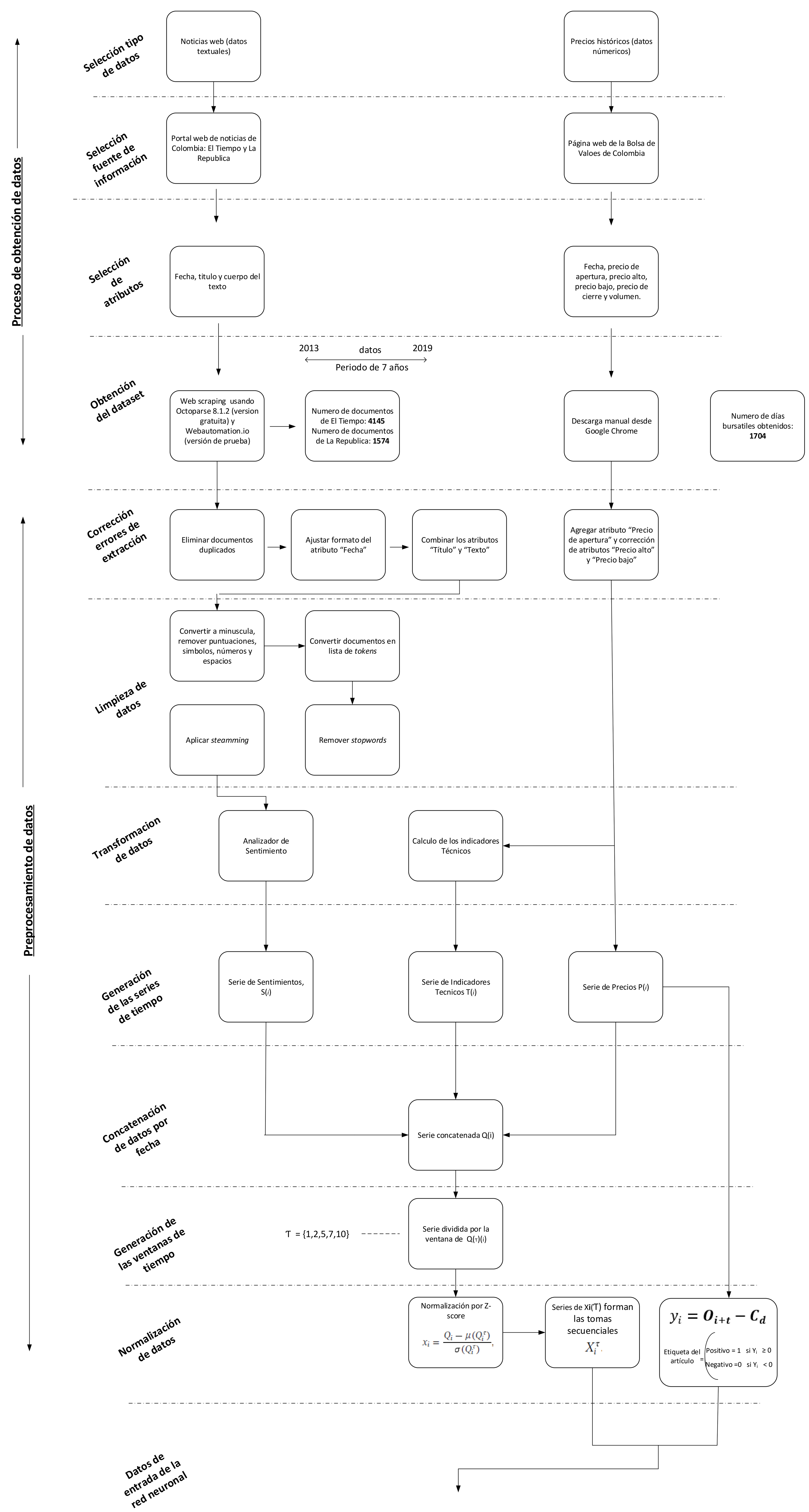
