3-8-2020

Juan Alberto Ortega Riveros David Eduardo Saavedra Hernandez

Universidad de los andes

Analítica de textos

Proyecto 1 de Inteligencia de Negocios

Contenido

[Comprensión del negocio y enfoque analítico 2](#_Toc34494554)

[Comprensión y preparación de los datos 2](#_Toc34494555)

[Perfilamiento de los datos 2](#_Toc34494556)

[Preparación de los datos 2](#_Toc34494557)

[Reading Data: 2](#_Toc34494558)

[Pre-processing 3](#_Toc34494559)

[Diccionarios de palabras 3](#_Toc34494560)

[Comprensión de los datos 4](#_Toc34494561)

[Modelado y evaluación 4](#_Toc34494562)

[Análisis de resultados 5](#_Toc34494563)

[Trabajo en equipo 5](#_Toc34494564)

[Retos enfrentados en el proyecto 5](#_Toc34494565)

[Organización de tareas 5](#_Toc34494566)

[Sustentación y evaluación del aporte individual 6](#_Toc34494567)

[Bibliografía 6](#_Toc34494568)

# Comprensión del negocio y enfoque analítico

|  |  |  |  |
| --- | --- | --- | --- |
| **Oportunidad/Problema Negocio** | | La oportunidad de negocio es la de automatizar la clasificación de reseñas a partir de los textos de reseñas. | |
| **Descripción del requerimiento desde el punto de vista de la minería de datos** | | Para esta oportunidad es necesario predecir si una reseña tiene sentimientos negativos o positivos hacia la película que se encuentra reseñando. Es por esto que es recomendable desarrollar una tarea de clasificación en la que se tomen parte de los datos para enseñar al algoritmo a identificar cuáles son positivas y cuales negativas para que luego de eso pueda predecir cuándo se agreguen nuevos datos. | |
| **Detalles de la actividad de minería de dato** | | | |
| **Tarea** | **Técnica** | | **Algorítmica y parámetros utilizados** |
| Clasificación | Lexicon based approach | | **Lexicon Based algorithm**: el algoritmo recibe 3 parámetros: los 2 diccionarios y las palabras limpias. Luego de esto hace match entre las palabras buenas y las malas. Entonces dada esta etiqueta asigna la clase con base en la mayor cantidad de palabras de cada sentimiento.  Positive and Negative List: Necesarios para etiquetar las palabras y con base en estos llevar a cabo la tarea de clasificación. |

# Comprensión y preparación de los datos

## Perfilamiento de los datos

El conjunto de datos en este caso está conformado de la siguiente manera: una columna con la reseña y otra columna con las clases que dice si el sentimiento es negativo o positivo. Son 201 filas todas con su correspondiente sentimiento. Sin embargo, el perfilamiento de los datos para este punto sin un pre-procesamiento sería muy poco diciente por lo que se decidió realizar dicho perfilamiento después de la preparación de los datos. En la sección de Comprensión de los datos

## Preparación de los datos

Para la preparación de los datos se utilizaron los siguientes componentes:

### Reading Data:

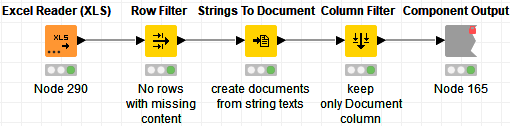


Ilustración 1

En este componente lo que hacemos es filtrar las tablas que no tengan la columna sentimientos. Luego lo que hacemos es crear documentos y dejamos solo ese documento para luego seguir al componente Pre-processing .

### Pre-processing

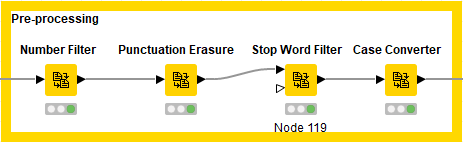


Ilustración 2

En este nodo quitamos los términos que son solo números y las “stop words” que no aportan al análisis de texto. Finalmente dejamos todas las palabras en minúsculas para estandarizar los datos.

### Diccionarios de palabras

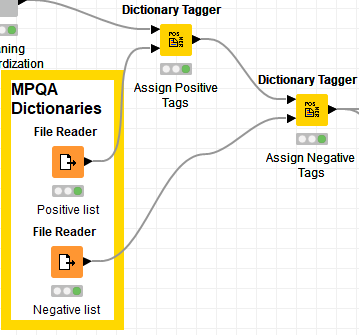


Ilustración 3

Ahora bien, lo que hacemos es etiquetar las palabras que nos encontramos en los documentos con diccionarios públicos de palabras positivas y negativas. Esto quiere decir que si una palabra se encuentra en algunos de los diccionarios se etiquetará en el documento como negativo o positivo dependiendo de en qué diccionario se encuentra y se verá como en la Ilustración 4 a través del nodo “Document Viewer”.

