
**FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES**

# APLICACIONES DEL TEXT MINING EN LA ACTUALIDAD CENTRADO EN EL ÁREA DE MARKETING

**Nombre: Álvaro López de Miguel**
**Director: Carlos Martínez de Ibarreta Zorita**

MADRID | Junio 2019

**RESUMEN**

La recolección de información siempre ha sido un pilar fundamental de cualquier empresa para tener éxito. Es necesario entender a nuestros consumidores para poder tener éxito a la hora de comercializar un producto o servicio.

Hoy en día se mueve mas información que nunca, aproximadamente 2,5 billones de bytes cada día (InfoChannel, 2018) y este número va en aumento hasta el punto de que se prevé que para el año 2020, cada persona creará 1,7MB de datos por segundo. No solo eso, sino que se prevé que el peso aproximado de los datos en todo el internet será de 40 zettabytes[1] en 2020 (Allahyari et al., 2017); un aumento muy drástico si tenemos en cuenta que era ´solo' 8 zettabytes en 2015. La mayoría de estos datos vienen en formato de vídeo debido al gran éxito de plataformas como YouTube y Netflix. Las fotos son otro formato de obtención de datos que está en constante crecimiento gracias a aplicaciones como Instagram y Snapchat. Aunque en este trabajo nos vamos a centrar en el formato tradicional (textos) que tampoco se queda atrás con respecto a los formatos mencionados anteriormente. Esto se puede observar fácilmente en la gran cantidad correos electrónicos y de artículos publicados cada día y fácilmente accesibles en virtud de buscadores como Google; Twitter es otra gran fuente de información que muchas empresas usan para monitorear a sus consumidores o público objetivo.

Tanta información se convirtió en un problema mas que una bendición para muchas empresas que no poseían los recursos ni los medios para recopilar estos datos, organizarlos y darles algún uso. Es por esto por lo que nació lo que hoy conocemos como *Data Mining.*

Palabras Clave: *text mining*, *data mining*, tecnología, marketing, corpus*, Asignación de Dirichlet Latente, Análisis de Semántica Latente*

[1] 1 ZB = 1 billón de terabytes = 10^21 bytes

**ABSTRACT**

The gathering of information has always been a fundamental pillar of any company in order to succeed. It is necessary to understand our consumers in order to be successful when marketing a product or service.

Nowadays, more information is being moved than ever, approximately 2.5 trillion bytes every day (InfoChannel, 2018) and this number is increasing to the point that it is expected that by the year 2020, each person will create 1.7MB of data per second. Not only that, but it is expected that the approximate weight of data across the internet will be 40 zettabytes by 2020 (Allahyari et al., 2017); a very drastic increase if we consider that it was only '8 zettabytes in 2015. Most of this data comes in video format due to the great success of platforms such as YouTube and Netflix. Photos are another format for obtaining data that is constantly growing thanks to applications like Instagram and Snapchat. Although throughout this research paper we will be focusing on the traditional format (texts) which is not far behind compared to the previously mentioned formats. And this can be easily seen in the large number of emails and articles published every day and easily accessible by search engines such as Google; Twitter is another great source of information that many companies use to monitor their consumers or target audience.

So much information almost became a problem rather than a blessing for many companies that did not have the resources or the means to collect this data, organize it and give it some use. This is why Data Mining was born.

Key words: text mining, data mining, technology, marketing, corpus, Latent Dirichlet Allocation, Latent Semantic Analysis

# ÍNDICE

# 1. Introducción

## 1.1. Objetivos

Mediante este trabajo pretendo cumplir los siguientes objetivos:

- Abordar el *text mining* como técnica de extracción de información y conocimiento, centrándonos en su uso en el área de marketing.

- Lograr un entendimiento de qué es el *text mining* y sus comienzos como herramienta sin profundizar en el aspecto técnico y matemático del asunto.

- Estudiar todas las diferentes aplicaciones que tiene esta herramienta tan útil en el mundo de la empresa.