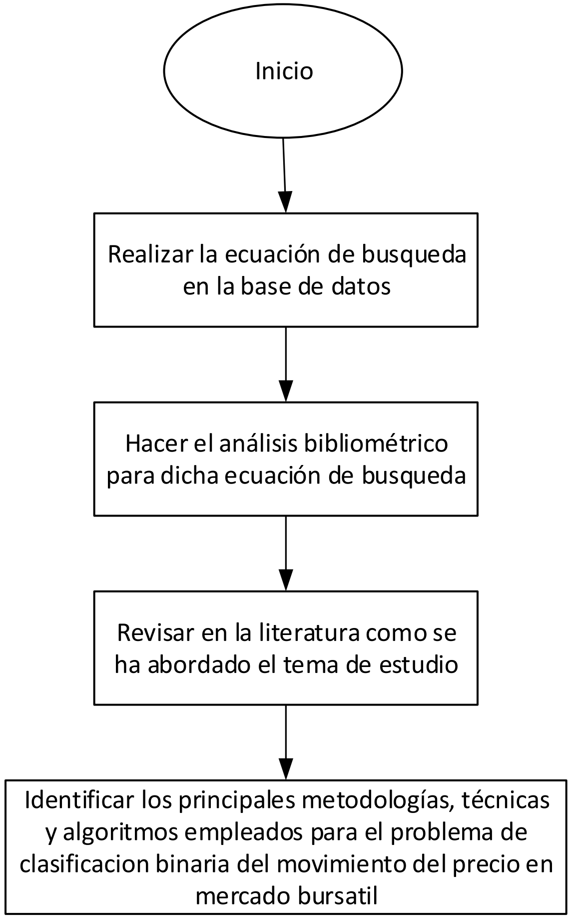
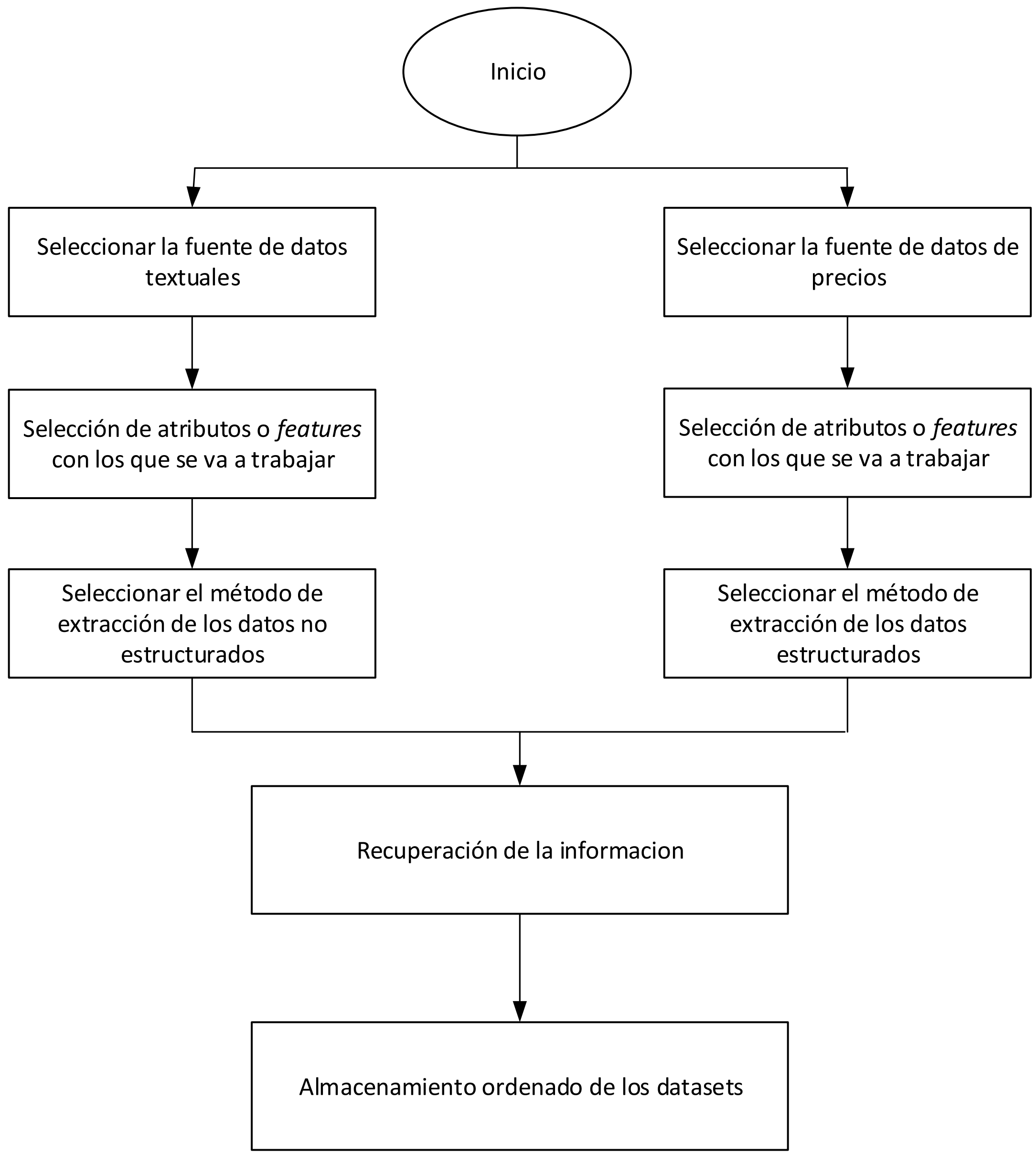
