Trabajo Fin de Máster

Aplicación de Técnicas de Minería de Textos en Inteligencia de Clientes

Jomayra Ramírez Figuereo

Máster en Técnicas Estadísticas

Curso 2018-2019

Resumen del trabajo final de máster

La minería de textos es un tema que en los últimos años ha interesado a un gran número de empresas de distintos sectores, debido a la creciente cantidad de datos textuales que se producen. Toda esta información se presenta como un problema a la hora de preguntarse cómo aprovechar toda esta información para determinar qué deben mejorar, eliminar o mantener, por lo que toda esta información se ha convertido en un recurso indispensable en la toma de decisiones y la definición de estrategias de *marketing*.

La minería de textos es el proceso de analizar colecciones de documentos de texto no estructurado[1], con el objetivo de capturar los temas y conceptos clave, descubriendo relaciones ocultas y tendencias existentes entre los textos sin necesidad de conocer las palabras exactas que los autores han utilizado para expresar dichos conceptos, realizado mediante el uso de un conjunto de herramientas de análisis.

Las técnicas de minería de textos han ido evolucionando hasta convertirse en un campo con gran peso dentro de la minería de datos, entendiéndose como minería de datos la extracción de información de patrones de grandes bases de datos. En Godoy-Viera (2015) analizan ambos conceptos y establecen diferencias según autores, por ejemplo, para Ronen Feldman y James Sanger en la minería de datos estos se guardan en formatos estructurados, y gran parte de su preprocesamiento se centra en la depuración y normalización de los datos, así como en crear un gran número de uniones de tablas. Mientras, en la minería de texto, el preprocesamiento se enfoca en reconocer y extraer características representativas para documentos en lenguaje natural, tales características pueden ser palabras clave relevantes, identificación de nombres de personas, organizaciones, etc.

El amplio abanico de aplicaciones en las que se pueden emplear la minería de textos ha provocado un gran interés por parte de la comunidad científica, por lo que existen muchos trabajos desarrollados en este campo. Debido al uso masivo e intensivo de los medios informáticos se disponen de las herramientas, métodos y algoritmos necesarios para analizar un gran volumen de textos y extraer información de utilidad. El desarrollo de métodos efectivos para lograr mejoras en estas tareas, continúa siendo un tema abierto de investigación, no obstante, el problema es complejo y es necesaria la continua investigación en la búsqueda de métodos y representaciones apropiadas.

El carácter interdisciplinario de la minería de textos se produce por incorporar el conjunto de técnicas de diferentes campos científicos, como la minería de datos, la lingüística, la estadística computacional y la informática.

[1] Por no estructurado se refiere a texto libre, generalmente en lenguaje natural. Desde el punto de vista del análisis cuantitativo de datos, el lenguaje natural que usamos los humanos para comunicarnos a menudo es incluido dentro de la categoría de “datos no estructurados”. Por ello, la mayoría de las subtareas incluidas en el procesamiento del lenguaje natural (PLN) incluyen la conversión de los “datos no estructurados” en “datos estructurados”.

Entre sus principales aplicaciones destaca:

- Extracción de información importante lo más rápidamente posible en grandes volúmenes de datos. Identificando hechos, datos puntuales y la relación existente entre ellos.
- Agrupación de documentos entre los que existe cierta similitud (*clustering*).
- Clasificación automática, con la cual se pretende asignar un documento a una clase o tema definido con anterioridad.
- Identificación de conceptos tratados en los documentos y creación de redes de conceptos. Esta función permitiría extraer los principales temas o ideas tratados en los documentos, extrayendo un conjunto de términos que son representativos del contenido de los documentos.
- Análisis de sentimientos o minería de opiniones. Se aplica para extraer y analizar opiniones sobre diversas marcas y productos, jugando un papel muy importante en la toma de decisiones de las empresas.

