Proyecto 1 Inteligencia de negocios

Daniel Perilla Ocampo - 201327313
Julio Alejandro Morales - 201327050
Julian David Mendoza Ruiz - 201730830

Comprensión del negocio y enfoque analítico

Oportunidad/problema de Negocio	Se desea entrenar un modelo de clasificación usando titulares de noticias financieras y poder a partir de ellos determinar si un titular no clasificado genera un sentimiento positivo o negativo.
Descripción de requerimiento desde el punto de vista de minería de datos	Se propone realizar una tarea clasificación a través de un procesamiento de análisis de sentimiento, con procesamiento de lenguaje natural con el fin de encontrar el sentimiento que produce un titular de noticias financiero. Esto con el fin de en un futuro poder predecir movimientos de inversión basados en el análisis de sentimiento.

Detalles de la actividad de minería de datos		
Tarea	Técnica	Algoritmo y parámetros utilizados (con la justificación respectiva)
Clasificación (Algoritmo supervisado)	Árbol de decisión	Algoritmo a utilizar: C4.5 Parámetros: sentimiento, títulos

Comprensión de los datos y preparación de los datos

Descripción de los datos Los datos para realizar este análisis fueron obtenidos de los siguientes repositorios:

- https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news?search=news+text

Abajo hay una tabla que explica la composición básica de lo que es la composición de los datos:

Nombre columna	Descripción
Sentimento	Indica el sentimiento del usuario. Puede ser “positivo” o “negativo” o “neutro
Titular	Es el titular de alguna noticia sobre compañías o acciones.

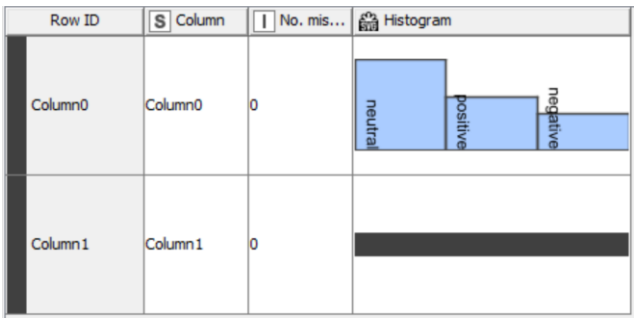
Datos	Filas	Columnas
#	4845	2

- **Limpieza de datos:**

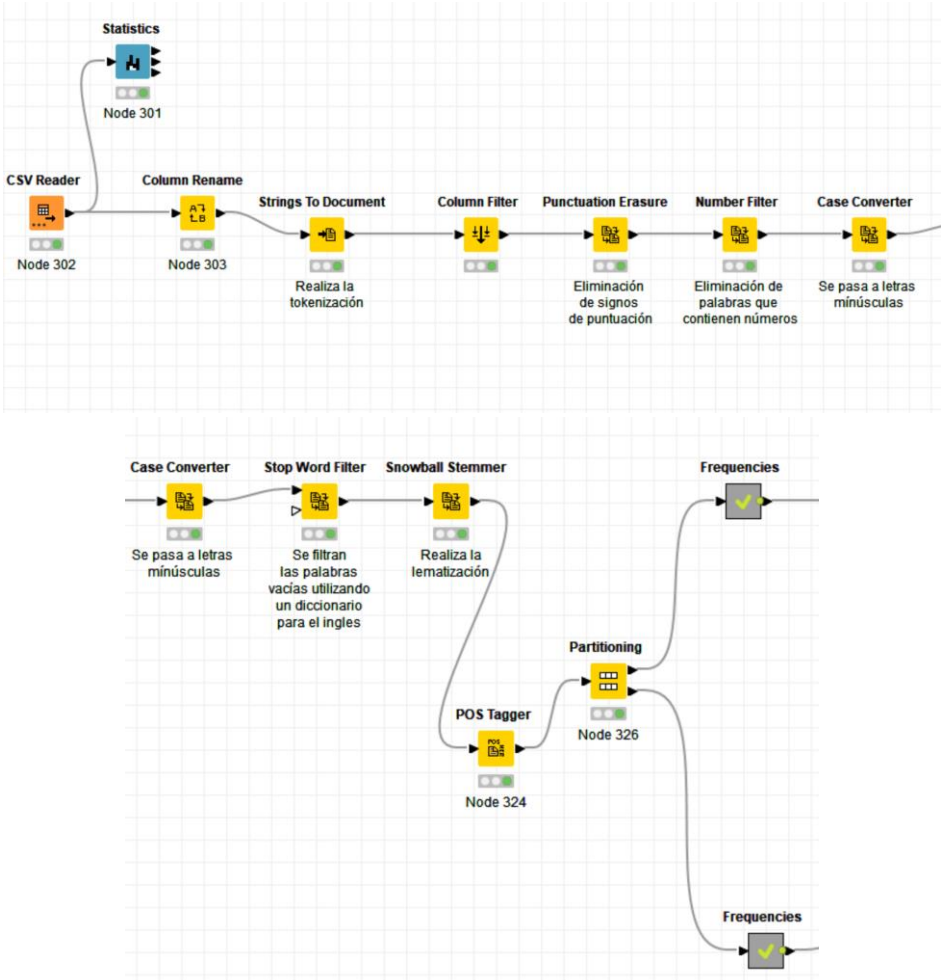
Afortunadamente, los datos que se analizaron no tenían huecos en términos de valores faltantes y no parecía que existiesen valores erróneos a pesar de carecer de diccionario. Por esa razón no fue necesario eliminar datos o modificar valores dentro de los datos de entrada.