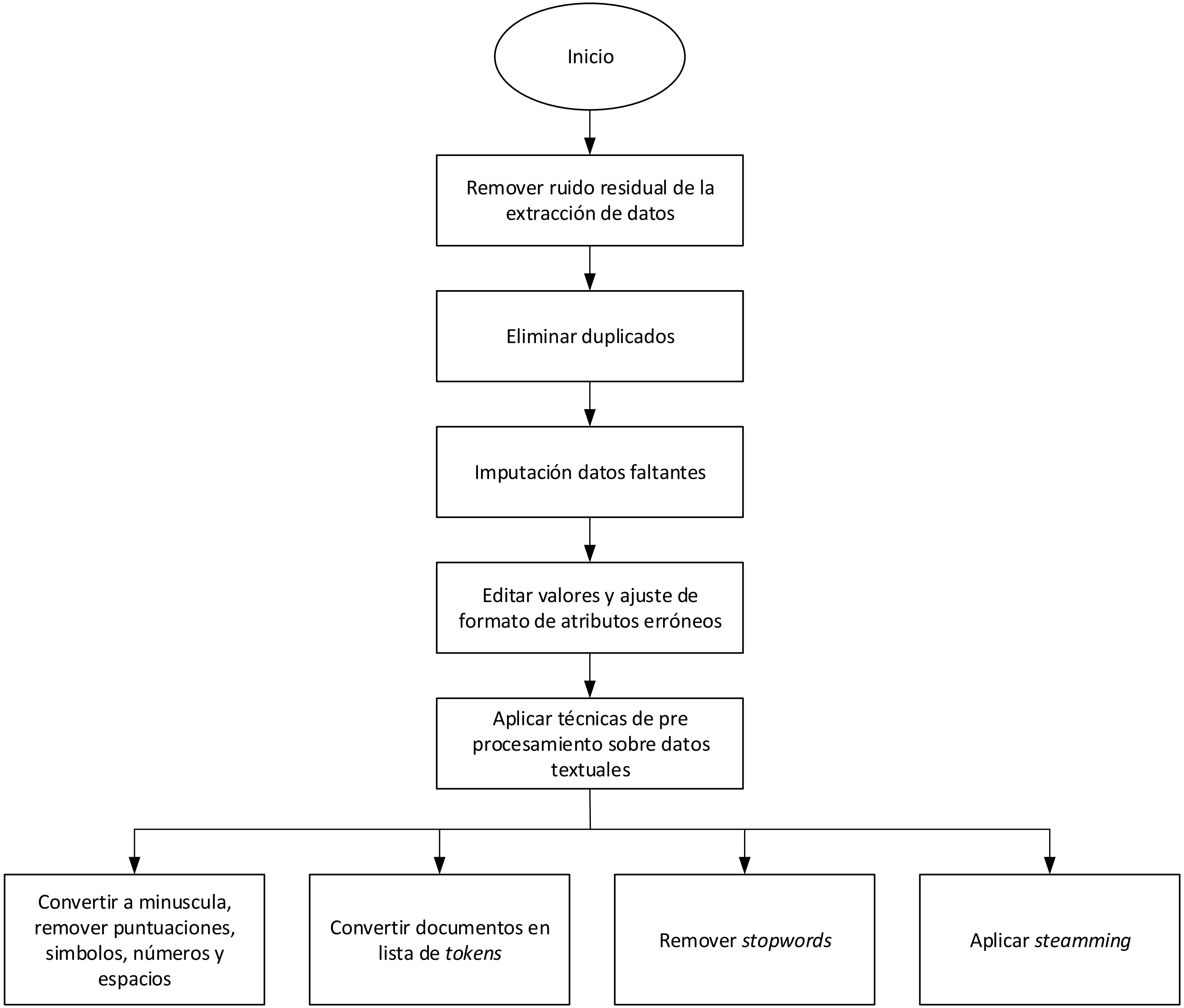