En este proyecto se presenta un análisis de técnicas de minería de textos a mensajes en idioma castellano, siendo la segunda lengua más hablada del mundo, con 442 millones de hablantes, sólo superada por el idioma chino y sus variantes. Es el quinto en extensión (31 países), por detrás del chino, el inglés, el árabe y el francés[2]. Existe una gran variedad de textos en castellano almacenado electrónicamente, lo que pone de manifiesto la necesidad de su tratamiento, considerando que la gran mayoría de las herramientas y ejemplos disponibles están en el idioma inglés.

Los datos a analizar tratan sobre encuestas realizadas a clientes. Cada encuesta está enfocada a conocer la satisfacción por medio de preguntas donde el cliente da puntuaciones y opiniones sobre el banco, sus productos o servicios, por ejemplo, encuestas enfocadas a clientes con: tarjetas de crédito, banca electrónica, banca móvil, entre otras. Estas encuestas se realizan por medio de correo electrónico, banca móvil y teléfono.

La empresa no permite la publicación de este proyecto por la confidencialidad de la información obtenida, ya que los datos tratan sobre opiniones de sus clientes y las técnicas utilizadas para tratarlos muestran informaciones que consideran sensibles.

1.1 Objetivos

El objetivo general es la implementación de técnicas de minería de textos para extraer información de las opiniones de los clientes sobre la entidad bancaria y sus productos. Identificando quejas y preferencias para la toma de decisiones y la implementación de mejoras.

[2] Puede consultarse en Ethnologue https://www.ethnologue.com/

Se tienen los siguientes objetivos específicos que permiten lograr el objetivo general:

- Estudiar las herramientas en R para el manejo de la minería de textos en castellano.
- Estudiar y comparar distintas ponderaciones de los términos para la representación de la colección de documentos.
- Aplicar un análisis de la fuerza del sentimiento a las opiniones de los clientes.
- Aplicar y comparar algoritmos de aprendizaje estadístico, que nos permita clasificar las opiniones.

Para llevar a cabo esta tarea, se presenta un resumen de lo realizado en cada uno de los capítulos

Capítulo 2: Descripción de los datos

Se explican los detalles sobre la recopilación de la información y se presenta un análisis exploratorio de las variables estructuradas.

A continuación, se describen las variables estudiadas:

- *encuesta_id*: identificador de la encuesta.
- *fecha*: fecha y hora final de la encuesta. Abarca desde noviembre 2017 hasta junio 2018.
- *cliente*: identificador de cada cliente.
- *tipo_encuesta*: los tipos de encuesta a estudiar, por ejemplo:
  - Banca Electrónica Particulares: encuesta para conocer la opinión y satisfacción del cliente que tiene Banca Electrónica.
  - Banca Móvil: encuesta para conocer la satisfacción del cliente con Banca móvil.
  - Encuesta Comercial: encuesta para conocer la opinión de los clientes que se le realiza alguna oferta comercial.

Entre las preguntas comunes por tipo de encuesta encontramos las siguientes:

- *puntuacion*: valoración del cliente sobre la recomendación que hace de la entidad (en un rango de 0-10).

  Se le pregunta al cliente ¿Recomendarías la entidad bancaria a algún conocido, familiar o amigo? Entonces,

- Si R =<6 Abre la siguiente pregunta ¿Qué podríamos hacer para mejorar tu recomendación? Esta respuesta corresponde con la variable *mejora_recomienda*

- Si R >6 ¿Cuál es el principal aspecto por el que nos recomiendas? Esta respuesta corresponde con la variable *motivo_recomienda*

- *insatisfaccion:* Esta variable responde a la pregunta ¿cuál es tu motivo de insatisfacción?, que se origina cuando el cliente otorga una puntuación igual o menor a 6 al pedirle que puntúe (en un rango de 0-10) la satisfacción que tiene con el producto o servicio ofrecido (que corresponde con el tipo de encuesta).

Capítulo 3: Análisis exploratorio de textos

El procesamiento de textos suele ser complejo. Su falta de estructura, su contenido y naturaleza heterogénea presentes en enormes bases de datos, hace que se requiera de técnicas que los trate convenientemente.

Con el objetivo de exponer el proceso de manipulación de textos y aplicar un análisis exploratorio, se realiza un ejemplo con la variable mejora_recomienda. Para ello, se utiliza principalmente el paquete tm ya que ofrece la cantidad de funcionalidades necesarias para gestionar y manipular documentos de texto en R

Las técnicas aplicadas son las siguientes:

- Creación del Corpus.

