Revista Venezolana de Computación
ISSN: 2244-7040
http://www.svc.net.ve/revecom
Vol. 6, No. 1, pp. 1-9, Junio 2019
Fecha recepción: 15/05/2019, Fecha aceptación: 16/11/2019

Predicción del Sentimiento de Mercado para los Crudos Pivotes Venezolanos con base en el Análisis de Noticias

Arlán Briceño[1], Aneriz Rodríguez[1], Haydemar Núñez[2]
arlan.briceno@gmail.com, aneriz.rodriguez@gmail.com, haydemar.nunez@ciens.ucv.ve

[1] Escuela de Computación, Universidad Central de Venezuela, Caracas, Venezuela
[2] Centro de Ingeniería de Software y Sistemas, Universidad Central de Venezuela, Caracas, Venezuela

Resumen: En este estudio se utilizan diversas técnicas de la minería de datos para emitir una recomendación basada en la clasificación del sentimiento de un grupo de noticias. La idea de esta recomendación es apoyar la toma de decisión sobre la dirección del precio futuro del petróleo crudo venezolano de exportación. Específicamente, se trabajó con minería de texto para clasificar noticias recientes del mercado petrolero a partir del aprendizaje de noticias previamente clasificadas. En este artículo se plantea y resuelve el problema de clasificación de noticias desde el punto de vista de un proyecto de minería de datos. El proceso de minería contempla la recolección de los datos, la limpieza y transformación de los mismos, aplicación de diferentes algoritmos de clasificación, y la fase de difusión y uso. Los resultados demuestran que el clasificador *Máquina de Vectores de Soporte* (*SVM*) ofrece el mejor desempeño para la correcta clasificación de las noticias.

Palabras Clave: Minería de Texto; Análisis de Términos más Frecuentes; Balanceo de Clases; Selección de Variables; Máquina de Vectores de Soporte.

Abstract: In this study, various data mining techniques are used to make a recommendation based on the sentiment classification of a newsgroup. The idea of this recommendation is to support decision making about the direction of the future price of Venezuelan export crude oil. Specifically, text mining was used to classify recent oil market news based on learning from previously classified news. This article discusses and solves the news classification problem from the point of view of a data mining project. The mining process contemplates the collection of data, the cleaning and transformation of it, the application of different classification algorithms, and the dissemination and use phase. The results show that the *Support Vector Machine* (*SVM*) classifier offers the best performance for the correct classification of news.

Keywords: Text Mining; Frequently Asked Terms Analysis; Class Balancing; Variables Selection; Support Vector Machine.

I. Introducción

El desarrollo de fórmulas para valoración de crudos es de gran importancia en la industria internacional del petróleo ya que estas permiten optimizar los ingresos por exportación de materia prima no renovable [1]. Un aspecto a destacar de este proceso es su complejidad, debido a la necesidad de análisis y ponderación (generalmente expuesto a críticas y revisiones de expertos) de una gran cantidad de aspectos técnicos, económicos y políticos en un contexto internacional, y a la insuficiencia de información pertinente.

En el caso de Venezuela, las fórmulas de precio usadas para los 4 crudos pivotes (*Santa Bárbara, Mesa 30, Merey 16 y Boscán*) consideran un factor de ajuste *K*, de estimación mensual, el cual refleja el comportamiento coyuntural del mercado y la estrategia de colocación de los crudos de exportación [2].

La estimación del factor *K* consiste en realizar un ajuste a las fórmulas de precios para los crudos pivote en los mercados de interés, a fin de evitar posibles distorsiones que pueden generarse debido a la disparidad de condiciones políticas, financieras, económicas y técnicas entre los países consumidores, productores y exportadores de crudos [2]. La respectiva estimación involucra un análisis global del mercado petrolero y la ponderación de factores cualitativos relevantes.

En este sentido, se desea determinar mensualmente el sentimiento del mercado para ajustar las fórmulas de precio de los crudos pivotes venezolanos *Santa Bárbara, Mesa 30, Merey 16 y Boscán*, en los mercados de la Costa Estadounidense del Golfo de México, Noroeste de Europa y Asia. La propuesta

para realizar esta tarea es utilizar la minería de datos [3][4] para el análisis de noticias económicas que, según los expertos, pueden influir en el mercado petrolero.

Este documento se encuentra estructurado de la siguiente forma: en la Sección II se describen algunos trabajos relacionados con el tema de interés. En la Sección III se detallan el contexto del problema y los aspectos que se consideran relevantes a este. Luego, en la Sección IV, se describe cómo fue aplicado el proceso de minería de datos para la construcción de modelos de predicción del sentimiento del mercado. Por último, se presentan las conclusiones derivadas de este trabajo.

II. Trabajos Previos

Las noticias de la web pueden ser usadas para rastrear con precisión no sólo varios fenómenos sociales sino también las tendencias financieras [5][6]. Los sistemas de los mercados financieros son complejos y, por lo tanto, las decisiones comerciales suelen basarse en información sobre una gran variedad de temas socioeconómicos y acontecimientos sociales.

Tradicionalmente los inversionistas, economistas y periodistas revisan publicaciones mensuales de datos macroeconómicos relacionados con las condiciones económicas. El gran problema con estos datos es que la información está disponible con un desfase que aumenta muy rápido. De hecho, los datos de un mes dado se publican generalmente a mediados del mes siguiente y suelen revisarse unos meses después. Ésta es la razón por la que en los últimos años se ha estudiado ampliamente un nuevo enfoque, basado en el aprendizaje automático, por su potencial para predecir la dirección de los mercados financieros.

Aplicando la teoría y las herramientas de la tecnología de la información y las comunicaciones (*TIC*), la estadística y la econometría, se han desarrollado sistemas capaces de recoger y analizar grandes cantidades de datos que están disponibles gratuitamente en la red. Por ejemplo, el análisis de sentimientos de la opinión expresada por los usuarios de las redes sociales es un campo de aplicación muy estudiado en los mercados financieros.

En este sentido, en [7] propusieron un método de predicción de la tendencia de los precios del petróleo con base en sentimientos. En el método propuesto se incluyen tres pasos principales, a saber, el análisis de sentimientos, el análisis de relaciones y la predicción de tendencias. En el análisis de sentimientos, este se extrae siguiendo un enfoque basado en diccionarios para capturar la información relevante en línea sobre los mercados petroleros y los factores conductores. En el análisis de relaciones, se aplica el test de causalidad de *Granger* para explorar cómo el sentimiento impacta el precio del petróleo. En la predicción de tendencias, el sentimiento se utiliza como una variable independiente importante, y se usan algunos modelos de predicción populares, como por ejemplo, regresión logística, máquina de vectores de soporte, árbol de decisión y redes neuronales, en particular el Perceptron Multicapa.