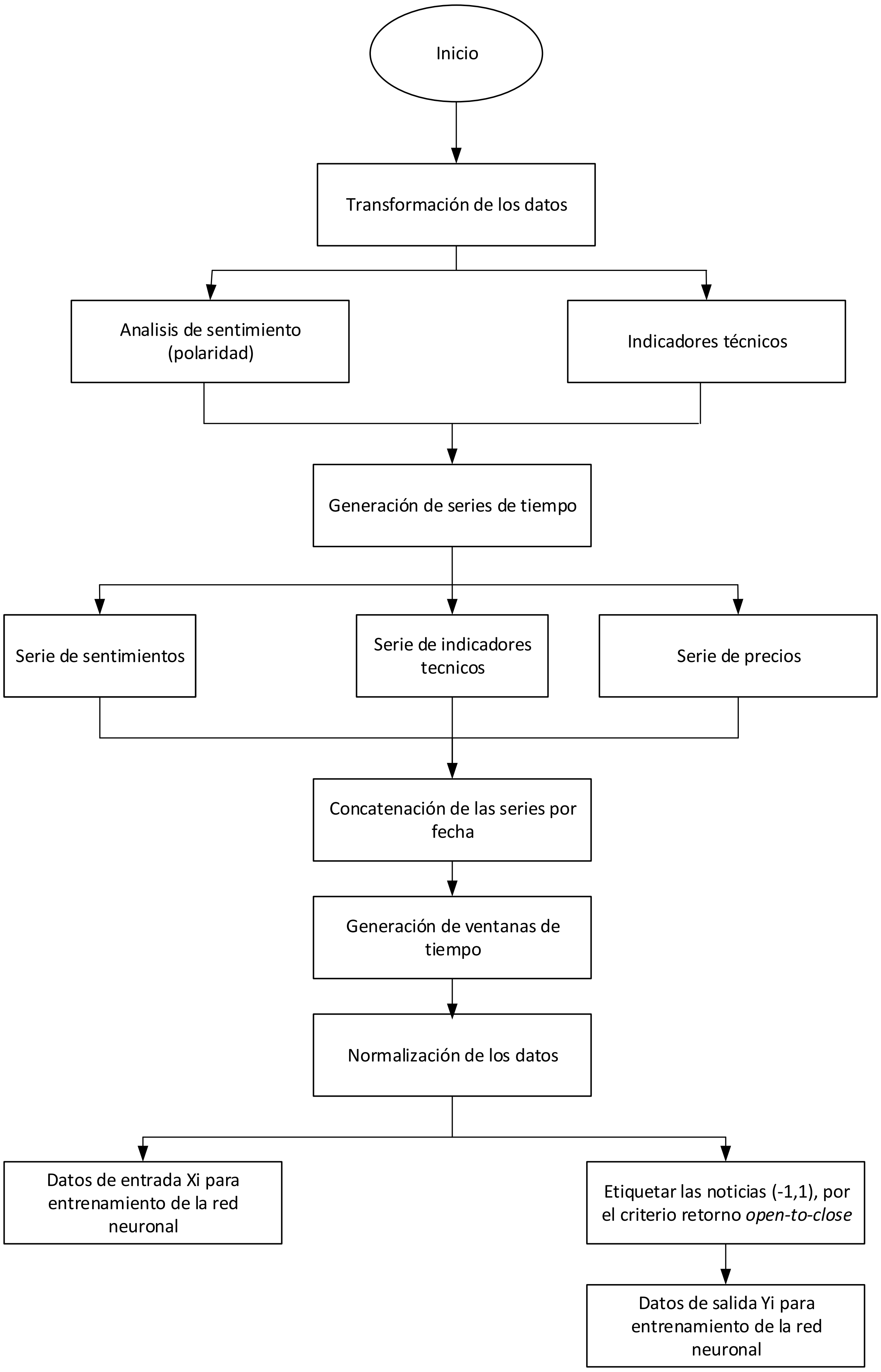