- Analizar las principales técnicas de *text mining* así como su implementación en la actualidad.

- Explicar los softwares mas populares para llevar a cabo la minería textual y como lo es implementado por distintas empresas o usuarios.

## 1.2. Metodología

Para lograr cumplir los objetivos expuestos en el apartado anterior, haré uso de numerosas fuentes de información. La mayoría consisten en trabajos de investigación publicados por expertos en la materia. Otros consisten en trabajos de fin de grado o máster que proporcionan información valiosa sobre *text mining* o sus diversas técnicas.

Y por último haré uso de páginas web de organizaciones fiables que facilitan conocimientos acerca de la minería textual y sus diferentes usos.

## 1.3. Partes del TFG

Este trabajo se puede dividir en tres partes fundamentales. En la primera se da a conocer *text mining* en general, desde la historia detrás de la creación de esta herramienta hasta sus distintas etapas.

La segunda parte es donde nos indagamos más profundamente en la minería de textos analizando las distintas técnicas que existen dentro de esta herramienta. Se valorarán las ventajas y desventajas de estas técnicas, así como su funcionamiento.

Por último, se analizarán las posibles aplicaciones de la minería textual proporcionando también ejemplos para demostrar el porqué de su uso.

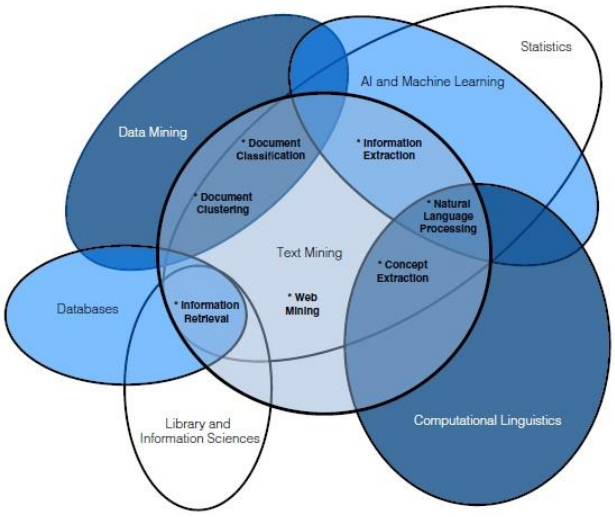
## 2. Text Mining

### 2.1. ¿Qué es text mining?

Existen muchas definiciones diferentes para referirse a la minería de textos. Una de estas posibilidades es la de que *text mining* se puede definir como el descubrimiento de textos o la exploración de textos en busca de información valiosa pero oculta (Kroeze, Matthee, Bothma, 2007).

Otra definición parecida, pero, en mi opinión, mas acorde con el ángulo empresarial y de marketing con el que voy a desarrollar este trabajo, es la definición aportada por Deloitte: la minería de textos es la técnica que consiste en usar tecnología avanzada para recopilar, almacenar y extraer información textual en busca de esas señales ocultas que pueden ayudarnos a realizar decisiones mas inteligentes en nuestro negocio o empresa (Delloite, 2013).

Esta práctica, comúnmente utilizada hoy en día, trata de encontrar tendencias en millones de documentos escritos que nos permiten recopilar y organizar conocimientos nuevos con el objetivo de lograr un incremento de beneficios, clientes y entendimiento de estos. La minería de textos es un procedimiento que interactúa con diferentes disciplinas de su entorno como ya hemos mencionado brevemente durante la introducción de este proyecto y continuaremos haciendo a lo largo de este trabajo. Esta interacción se puede observar claramente a través de el siguiente diagrama de Venn.