En [8] analizaron, en un periodo de dos años, la relación entre los precios diarios del crudo *WTI* (*West Texas Intermediate*) y las múltiples variables extraídas de Twitter, Google Trends, Wikipedia y la base de datos *Global Data on Events, Language, and Tone* (*GDELT*). Se aplicó un análisis basado en semántica para estudiar el sentimiento, la emotividad y la complejidad del lenguaje utilizado. Los modelos *ARIMAX* se utilizaron para hacer predicciones y confirmar valores de variables.

En [9] propusieron un método de predicción del precio del crudo basado en la minería de texto en línea, con el objetivo de captar los antecedentes más recientes de las fluctuaciones de precios en el mercado. Específicamente, se trata de un intento inicial de aplicar técnicas de aprendizaje profundo para la predicción del precio del petróleo crudo y extraer patrones ocultos dentro de los medios de comunicación en línea utilizando una red neuronal convolucional (*CNN*). Si bien las características del sentimiento de las noticias y las características extraídas por el modelo de *CNN* revelan relaciones significativas con el cambio de precios, es necesario agruparlas según su tópico, usando el modelo de tópicos *Latent Dirichlet Allocation* (*LDA*), para distinguir los efectos sobre el precio de varios tópicos de noticias en línea y obtener una mayor precisión en la predicción.

Recientemente, en [10] propusieron un modelo de predicción del consumo de petróleo basado en datos en línea que utiliza las tendencias de Google, el cual refleja varios factores relacionados basados en una enorme cantidad de resultados de búsqueda. Este modelo implica dos pasos principales, el análisis de relaciones y la mejora de la predicción. En primer lugar, se realizan pruebas de cointegración y el test de causalidad de *Granger* con el fin de probar estadísticamente el poder predictivo de las tendencias de Google, en términos de tener una relación significativa con el consumo de petróleo. En segundo lugar, las tendencias efectivas de Google se introducen en los métodos populares de predicción para pronosticar tanto las tendencias como los valores del consumo de petróleo.

III. Planteamiento del Problema

Partiendo de los datos de precios de crudos y productos refinados, flujos de crudos y productos a nivel mundial, y las características del sistema de refinación de cada región estudiada; se procede a identificar y analizar el comportamiento de los factores cuantitativos y cualitativos que influyen en la formación de precios del crudo [11].

Los aspectos principales que regularmente se consideran para la estimación del factor *K* varían mensualmente dependiendo de las condiciones y acontecimientos específicos del mercado [1], y pueden obtenerse mediante suscripciones a servicios internacionales de información petrolera. Estos aspectos son agrupados en 4 categorías las cuales son: *refinación, crudo, estacionalidad y asfaltos* [2].

En la categoría *refinación* se considera el volumen de crudo a consumir (porcentaje de procesamiento en refinación); temporada de mantenimiento (capacidad incorporada o desincorporada para el mes en estudio); cierre de refinerías; adecuaciones

en sistemas de refinación; paradas de refinación no programadas; situaciones de fuerza mayor en refinación; oferta y demanda de productos refinados principales (naftas, gasolinas, destilados medios y fueloil); márgenes de refinación; craqueo, inventarios y arbitraje de productos refinados entre cuencas.

Con respecto a la categoría *crudo*, se considera la incorporación o desincorporación de volúmenes de crudo de competidores, programas de exportación de los crudos competidores, inventarios de crudos, situaciones de fuerza mayor para la exportación de crudos, almacenamiento flotante de crudos, expectativas de disponibilidad de oferta de crudos, expectativas de demanda de crudos, situación de oleoductos y arbitrajes de crudos entre cuencas.

En cuanto a la categoría *estacionalidad*, se consideran temporadas de gasolina o diésel para calefacción, generación termoeléctrica durante el verano en Arabia Saudita, monzones en Asia y huracanes en costa este de los Estados Unidos de América.

Para los *asfaltos* se consideran precios, temporada de asfaltado por región, expectativas de demanda, expectativas de oferta o disponibilidad, inventarios, mantenimiento o cierres no programados de refinerías asfalteras y arbitraje entre cuencas.

Como se puede notar, son muchos los aspectos que se deben considerar para estimar el factor de ajuste *K*. El proceso de estimación del factor *K*, como se aborda en este trabajo, consiste en realizar un análisis de las noticias agrupadas según las categorías mencionadas anteriormente y según los mercados de interés, para extraer el sentimiento de mercado petrolero.

IV. Proceso de Minería de Datos

A continuación se describen los pasos de la minería de datos [12] que se siguieron para la resolución del problema:

A. Recopilación e Integración

En la etapa de recopilación e integración de datos se tiene el problema planteado, los objetivos que se desean lograr, y la fuente de datos que se utilizará. Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde el punto de vista del negocio, para luego usar este conocimiento y definir el problema en términos de minería de datos, así como un plan preliminar que permita alcanzar los objetivos propuestos.

1) Objetivo del Negocio: Determinar mensualmente el sentimiento del mercado para ajustar el factor *K* de las fórmulas de precio de los crudos pivotes venezolanos *Boscán, Mesa 30, Merey 16 y Santa Bárbara*, en los mercados de la Costa Estadounidense del Golfo de México, Noroeste de Europa y Asia.

2) Fuentes de Datos: Los datos son obtenidos por suscripción a los siguientes servicios globales de información y precios sobre energía: *Argus* (https://www.argusmedia.com) y *S&P Global Platts* (https://www.spglobal.com) (fuentes externa de datos). Los expertos en analizar el mercado petrolero recopilan manualmente las noticias y luego las clasifican según el mercado destino, categoría de noticia y sentimiento. En la sección *Conociendo los Datos* se describe con mayor detalle el conjunto de datos.

3) Criterios de Éxito del Negocio: Se desea tener una precisión igual o superior a la que están proporcionando los expertos, 75% para al menos un crudo pivote en un mercado y, a su vez, reducir el tiempo que tardan los expertos en obtener el análisis completo del mercado.