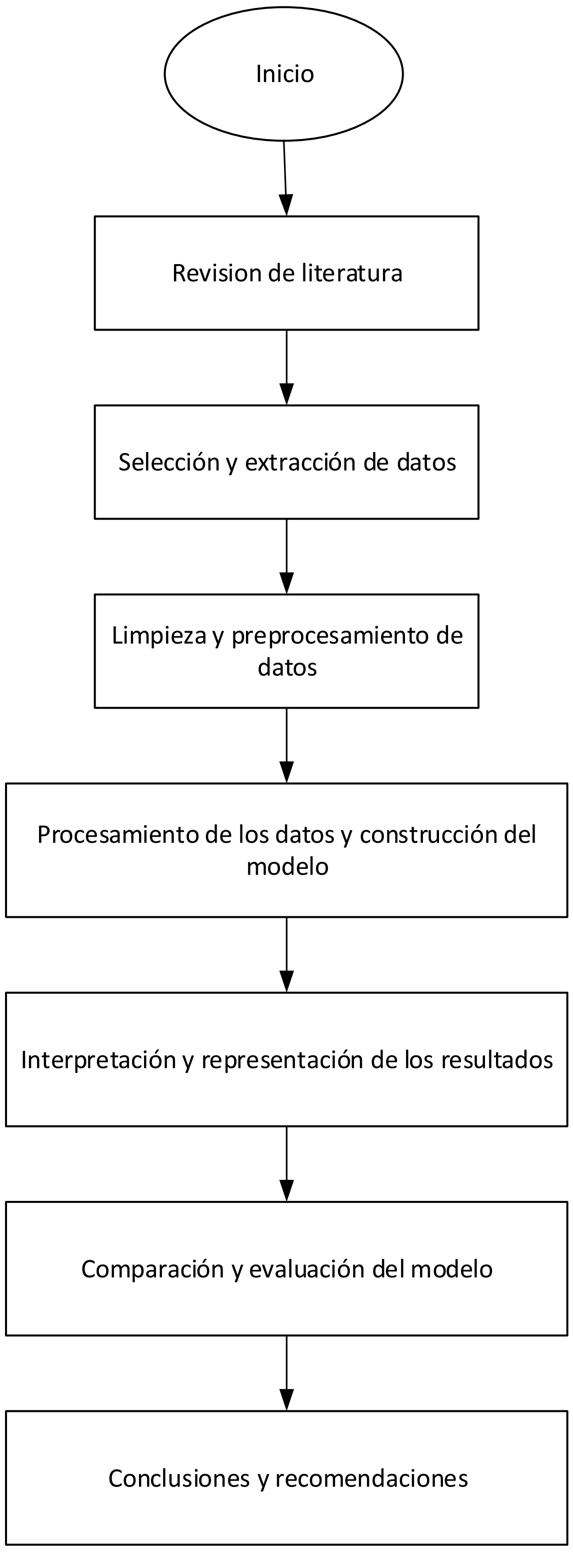
Proceso de obtención de datos
Preprocesamiento de datos
Selección tipo de datos
Noticias web (datos textuales)
Precios históricos (datos númericos)
Selección fuente de información
Portal web de noticias de Colombia: El Tiempo y La Republica
Página web de la Bolsa de Valoes de Colombia
Selección de atributos
Fecha, titulo y cuerpo del texto
Fecha, precio de apertura, precio alto, precio bajo, precio de cierre y volumen.
2013
datos
2019
Periodo de 7 años
Obtención del dataset
Web scraping usando Octoparse 8.1.2 (version gratuita) y Webautomation.io (versión de prueba)
Numero de documentos de El Tiempo: **4145** Numero de documentos de La Republica: **1574**
Descarga manual desde Google Chrome
Numero de días bursatiles obtenidos: **1704**
Corrección errores de extracción
Eliminar documentos duplicados
Ajustar formato del atributo "Fecha"
Combinar los atributos "Título" y "Texto"
Agregar atributo "Precio de apertura" y corrección de atributos "Precio alto" y "Precio bajo"
Limpieza de datos
Convertir a minuscula, remover puntuaciones, simbolos, números y espacios
Convertir documentos en lista de *tokens*
Aplicar *steamming*
Remover *stopwords*
Transformacion de datos
Analizador de Sentimiento
Calculo de los indicadores Técnicos
Generación de las series de tiempo
Serie de Sentimientos, S(i)
Serie de Indicadores Tecnicos T(i)
Serie de Precios P(i)
Concatenación de datos por fecha
Serie concatenada Q(i)
Generación de las ventanas de tiempo
Τ = {1,2,5,7,10}
Serie dividida por la ventana de Q(τ)(i)
Normalización de datos
Normalización por Z-score
$x_i = \frac{Q_i - \mu(Q_i^\tau)}{\sigma(Q_i^\tau)},$
Series de Xi(T) forman las tomas secuenciales
X_i^τ
$y_i = O_{i+t} - C_d$
Etiqueta del artículo = Positivo = 1 si $Y_i \geq 0$ Negativo =0 si $Y_i < 0$
Datos de entrada de la red neuronal