Fuente: Talib, R., Hanif, M. K., Ayesha, S., Fatima, F. (2016)

### 2.2. Data mining vs Text mining

La diferencia entre el *text mining* y *data mining* es los datos que son recopilados y analizados. La minería de datos se centra en cualquier tipo de información, desde los millones de emails enviados diariamente hasta vídeos o imágenes en redes sociales. Por otro lado, la minería de textos es la aplicación de la lingüística computacional y del procesamiento de textos para la extracción de conocimiento nuevo (Ana Isabel Valero Moreno, 2017). Según Ana Isabel Valero, "relacionada con la minería de datos podríamos decir que la minería textual es su `hermana pequeña´".

Cuando hablamos de *data mining*, la información que obtenemos de entrada se intuye, pero no está expresada claramente. Es desconocida, está oculta y es difícil de extraer sin recurrir a procesos automáticos o computacionales. Por otro lado, los textos que analizamos mediante herramientas de *text mining* no están ocultadas de la misma manera. Una persona puede encontrar estos textos y comprenderlos sin uso de ninguna tecnología, pero, la tecnología es necesaria para encontrar la información exacta que buscamos, dado que un humano no es capaz de leer y procesar la gran cantidad de documentos que se encuentran a nuestro alcance (Witten). La minería de textos tiene como objetivo extraer la información que se desea analizar sin un humano como intermediario.

Otra similitud que encontramos cuando relacionamos estas dos prácticas es que la información debe ser accionable. Esto significa que los resultados o conocimiento que obtengamos tras la aplicación de *text* o *data mining* tenga el potencial de ser útil para nuestro fin u objetivo. Respecto a *data mining*, esto es relativamente sencillo ya que aquellos modelos accionables serán los que nos permiten extraer información o predicciones importantes de la misma fuente. Para muchas técnicas de *text mining*, es muy difícil determinar que información es potencialmente útil ya que depende más del individuo que está realizando el estudio. Es por ello por lo que, cuando nos referimos a información accionable dentro de la minería de textos, nos centramos más en si esta información es comprensible en el sentido de que nos ayuda a entender los datos que se nos presentan. Esto es más necesario cuando la información extraída es para uso humano, aunque también es importante saber que "esta característica es menos aplicable a *text mining* ya que, a diferencia de la minería de datos, los datos de entrada en sí son comprensible" (Witten).

### 2.3. Historia y evolución del *text mining*

Ya en los años 80 se observaron los primeros esfuerzos de lo que hoy conocemos como minería de textos, aunque en ese momento se necesitaba de un gran esfuerzo humano ya que los avances tecnológicos todavía no se habían aplicado en este campo. Con el paso de los años, hemos visto como esta técnica ha avanzado hasta el punto en el que existe poca presencia humana dentro de este proceso. Aparte de esto, cada vez

es más importante la minería de textos multilingüe, técnica que se centra en obtener información en otros idiomas.

Fue en 1977 cuando "el sistema THOMAS, ilustró cómo las palabras o las frases clave podían utilizarse para guiar a los usuarios en el descubrimiento de documentos de referencia útiles. Las frases clave son un tipo especialmente útil de información abreviada. Sin embargo, tales frases se eligen con frecuencia manualmente, bien por los autores o por indizadores profesionales. Condensan documentos en unas pocas palabras y frases, ofreciendo una descripción breve y precisa de los contenidos de un documento" (Minería de textos, 2010).

Se produjo una especial atención hacia la minería de datos y, por consecuencia, a la minería de textos, tras los atentados del 11 de septiembre de 2001 en Nueva York (Valero Moreno, 2017). Esto fue debido a que el ejército y gobierno americano usaron técnicas de minería de datos en su lucha contra al terrorismo; lucha que se publicitó mucho por los distintos medios de información.

### 2.4. Etapas de la minería de textos

La mayoría de los errores que se producen dentro de la minería de textos no viene del tipo de software usado sino del pobre entendimiento del problema con el que nos encontramos (*Data Mining for Business Analytics*, 2018). Es por ello por lo que existen una serie de pasos que debemos cumplir detenidamente antes de plantearnos ningún tipo de algoritmo. Estos pasos son los siguientes:

1. **Desarrollar un entendimiento del propósito de nuestro proyecto.** Este paso consiste en determinar lo que se hará con los resultados que obtengamos, a quien le afectará y si se trata de solo una toma o un estudio continuo.