- **Transformaciones:**

Como se estuvo trabajando con datos en forma de textos fue necesario hacer una serie de transformaciones para que pudiéramos aplicar minería de datos a los datos de entrada. Gracias a que los datos de entrada son textos. Lo primero que se tuvo que hacer es convertir los datos en vectores que estaban compuestos de unos y ceros dependiendo de si dentro de su cuerpo tenían determinadas palabras. Ya luego de esta transformación

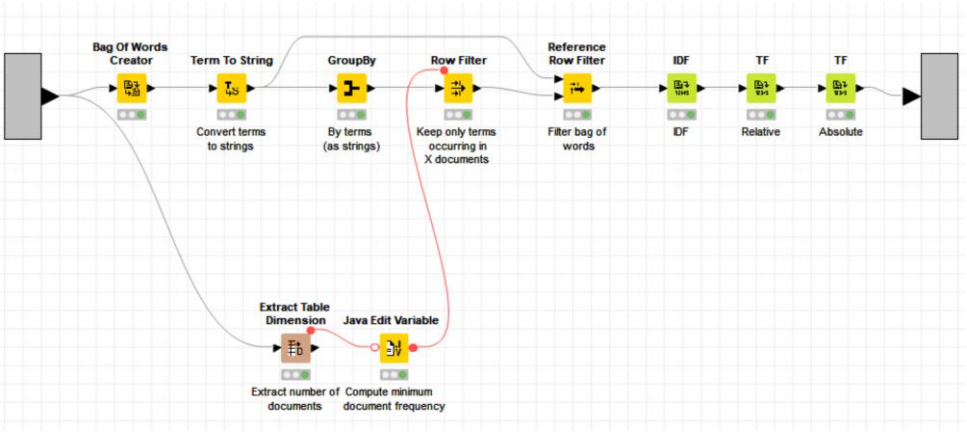


Como se puede observar en la tabla anterior, el dataset que se utilizó para el estudio no tiene ningún valor faltante en ninguna de las columnas. Además, la clase “neutral” tiene considerablemente más valores que el resto por lo que esto afectará la precisión a la hora de entrenar el modelo de clasificación

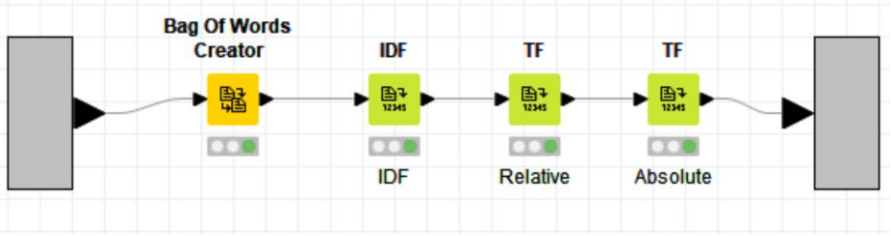


Los nodos dentro del flujo que van desde “Column rename” hasta los meta-nodos “Frequencies” corresponden a la transformación de los titulares de tipo string a un documento ya preparado para hacer minería de texto. En particular, se crea la columna de “document” a partir de los titulares que son de tipo string y luego se elimina esta columna pues no se utiliza. Después se elimina la puntuación y los números de cada documento además de cambiar todas las letras a minúsculas. Cabe aclarar que en este punto se crea la columna documento preprocesado, la cual se utilizará para realizar las transformación por el camino del flujo de los datos de entrenamiento.