Inicio
Realizar la ecuación de busqueda en la base de datos
Hacer el análisis bibliométrico para dicha ecuación de busqueda
Revisar en la literatura como se ha abordado el tema de estudio
Identificar los principales metodologías, técnicas y algoritmos empleados para el problema de clasificacion binaria del movimiento del precio en mercado bursatil


Inicio
Seleccionar la fuente de datos textuales
Seleccionar la fuente de datos de precios
Selección de atributos o *features* con los que se va a trabajar
Selección de atributos o *features* con los que se va a trabajar
Seleccionar el método de extracción de los datos no estructurados
Seleccionar el método de extracción de los datos estructurados
Recuperación de la informacion
Almacenamiento ordenado de los datasets


Inicio
Remover ruido residual de la extracción de datos
Eliminar duplicados
Imputación datos faltantes
Editar valores y ajuste de formato de atributos erróneos
Aplicar técnicas de pre procesamiento sobre datos textuales
Convertir a minuscula, remover puntuaciones, simbolos, números y espacios
Convertir documentos en lista de *tokens*
Remover *stopwords*
Aplicar *steamming*


Inicio
Transformación de los datos
Analisis de sentimiento (polaridad)
Indicadores técnicos
Generación de series de tiempo
Serie de sentimientos
Serie de indicadores tecnicos
Serie de precios
Concatenación de las series por fecha
Generación de ventanas de tiempo
Normalización de los datos
Datos de entrada Xi para entrenamiento de la red neuronal
Etiquetar las noticias (-1,1), por el criterio retorno *open-to-close*
Datos de salida Yi para entrenamiento de la red neuronal


Inicio
Revision de literatura
Selección y extracción de datos
Limpieza y preprocesamiento de datos
Procesamiento de los datos y construcción del modelo
Interpretación y representación de los resultados
Comparación y evaluación del modelo
Conclusiones y recomendaciones

Iterative Approach toModelBuilding. The development of amodel of this kind to describe the dependence structure in an observed time series is usually best achieved by a three-stage iterative procedure based on identification, estimation, and diagnostic checking.

