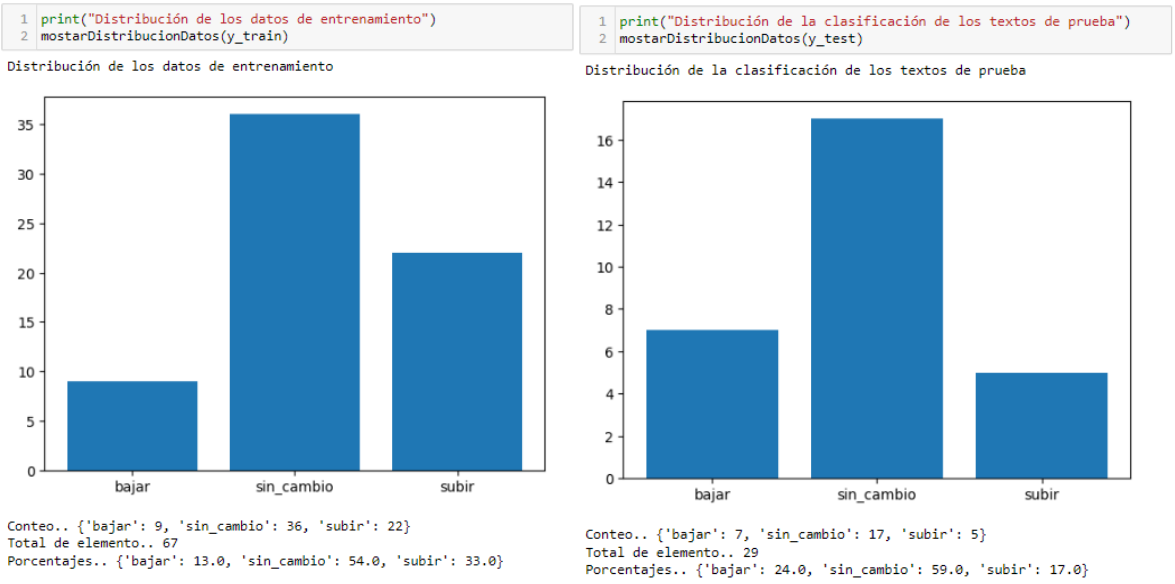
Para construir el clasificador, se utilizó el 70% de los datos para entrenarlo y el 30% para probarlo, figura 21. Se puede apreciar en el gráfico 2 (A) y (B) que las distribuciones de estos conjuntos son muy similares al mostrado en el gráfico 1 y estadísticamente iguales de acuerdo con la prueba de chi cuadrada con un p-valor del 5% como se muestra en el anexo II.

```
%%time
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test =  train_test_split(X,y, test_size=0.30, random_state=42)

len(X_train), len(X_test), len(y_train), len(y_test)
```

```
Wall time: 998 µs

(67, 29, 67, 29)
```

Figura 21 *Separación de los conjuntos de entrenamiento y prueba (70-30%).*

Fuente: Elaboración propia.



(A) (B)

Gráfico 2 *Distribuciones de las clasificaciones de la TII del conjunto de entrenamiento (A) y prueba (B).*

Fuente: Elaboración propia.

4.3.7 Aumentar los datos artificialmente

Se consideró que el tamaño del conjunto de entrenamiento era pequeño y ante la imposibilidad de obtener más minutas del COPOM, se procedió a incrementar su tamaño artificialmente a 10,000 observaciones [24]. Para ello se utilizó muestreo con reemplazo, gráfico 3, buscando que la probabilidad de cada documento fuese uniforme. Esto permitió que se respetara la distribución de la clasificación original como se presenta en el gráfico 4. Para más información, véanse los anexos II y III respectivamente para conocer los resultados de las pruebas de chi cuadrada.

A continuación, se procedió a reemplazar el 30% de los textos del conjunto de entrenamiento con el glosario contenido en el archivo COPOM_sinonimos_02.gil que contiene 552 términos que se obtuvieron de la Internet y del diccionario de términos que proporciona el Banxico en su página web [23] como se presenta en la figura 22.

Con el nuevo conjunto de entrenamiento y de pruebas, se procedió a almacenarlos en archivos JSON para posteriormente ser utilizados en el proceso de construcción del modelo de aprendizaje, como se muestra en las figuras 23 y 24.

```
#
# Generar nuevos artificialmente documentos e incorporarlos a X_train & y_train
#

n=10_000 # número de nuevos documentos a generar con random insert
idx=np.random.randint(0,len(X_train), n)

# Pintar la distribución de los valores
import matplotlib.pyplot as plt
plt.hist(idx)
```

```
(array([1087., 1049.,  894., 1054.,  888., 1048., 1061.,  872., 1035.,
        1012.]),
 array([ 0. ,  6.6, 13.2, 19.8, 26.4, 33. , 39.6, 46.2, 52.8, 59.4, 66. ]),
 <BarContainer object of 10 artists>)
```

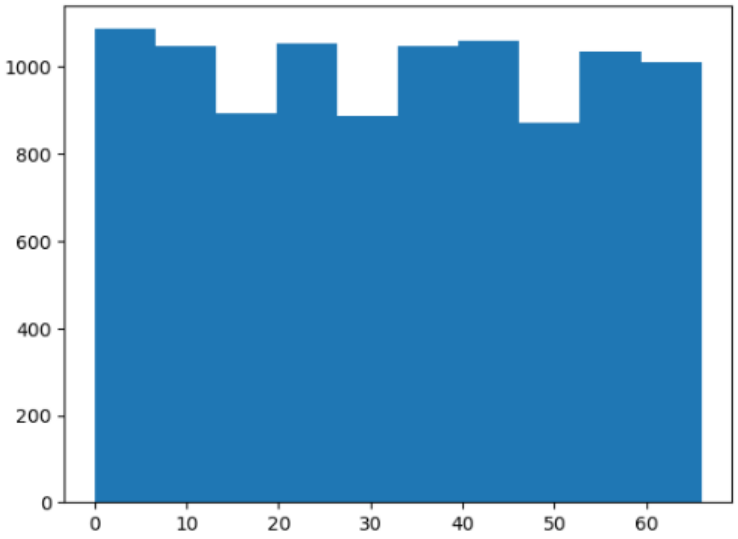
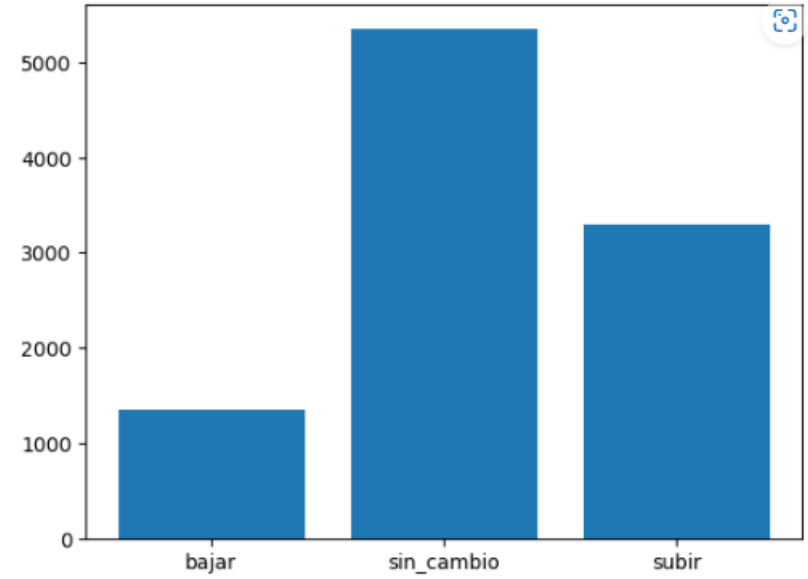


Gráfico 3 *Muestreo con reemplazo del conjunto de entrenamiento.*

Fuente: Elaboración propia.

```
print("Distribución de la clasificación de los textos antes de ser aumentados")
mostarDistribucionDatos([y_train[i] for i in idx])
```

```
Distribución de la clasificación de los textos antes de ser aumentados
```




```
Conteo.. {'bajar': 1360, 'sin_cambio': 5342, 'subir': 3298}
Total de elemento.. 10000
Porcentajes.. {'bajar': 14.0, 'sin_cambio': 53.0, 'subir': 33.0}
```