2. **Obtener el *dataset* que será usado en nuestro análisis.** Se trata de buscar una muestra representativa de lo que estamos tratando de averiguar dentro de una o diversas bases de datos. Las diferentes bases de datos se pueden clasificar en internas o externas. Internas son aquellas que se encuentran en nuestra posesión y que clasifica datos nuestros como puede ser la lista de clientes de una empresa. Bases de datos externas son lo contrario a esto.

3. **Explorar y limpiar nuestra data.** En esta fase nos centramos en verificar que los datos que vamos a analizar están en buenas condiciones, es decir, que todos los datos usan las mismas unidades de medición y tiempo. También hemos de asegurarnos de que no faltan datos importantes y si faltan, que hacer para incorporarlos o averiguar como puede afectar la no inclusión de estos datos en nuestro análisis final en el caso de que no podamos acceder a los datos en cuestión. Se suelen realizar gráficas para revisar los datos como, por ejemplo, un diagrama de dispersión que nos ayude a identificar la relación entre las distintas variables (Valero Moreno, 2017).

4. **Reducir la dimensión de nuestros datos.** No siempre es necesario, pero existen ocasiones en el que tenemos variables que no son necesarias o que, por lo contrario, nos falta alguna variable. Por ello es importante revisar nuestros datos y asegurarnos que todas nuestras variables aportan algo a nuestro análisis.

5. **Determinar que aplicación de *text mining* vamos a realizar**. Existen diferentes fines para los que usamos *text mining*. Podemos estar realizando una predicción, clasificación, *clustering*, etc. Estas diferentes aplicaciones las desarrollaremos detalladamente a lo largo de este trabajo.

6. **División de los datos (aprendizaje supervisado).** Si se trata de un aprendizaje supervisado[2] como puede ser la predicción, se debe dividir nuestros datos aleatoriamente en tres partes: training, validación y *dataset* de prueba.
    6.1.1. Training consiste en los datos que utilizamos para "entrenar" a nuestro modelo. Nuestro algoritmo aprende mediante estos datos.
    6.1.2. El compuesto de datos de validación sirve para ajustar los parámetros de nuestro modelo.
    6.1.3. Los datos de prueba son los que sirven para evaluar nuestro modelo final y solo se puedo usar una vez se hayan pasado por los dos procesos anteriores (Shah, 2017).

7. **Elección de la técnica que se desea utilizar.** Existen diferentes técnicas de minería de textos en las que profundizaré mas adelante. Este paso consiste en elegir una de ellas. Algunas de estas técnicas son el modelo Booleano, LSA (*Latent Semantic Analysis*), LDA (*Asignación de Dirichlet Latente*).

8. **Usar algoritmos para llevar a cabo nuestro proyecto.** Este es un proceso bastante repetitivo en el que se prueban diferentes variables del mismo algoritmo hasta que se da con el adecuado.

9. **Interpretar los datos del algoritmo.** Durante este paso, elegiremos el algoritmo que mejor nos va a proporcionar la información que estamos buscando. Probaremos este algoritmo con nuestra *dataset* de prueba (datos de prueba) que explicamos anteriormente.

10. **Implementar el modelo.** Este último paso consiste en poner en funcionamiento a nuestro modelo.

Anteriormente hemos descrito los pasos que debemos tomar para lograr un modelo de *text* o *data mining* exitoso, desde entender el propósito de nuestro algoritmo hasta

[2] "El aprendizaje supervisado permite buscar patrones en datos históricos relacionando todos campos con un campo especial, llamado campo objetivo. Por ejemplo, los correos electrónicos se etiquetan como "spam" o "legítimo" por parte de los usuarios.
Por otro lado, el aprendizaje no supervisado usa datos históricos que no están etiquetados. El fin es explorarlos para encontrar alguna estructura o forma de organizarlos" (González, 2014).

su desarrollo. A continuación, veremos las distintas etapas por las que pasa un algoritmo de *text mining* desde que recoge la información hasta su presentación final.