1. By *identification* we mean the use of the data, and of any information on how the series was generated, to suggest a subclass of parsimonious models worthy to be entertained.

2. By *estimation* we mean efficient use of the data to make inferences about the parameters conditional on the adequacy of the model entertained.

3. By *diagnostic checking* we mean checking the fitted model in its relation to the data with intent to reveal model inadequacies and so to achieve model improvement.

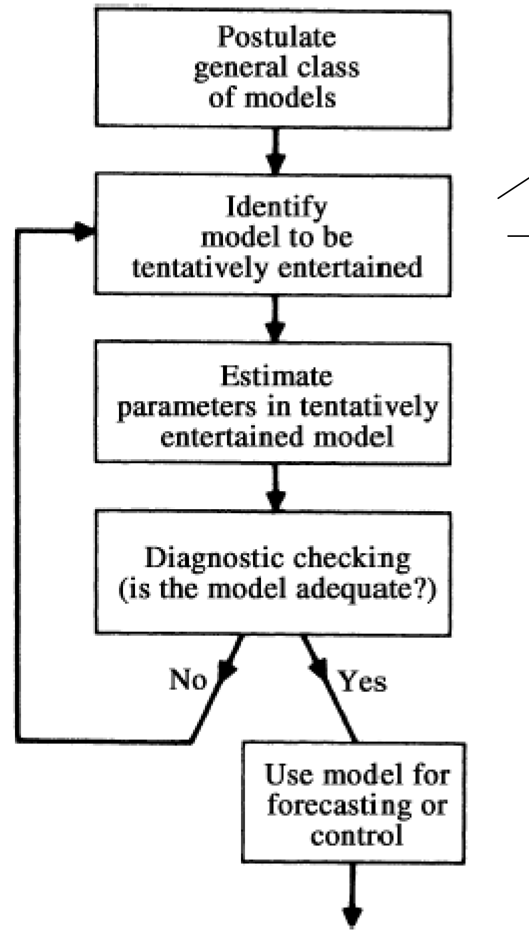


FIGURE 1.7 Stages in the iterative approach to model building.

Tecnicas de identificación:
ACF y PACF

The specific aim here is to obtain some idea of the values of p, d, and q needed in the general linear ARIMAmodel and to obtain initial estimates for the parameters. The tentative model specified provides a starting point

Autocorrelacion

the autocorrelation function will not die out quickly and will fall off slowly and very nearly linearly. A similar argument may be applied if more than one of the roots approaches unity.
Therefore, a tendency for the autocorrelation function not to die out quickly is taken as an indication that a root close to unity may exist.

Integrated (differencing)

For the reasons given, it is assumed that the degree of differencing d, necessary to achieve stationarity, has been reached when the autocorrelation function of $w_t = \nabla^d z_t$ dies out fairly quickly. In practice, d is normally 0, 1, or 2, and it is usually sufficient to inspect the first 20 or so estimated autocorrelations of the original series, and of its first and second differences, if necessary.

Prepare your data by using transformations (e.g. square roots or logarithms) to stabilize the variance and differencing to remove remaining seasonality or other trends.
Identify any processes that appear to be a good fit for your data.
Find which model coefficients provide the best fit for your data. This step is computationally complex and usually performed by a computer. Akaike's Information Criterion (AIC) is one option: if you compare two models, the one with the lower AIC is usually the "better" model.
Test the models' assumptions to see how well the model holds up to closer scrutiny. If your chosen model is inadequate, repeat steps 2 and 3 to find a potentially better model.
Compute forecasts on your chosen model with computer software.