Gráfico 4 *Distribución de la clasificación de las 10 mil observaciones generadas.*

Fuente: Elaboración propia.

```
In [23]:
%%time
import numpy as np

dic_sinonimos=f.obtenerDiccionarioSinonimosCopom('../dat/gil/COPOM_sinonimos_02.gil')

# Copiar el contenido en listas temporales
X_train_augmented=[]
y_train_augmented=[]


for ii,i in enumerate(idx):
    texto=X_train[i] # Tomar el documento a modificar
    clasif=y_train[i] # Tomar la clasificación del cambio de la TII

    # Aplicar los sinónimos y las definiciones al texto
    cnt=-1 # No se modificó el texto
    if np.random.rand()<0.3:
        cnt,texto=f.incorporar_sinonimos(texto, dic_sinonimos, probabilidad_de_cambio=0.30) #Obtener el nuevo texto con los

    # Limpiar el nuevo texto
    texto=limpiarTexto(texto.lower()) # Limpiar los caracteres

    # agregar el nuevo texto con los sinónimos y su valor de clasificación
    X_train_augmented.append(texto)
    y_train_augmented.append(clasif)

    if ii%100==0 and ii>0:
        print("> avance=",np.round(100*ii/n,0)," ii=",ii+1," n=",n, "; i=",i, "; cnt=",cnt)


> avance= 82.0  ii= 8201  n= 10000 ; i= 1 ; cnt= -1
> avance= 83.0  ii= 8301  n= 10000 ; i= 33 ; cnt= 98
> avance= 84.0  ii= 8401  n= 10000 ; i= 15 ; cnt= -1
> avance= 85.0  ii= 8501  n= 10000 ; i= 55 ; cnt= -1
> avance= 86.0  ii= 8601  n= 10000 ; i= 37 ; cnt= 49
> avance= 87.0  ii= 8701  n= 10000 ; i= 54 ; cnt= 104
> avance= 88.0  ii= 8801  n= 10000 ; i= 65 ; cnt= -1
> avance= 89.0  ii= 8901  n= 10000 ; i= 37 ; cnt= -1
> avance= 90.0  ii= 9001  n= 10000 ; i= 42 ; cnt= -1
> avance= 91.0  ii= 9101  n= 10000 ; i= 7 ; cnt= -1
> avance= 92.0  ii= 9201  n= 10000 ; i= 60 ; cnt= -1
> avance= 93.0  ii= 9301  n= 10000 ; i= 32 ; cnt= -1
> avance= 94.0  ii= 9401  n= 10000 ; i= 3 ; cnt= -1
> avance= 95.0  ii= 9501  n= 10000 ; i= 54 ; cnt= -1
> avance= 96.0  ii= 9601  n= 10000 ; i= 65 ; cnt= -1
> avance= 97.0  ii= 9701  n= 10000 ; i= 64 ; cnt= -1
> avance= 98.0  ii= 9801  n= 10000 ; i= 56 ; cnt= -1
> avance= 99.0  ii= 9901  n= 10000 ; i= 46 ; cnt= -1
Wall time: 1h 42min 51s
```

Figura 22 *Aplicación de los sinónimos y términos al conjunto de los 10 mil textos generados.*

Fuente: Elaboración propia.

```
In [27]:
%%time
tmp=construirJson(X_train_augmented, y_train_augmented, nombre="COPOM_Entrenamiento")
len(tmp), len(X_train_augmented), len(y_train_augmented)

> Avance:8400, 84.0%
> Avance:8500, 85.0%
> Avance:8600, 86.0%
> Avance:8700, 87.0%
> Avance:8800, 88.0%
> Avance:8900, 89.0%
> Avance:9000, 90.0%
> Avance:9100, 91.0%
> Avance:9200, 92.0%
> Avance:9300, 93.0%
> Avance:9400, 94.0%
> Avance:9500, 95.0%
> Avance:9600, 96.0%
> Avance:9700, 97.0%
> Avance:9800, 98.0%
> Avance:9900, 99.0%
Wall time: 1h 53min 10s
Out[27]: (621806603, 10000, 10000)

In [28]:
%%time
#
# grabar el contenido en un archivo JSON
#
with open('../dat/JSON/COPOM_Entrenamiento.json', 'w', encoding='utf-8') as f_out:
    f_out.write(tmp)
f_out.close()

Wall time: 2.33 s
```

Figura 23 *Almacenamiento de los 10 mil textos generados.*

Fuente: Elaboración propia.

```
tmp=construirJson(X_test, y_test, nombre="COPOM_Prueba")
len(tmp), len(X_train_augmented), len(y_train_augmented)
```

```
> Avance:0, 0.0%

(1711992, 10000, 10000)
```

```
%%time
#
# grabar el contenido en u archivo JSON
#
with open('../dat/JSON/COPOM_Prueba.json', 'w', encoding='utf-8') as f_out:
    f_out.write(tmp)
f_out.close()
```

```
Wall time: 10.9 ms
```

Figura 24 *Almacenamiento de los datos de prueba.*

Fuente: Elaboración propia.

4.4 Modelado

Para construir el modelo de aprendizaje computacional, se procedió a abrir el archivo COPOM_Entrenamiento.json y almacenar los textos y su clasificación en las variables X_train y y_train como se muestra en las figuras 25 y 29, para ser procesados con EvoMSA.

```
%%time
#
# abrir el archivo JSON
# Paso 100. Obtener los datos
#
import json
strFile='../dat/JSON/COPOM_Entrenamiento_010K.json'

with open(strFile, encoding='utf8') as json_file:
    data=json.load(json_file)
json_file.close()
```

```
Wall time: 4.58 s
```

```
%%time
# Ontemer los datos de COPOM.JSON
#X_train=[f.limpiarCaracteres(d['argumentos']) for d in data["COPOM_Entrenamiento"]]
X_train=[d['argumentos']      for d in data["COPOM_Entrenamiento"]]
y_train=[d['movimiento_tii']  for d in data["COPOM_Entrenamiento"]]

len(X_train), len(y_train)
```

```
Wall time: 2.99 ms

(10000, 10000)
```

Figura 25 *Apertura del archivo de entrenamiento con los 10 mil textos.*

Fuente: Elaboración propia.

4.4.1 Características de EvoMSA

EvoMSA es el acrónimo de Evolutionary Multilingual Sentiment Analysis y tiene el propósito de facilitar los análisis de sentimiento a documentos de texto de naturaleza diversa escritos en diversos idiomas[8] gracias a la bolsa de palabras que incorpora. Así mismo, cuenta con algoritmos previamente entrenados que facilitan la construcción de las clasificaciones semánticas mediante la función (1).

$$g: \mathbb{R}^d \rightarrow \mathbb{R}^c \quad \ldots (1)$$

Siendo g la función que opera en un espacio vectorial multidimensional que se encuentran en el conjunto de los números reales $\mathbb{R}$ que se compone de las diferentes partes de la biblioteca de EvoMSA y que considera la cantidad d de los tokens que se pueden extraer de los documentos que son de interés por clasificar en los c grupos definidos.

La implementación computacional de (1) se presenta en la figura 26 la cual indica la pila de los 6 elementos que componen a la función g y que para los efectos de este proyecto aplicativo solamente se emplean los módulos TR B4MSA, HA B4MSA, FT Fast Text y EvoDAG y que se describen a continuación [21, 22, 25].

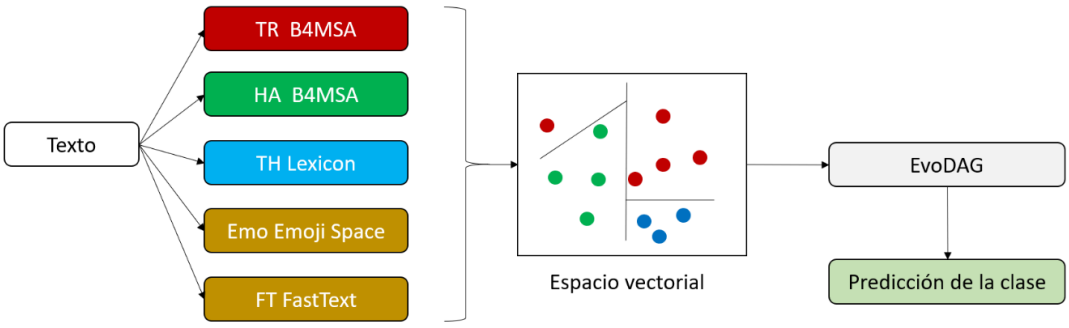


Figura 26 *Componentes de EvoMSA.*

Fuente: Documentación de EvoMSA.