Como ya expliqué anteriormente, la minería de textos está relacionada con diferentes disciplinas. Las siguientes etapas o técnicas que encontramos dentro del proceso de *text mining* están relacionadas especialmente con la recuperación de información y la lingüística computacional. Por ello, antes de profundizar en cada técnica, procederé a explicar de manera mas detallada la relación de estas dos disciplinas con el *text mining*.

La extracción de información trata de la identificación y extracción de datos y relaciones relevantes de textos desestructurados; el proceso de creación de datos estructurados a partir de texto sin estructura alguna (Miner et al., 2012). Esto es una parte importante de la minería de textos, pero también es primordial entender la diferencia entre estás dos disciplinas ya que, aunque comparten similitudes, no son lo mismo. La única finalidad de la recuperación de información es obtener aquellos documentos útiles que satisfagan las necesidades del cliente, mientras que la minería de textos, a parte de necesitar una pregunta concreta para realizar el estudio, se encarga de analizar esta nueva información y encontrar un patrón que se repita.

Por otro lado, la lingüística computacional tiene como objetivo el estudio gramatical y sintáctico de los documentos en formato electrónico, usando técnicas para procesarlos con el fin de que puedan ser comprensibles para un ordenador (Ana Isabel Valero, 2017). La lingüística computacional, por tanto, es una parte primordial dentro de la minería de textos, pero esto no significa que tengan el mismo objetivo; *text mining* se extiende mucho mas allá del análisis gramatical y sintáctico de textos y espero poder aclararlo a lo largo de este trabajo.

Una vez ya aclaradas las disciplinas que toman parte en las etapas de del *text mining*, procederé a profundizar en estas distintas etapas.

- Pre-procesamiento de los textos: este proceso consiste en configurar el texto que se pretende analizar para convertirlo en un formato simple y fácil de analizar. Para lograr esto, se divide el texto en secuencias de letras delimitadas solo por espacios y signos de puntuación. Aparte de esto, se eliminan palabras que no aportan significado alguno al mensaje que se quiere transmitir; este tipo de palabras suelen ser preposiciones artículos y conjunciones. Una vez realizado esto, se procede a agrupar las palabras del texto en cuestión dentro de su familia léxica. También es recomendable tener en cuenta aquellas frases que se repiten a menudo.

- Identificación de los nombres propios: es crucial poder destacar aquellas palabras dentro del texto que son nombres propios como pueden ser personas, empresas, países, etc. La lingüística computacional tiene especial importancia en esta etapa.

- Representación de los documentos mediante un modelo: Es esencial representar la información que vamos a analizar en un modelo. Como indica Ana Isabel Valero en su trabajo Técnicas Estadísticas en la Minería de Textos, el modelo comúnmente utilizado en *text mining* es el vectorial. Este modelo se encarga de caracterizar el documento o documentos que se pretenden analizar por medio de términos que lo representen. Estos términos se pueden sacar directamente del texto en cuestión o se puede realizar mediante una herramienta informática. En esta etapa es donde se presenta la recuperación de información ya que esta disciplina se encarga de calcular la distancia que existe entre los vectores (cada vector representa uno de los documentos) y el vector que realiza la búsqueda de información. De esta manera podemos obtener los documentos con la información más semejante al vector de búsqueda (Bengio, Ducharme, Vincent, Jauvin, 2003).

- Categorización automática: esta etapa trata de agrupar los documentos en categorías preestablecidas. Cuando se establecen estas categorías, no se utiliza ninguna información aparte del contenido en sí del documento. Encontramos dos formas de categorización. La primera es *singlelabel,* la cual, como su nombre indica, adjudica cada documento en una sola categoría. Mientras que *multilabel,* asigna cada documento a dos o más categorías. Existen dos maneras distintas para determinar si un documento pertenece a una categoría o no. La primera es *hard categorization* que directamente decreta si el documento en cuestión pertenece a una categoría o no. La segunda es *ranking categorization* que calcula la probabilidad que tiene un documento de pertenecer a una categoría u otra.