4) Evaluación de la Situación:

- El negocio estima con esfuerzo propio el factor *K* obteniendo un margen de error aceptable.
- Cuenta con personal propio para realizar las actividades de los analistas del mercado petrolero.
- Dispone del acceso necesario a los principales proveedores globales de información sobre energía.
- Dispone de acceso a un software propietario de programación lineal para estimar precios de crudos vía valor bruto en productos.
- Cuenta con el asesoramiento de expertos en el negocio petrolero.
- Ausencia de un método para validar los juicios de experto en el análisis y la ponderación de factores cualitativos.

Esta situación es la que lleva a la búsqueda de una solución por medio de la minería de datos, a fin de obtener una herramienta de predicción que permita apoyar la toma de decisiones en cuanto al factor de ajuste *K* de la fórmula de precio de los crudos, así como para reducir el tiempo de análisis de los expertos.

5) Objetivos de la Minería de Datos: Los objetivos desde el punto de vista de la minería de datos son:

- Predecir, para cada noticia, la dirección en la que influye sobre el factor de ajuste *K*.
- Tarea de Minería de Datos: Clasificación.

6) Criterios de Éxito de la Minería de Datos: Desde la visión de la minería de datos, el criterio de éxito es obtener un rendimiento de predicción superior al 75%.

B. Preparación de los Datos

En este contexto, la preparación de los datos consiste en conocer el significado de todo el conjunto, seleccionar los datos adecuados para el estudio y realizar las transformaciones necesarias para convertir los documentos a una representación estructurada para poder aplicar los algoritmos de aprendizaje. Por último, aplicar técnicas de selección de variables para la reducción de la dimensionalidad.

1) Conociendo los Datos:

- El cuerpo (corpus) de entrada está constituido por 779 documentos de noticias que pueden influir en el sentimiento del mercado petrolero venezolano.
- Cada noticia se acompaña de los atributos *Fecha*, *Mercado*, *Categoría* y el *Sentimiento* (clase) asociados a cada

crudo pivote venezolano, cuya descripción se muestra a continuación:

- *Fecha*: denota el mes y año de la noticia.
- *Mercado*: los distintos mercados en los que influye la noticia. Estos pueden ser *CEGM* (Costa Estadounidense del Golfo de México); *NOE* (NorOeste de Europa); *Asia*.
- *Categoría*: las distintas categorías en las que se agrupan las noticias. Estas son *Refinación*; *Crudo*; *Estacionalidad*; *Asfaltos*.
- *Sentimiento Crudo* x: valor nominal en el conjunto $\{S : Sube, N : Neutro, B : Baja\}$ que denota si la noticia influye en el alza (S) del precio del crudo x, o es una noticia neutral (N) que no afecta su precio, o la noticia influye a la baja (B) del precio. Este valor fue proporcionado por los expertos.

Una muestra de estos datos se puede ver en la Tabla I, y un resumen se muestra en la Tabla II.

2) Selección de los Datos: Se tomó todo el conjunto de datos disponible, Dic2015 - Mar2017, para un total de 16 meses de información.

3) Transformación de los Datos: Esta etapa consiste en aplicar todas las transformaciones necesarias para lograr una representación estructurada a partir de los textos, siguiendo el modelo vectorial basado en bolsa de palabras. Los pasos que se aplicaron se mencionan a continuación:

- Traducción por parte de los expertos de algunas noticias desde el idioma inglés al idioma español.
- Corrección ortográfica de palabras.
- Corrección semántica de oraciones.
- Separación de noticias en *tokens* (tokenización).
- Eliminación de caracteres no imprimibles, acentos, números, símbolos, signos de puntuación y palabras vacías (*stopwords*).
- Adicionalmente, los expertos aportaron una lista de 500 palabras vacías aproximadamente, lo que permitió reducir la dimensionalidad de la matriz documento-término de 37.133 a 25.876 tokens, lo que representa una reducción del $30,32\%$. Estos tokens son la unión de palabras, bigramas y trigramas. Luego, se consideró solo analizar aquellos tokens que aparecieran en al menos el 1% del conjunto de noticias y en a lo sumo el 90% de las mismas. Esto se hizo con la idea de evitar aquellos tokens que sean específicos de una noticia ($< 1\%$) o comunes a todas las noticias ($> 90\%$). Esta restricción permitió reducir la dimensionalidad de la matriz de 25.876 a 615 tokens, lo que representa una reducción del $97,62\%$.
- Conversión a minúsculas.
- Lematización (*stemming*): aplicación del algoritmo *Porter Stemmer* para el español.
- Ejecución de 89 *expresiones regulares* para homologar expresiones escritas de diferentes maneras, como por ejemplo:
  - betun $\rightarrow$ bitumen
  - grado liviano $\rightarrow$ crudo liviano
  - usa $\rightarrow$ eeuu

Finalmente, para construir una representación estructurada de las noticias se construye una matriz documento-término donde cada elemento $TfIdf_{ij}$ se calcula de la siguiente manera:

$$TfIdf_{ij} = TfN_{ij} * Idf_i \tag{1}$$

$$TfN_{ij} = \frac{F(t_i, d_j)}{T(d_j)} \tag{2}$$

$$Idf_i = log_2(\frac{m}{D(t_i)}) \tag{3}$$

donde:

- m: cantidad de documentos
- n: cantidad de términos
- t_i: término i para $i = 1, 2, ..., n$
- d_j: documento j para $j = 1, 2, ..., m$
- $TfIdf_{ij}$: métrica $TfIdf$ del t_i en el d_j
- TfN_{ij}: frecuencia relativa o normalizada del t_i en el d_j
- $F(t_i, d_j)$: frecuencia absoluta o cantidad de veces que aparece el t_i en el d_j
- $T(d_j)$: cantidad de términos en el d_j
- Idf_i: componente Idf del t_i
- $D(t_i)$: cantidad de documentos que contienen el t_i

4) Selección de Variables: Para la selección de variables o tokens de la matriz de documento-término, se utilizaron tres métodos de selección de atributos por ranqueo (*Chi Cuadrado, Gain Ratio* e *Information Gain*) [13], los cuales seleccionaron exactamente las mismas variables para cada crudo, quedando distribuida la cantidad de atributos seleccionados de la siguiente manera:

- 404 para el crudo *Merey 16*.
- 401 atributos para el crudo *Boscán*.
- 370 atributos para el crudo *Mesa 30*.
- 351 atributos para el crudo *Santa Bárbara*.