[8] Los idiomas incorporados en EvoMSA son: alemán (de), árabe (ar), catalán (ca), chino (zh), coreano (ko), español (es), francés (fr), hindi (hi), indonesio (in), inglés (en), italiano (it), japonés (ja), neerlandés -Países Bajos- (nl), polaco (pl), portugués (pt), ruso (ru), tagalo -Filipinas- (tl) y turco (tr) [25]

TR B4MSA (TRainning set Base Multilingual Sentimen Analysis), realiza la limpieza de los textos del COPOM proporcionados mediante un tamizado basado en el TF-IDF y que son incorporados en una matriz de entrada de más de 4 millones de tokens que se emplearon para entrenar previamente a EvoMSA. Así mismo, este módulo construye un espacio vectorial multidimensional empleando una máquina de soporte vectorial lineal (SVC) que separa los d documentos y las c clasificaciones proporcionadas.

HA B4MSA (Human Annotated Base Multilingual Sentiment Analysis) incorpora las c clasificaciones de las minutas del COPOM sobre las que operará la función (1). Esta funcionalidad se puede complementar con el módulo TH Lexico (Thumbs Up-Down) la cual incorpora una valoración positiva o negativa al token que se esté procesando al momento de construir el clasificador.

El módulo *Emo Emoji Space*, contiene una colección de emoticonos que ya cuentan con una preclasificación de su significado. Este módulo es de gran valor cuando se procesan documentos que los contienen. En el caso de este proyecto, no se emplea este módulo.

Finalmente, *FT Fast Text* proporciona una representación semántica de los tokens que son empleados en la clasificación y proporciona una probabilidad de que estos se encuentren en los conjuntos de entrenamiento d y de clasificación c.

Cuando el espacio multivectorial de entrenamiento de las minutas del COPOM es construido, se procede a utilizar el módulo EvoDAG (Evolutionary Dynamic Algorithm Genetic) que se ocupa en construir el modelo de predicción basado en programación evolutiva. Ésta considera diversas funciones de clasificación que son combinadas para encontrar al mejor clasificador disponible.

4.4.2 Entrenando a EvoMSA con la función TextRepresentation

El modelo de aprendizaje computacional a utilizar será un modelo de clasificación que fue construido por investigadores del INFOTEC [21]. Se empleará la función *TextRepresentation(.)* por ser un clasificador que relaciona un texto con su respectiva etiqueta [26] como se exhibe en la figura 26.

La idea fundamental en la implementación de *TextRepresentation(.)* es la de relacionar en un espacio vectorial los tokens de los textos que se desean clasificar. En la transformación de los tokens a valores numéricos, se emplea la técnica IDF para valorar la importancia de estos.

Inicialmente se consideran M conjuntos de datos que fueron etiquetados con antelación con el propósito de encontrar la colección de clasificadores $(\mathrm{c}_1, \mathrm{c}_2, \dots, \mathrm{c}_\mathrm{M})$ que son construidos con la BoW que incorpora la biblioteca y con el empleo de clasificador basado en una máquina de soporte vectorial lineal, donde cada clase se encuentra normalizada en el rango de $0 < \mathrm{c}_\mathrm{i} < 1$.

Los M documentos son etiquetados en un número finito de K categorías propuestas por el usuario de la función (1). En el proceso se construyen un nuevo conjunto de datos de K para poder hacer matricialmente equivalentes los conjuntos M y K.

La forma en que matemáticamente se relacionan los elementos es mediante una transformación lineal que se aplica a la función (1), donde se asume que los tokens t extraídos de cada documento x_i pueden construir K espacios vectoriales ortogonales v_t que inicialmente mapean al espacio de los M documentos originalmente proporcionados, de tal manera que se obtenga para cada c_i deseado para M una función de decisión de la forma $\mathrm{x}'_\mathrm{i} \in \mathbb{R}^\mathrm{M}$.

Dado que se aplica la transformación TF-IDF a los M documentos, implica que hay un proceso de ponderación de la relevancia de cada uno de los tokens que se encuentran en cada uno de los vectores v_t considerando la totalidad del espacio multivectorial K que es cuantificado con $\frac{\sum_\mathrm{t} \mathrm{v}_\mathrm{t}}{\|\sum_\mathrm{k} \mathrm{v}_\mathrm{t}\|}$.

Por otra parte, se hace necesaria la estimación de pesos que ayuden a separar cada espacio vectorial y que son estimados numéricamente. Dichos pesos denotados por w_i van acompañados de un sesgo promedio del espacio vectorial valorado y que tiene un valor $\mathrm{w}_{\mathrm{i}0}$. Estos elementos, con base en [26], se relacionan mediante el producto punto que se exhibe en la función (2).

$$\mathrm{x}'_\mathrm{i} = \mathrm{w}_\mathrm{i} \cdot \frac{\sum_\mathrm{t} \mathrm{v}_\mathrm{t}}{\|\sum_\mathrm{k} \mathrm{v}_\mathrm{t}\|} + \mathrm{w}_{\mathrm{i}0}$$

$$x_i' = W \cdot \frac{\sum_t v_t}{\|\sum_k v_t\|} + w_{i0}$$

$$x' = \sum_t u_t + w_0 \ \ldots (2)$$

Siendo $u_t = \frac{1}{\|\sum_k v_t\|} W v_t$, donde $W \in \mathbb{R}^{M \times d}$ y $w_0 \in \mathbb{R}^M$.

Ya calculados cada uno de los vectores x' se procede a normalizarlos al aplicar la norma uno como se muestra en la función (3). De esta manera se han construido los espacios vectoriales ortogonales para cada uno de los M documentos con las K clasificaciones.

$$x = \frac{x'}{\|x'\|} \ldots (3)$$

```
%%time
s=''
for texto, clase in zip(X_train, y_train):
    s = s + '{'
    s = s + '"text" :' + '"' + texto + '"' + ","
    s = s + '"klass":' + '"' + clase + '"'
    s = s + '},'

s=s[0:len(s)-1]
str_dato='{"datos":[' + s + ']}'
```

```
Wall time: 1h 26min 25s
```

```
import json
datos=json.loads(str_dato)
```

```
%%time
#
# grabar el contenido en un archivo JSON
#
with open('../dat/JSON/COPOM_datos_entrenamiento_010K.json', 'w', encoding='utf-8') as f_out:
    f_out.write(str_dato)
f_out.close()
```

```
Wall time: 1.9 s
```

Figura 27 *Construcción de las entradas para el modelo EvoMSA.*

Fuente: Elaboración propia.

Con los datos que se procesaron, figura 25, se procedió a utilizar la biblioteca EvoMSA, en su segunda versión [22] como se presenta en la figura 27 donde los datos de entrenamiento se almacenaron en el archivo COPOM_datos_entrenamiento_010k.json. Posteriormente, se entrenó el modelo y

se almacenó en la variable evo que fue serializado para no perder el trabajo empleado en su construcción como se presenta en las figuras 28 y 29.

```
from EvoMSA.evodag import TextRepresentations
```

```
%%time
text_repr =  TextRepresentations(lang='es') #, dataset=True )
#text_repr =  TextRepresentations(lang='es', dataset=True )
```

```
Wall time: 28.8 s
```

```
%%time
evo=text_repr.fit(list(datos["datos"]))
```

```
Wall time: 1h 4min 59s
```

Figura 28 *Código que construye el clasificador con EvoMSA.*

Fuente: Elaboración propia.

```
import pickle

# Almacenar el objeto a serializar
miFile = open('Evo_entrenado_10K', 'wb')

pickle.dump(evo, file=miFile)

# close the file
miFile.close()
```

Figura 29 *Apertura del archivo de entrenamiento con los 10 mil textos.*

Fuente: Elaboración propia.