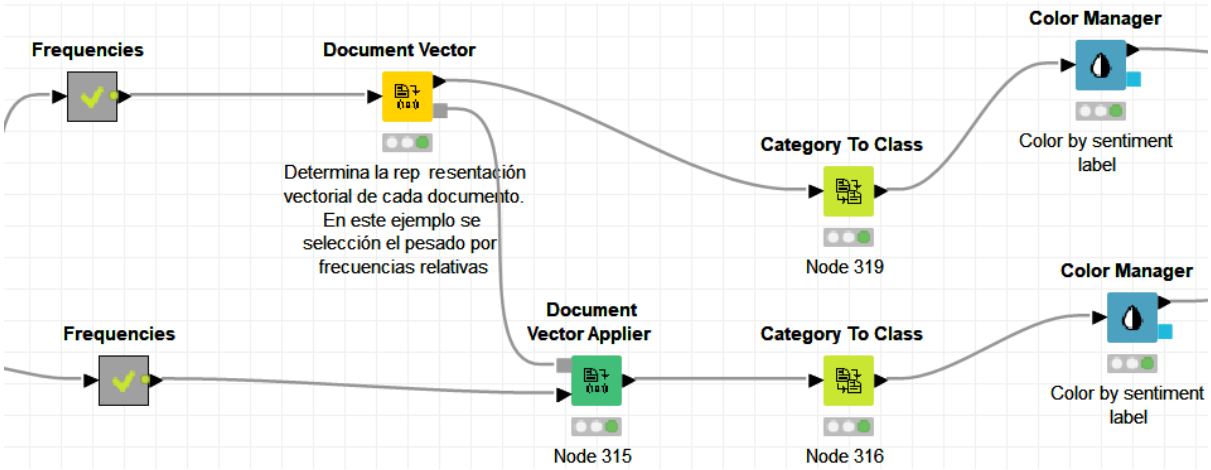
Para poder realizar minería de texto correctamente se eliminan las “stop words”, palabras que en general no aportan información útil dentro del documento, se hace lematización para que las palabras conjugadas no sean tomadas como distintas y se asigna un POS Tag.



Para preparar los datos (se utiliza la columna de “preprocessed document”) que se utilizarán en el entrenamiento del modelo se procesa el documento y se crea una representación de bolsa de palabras por cada uno. Con Term to string se crea una columna con las palabras de cada documento (ya procesado) y luego se agrupan para saber cuántas veces aparecen dentro de todos los documentos. Para agilizar el procesamiento de los datos se filtran los términos que no hayan aparezcan en x documentos (el mejor resultado fue 20). Finalmente se filtran de la bolsa de palabras los términos que no cumplen los requisitos del filtro anterior y se obtiene el TF-IF, una medida de la importancia de una palabra dentro de su documento.



Este flujo de frecuencias es para los datos de prueba, solo se crea la bolsa de palabras y se obtiene el TF-IF. En este flujo la transformación se hace con respecto al documento de la columna “document”.



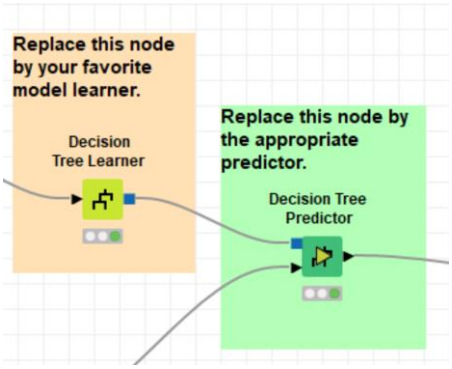
En la recta final de este proceso de preparación de datos, una vez ya se ha obtenido la frecuencia relativa de los términos (TF) y se tiene la bolsa de palabras respectiva tanto de los datos de entrenamiento como de los de prueba se procede a lo siguiente:

- A partir de los datos de entrenamiento se genera un modelo de Vector de Documentos, esto es un modelo de la representación vectorial de los documentos que entran como parámetro y se hace el pesado de este vector a partir de las palabras de la bolsa de palabras como variables binarias.
- Por otro lado, para los datos de prueba, si se genera efectivamente un Vector de Documentos a partir del set de documentos original casi sin procesar, que se alimenta y aplica como base el modelo creado anteriormente.

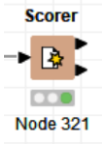
Finalmente, se toma el modelo de Vector de Documentos generado para los datos de entrenamiento, se le agrega nuevamente la columna de la variable objetivo (la de los sentimientos), se le asigna un color diferente a cada posible valor de clasificación y esto se le da como entrada al nodo de aprendizaje para generar el árbol de decisiones.

Por su lado, al Vector de Documentos generado para los datos de prueba, se hace algo similar que, a los anteriores, se les agrega la variable objetivo, la identificación por colores y ya se pasan como entrada al modelo final dónde serán clasificados de acuerdo al árbol de decisiones generado.

Modelado y evaluación.



La tarea de clasificación se concreta en estos dos nodos. El flujo con los datos de entrenamiento se conecta al nodo “Decision Tree Learner” donde el modelo es entrenado para poder predecir el sentimiento de un título teniendo en cuenta los títulos clasificados anteriormente. Se utilizó el índice de GINI sin “pruning” ya que nos resultó en la mayor precisión una vez se utilizó el modelo entrenado con los datos de prueba mediante el nodo “Decision Tree Predictor”.