Otro de los métodos utilizados para la selección de variables fue el *análisis de términos más frecuentes*. Para aplicar éste método, primero se totalizó cada término según su frecuencia de ocurrencia en la colección de documentos de noticias *Tf-Idf*, luego se ordenan los términos descendentemente según la totalización del paso anterior, y por último, para cada crudo se seleccionan los n primeros términos más frecuentes, donde n denota la cantidad de atributos elegidos mediante los métodos por ranqueo. Con éste método, los atributos seleccionados difieren de los atributos elegidos por los métodos de ranqueo. Los autores consideran que esta diferencia se debe a la propia naturaleza de ambas familias de métodos, es decir, mientras que los métodos por ranqueo utilizados se basan en la entropía de los atributos para determinar su relevancia, en el análisis de términos más frecuente se usa la frecuencia relativa de cada término con respecto a la colección de documentos.

5) Construcción de Nuevos Atributos: Se realizó una transformación en el atributo clase, cambiando su tipo de dato real en el rango $[-1, 1]$ a nominal. Así, se cambiaron los valores negativos al valor *B*, los valores de 0 a *N*, y los valores positivos al valor *S*.

Tabla I: Muestra de los Datos

FECHA	MERCADO	CATEGORÍA	NOTICIA	SANTA_BARBARA	MESA	MEREY	BOSCAN
01/12/2015	CEGM	Refinación	Se espera que el mantenimiento de crudo REFQ4 sea relativamente ligero en el futuro y se mantendrá así durante los primeros meses del primer trimestre de 2016. Dada la reciente recuperación de margen esperamos que CEGM continúe aumentando las carreras hasta muy por encima del 90%	B	B	B	B
01/12/2015	CEGM	Crudo	Se espera que el crudo de WTS Midland regrese a una prima contra WTI Cushing una vez que las operaciones de reanudación del oleoducto y mantenimiento local disminuyan. Esto aumentará los precios de la fórmula venezolana	N	N	N	N
01/12/2015	CEGM	Estacionalidad	Una abundancia de fondos de torre de vacío baratos (VTBs) en el USG agregó presión a la baja a los mercados de asfalto en masa. La gran cantidad compradores que buscan recibir entregas finales para terminar la temporada de pavimentación estaban retrasando las compras con la esperanza de pagar precios más bajos la próxima semana. Barcazas del Golfo de Estados Unidos se escucharon vendiendo tan bajo como $210-220/st en la parte posterior de la demanda débil una disminución de $20/st de la semana pasada. Una importante refinería de la costa del Golfo anunció que se estaría manteniendo en unidades no reveladas lo que podría aumentar el exceso de VTBs en la región y establecer el escenario para más debilitamiento a corto plazo en la mayoría de los mercados	N	N	N	B
01/01/2016	NOE	Asfalto	En Alemania Francia y Benelux se acordó mantener los precios domésticos a los niveles de enero. La actividad permanece débil como es usual en la época	N	N	N	B

Tabla II: Resumen de los Datos

VARIABLE	TIPO	AUSENCIAS	MODA	VALORES
Fecha	Fecha	0	01/03/17	01/12/2015...01/03/2017
Mercado	Nominal	0	Asia	CEGM, NOE, ASIA
Categoría	Nominal	0	Crudo	REFINACIÓN CRUDO ESTACIONALIDAD ASFALTO
Noticia	Nominal	0	NA	-
Santa Bárbara	Nominal	0	N	S, N, B
Mesa	Nominal	0	N	S, N, B
Merey	Nominal	0	N	S, N, B
Boscán	Nominal	0	N	S, N, B

6) Análisis Exploratorio: Con el análisis exploratorio se busca un entendimiento básico de los datos y de las relaciones existentes entre las variables. Para lograrlo, primero se deben organizar y preparar los datos de manera que puedan aplicarse métodos estadísticos para la detección de fallas en el diseño y recolección de los mismos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos (*outliers*) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad). Es importante señalar que el análisis exploratorio se llevó a cabo con todas las palabras y haciendo uso de los procedimientos gráficos descritos en [14] para manipular, resumir y visualizar características del texto.

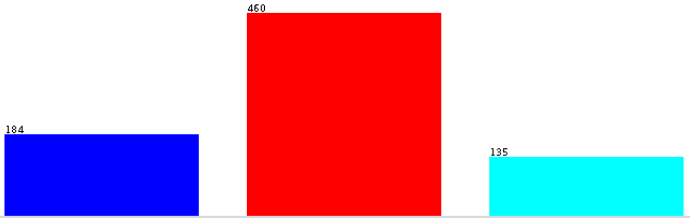

Figura 1: Histograma de Clase para *Santa Bárbara*

La Figura 1 muestra la distribución de las clases para el crudo *Santa Bárbara*, con cada clase descrita por un color: Azul = *B* (Baja), Rojo = *N* (Neutro), Cyan = *S* (Sube). En la Figura 2 se aplicó el análisis de términos más frecuentes para conocer cuáles eran los términos más frecuentes en el conjunto de noticias. Específicamente, se utilizaron las medidas *Tf* y *TfIdf* para graficar los 10 primeros términos más frecuentes.

La Figura 3 muestra los términos con una correlación mayor o igual al 95% que están asociados al mercado de *Asia*. Mientras mayor sea la correlación, más gruesa es la línea entre los términos.

También se aplicó el análisis de términos más frecuentes para pares de palabras (*bigramas*). Los resultados se muestran en