4.5 Evaluación de los resultados

Para valorar la precisión del clasificador construido con EvoMSA 2.0, se utilizó el conjunto de pruebas que se almacenó en el archivo COPOM_Prueba.json, separando los argumentos del clasificador y proporcionándolos al modelo serializado para que entregue los resultados como se presenta en la figura 30.

Ya con los resultados del modelo, se procedió a construir una matriz de confusión y estimar las pruebas de precisión que se presentan en el gráfico 5. Se

puede apreciar que, de los 29 registros de prueba, sólo 24 pudieron clasificarse correctamente, representando el 83.45% (precisión_score).

Por otra parte, cuando se valoró la ratio de los verdaderos positivos (recall_score) nos indica que el modelo tuvo una exactitud de 82.76%. Para finalizar, se estimó el valor predictivo positivo, que se calcula como el cociente entre los valores clasificados correctamente respecto a todas las predicciones positivas que el modelo proporcionó (f1_score) arrojando un resultado de 82.11%.

```
%%time
#
# abrir el archivo JSON
# Paso 100. Obtener los datos
#
import json
strFile='../dat/JSON/COPOM_Prueba.json'

with open(strFile, encoding='utf8') as json_file:
    data=json.load(json_file)
json_file.close()
```

```
Wall time: 90.8 ms
```

```
%%time
# Ontemer los datos de COPOM.JSON
#X_train=[f.limpiarCaracteres(d['argumentos']) for d in data["COPOM_Entrenam
X_test=[d['argumentos']      for d in data["COPOM_Prueba"]]
y_test=[d['movimiento_tii']  for d in data["COPOM_Prueba"]]

len(X_test), len(y_test)
```

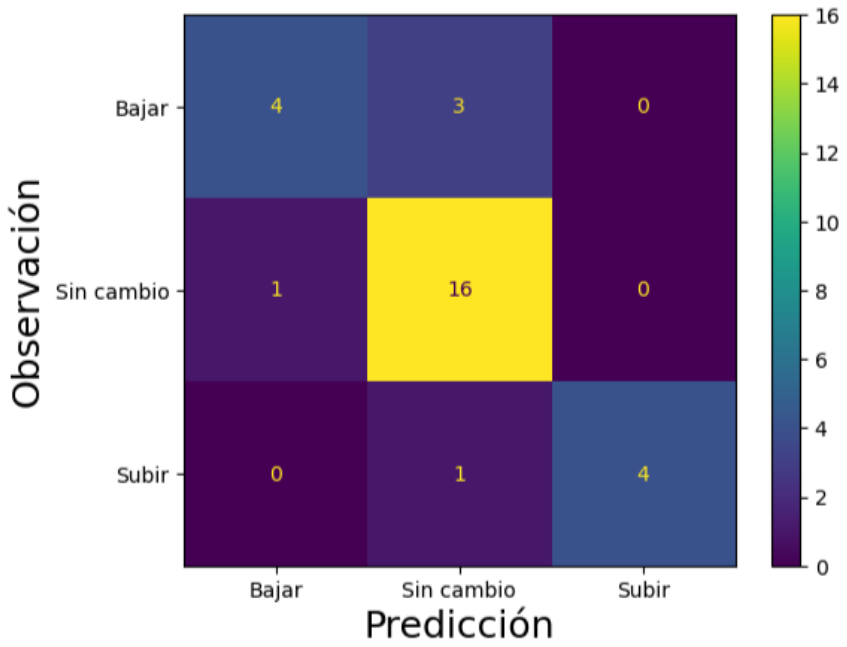
```
Wall time: 2.99 ms

(29, 29)
```

```
%%time
lst_pred=[]
lst_obs=[]

# Clasificar varios textos
for argumento,y_obs in zip(X_test, y_test):
    #y_pred=evo.predict([argumento])
    y_pred=evo.predict([argumento]).tolist()

    lst_pred.append(y_pred)
    lst_obs.append(y_obs)

```

```
Wall time: 13.9 s
```

Figura 30 *Asignando los datos de prueba al clasificador.*

Fuente: Elaboración propia.




```
from sklearn.metrics import precision_score, recall_score, f1_score
```

```
print("precision_score..",precision_score(lst_obs, lst_pred, average='weighted'))
print("recall_score.....",recall_score(   lst_obs, lst_pred, average='weighted'))
print("f1_score.........",f1_score(       lst_obs, lst_pred, average='weighted'))


```

```
precision_score.. 0.8344827586206898
recall_score..... 0.8275862068965517
f1_score......... 0.8211659935797865
```

Gráfico 5 *Matriz de confusión y pruebas de precisión con los datos aumentados.*

Fuente: Elaboración propia.

4.5.1 Valorar a EvoMSA sin ampliar los datos

Durante el proceso de construcción, pruebas y valoración de los resultados del modelo de clasificación con los datos aumentados, surgió la pregunta. ¿El modelo con EvoMSA será capaz de proporcionar resultados similares si únicamente emplea su propia BoW? Para saber la respuesta se procedió a realizar el entrenamiento y las pruebas como se detalla a continuación.

Se procedió a abrir el archivo COPOM.JSON original y se definieron los conjuntos de entrenamiento de pruebas con un 70% y 30% de los datos como se

muestra en las figuras 31 a 33. Posteriormente se entrenó el modelo y se almacenó como un archivo serializado, figura 34.

```
In [1]:
import funcionesCOPOM_20230115_0835 as f
```

```
In [1]:
%%time
#
# abrir el archivo JSON
# Paso 100. Obtener los datos
#
import json
with open('../dat/JSON/COPOM.json', encoding='utf8') as json_file:
    data=json.load(json_file)
json_file.close()
```

```
Wall time: 63.8 ms
```

```
In [2]:
# Eliminar el registro #62 que tiene basura en los textos
del data["COPOM"][62] # Eliminarlo por contener caracteres no legibles desde el origen
```

Figura 31 *Apertura del archivo con las 98 minutas del COPOM.*

Fuente: Elaboración propia.

```
In [10]:
%%time
#
#Tomar los argumentos y las variaciones de la TII
#
# Limpiar los textos de aquellos caracteres que dificultan su procesamiento
#

X=[d["argumentos"].lower() for d in data["COPOM"]]
y=[d["Descripcion_Cambio_TII"].lower() for d in data["COPOM"]]

len(X),len(y)
```

```
Wall time: 61.6 ms
```

```
Out[10]: (96, 96)
```

Figura 32 *Construcción de los vectores con los documentos y sus clasificaciones.*

Fuente: Elaboración propia.

```
In [18]:
%%time
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test =  train_test_split(X,y, test_size=0.30, random_state=42)

len(X_train), len(X_test), len(y_train), len(y_test)
```

```
Wall time: 998 µs
```

```
Out[18]: (67, 29, 67, 29)
```

Figura 33 *Construcción de los conjuntos de entrenamiento y de prueba.*

Fuente: Elaboración propia.

```
In [21]:
from EvoMSA.evodag import TextRepresentations

In [22]:
%%time
text_repr =  TextRepresentations(lang='es') #, dataset=True )

Wall time: 28.5 s

In [23]:
%%time
evo=text_repr.fit(list(datos["datos"]))

Wall time: 16.3 s

In [24]:
# Serializar el objeto entrenado para ser utilizado posteriormente
import pickle

# Almacenar el objeto a serializar
miFile = open('Evo_entrenado', 'wb')

pickle.dump(evo, file=miFile)

# close the file
miFile.close()
```

Figura 34 *Construcción del modelo de clasificación con EvoMSA con las 98 minutas sin preprocesar.*

Fuente: Elaboración propia.

Al valorar la precisión del modelo, se observó que éste obtuvo con los datos de entrenamiento una eficiencia media de 64.45% (precisión=66.19%, recall=67.16 y f1-score=60.02%), mientras que con los datos de prueba se promedió un 58.27% (precisión=54.15%, recall=65.52 y f1-score=55.15%) como se presentan en los gráficos 6 y 7. Esto indica que existe una diferencia en el desempeño del modelo ante una variación tan significativa en la cantidad de datos que se le proporcionan, de 10 mil con el aumento de los datos versus los 98 documentos originales.