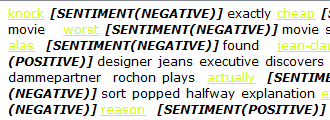


Ilustración 4

## Comprensión de los datos

Después de nuestro pre-procesamiento pudimos sacar algunos datos después de filtrar las palabras que no se les hubiera asignado sentimiento positivo o negativo dados los diccionarios públicos de Knime. Con eso pudimos sacar los siguientes gráficos:

Ilustración 5

En la Ilustración 5 podemos ver la participación de los principales 6 sentimientos que encontramos según los diccionarios y las veces que aparecieron en los textos.

Además de esto en la Ilustración 6 podemos ver la cantidad de etiquetas de sentimientos negativos y positivos que hay en todos los datos según los diccionarios de sentimientos. Con esta gráfica podemos ver un poco más sobre que hay más cantidad de palabras asociadas a sentimientos positivos que asociadas a sentimientos negativos.

Ilustración 6

La información anterior nos ayuda a tener una idea general de los datos. Sin embargo, con base en ella es difícil de tomar decisiones que nos ayuden a mejorar el algoritmo.

# Modelado y evaluación

Luego de la preparación de los datos nos encontramos con el WorkFlow consignado en la Ilustración 7. En este vemos el proceso completo después del procesamiento para poder llegar a un modelo de clasificación para futuros datos:

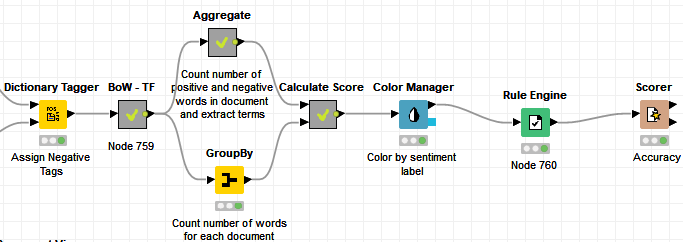


Ilustración 7: WorkFlow Completo

# Análisis de resultados

Los resultados obtenidos se muestran en la Ilustración 8: Resultados. Donde podemos ver una efectividad de un 64,5%.

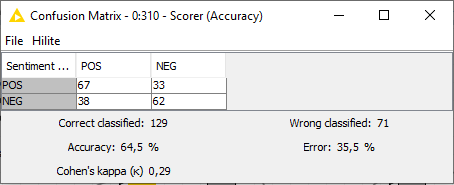


Ilustración 8: Resultados

En este caso al tratarse de esta tarea de clasificación no podemos mostrar ejemplos prácticos para el negocio. Sin embargo, podemos afirmar que los nuevos datos serán clasificados correctamente con un 64,5% de efectividad. Ahora bien, según Ogneva en la clasificación de sentimientos los humanos solo están de acuerdo el 79% de las veces (Ogneva,2020). Esto quiere decir que en términos de negocio un algoritmo que acierte el 64,5% de las veces puede ser beneficioso dado que se trata de algo automático. De esta forma podemos estar satisfechos con los resultados obtenidos en nuestro proyecto de minería de datos.

Además de lo anterior encontramos las siguientes tablas como resultado de nuestro trabajo:

|  |  |
| --- | --- |
| row ID | count |
| plot[NEGATIVE(SENTIMENT)] | 98 |
| little[NEGATIVE(SENTIMENT)] | 91 |
| bad[NEGATIVE(SENTIMENT)] | 79 |
| comedy[NEGATIVE(SENTIMENT)] | 50 |
| hard[NEGATIVE(SENTIMENT)] | 45 |

|  |  |
| --- | --- |
| row ID | count |
| actually[POSITIVE(SENTIMENT)] | 62 |
| love[POSITIVE(SENTIMENT)] | 61 |
| real[POSITIVE(SENTIMENT)] | 59 |
| star[POSITIVE(SENTIMENT)] | 49 |
| hollywood[POSITIVE(SENTIMENT)] | 48 |