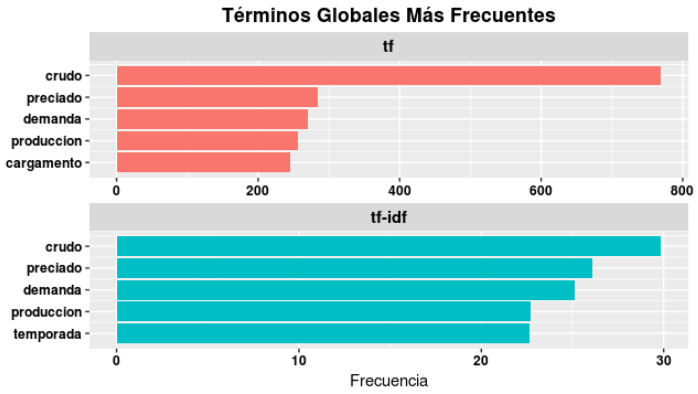

Figura 2: Términos más Frecuentes según *Tf* y *TfIdf*

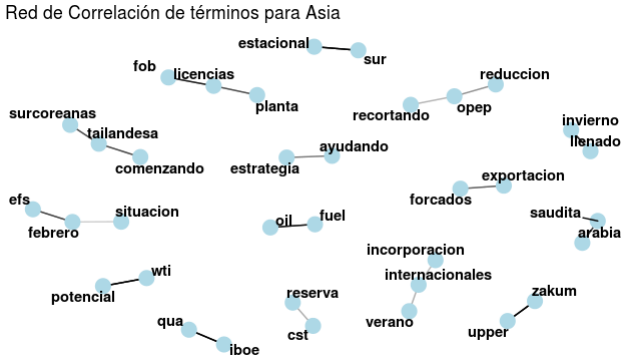

Figura 3: Red de Correlación, Mercado de Asia

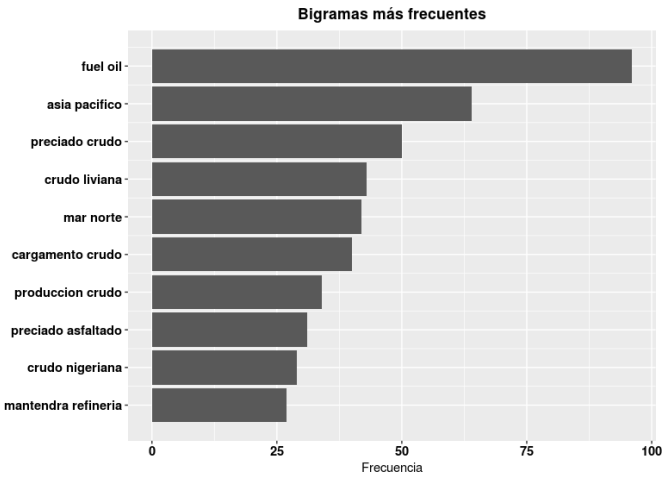

Figura 4: Bigramas más Frecuentes

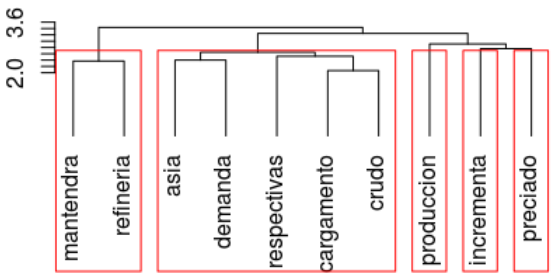

Figura 5: Dendrograma

la Figura 4. La Figura 5 muestra las relaciones más fuertes entre las palabras más frecuentes entre todas las noticias.

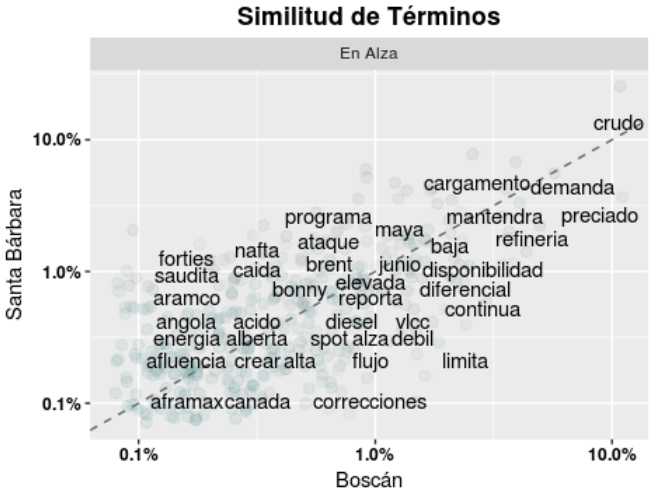

Figura 6: Similitud de Términos entre *Santa Bárbara* y *Boscán*

Para visualizar cuáles son los términos que están mayormente asociados a un crudo dado en comparación con otro crudo, se emplea el gráfico de similitud por proporción de palabras. La Figura 6 muestra este gráfico para los crudos Santa Bárbara y Boscán, agrupado por sentimiento al alza.

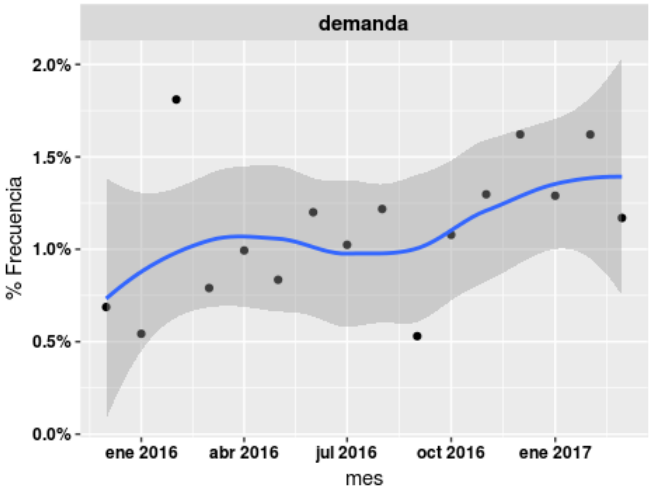

Figura 7: Variación de Frecuencia en el Tiempo

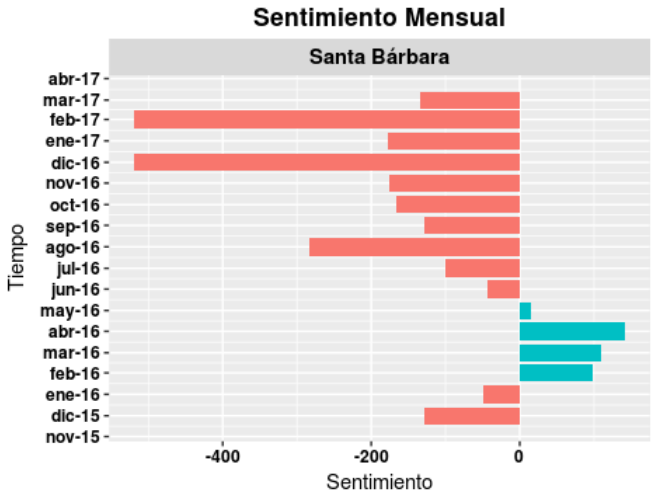

Figura 8: Sentimiento Mensual - Santa Bárbara

La Figura 7 ilustra la variación de la frecuencia en el tiempo para el término *demanda*, y un gráfico sobre el sentimiento mensual para el crudo *Santa Bárbara* se muestra en la Figura 8.