- Preprocesamiento y limpieza de textos:

    - Conversión a minúsculas.
    - Eliminamos las puntuaciones.
    - Eliminando números.
    - Removemos tildes.
    - Se eliminan los stopwords.
    - Se eliminan espacios vacíos.
    - Stemming.

- Creación de Matriz de Documentos-Término.

  Una de las características principales de la minería de textos es transformar un conjunto de datos textuales en un conjunto de datos estructurados, colocándolos posteriormente en una base de datos tradicional para su posterior análisis.

En esta sección se presenta:

  - Esquema de ponderación TF-IDF e Importancia de términos.
  - Técnicas de reducción de la dimensión de la matriz de documentos-términos.

- Palabras frecuentes y métodos gráficos.
- Correlación entre los términos.
- Tokenización por bi-gramas.

Capítulo 4: Clasificación de sentimientos no supervisado-enfoque basado en recursos léxicos

En este capítulo se presenta una aproximación no supervisada basada en recursos léxicos, para extraer la fuerza o grado del sentimiento a la variable *motivo de insatisfacción. Además,* se analiza por grupos de encuestas ya que las respuestas de esta variable dependen de la variable *tipo de encuesta*.

El método aplicado trabaja a nivel de términos, utilizando un diccionario léxico afectivo con un listado de palabras clasificadas en positivas y negativas. Además, se trata la negación, los intensificadores[3] como problemáticas que se presentan alrededor de la clasificación de la polaridad de opiniones.

Para el análisis de sentimientos se utiliza la función polarity() del paquete qdap. Esta función estima el sentimiento de una cadena de texto, tomando en cuenta la relación que hay entre una palabra y las que están a su alrededor en un intervalo especificado. Las palabras que componen la cadena de texto se analizan por medio de diccionarios: palabras positivas y negativas y términos negadores e intensificadores.

Para llevar a cabo esta tarea se introduce el problema a abordar con las dificultades que conlleva, y se presenta un resumen de los trabajos más relevantes que podemos encontrar en la literatura. Seguido, se presenta la metodología y las herramientas a utilizar para su implementación en R. Finalmente se analizan los resultados obtenidos mediante un análisis exploratorio.

Capítulo 5: Clasificación supervisada

En este capítulo se aplican técnicas de clasificación supervisada utilizando las variables *mejora recomienda* y *motivo recomienda*, para determinar si la opinión proviene de un cliente que da una puntuación igual o mayor que 7 al recomendar la entidad. Se obtiene un resultado

[3] En el contexto de este trabajo englobará bajo el término "intensificadores" tanto aquellas palabras que aumentan la intensidad del significado como las que lo disminuyen.

binario, representado como “0” si la puntuación es igual o mayor que 7 y “1” si el cliente otorgó una puntuación menor que 7.

Incluimos en el modelo una variable independiente con los *score* de polaridad de las opiniones, calculados con la función polarity(), utilizada en el capítulo 4. Al final del capítulo se comparan los resultados obtenidos con este modelo (agregando esta variable numérica) y con un modelo creado solo con variables de textos.

Los pasos realizados son los siguientes:

- Procesamiento del conjunto de datos y extracción de las características.
- Se crea conjuntos de entrenamiento y prueba: se define un conjunto de entrenamiento del 80% de los datos para enseñar al computador como clasificar y el 20% para evaluación.
- Se explican las distintas métricas para analizar y comparar los algoritmos.
- Aplicación de técnicas de clasificación automática de textos, entre estas:
    - Regresión logística.
    - Árbol de decisión.
    - Bosques aleatorios o random forest.
    - Algoritmo K-vecinos más cercanos.
    - Support Vector Machines (SVM).
    - Algoritmo Naive Bayes.
- Comparación de los algoritmos de clasificación.

Se termina el proyecto con una exposición de las conclusiones obtenidas y las mejoras que se podrían tomar en cuenta para trabajos futuros.