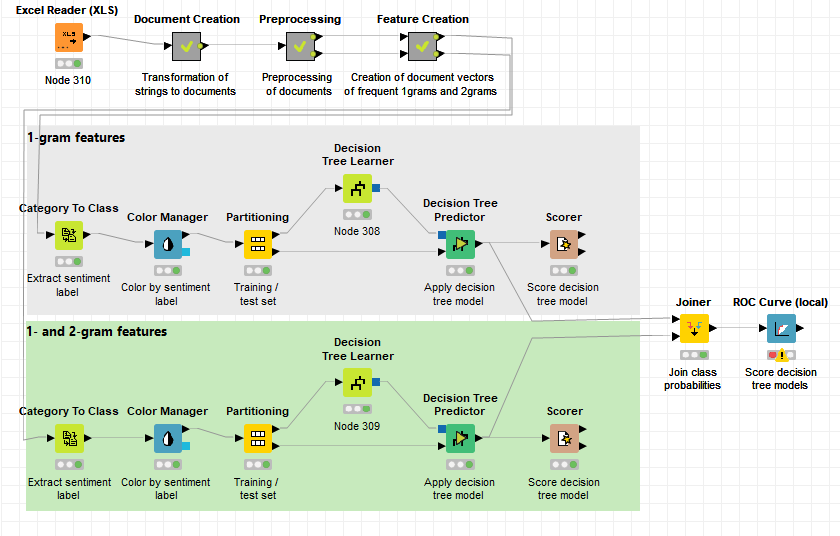
Las anteriores palabras son las palabras más relevantes para el negocio debido a que son con las que se tiene más certeza a la hora de clasificar las reseñas.

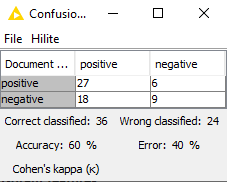
# Trabajo en equipo

## Retos enfrentados en el proyecto

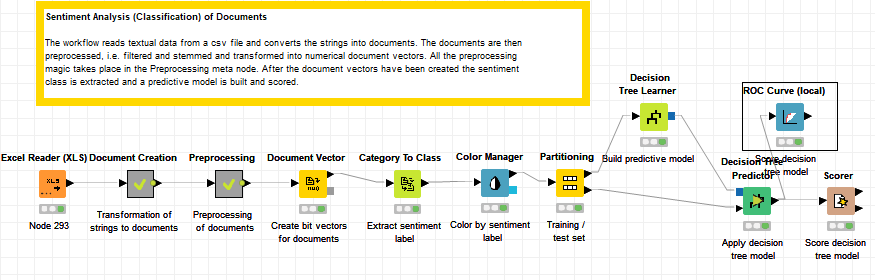
El principal reto se basó en encontrar la forma de pre-procesar los datos para de esta manera utilizar alguna técnica para clasificarlos de acuerdo con su sentimiento. Para solucionar este problema se intentó buscar en Google ejemplos de workflows de análisis de sentimientos en textos con la herramienta Knime. Por suerte todos estos ejemplos ya tenían incluido el preprocesamiento de los textos, incluyendo la creación de la bolsa de palabras y la tokenización de ellas. Luego de ello, se intentó ajustar los ejemplos con los datos proporcionados para el proyecto, no teniendo buenos resultados con dos de los tres ejemplos utilizados.

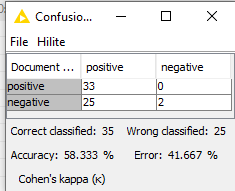
Uno de ellos fue con la creación de n-gramas y árboles de decisión, algoritmo que prometía tener un buen desempeño en la tarea de clasificación al crear conjuntos de n palabras, como por ejemplo “not bad” o “good movie”. Sin embargo, aún modificando parámetros para mejorar su eficacia, solo se logró un 60%.





Otro ejemplo utilizado fue uno en el que se basaba en árboles de decisión, pero solo que palabras individuales. Sin embargo, no pudimos aumentar su eficiencia a un número mayor a 60%.





## Organización de tareas

En la siguiente tabla se encontrará la dedicación en horas de cada una de las tareas que fueron necesarias para llevar a cabo el proyecto.

|  |  |  |
| --- | --- | --- |
| Tareas | Juan Ortega | David Saavedra |
| Diseño del WorkFlow | 3 horas | 1 hora |
| Perfilamiento de datos | 1 hora | 2 horas |
| Informe | 1 hora | 2 horas |
| Tutorial |  | 0.5 horas |

## Sustentación y evaluación del aporte individual

En la siguiente parte se repartirá 100 puntos entre los integrantes del grupo de acuerdo a lo desarrollado por cada uno y una breve justificación. Ahora bien, en los siguientes se destacan las actividades que cada uno realizó de manera significativa, sin embargo, ambos estuvimos presentes en todas las actividades solo que no con la misma dedicación.

Juan Alberto Ortega Riveros (50 puntos): Fue el principal promotor del proyecto que se encargó de gran parte de la investigación sobre la minería de textos. Además de esto, fue quien encontró el algoritmo actual que nos llevó a la efectividad de casi el 65%.

David Eduardo Saavedra Hernandez (50 puntos): Fue el encargado de la documentación completa del proyecto y del perfilamiento de los datos.

# Bibliografía

* Ogneva, M. (2020). How Companies Can Use Sentiment Analysis to Improve Their Business. Retrieved 7 March 2020, from <https://mashable.com/2010/04/19/sentiment-analysis/>