7) Vista Minable: El último paso en la preparación de los datos es la construcción de la vista minable. En este estudio, se construyeron cuatro vistas minables, una por cada uno de los crudos. Debido a que las clases estaban desbalanceadas, se utilizó el método *SMOTE* [15] para generar muestras sintéticas de las clases minoritarias.

A continuación se resumen los pasos generales que se siguieron para construir una vista minable con métrica *TfIdf*.

- Calcular la matriz de documento-término DT_{mxn} con métrica $TfIdf_{ij}$.
- Eliminar atributos redundantes. Resultado $DT_{mxn'}$. Los atributos redundantes son aquellos que están altamente correlacionados, aplicando la correlación de Pearson mayor a 0.75.
- Balancear las clases minoritarias. Resultado $DT_{m'xn'}$. El balanceo se realizó aplicando el algoritmo *SMOTE*. Las clases minoritarias para algunos crudos pivotes estaban por debajo del 15% y para otros crudos estaban por debajo del 19%.
- Seleccionar atributos por importancia o frecuencia. Resultado $DT_{m'xn''}$.

C. Minería de Datos

En este estudio se buscó predecir, para cada noticia, la dirección (S, N, B) en la que influye sobre el factor de ajuste K. Para lograrlo, es necesario contar con un conjunto de noticias previamente clasificadas y aprender de ellas. A continuación se presentan las distintas máquinas de aprendizaje utilizadas para la tarea de clasificación de noticias, las cuales fueron entrenadas haciendo uso de la herramienta *Weka*:

- Bayes Ingenuo con Estimador Núcleo
- K-Vecinos para $K = 3, 5, 7$

- RIPPER (JRip) con $F = 3$, $N = 2$, $O = 2$, $S = 1$ y *Prunning* activado
- C4.5 (J48) con $C = 0.25$ y $M = 2$
- Random Forest con $P = 100$, $l = 100$, $slots = 1$, $K = 0$, $M = 1$, $V = 0.001$ y $S = 1$
- Máquina de Vectores de Soporte (SVM) con K=*NormalizedPolyKernel* y *BuildCalibrationModels* activado

La aplicación de estos algoritmos se hizo con validación cruzada de K-particiones (*K-fold cross validation*) con $k = 10$. Considerando que para el negocio es más importante reconocer cuando una noticia influye en el alza de los precios, entonces la clase positiva es *S*. Para seleccionar la máquina de aprendizaje con los mejores resultados de clasificación, se toma aquella que presente las siguientes métricas más altas en función de la clase positiva:

- Instancias Correctamente Clasificadas (ICC).
- Tasa de Verdaderos Positivos (TVP).
- Precisión.
- Sensibilidad.
- Medida-F.
- G-Media.

Es importante señalar que para cada crudo, las máquinas de aprendizaje se aplicaron utilizando el conjunto de atributos seleccionados por los métodos de ranqueo (*Chi Cuadrado, Information Gain y Gain Ratio*) así como el conjunto de atributos seleccionados por el análisis de términos más frecuentes con *TfIdf*.

D. Evaluación

Las Tablas III y IV muestran los resultados para cada una de las máquinas de aprendizaje. Se observa que la Máquina de Vectores de Soporte (*SVM*) es la que exhibe, en general, los mejores valores de las métricas de rendimiento. Tomando en cuenta que se utilizó para cada crudo dos conjuntos de atributos como consecuencia de los métodos de selección de variables, el modelo de predicción se construyó de la siguiente manera:

- Para los crudos *Boscán, Merey 16* y *Santa Bárbara*, la Máquina de Vectores de Soporte (*SVM*) se entrenó con el conjunto de atributos seleccionados por los métodos de ranqueo (*Chi Cuadrado, Gain Ratio, Information Gain*).
- Para el crudo *Mesa 30*, la Máquina de Vectores de Soporte (*SVM*) se entrenó con el conjunto de atributos seleccionados mediante el análisis de términos más frecuente con *TfIdf*.

La Tabla III muestra las métricas de las máquinas de aprendizaje usando los atributos seleccionados por los métodos de ranqueo para el crudo *Boscán*, mientras que la Tabla IV muestra las métricas de las máquinas de aprendizaje usando los atributos seleccionados mediante el análisis de términos más frecuente con *TfIdf* para el crudo *Boscán*.

Aunque en la literatura puede parecer que *Random Forest* obtiene, en general, un mejor desempeño en tareas de clasificación, nuestros resultados demuestran que esto no es así cuando los datos son textos.

E. Difusión y Uso

El modelo de predicción recibe como entrada un grupo de noticias a clasificar organizado como sigue: *MERCADO, CATEGORIA* y *NOTICIA*. A este grupo se le aplica la Máquina de Vectores de Soporte (*SVM*) para asignarle una clase (*S,N,B*) o sentimiento de mercado a cada una de las noticias. Luego que todas las noticias hayan sido clasificadas, el modelo emite una recomendación sobre si se debe subir, mantener o bajar el valor del factor *k*, según el siguiente criterio:

Para Todo $c_i \in C$ **Hacer**
  Si $T(c_i) > 1,5 * mg$ **Entonces**
    $R(c_i)$
  Fin Si
Fin Para

donde:

- $C = \{c_i / c_1 = S, c_2 = N, c_3 = B\}$.
- $T(c_i)$: cantidad de noticias clasificadas como c_i
- $mg = \sqrt[3]{T(c_1) * T(c_2) * T(c_3)}$ (media geométrica)
- $R(c_i)$: emitir recomendación según c_i

Adicionalmente, el modelo calcula la probabilidad de que cada noticia pertenezca a una clase y genera una gráfica que muestra la tendencia del sentimiento del mercado como lo ilustra la Figura 9. En la actualidad, el sistema se encuentra en período de prueba.

V. Conclusión

Los resultados demuestran que tanto los objetivos como los criterios de éxito desde el punto de vista del negocio y de la minería de datos fueron logrados. En este sentido, se superó ampliamente el objetivo del negocio ya que se logró predecir con una precisión cercana al 90% el sentimiento del mercado para los 4 crudos pivote en los 3 mercados de interés. Cuando los expertos no lograron un acuerdo sobre el sentimiento de alguna noticia, estos utilizaron la clasificación hecha por el modelo para decidir sobre el sentimiento de la noticia siempre y cuando la probabilidad de que una noticia perteneciera a una clase superara el 90%. Así, se logró reducir en un 50% aproximadamente el tiempo de análisis del mercado mediante el uso del modelo anteriormente descrito para la predicción y recomendación. Adicionalmente, se logró el objetivo de la minería de datos mediante la clasificación de las noticias con una precisión cercana al 90%.