4.5.2 Valorar los modelos de EvoMSA con Bootstrap

Se pudo apreciar en la sección 4.5.1 que hay una diferencia en la precisión de la clasificación cuando se emplearon los datos de prueba, gráficos 5 y 7, entre los modelos que fueron entrenados con el aumento de los datos versus los que no fueron aumentados, señalando nominalmente que ambos modelos no proporcionan los mismos resultados. Sin embargo, para verificar esta suposición, se procederá a

valorarlos nuevamente con los 29 datos de prueba mediante una selección de 500 subconjuntos elegidos con Bootstrap [27, 30].

Las hipótesis que se valorarán serán si ambos modelos proporcionan los mismos resultados; o bien, si estos son diferentes. Para este efecto se construirá un intervalo de confianza con un nivel de $\alpha = 5\%$ de tal forma que:

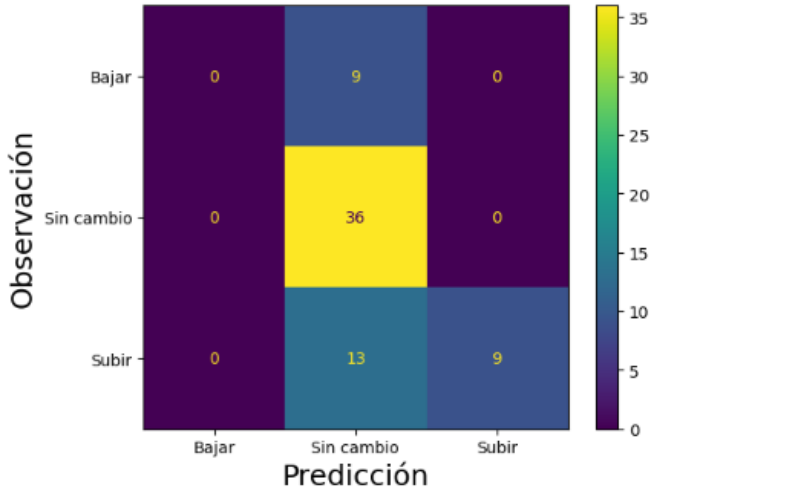
H_0: Los modelos SI proporcionan los mismo resultados

H_a: Los modelos NO proporcionan los mismo resultados

In [48]:
```
%%time
lst_pred=[]
lst_obs=[]

# Clasificar varios textos
for argumento,y_obs in zip(X_train, y_train):
    y_pred=evo.predict([argumento]).tolist()

    lst_pred.append(y_pred)
    lst_obs.append(y_obs)
```

```
CPU times: total: 1min 28s
Wall time: 28 s
```

In [49]:
```
matriz_confusion(lst_obs,lst_pred)
```



In [50]:
```
#
# Resultados con los datos de entrenamiento
#

from sklearn.metrics import precision_score, recall_score, f1_score

print("precision_score..",precision_score(lst_obs, lst_pred, average='weighted'))
print("recall_score.....",recall_score(   lst_obs, lst_pred, average='weighted'))
print("f1_score.........",f1_score(       lst_obs, lst_pred, average='weighted'))


```

```
precision_score.. 0.6618630983015955
recall_score..... 0.6716417910447762
f1_score......... 0.602218830350649
```

Gráfico 6 *Desempeño del clasificador con los datos de entrenamiento y sin un preprocesamiento de limpieza de los datos originales ni aumentar los datos.*

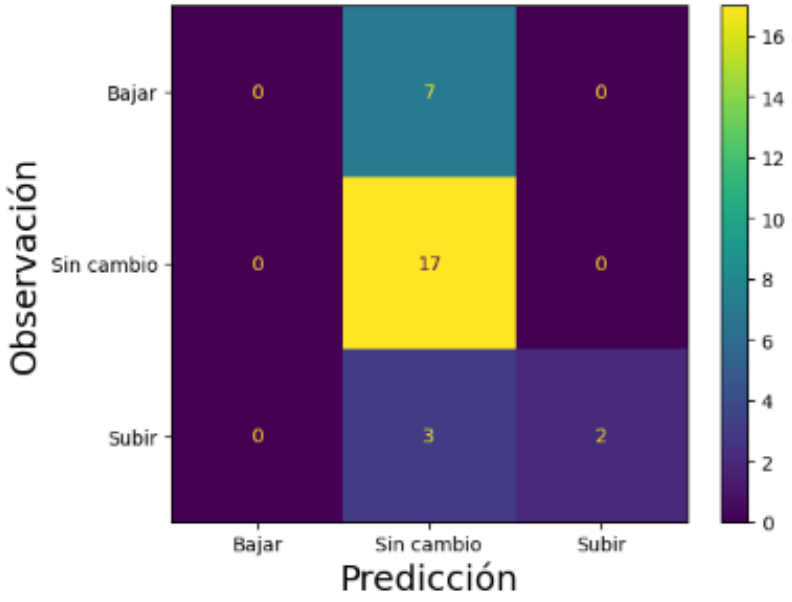
Fuente: Elaboración propia.

In [51]:
```
%%time
lst_pred=[]
lst_obs=[]

# Clasificar varios textos
for argumento,y_obs in zip(X_test, y_test):
    y_pred=evo.predict([argumento]).tolist()
    lst_pred.append(y_pred)
    lst_obs.append(y_obs)
len(lst_pred), len(lst_obs)
```

```
CPU times: total: 38.5 s
Wall time: 12.2 s
```

Out[51]: (29, 29)

In [52]:
```
matriz_confusion(lst_obs,lst_pred)
```



In [53]:
```
#
# Resultados con los datos de prueba
#

from sklearn.metrics import precision_score, recall_score, f1_score

print("precision_score..",precision_score(lst_obs, lst_pred, average='weighted'))
print("recall_score.....",recall_score(   lst_obs, lst_pred, average='weighted'))
print("f1_score.........",f1_score(       lst_obs, lst_pred, average='weighted'))


```

```
precision_score.. 0.541507024265645
recall_score..... 0.6551724137931034
f1_score......... 0.5515002239140171
```

Gráfico 7 *Desempeño del clasificador con los datos de prueba y sin un preprocesamiento de limpieza de los datos originales ni aumentar los datos de entrenamiento.*

Fuente: Elaboración propia.

En la figura 35, se presenta la construcción de dos intervalos de confianza para comparar las clasificaciones de los 29 datos de prueba observados (variable lst_y) donde, i) resultados del modelo entrenado sin el incremento de los datos (variable lst_pred) y ii) resultados del modelo entrenado con los datos incrementados (variable lst_pred_10k).

Se observó que, en el primer caso, la precisión esperada para el modelo sin el incremento de los datos de entrenamiento rondó entre el 68.13-69.64% y en el segundo modelo alcanzó 71.41-72.93%. Estos resultados indican que se rechaza la hipótesis nula en favor de la hipótesis alternativa.

In [56]:
```
%time
# asignar los datos de prueba a los modelos
hy_1=evo_10k.predict(X_test) # Calcular las predicciones del modelo con los datos aumentados
hy_2=evo.predict(X_test) # Calcular las predicciones del modelo sin aumentar los datos
```

```
CPU times: total: 0 ns
Wall time: 0 ns
```

In [60]:
```
%%time

import matplotlib.pyplot as plt

S = np.random.randint(low=0, high=len(y_test), size=(500,len(y_test))) # construir los vectores del remuestreo

B=[] # almacenar los resultados de la prueba macro-f1
p_valor=0 # acumulador para estimar p-value

# calcular la precisión com macro-f1 del modelo con los datos aumentados
for s in S: # recorrer los subconjuntos construidos aleatoriamente de y_test
    lst_y=[] # clasificación observada
    lst_pred_10k=[] # clasificación estimada por el modelo con los datos aumentados
    lst_pred    =[] # clasificación estimada por el modelo sin los datos aumentados
    for i in s:
        lst_y.append(y_test[i])
        lst_pred_10k.append(hy_1[i])
        lst_pred.append(hy_2[i])
```