Se utilizó el nodo “Scorer” para obtener las métricas de rendimiento del modelo.

Análisis de resultados.

A continuación, se pueden ver las 2 tablas de resultados asociadas con el nodo “Scorer”. La matriz de confusión muestra la relación entre la predicción de los valores (denominado por I al comienzo de las columnas) y los valores en realidad (mostrados como cuadros negros en las filas).

Matriz de confusión:

Row ID	I neutral	I positive	I negative
neutral	783	70	11
positive	282	84	43
negative	113	23	45

De la matriz de confusión se pueden sacar las siguientes conclusiones:

- Se predijeron correctamente 783 datos neutrales.
- Se predijeron correctamente 84 datos positivos.
- Se predijeron correctamente 45 datos negativos.

Estadísticas de precisión:

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...
neutral	783	395	195	81	0.906	0.665	0.906	0.331	0.767
positive	84	93	952	325	0.205	0.475	0.205	0.911	0.287
negative	45	54	1219	136	0.249	0.455	0.249	0.958	0.321
Overall	?	?	?	?	?	?	?	?	?

De estos datos lo que es importante decir son los datos de Recall y Precisión. El Recall habla de la proporción de datos entre los datos verdaderos escogidos sobre la totalidad de los datos verdaderos. La Precisión por su lado se refiere a la proporción entre los datos verdaderos escogidos y la totalidad de los datos.

Gracias a los resultados obtenidos se puede determinar que el Recall para los datos de neutral es muy bueno (0.906) sin embargo para lo que son los datos negativos y positivos no se tienen buenos resultados (0.249 y 0.205 respectivamente). Esto se dice sucede no solo a las fallas en el modelo, sino también a la cantidad de datos de tipo neutro.

Para el negocio en sí, se puede decir que esto es una base para un posterior análisis a la realización de un sistema de inversión de ayuda o automático en la inversión de acciones. Con esto se puede aproximar la manera mediante la cual los títulos de los periódicos influenciar el valor de las acciones en el día.

Trabajo en equipo:

Para el desarrollo del trabajo en equipo se tuvo que tener en cuenta los distintos horarios de cada uno de los integrantes además de las tareas a realizar. Por este motivo, se implementó en primera instancia, se configuró los horarios de trabajo en doodle.

Luego de establecer el horario de trabajo, se definió el trabajo a realizar dependiendo del tema escogido, en nuestro caso el de la relación entre los titulares de las revistas de negocio con la valoración de las acciones en las bolsas de valores.

Gracias a que en este proyecto se hizo uso de datos anotados, lo que teníamos que hacer era realizar la preparación de los datos de manera efectiva, para luego aplicar las técnicas de minería de datos que ya conocíamos.

Descripción de los retos enfrentados en el proyecto y las formas planteadas para resolverlos.

Reto	Solución	Encargado	Tiempo
Preparación de los datos (Teniendo en cuanto una relación de análisis de diccionario generado manualmente)	Como grupo decidimos que lo mejor era preguntarle a la profesora cuál era la mejor manera para preparar los datos teniendo en cuenta nuestras consideraciones del trabajo de minería de datos. (La profesora nos dió las herramientas necesarias para realizar la preparación de los datos de una	Equipo	4 h

	manera más sencilla y que derivó en datos más fáciles de analizar)		
Creación e implementación del modelo	Después de una búsqueda exhaustiva se pudo encontrar una página en knime que nos sirvió de guía para la creación y implementación de nuestro propio flujo	Equipo	6 h
Creación de documentos (Word, PP y Video)	Ya con el modelo implementado, se realizó lo que es la documentación para la entrega, en esta se tiene en cuenta el word, el powerpoint y el video	Equipo	2 h

Sustentación y evaluación del aporte individual

Durante el desarrollo del proyecto, se decidió realizar reuniones entre los miembros del equipo con el fin de socializar el desarrollo del proyecto, revisar las actividades realizadas, identificar tareas faltantes y administrar las responsabilidades establecidas.

A continuación, se presenta la tabla de resumen que establece la organización de los integrantes del equipo frente a las actividades requeridas para el proyecto. Los puntos obtenidos por cada integrante se asignaron de manera equitativa teniendo en cuenta el cumplimiento de cada estudiante con la tarea asignada, la disposición frente a la asignación de dichas tareas y el tiempo requerido para finalizarlas.

Integrante	Justificación	Tiempo Utilizado	Puntos
Julio Morales	· Comprensión y preparación de los datos. · Modelado y Evaluación. · Análisis de resultados.	12 h	33.3
Julián David Mendonza	· Comprensión y preparación de los datos. · Modelado y Evaluación. · Análisis de resultados	12 h	33.3
Daniel Perilla	· Comprensión y preparación de los datos. · Modelado y Evaluación. · Análisis de resultados	12 h	33.3