MARCO DE REFERENCIA

ANALÍTICA DE TEXTO

Juan Pablo González | Tópicos especiales en Telemática | 20 de noviembre

Uno de los enfoques más importantes referentes al Big Data tiene que ver con el procesamiento y el uso que se les dan a los datos. En esta ocasión contamos con una base de datos de artículos o noticias, con estos datos pudimos experimentar técnicas de limpieza de datos como son la eliminación de stop-words y el lemming. Sin embargo, una vez tenemos estos datos ¿Qué podemos hacer con ellos?

El crecimiento de los datos es realmente algo del día a día, en cada segundo son creados y procesados millones de datos alrededor del mundo. Sin embargo, no todos los datos creados son relevantes ni es justificable analizarlos. Es interesante en este caso ver las diferentes técnicas de analítica de texto existentes y enfocarnos en una de estas para obtener información relevante de los artículos que tenemos en nuestro dataset.

Si hablamos de analítica de texto, se tienen en cuenta un montón de conceptos teóricos relacionados con procesamiento de lenguaje natural, sintaxis, gramática, semántica, lenguajes, etc. El propósito que tenemos con este proyecto, más que entender estos conceptos es apropiarnos de una técnica de analítica y ponerla en práctica.

Refiriéndonos a una parte de la analítica que sí nos interesa, la minería de texto es de gran importancia en nuestro proyecto. Con este término hablamos de técnicas que buscan inferir información estructurada de calidad tomada de un gran volumen de texto sin estructura (Wachsmuth, 2015). Esta tarea la facilitamos con la limpieza de los datos anteriormente mencionada.

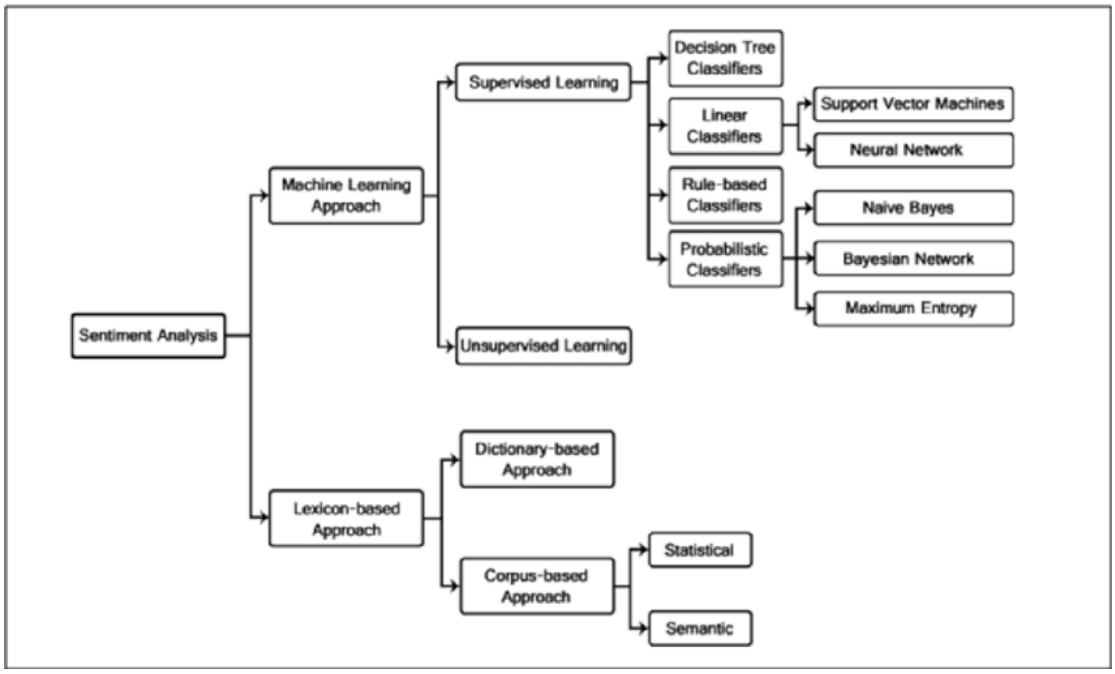
Teniendo en cuenta el termino de minería de texto, ahora vamos a referirnos a un término similar: Minería de opinión, o mejor conocido como Análisis de Sentimiento. Cuando hablamos de lenguaje natural nos referimos por obvias razones a todo aquello que un humano dice o escribe. Y es claro que cualquier cosa que escribimos o decimos tiene un significado, un propósito, tiene un medio y en la mayoría de los casos tiene un destino. Pero también podemos determinar si lo que decimos es positivo, negativo, o es neutro.

Las redes sociales son lugares donde la gente suele expresar opiniones, sentimientos, mensajes a otras personas y más. Usualmente la gente comparte su forma de pensar en redes como Twitter de una forma rápida, donde pueden ser atendidos por millones de personas y lograr influencia de esta forma fácilmente en cuestión de segundos. Se afirma que por medio del análisis de los mensajes enviados en Twitter sobre cualquier tema se podría obtener la misma información que se consigue por medio de una encuesta de opinión, de una forma más inmediata y menos costosa (Sande, 2018).

Podemos definir a la técnica de Análisis de sentimiento como es el proceso de determinar el tono emocional que hay detrás de una serie de palabras. Con el propósito de entender las actitudes, opiniones y emociones expresadas en una mención (Bannister, 2015). Ahora volviendo a nuestro proyecto, podemos inferir por medio de esta técnica si una noticia es positiva (buena) o negativa (mala).

Para realizar un Análisis de sentimiento se tienen varias alternativas o técnicas. Teóricamente se reúnen en dos tipos: Por medio de modelos de Machine Learning o basándonos en características léxicas. (Baviera, 2017)

Para la alternativa del Machine Learning se tienen un montón de modelos que pueden ser entrenados, desde simples arboles de decisión hasta el uso de probabilidades por medio del teorema de Bayes. Es claro que para apropiarnos de estas alternativas es necesario tener un buen volumen de datos que estén ya clasificados y con los cuales se pueda entrenar el modelo, para posteriormente ingresar datos nuevos que el mismo modelo pueda clasificar. Además de esto es necesario tener suficiente poder de computo para crear el modelo y entrenarlo.



(Baviera, 2017)

A pesar de que la alternativa de Machine Learning podría ser mucho más precisa, también se logra una buena precisión por medio del análisis léxico de las palabras y oraciones. Estas técnicas básicamente consisten en relacionar palabras con sentimientos, por ejemplo: Cuando en un Tweet se dice algo como “Amo visitar Medellín” se pueden analizar palabras claves como “Amo” y relacionar directamente este Tweet como positivo, algo que para hacerlo por medio de Machine Learning se podría hacer entrenando un modelo para que logre interpretar “Amo” como algo positivo.

Muchas veces la importancia de los datos viene de la capacidad de inferir información importante de estos. Es fundamental comprender entonces el impacto que puede tener un análisis de sentimientos, por ejemplo, en las opiniones encontradas acerca de un producto que se compra en un supermercado, un candidato político, un evento o lugar turístico, etc.

Bibliografía

Bannister, K. (2015). *Entendiendo el análisis de sentimiento: qué es y para qué se usa*. Obtenido de Brandwatch: https://www.brandwatch.com/es/blog/analisis-de-sentimiento/

Baviera, T. (2017). *Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength.* Universitat Politècnica de València.

Sande, J. S. (2018). *Análisis de Sentimientos en Twitter.* Universidad Abierta de Cataluña.

Wachsmuth, H. (2015). *Text Analysis Pipelines: Towards Ad-hoc Large-Scale Text Mining.* Springer.