Los autores consideran que los aportes de este trabajo están relacionados con la aplicación de técnicas de aprendizaje automático en el contexto económico petrolero venezolano, y cómo pueden ser utilizadas para apoyar la toma de decisiones en un sector tan importante para el país.

Como el modelo propuesto fue entrenado a partir de juicios de expertos y este juicio cambia en función de las condiciones del mercado petrolero, es necesario re-entrenar el modelo cada

Tabla III: Métricas basadas en Métodos de Ranqueo

Crudo Pivote	Boscán	Métodos de Selección de Atributos	Chi Cuadrado, Information Gain y Gain Ratio *
# Atributos	401	# Instancias	1217

Métrica		Bayes Ingénuo (Estimador Núcleo)	SMV (SMO) (NormalizedPolyKernel) (BuildCalibrationModels)	KNN (K=3)	KNN (K=5)	KNN (K=7)	RIPPER (Jrip)	C4.5 (J48) (BinarySplit)	Random Forest
ICC		87,35%	92,19%	70,99%	67,38%	62,12%	74,94%	81,35%	89,65%
IIC		12,65%	7,81%	29,01%	32,62%	37,88%	25,06%	18,65%	10,35%
TVP	B	92,70%	90,40%	98,50%	95,90%	92,70%	65,20%	77,80%	84,20%
	N	80,00%	92,10%	39,80%	35,70%	28,20%	87,70%	84,80%	98,00%
	S	94,60%	94,30%	96,50%	92,70%	89,20%	62,90%	79,00%	80,60%
TFP	B	6,90%	2,90%	26,50%	31,30%	35,30%	5,50%	5,50%	0,80%
	N	4,10%	7,00%	1,10%	1,80%	1,80%	30,40%	18,40%	17,20%
	S	7,40%	2,70%	12,60%	12,30%	15,50%	6,30%	6,40%	0,70%
Precisión	B	84,10%	92,50%	59,20%	54,50%	50,60%	82,30%	84,70%	97,60%
	N	94,30%	91,80%	97,00%	94,30%	92,90%	71,10%	79,70%	82,90%
	S	81,60%	92,50%	72,70%	72,50%	66,70%	77,60%	81,10%	97,70%
Sensibilidad	B	92,70%	90,40%	98,50%	95,90%	92,70%	65,20%	77,80%	84,20%
	N	80,00%	92,10%	39,80%	35,70%	28,20%	87,70%	84,80%	98,00%
	S	94,60%	94,30%	96,50%	92,70%	89,20%	62,90%	79,00%	80,60%
Medida-F	B	88,20%	91,40%	74,00%	69,50%	65,50%	72,80%	81,10%	90,40%
	N	86,60%	92,00%	56,50%	51,80%	43,30%	78,50%	82,20%	89,90%
	S	87,60%	93,40%	82,90%	81,30%	76,40%	69,50%	80,10%	88,30%
G-Media		88,86%	92,25%	72,32%	68,21%	61,55%	71,12%	80,48%	87,29%

* Los 3 métodos seleccionaron los mismos atributos

Tabla IV: Métricas basadas en TfIdf

Crudo Pivote	Boscán	Métodos de Selección de Atributos	Término más Frecuente
# Atributos	401	# Instancias	1217

Métrica		Bayes Ingénuo (Estimador Núcleo)	SMV (SMO) (NormalizedPolyKernel) (BuildCalibrationModels)	KNN (K=3)	KNN (K=5)	KNN (K=7)	RIPPER (Jrip)	C4.5 (J48) (BinarySplit)	Random Forest
ICC		82,09%	91,13%	76,58%	70,50%	68,69%	74,28%	79,21%	88,58%
IIC		17,91%	8,87%	23,42%	29,50%	31,31%	25,72%	20,79%	11,42%
TVP	B	90,60%	90,10%	98,00%	94,40%	92,40%	64,30%	77,20%	82,70%
	N	72,10%	91,60%	53,40%	47,30%	45,70%	86,60%	81,40%	97,70%
	S	90,50%	91,40%	94,60%	85,70%	83,80%	63,20%	77,50%	78,70%
TFP	B	11,00%	3,40%	24,50%	30,60%	32,10%	6,70%	5,60%	0,90%
	N	6,80%	8,70%	1,80%	4,70%	4,60%	30,40%	19,00%	18,90%
	S	8,50%	2,30%	6,50%	6,70%	7,80%	6,00%	8,80%	0,80%
Precisión	B	76,40%	91,10%	61,00%	54,70%	52,90%	78,90%	84,30%	97,30%
	N	90,00%	90,00%	96,10%	89,50%	89,50%	70,80%	78,50%	81,50%
	S	78,70%	93,20%	83,50%	81,80%	79,00%	78,70%	75,50%	97,30%
Sensibilidad	B	90,60%	90,10%	98,00%	94,40%	92,40%	64,30%	77,20%	82,70%
	N	72,10%	91,60%	53,40%	47,30%	45,70%	86,60%	81,40%	97,70%
	S	90,50%	91,40%	94,60%	85,70%	83,80%	63,20%	77,50%	78,70%
Medida-F	B	82,90%	90,60%	75,20%	69,20%	67,30%	70,90%	80,60%	89,40%
	N	80,10%	90,80%	68,70%	61,90%	60,50%	77,90%	79,90%	88,90%
	S	84,20%	92,30%	88,70%	83,70%	81,40%	70,10%	76,50%	87,00%
G-Media		83,93%	91,03%	79,11%	72,60%	70,73%	70,60%	78,68%	85,99%

vez que las condiciones varían. En otras palabras, el modelo debe re-entrenarse en épocas de alta volatilidad de los precios porque existe un cambio en las condiciones que desestabiliza el mercado.

Referencias

[1] T. Rifai, *The Pricing of Crude Oil*. Economic and Strategic Guidelines for an International Energy Policy, Praeger Publishers, 1975.

[2] A. Briceño y J. Suárez, *Informe de Posicionamiento sobre las Actividades de la Oficina para la Determinación de los Precios de Crudo de Exportación (ODPCE) del Ministerio del Poder Popular de Petróleo*. Dirección Ejecutiva de Automatización, Informática y Telecomunicaciones, Petróleos de Venezuela, S.A. 2016, Sin publicar.

[3] C. Aggarwal, *Data Mining. The Text Book*, Springer, 2015.