In [89]:
```
S = np.random.randint(low=0, high=len(y_test), size=(500,len(y_test))) # construir los vectores del remuestreo

cnt1=[] # contador de resultados entre y_test vs y_evo_sin_incremento_de_datos
cnt2=[] # contador de resultados entre y_test vs y_evo_con_incremento_de_datos

for s in S:
    for i in s:
        # contador de los resultados entre las obs y modelo sin aumento de datos
        if lst_y[i]==lst_pred[i]:
            cnt1.append(1)
        else:
            cnt1.append(0)

        # contador de los resultados entre las obs y modelo con aumento de datos
        if lst_y[i]==lst_pred_10k[i]:
            cnt2.append(1)
        else:
            cnt2.append(0)

# Calculando las probabilidades
p1=np.mean(cnt1)
p2=np.mean(cnt2)

# Calculando los intervalos de confianza
alpha=0.05
z = norm().ppf(1 - alpha / 2)

se1 = np.sqrt((p1*(1-p1))/len(cnt1))
se2 = np.sqrt((p1*(1-p1))/len(cnt2))
#se3 = np.sqrt((p1*(1-p1))/len(cnt3))

C1= (np.round(p1-z*se1, 6), np.round(p1+z*se1,6))
C2= (np.round(p2-z*se2, 6), np.round(p2+z*se2,6))

print("1) Intervalo entre y_test vs y_evo_sin_incremento_de_datos.....................",C1)
print("2) Intervalo entre y_test vs y_evo_con_incremento_de_datos.....................",C2)

```

```
1) Intervalo entre y_test vs y_evo_sin_incremento_de_datos..... (0.681361, 0.696432)
2) Intervalo entre y_test vs y_evo_con_incremento_de_datos..... (0.714189, 0.729259)
```

Figura 35 *Construcción de los intervalos de confianza con un 5% de nivel de confianza para los modelos construidos con EvoMSA.*

Fuente: Elaboración propia.

Conclusiones

Conclusiones

Desde la perspectiva de la praxis financiera, las variaciones de la TII son relevantes para el funcionamiento de este sector económico, ya que ésta es la tasa mínima que una empresa debería pagar por el costo del dinero, por lo que pronosticar su potencial tendencia es de gran relevancia para los tomadores de decisión financiera.

- El tema de investigación elegido se encuentra alineado en una tendencia prometedora dentro de los sectores académicos y profesionales relacionados con las actividades financieras en todo el mundo, debido a que este sector industrial tiene un alto impacto en el buen funcionamiento de las sociedades modernas, por lo que se podría decir que este trabajo se clasifica en los relacionados con el “Pronóstico financiero de variables”.
- Resultó curioso el observar que, pese a que el 42% de las aplicaciones publicadas se abocan a pronosticar diversas variables financieras, y emplean el 69% de las fuentes de datos (31% para predecir y 28% en el empleo de los corpus financieros). En ningún momento se identificaron aplicaciones para la tasa de referencia publicadas. Aunque se descubrió la existencia de un próximo documento que versa sobre el tema con un enfoque econométrico y para la LIBOR del Banco de Inglaterra [29].
- Esta ausencia de investigación para el pronóstico de las tasas de referencia podría deberse a la escaza información disponible, ya que en comparación al mercado de capitales que genera y acumula miles de GB o TB de información anualmente, los bancos centrales tienen una capacidad limitada debido a que las reuniones de política monetaria se realizan entre 9 a 15 veces al año y el volumen agregado de los documentos publicados se mide en KB o MB en el mejor de los casos.
- Pese a la nula investigación que rodea la estimación de la TII, este trabajo puede ser considerado como innovador por allanar en un tema financiero que es relevante para los analistas y tomadores de decisión económico-financiera por su alta relevancia en el desempeño empresarial enmarcado en una economía moderna.

- Desde la perspectiva estadística con la que se valoraron los resultados de los modelos con ampliación o no de los datos de entrenamiento, se observó que sí hay una diferencia significativa de entrenar a EvoMSA con datos aumentados y reemplazando diversas palabras con sinónimos o términos que son empleados en la jerga alcanzando predicciones superiores al 80% del movimiento de la tasa de referencia en México, por lo que se sugiere incluir esta práctica dentro del preprocesamiento de los datos para usos futuros de este modelo.

A partir de la experiencia obtenida en el desarrollo de este proyecto, se ha considerado que serían de interés académico y profesional las siguientes líneas de investigación dentro del análisis de sentimiento financiero, o la clasificación de este tipo de documentos:

1. Ahondar en el análisis automatizado de los estados financieros de empresas de sectores industriales.
2. Aplicar el análisis de sentimiento, o la clasificación de los documentos de estabilidad financiera que publica el Banxico para valorar la similitud entre estos.
3. Valorar la inmediatez con la que los precios asimilan la nueva información del entorno para identificar oportunidades de arbitraje financiero bajo el supuesto de la existencia de un rezago de los agentes económicos para identificarla.
4. Valorar la velocidad con la que las noticias son asimiladas en los precios de los activos de deuda o capital.
5. Construir un glosario de términos económico financiero que permita mejorar la precisión de los modelos basados en el PLN.

Considerando las capacidades de procesamiento que tiene EvoMSA, se considera que las potenciales líneas de investigación enfocadas en el procesamiento de información económica y financiera podrían ser.

6. Valorar el impacto de incorporar este modelo de clasificación del movimiento de la TII en las metodologías de análisis de sensibilidad de los portafolios de inversión.
7. Modelar el comportamiento de otras variables económicas y financieras a partir de las variaciones en la tasa objetivo.
8. Estudiar las causas por las que el contenido de la BoW de EvoMSA arrojó predicciones que en el mejor de los casos rondaba el 58% de precisión, ya que, cualitativamente hablando, resulta curioso que al utilizar los 552 términos financieros empleados en las minutas del COPOM provoque que este modelo incremente su precisión en más de un 80%.

Comentarios a las observaciones del tercer lector

Entre los comentarios recibidos, se formularon observaciones acerca de cómo se podría reformular el presente proyecto si no se contase con las minutas del COPOM, empleando la minería de textos y el reconocimiento de entidades. Por otra parte, se preguntó cómo se procesaría la información sin emplear el modelo EvoMSA para formular un pronóstico que fuese útil al tomador de decisiones. Para responder a estas observaciones, se formularon y respondieron las siguientes cinco preguntas.

1. Proporcionar un bosquejo de las características de la minería de textos y del reconocimiento de entidades como un medio de extracción de información semiestructurada.

La *minería de textos* es una técnica que se aplica manualmente para extraer la información de documentos y expresarlos mediante vectores que relacionan a cada documento con los términos más relevantes, de acuerdo con el contexto del estudio que se esté abordando. Los elementos para considerar estarían en función de la BoW proporcionada, descartando aquellos que no aportan valor, lematizándolos y extrayendo sus raíces. Estos pasos fueron aplicados en el presente documento.

El *Reconocimiento de Entidades Nombradas* (Named Entity Recognition-NER) extrae la información de un documento al identificar sus tokens más

relevantes. Ésta aplica una clasificación que se realizó previamente mediante un análisis semántico a otros textos. Las principales tipificaciones que se aplican son: persona, gpe (geographic position entity), tiempo, organización, hecho y localización. La ventaja de esta forma de procesamiento radica en que al aplicarla se utiliza un modelo previamente entrenado, lo que simplifica el preprocesamiento de los documentos para la construcción de sus respectivos vectores al descartar los términos que sean poco relevantes en el PLN.

Complementando lo anterior, sería posible emplear técnicas de agrupamiento de documentos basados en K-NN, PCA, redes de clasificación Naive o reglas de asociación que permitan separar los documentos con la información económica y financiera relacionada con el movimiento de la tasa de referencia en carpetas.

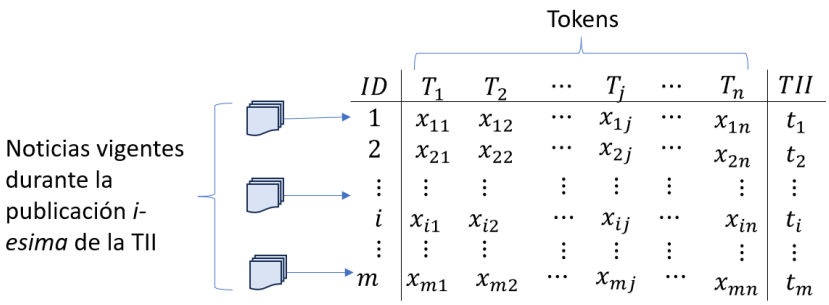
2. ¿Cuáles serían las fuentes de información a emplear si no se contase con las minutas del COPOM?

Para los propósitos de este proyecto aplicativo, es necesario contar con dos tipos de fuentes de información. Las noticias sobre la situación económica de México y de los EE. UU. Esta información se extraería de los periódicos mexicanos “El Economista” y “El Financiero” en sus versiones electrónicas.

La segunda fuente de información debe proporcionar el valor de la tasa de referencia. Éste se obtendría de la serie de tiempo que proporciona el Banxico en su página web, como se realizó en el presente trabajo.

3. ¿Cómo se arreglarían los datos para el entrenamiento del modelo de aprendizaje?

Considerando que las publicaciones de la TII se programan al inicio de cada año, se agruparían las noticias para el período de tiempo en que la tasa de referencia (t_i) fuese vigente para el conjunto de noticias del período (x_{ij}). De esta forma se relacionarían en un solo vector. Al repetir este proceso se construirá una matriz con la que se alimentará al modelo de aprendizaje computacional como se presenta en la siguiente imagen.



Hay que señalar que la relación token-noticia se expresa con el valor x_{ij} que sería ponderado mediante el método TF-IDF para cada período i señalado en la columna ID.

4. ¿Cuál sería el modelo de aprendizaje computacional que podría sustituir a EvoMSA?

El modelo de aprendizaje computacional que se podría emplear se basaría en una red neuronal profunda como el LSTM debido a que ésta tiene la característica de no olvidar los datos previos con los que fue alimentada. Esta red neuronal se alimentaría con los valores que se producirían en el inciso 3.

5. ¿Es posible producir un pronóstico de la TII con una ventana de tiempo que resultase de utilidad al tomador de decisiones?

El modelo propuesto arroja como resultado si la tasa de interés se *mantendrá sin cambio, subirá o bajará*. Por lo que adaptarlo para que arroje un pronóstico numérico es factible si los resultados se acotasen a variaciones en puntos base como *0, ±5, ±10, ±15, ±20 y ±25* por ser los valores más observados históricamente.

Con relación a una ventana de tiempo, estas ya están definidas y tienen fechas que son publicadas en la página del Banxico. Las reuniones del COPOM se realizan en promedio una vez cada 32 días naturales, por lo que los gestores financieros pueden valorar sus escenarios a partir de la información que está disponible en el mercado y con ello pronosticar la decisión esperada que podría tomar el Instituto Central.

Bibliografía

[1] S. B. Kumar y V. Ravi, «A survey of the application of text mining in financial domain,» Knowledge-Base Systems, nº 114, pp. 128-147, 2016.

[2] Banco de México, «Banco de México,» Banco de México, [En línea]. Available: https://www.banxico.org.mx/publicaciones-y-prensa/minutas-de-las-decisiones-de-politica-monetaria/minutas-politica-monetaria-ta.html.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safeai, E. D. Trippe, J. B. Gutierrez y K. Kochut, «A Brief Survey of text Mining: Classification, Clusterin and Extraction Techniques.,» July 2017.

[4] R. P. Schumaker, Y. Zhang, C.-N. Huang y H. Chen, «Evaluating sentiment in financial news articles,» Elsevier, nº 53, pp. 458-464, 2012.

[5] S. W. K. Chan y M. W. C. Chong, «Sentiment analysis in financial text,» Elsevier, nº 94, pp. 53-64, 2017.

[6] C. Kearney y S. Liu, «Textual sentiment in finance: A survey of methods and models,» Elsevier, nº 33, pp. 171-184, 2014.

[7] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev y D. Trajanov, «Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers,» IEEE Access, vol. 8, pp. 131662 - 131682, 2020.

[8] Anzaldo, G. y Benavides, G. «Expectativas en las tasas de interés y noticias de política monetaria de EE. UU.,» Revista Mexicana de Economía y Finanzas (Nueva Época), vol. 15, nº 1, pp. 17-35, 2020.

[9] E. Fama, «The behavior of stock prices,» Journal of Business, nº 38, pp. 34-106, 1964.

[10] E. Fama, «Efficient Capital Markets: A view of Theory and Empirical Work,» Journal of Finance, 1970.

[11] L. Bachelier, «Théorie de la spéculation,» Annales scientifiques de l'ENS, pp. 17-67, 1900.

[12] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia y E. A. Villaseñor, «A case study of Spahish text transformations for twitter sentiment analysis,» Elsevier, nº 81, pp. 457-471, 2017.

[13] A. Bouziane, D. Bouchiha, N. Doumi y M. Malki, «Question Aneswering Systems: Survey and Trends,» Procedia Computer Science, nº 73, pp. 366-375, 2015.

[14] H. Pandya y B. Bhatt, «Question Answering Survey: Directions, Challenges, Datasets, Evolution Matrices,» 2021.

[15] B. Fu, Y. Qiu, C. Tang, H. Yu y J. Sun, A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges, Beijing: Alibaba Group, Institute of Computing Technology, Chinese Academy of Science, 2020.

[16] H. Markowitz, «Portfolio Selection», Journal of Finance, nº 7, pp. 77-91, 1952.

[17] S. F. Castro-Enciso, «Creación de Portafolios de Inversión utilizando Algoritmos Evolutivos Multiobjetivo», Tesis de maestría, CDMX, 2005.

[18] J. Han y M. Kamber, Data minig: Concepts and Techniques, M. K. Publishers, Ed., San Francisco, CA., 2006.

[19] MLOps, «MLOps,» Samadrita Ghosh, 14 11 2022. [En línea]. Available: https://neptune.ai/blog/machine-learning-project-with-less-data. [Último acceso: 20 11 2022].

[20] Real Academia de la Lengua Española, [En línea]. Available: https://www.20minutos.es/noticia/4698766/0/10-palabras-mas-largas-diccionario-rae/.

[21] Mario Graff, Sabino Miranda-Jiménez, Eric S. Téllez, Daniela Moctezuma, «EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis» Computational Intelligence Magazine, volumen 15, issue 1, pp 76-88, Feb 2020, DOI 10.1109/MCI.2019.2954668, https://ieeexplore.ieee.org/document/8956106

[22] Documentación en Internet sobre EvoMSA versión 2.0, INFOTEC-INGEOTEC, https://evomsa.readthedocs.io/en/docs/v2.html#v2

[23] Diccionario de Banxico educa. http://educa.banxico.org.mx/recursos_banxico_educa/glosario.html

[24] Shorten, C., Khosgoftaar, T.M., Furht, B., «Text Data Augmentation for Deep Learning», Journal of Big Data, Springer Open, Open Access, https://doi.org/10.1186.s40537-021-00492-0, (2021) 8:101

[25] Documentación en Internet sobre EvoMSA versión 2.0, INFOTEC-INGEOTEC, https://evomsa.readthedocs.io/en/docs/v2.html

[26] Documentación en Internet sobre la función TextRepresentation de EvoMSA versión 2.0, INFOTEC-INGEOTEC, https://evomsa.readthedocs.io/en/docs/text_repr.html

[27] Documentación en Internet sobre el uso de la técnica Bootstrap para comparar el desempeño de modelos, google colab, https://colab.research.google.com/drive/1GFjk1gJtlPV4iw-SJN_2VY1aVcv5rbfp?usp=sharing#scrollTo=ZZldn_Xy12SP

[28] Anzaldo, G. «La aplicación de un modelo de inversión difuso de entropía bursátil caso: un portafolio de inversión para la industria de la construcción», Tesis de doctorado, Huixquilucan, Edo. Mex., 2012

[29] Benavides. G, Colla, E. «El Poder de Publicaciones de Minutas de Política Monetaria. El Caso del Reino Unido», Próximo a publicarse, 2023.