[4] S. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Second Edition. Springer. 2015.

[5] H. Choi and H. Varian, *Predicting the Present with Google Trends*, Technical report, 2009.

[6] T. Preis, H. Moat, and H. Stanley, *Quantifying Trading Behavior in Financial Markets using Google Trends*, Scientific Report, 3, 1684, 2013.

[7] L. Jian, X. Zhenjing, Y. Lean, and T. Ling, *Forecasting Oil Price Trends with Sentiment of Online News Articles*, Procedia Computer Science, vol. 91, pp. 1081-1087, https://doi.org/10.1016/j.procs.2016.07.157, 2016.

[8] E. Mohammed, F. Andrea, B. Elisa, and A. Peter, *Using Four Different Online Media Sources to Forecast Crude Oil Price*, 2017.

[9] L. Xuerong, S. Wei, and W. Shouyang, *Text-based Crude Oil Price Forecasting: A Deep Learning Approach*, International Journal of Forecasting, vol. 35, no. 4, pp. 1548-1560, https://doi.org/10.1016/j.ijforecast.2018.07.006, 2019.

[10] Y. Lean, Z. Yaqing, T. Ling, and Y. Zebin, *Online Big Data-driven Oil Consumption Forecasting with Google Trends*, International Journal of Forecasting, vol. 35, no. 1, pp. 213-223, https://doi.org/10.1016/j.ijforecast.2017.11.005, 2019.

[11] J. R. Zanoni, *El Precio del Petróleo sus Determinantes y su Fijación por la OPEP*, Ediciones FaCES/UCV, 1981.

[12] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0, Step-by-Step Data Mining Guide*, 2000.

[13] P. Sethi and M. Jain, *A Comparative Feature Selection Approach for the Prediction of Healthcare Coverage*, ICISTM 2010, CCIS 54, pp. 392-403. Springer-Verlag Berlin Heidelberg, 2010.

[14] J. Silge and D. Robinson, *Text Mining with R. A Tidy Approach*, Publisher: O'Reilly Media, 2017.

[15] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, https://doi.org/10.1613/jair.953, 2002.

Clasificador de Noticias del Mercado Petrolero

Se recomienda mantener el valor del factor k para el crudo Boscán

NOTICIAS CLASIFICADAS

Show 5 entries Search:

MERCADO	CATEGORÍA	NOTICIA	Sentimiento	Prob_Baja	Prob_Neutro	Prob_Alza
CEGM	Refinación	Se espera que el mantenimiento de crudo REFQ4 sea relativamente ligero en el futuro y se mantendrá así durante los primeros meses del primer trimestre de 2016. Dada la reciente recuperación de margen esperamos que CEGM continúe aumentando las carreras hasta muy por encima del 90%	Baja	99.71	0.29	0.00
CEGM	Refinación	Los grados de Permian se mantuvieron con descuentos contra el WTI de Midland debido principalmente al intensivo mantenimiento de las refinerías en la zona central de EEUU. A pesar de las caídas en los números de taladros en los EEUU la producción sigue creciendo de acuerdo con los datos del gobierno de dicho país debido a que las áreas más productivas se han convertido en el foco de las operaciones. El número de taladros en Permian ha caído en cada una de las seis últimas semanas.	Neutro	11.14	88.85	0.01
CEGM	Crudo	Se espera que el crudo de WTS Midland regrese a una prima contra WTI Cushing una vez que las operaciones de reanudación del oleoducto y mantenimiento local disminuyan. Esto aumentará los precios de la fórmula venezolana.	Neutro	0.00	99.69	0.31
CEGM	Crudo	La comercialización de crudo LLS para diciembre comenzó la sesión con una prima de 85 centavos por barril sobre el WTI. El Mars para diciembre se comercializó con descuentos de $4/barril a $3.75/barril fortaleciéndose junto al LLS después de haber sido comercializado un día antes a -4.20/-4.10 $/b. Ambos crudos el LLS y el Mars estuvieron lo más fuertes que han estado desde el 27 de octubre durante la segunda sesión de los cargamentos de las primeras ventanas de diciembre.	Neutro	0.22	99.78	0.00
CEGM	Crudo	La prima del crudo de Alaska Slope Norte (ANS) con respecto al WTI CMA Nymex de diciembre se fortaleció en unos 13 centavos por barril después del fin de semana para reflejar el descuento de $2.03/barril contra el CMA ICE Brent de Diciembre establecido la última vez que fue comercializado este crudo para entrega en diciembre en la Costa Oeste de los EEUU el 27 de Octubre.	Neutro	0.21	99.79	0.00

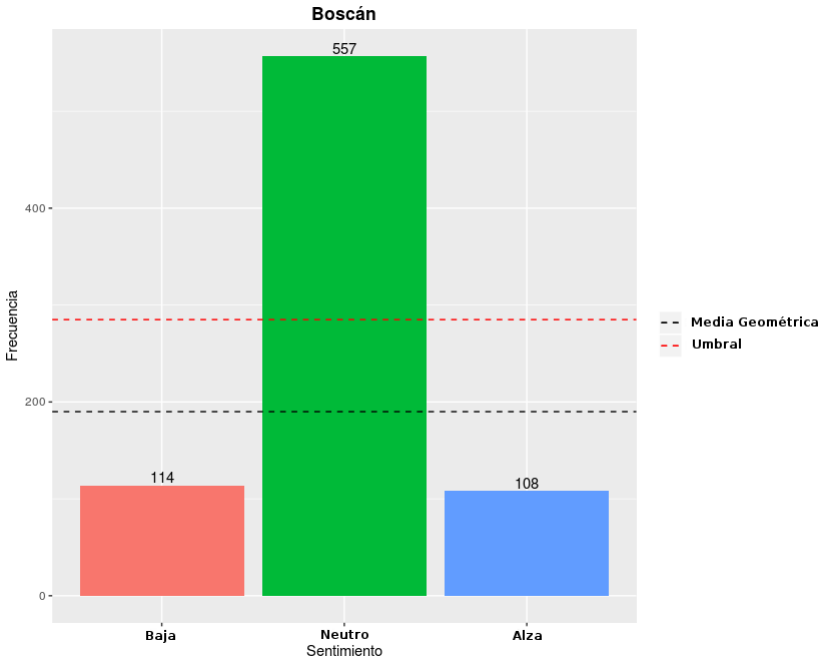

Figura 9: Sistema de Recomendación sobre el Sentimiento del Mercado Petrolero