[30] Graff, M. Aprendizaje Computacional. Comparación de Algoritmos. https://ingeotec.github.io/AprendizajeComputacional/capitulos/13Comparacion/

[31] Documento en Internet sobre definiciones financieras.
https://economipedia.com/definiciones/activo-financiero.html

[32] Documento en Internet sobre definiciones financieras.
https://economipedia.com/definiciones/arbitraje.html

[33] Documento en Internet sobre El Decreto de Ley del Banco de México. Capítulo I. De la Naturaleza, las Finalidades y las Funciones, 1993.
https://www.banxico.org.mx/marco-normativo/marco-juridico/ley-del-banco-de-mexico/%7B9BCADA4D-1CFD-92F0-DEC1-634AC4F7BB12%7D.pdf

[34] Documento en Internet sobre la misión, visión y los objetivos del Banxico.
https://www.banxico.org.mx/conociendo-banxico/mision-vision-objetivos-banco.html

[35] Documento en Internet sobre definiciones financieras.

https://economipedia.com/?s=tasa+de+inter%C3%A9s

Anexos

Anexo I

A continuación, se presentan las fuentes de información más utilizadas en las aplicaciones de PLN en finanzas.

1. **Fuentes para noticias empresariales** [1, 4, 6]. Amazon news, Amazon reviews, Bloomberg.com, comtex.com, Dow Jones newswires, Financial news, Financial reports, German ad hoc announcements, Google news, Hang Seng index, Reuters, S&P news, PRNewsWire (newswire.com), Yahoo news (biz.yahoo.com) , LiveJournal site, News headlines, Quamnet.com, Macro news, News storing & commentaries, US financial news, Índices accionarios, Shangai stock Phishing, Stock news, S&P500 Index, Taiwan Stock Exchange Financial Price Index, Yahoo finance (finance.yahoo.com), Tehran tock phishing, Yahoo stock, Corporate Annual reports, Corporate news, The US corporate filling, Top 100 e-commerce Companies, Gobierno. Banxico. Minutas del COPOM, Gobierno.FED. Federal Open Market Committee, (https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm), e-mails de las empresas, Movie reviews, URL's, Textual data of the web, Theorical study, US message postings
2. **Fuentes para redes sociales** [1, 3, 7]. Facebook, Tweeter/Tweets, Social media news
3. **Fuentes para corpus financieros** [5, 6, 7]. DICTION, GI/Harvard, Finance-specific, otros
4. **Fuentes para ciberseguridad** [1, 5, 6]. Antiphising search, Millermiles, Monkey.org, Page pool, phish tank, Phishing corpus, SpamAssassin, Spambase, University of Alabama Phishing corpus, 3Sharp, Annexia spam archive, Chinese corpus, CMU, Ling-spam, SpamAssassin, PU corpus, API call, Android OS malware, DARPA 99, Kingsoft internet security laboratory, Kingsoft anti-virus lab, Web pages, Windows API calls, Log entires, System call, System calls of DARPA 99, System call sequence of DARPA 99, Reportes financieros de las empresas, US securities and Exchange Commision releases, TCE Threating emails, PU 1.

Anexo II

A continuación, se presenta en el cuadro 6 las pruebas de chi cuadrada para analizar la igualdad estadística entre las diversas distribuciones de las clasificaciones de la variación de la TII del gráfico 2 A y B. Se puede apreciar que no fue posible rechazar la hipótesis nula en ninguno de los casos.

Ho: Las distribuciones obtenidas son estadísticamente iguales a la descrita en los datos originales
Ha: Las distribuciones obtenidas son estadísticamente distintas a la descrita en los datos originales

Figura	Datos a procesar	Bajar	Sin_cambio	Subir	Suma	chi-cuadrada	p-value	Conclusión
21	Originales	17%	55%	28%	100%	1.00	0.61	Ho
23-A	(A) Datos de entrenamiento	13%	54%	33%	100%	0.99	0.61	Ho
23-B	(B) Datos de prueba	24%	59%	17%	100%	0.95	0.62	Ho
24	10 mil datos generados	14%	53%	33%	100%	0.99	0.61	Ho

Cuadro 6 *Pruebas de chi cuadrada de las distribuciones de las clasificaciones de las variaciones de la TII.*

Fuente: Elaboración propia.

Anexo III

A continuación, se presenta en el cuadro 7 la prueba de chi cuadrada para analizar la igualdad estadística entre los 10 mil datos generados artificialmente, (obs) versus la distribución uniforme de estos (Esperada). Se puede apreciar que no fue posible rechazar la hipótesis nula, por lo que se asume que ambas distribuciones son uniformes.

Ho: Las distribuciones obtenidas son estadísticamente iguales a la descrita en los datos originales
Ha: Las distribuciones obtenidas son estadísticamente distintas a la descrita en los datos originales

Intervalo	Obs	Esperada
1	1,087	1,000
2	1,049	1,000
3	894	1,000
4	1,054	1,000
5	888	1,000
6	1,048	1,000
7	1,061	1,000
8	872	1,000
9	1,035	1,000
10	1,012	1,000
Suma	**10,000**	**10,000**

	Valor
chi-cuadrada	0.00
p-value	1
Conclusión	Ho

Cuadro 7 *Prueba de chi cuadrada sobre la uniformidad de los 10 mil datos generados.*

Fuente: Elaboración propia.

Anexo IV

En esta sección se reconoce la ayuda recibida de profesores y colegas que me acompañaron durante mi estancia en la MCDI. Mi gratitud por sus valiosas aportaciones a mi educación.

Análisis de Algoritmos y Estructuras para Datos Masivos. ¡Gracias Dr. *Eric Sadit* Téllez Ávila!

Análisis Exploratorio de Datos. ¡Gracias Dr. *José* Ortíz Bejar!

Aprendizaje Computacional. ¡Gracias Dr. *Mario* Graff Martínez!

Cómputo de Alto Rendimiento. ¡Gracias Dra. *Magali* Arellano Vázquez!

Cómputo Evolutivo. ¡Gracias Dr. *Miguel Ángel* Porta García!

Estadística. ¡Gracias Dr. *Dagoberto* Armenta Medina!

Inteligencia de Negocios. ¡Gracias Dra. *Fabiola* Ramírez Escobedo!

Procesamiento de Información. ¡Gracias Dr. *José Luis* Jiménez Márquez!

Seminario de Proyectos I. ¡Gracias Dra. *Fabiola* Ramírez Escobedo!

Matemáticas para la Ciencia de Datos, Modelos Avanzados en Ciencia de Datos, Seminario de Proyectos II y III. ¡Gracias Dr. *Juliho David* Castillo Colmenares!

Tratamiento Digital de Imágenes. ¡Gracias Dr. *Edgar* González Fernández!

Compañera de cursos y de proyectos. ¡Gracias Lic. MAC. *Dulce María* Reyes Lucas!

Compañero de cursos y de proyectos. ¡Gracias Lic. MAC. *Ismael* Medina Muñoz!

Personal administrativo del INFOTEC y Coordinación de la MCDI. ¡Gracias por su apoyo, paciencia y amabilidad en la atención que siempre me prestaron!

Al **tercer lector** por ofrecer sus comentarios a documento.

Anexo V

En la década de los años setenta del siglo XX, el economista norteamericano John B. Taylor propuso una ecuación con la que se pretendía estimar el valor de tasa de referencia publicada por el Sistema de la Reserva Federal de los EE. UU (FED). Esta ecuación incorpora los valores actuales y deseados de las variables económicas que son afectadas por la tasa de referencia como son, la inflación, la producción industrial y el nivel que la tasa de equilibrio interbancaria tiene al momento de hacer la estimación.

Esta ecuación es un insumo relevante en los preparativos de las sesiones del COPOM de la FED por incorporar los dos principales objetivos que debe optimizar en la economía real. El control de la inflación y la promoción del empleo. En (4) se presenta su forma.

$$i_t = \pi_t + r_t^* + a_\pi(\pi_t - \pi_t^*) + a_y(y_t - y_t^*) \,..(4)$$

Siendo i_t el valor de la tasa objetivo, π_t la inflación como deflactor del PIB, r_t^* la tasa de interés real de equilibrio, π_t^* Es la tasa de inflación deseada, y_t es el logaritmo del PIB, y_t^* es el logaritmo del PIB potencial.

Para estimar el componente de la inflación, los analistas buscan optimizar $a_\pi(\pi_t - \pi_t^*)$, mientras que para el empleo consideran $a_y(y_t - y_t^*)$. π_t y r_t^* representan el entorno económico y financiero al momento en que se está realizando la valoración de la i_t. Los coeficientes a_π y a_y tienen el valor de 0.50, propuestos por Taylor, por ser igual de importantes para el Sistema de la Reserva Federal de los EE. UU. (FED).

Índice de términos