- Relación entre términos y conceptos: esta etapa solo consiste en extraer los términos y conceptos más importantes de un texto y determinar la relación entre ellos.

## 3. Técnicas del *text mining*

En este apartado de mi trabajo voy a explicar las distintas técnicas más usadas en la minería de textos. Todas ellas tienen el mismo objetivo que trata de extraer e identificar información que suele estar oculta o difícil de desenterrar de la complicada y abundante red informática que conocemos como ciberespacio.

### 3.1. Latent Semantic Analysis (LSA)

Este modelo, también conocido como análisis de semántica latente en español, supone que existen ciertas palabras que tienen alguna relación difícil de observar para procesos automáticos encargados de realizar estudios de *text mining*. Un ejemplo puede ser motor y embrague los cuales están relacionados al ser los dos partes de un vehículo motorizado. Como hemos visto en el anterior ejemplo, estas dos palabras que en principio parecen ser completamente independientes, tienen una relación por un tema subyacente difícil de observar.

LSA consiste en un modelo para extraer y representar el significado contextual y semántico de palabras por medio de cálculos estadísticos aplicados a grandes cuerpos de texto (Landauer, Foltz, Laham,, 1998). LSA, junto a LDA, son las dos técnicas más usadas en el *text mining*. Pasaremos a definir la última más adelante.

El análisis de semántica latente tiene muchas aplicaciones distintas pero las más conocidas son las de corrección de textos y la medición de la coherencia y cohesión de un texto. Originalmente este modelo era utilizado para la recuperación de textos, pero finalmente se consideró esta técnica como apta para la adquisición y representación de información.

**¿Cómo funciona?**

El principio del LSA consiste en procesar un texto de gran tamaño conocido generalmente como corpus lingüístico. Este corpus se ordena en una matriz de frecuencias en la que las filas son las palabras y las columnas son los distintos párrafos o frases que hallamos en el corpus. Cabe destacar que dentro de este corpus encontraremos miles de palabras, frases y párrafos. Al haber ordenado el documento en una matriz, podemos observar claramente el número de veces que se repite una palabra. Para que esta matriz nos pueda aportar la mayor cantidad de información importante, procedemos a realizar una ponderación con el objetivo de restarle importancia a aquellas palabras que se repiten en numerosas ocasiones como pueden ser determinantes y conjunciones mientras que se aumenta la importancia de aquellas palabras que no se repiten o se repiten de manera moderada. Esto se realiza debido a que son las palabras que no se repiten a menudo las cuales nos podrás propiciar la información mas relevante e importante en cuanto al corpus (Botana, 2010).

El siguiente paso es aplicar un algoritmo que pretende disminuir las dimensiones de la matriz para que sea una cifra más asequible. Este algoritmo es conocido como Descomposición en Valores Singulares o *Singular Value Decomposition (SVD)*.

Según Ana Isabel Valero, la ventaja de representar el lenguaje vectorialmente es que los vectores permiten comparaciones por medio de distancias euclídeas, cosenos y otras medidas. Aparte de esto, por medio de la matriz que ya tenemos, podemos introducir nuevos vectores que representan documentos introducidos posteriormente. Estos documentos son conocidos como pseudodocumentos y es todo aquel documento que hemos decidido añadir al espacio semántico pero que no forman parte del corpus inicial. La técnica del LSA nos permite añadir estos a la matriz ya simplificada sin necesidad de tener que volver a empezar desde el principio.

### 3.2. Modelo Booleano

El modelo Booleano es conocido por ser el modelo más sencillo, pero a la vez el más útil. Esta técnica consiste en establecer una serie de palabras clave las cuales son consideradas como las más importantes del texto y en las que se debe centrar nuestra

búsqueda. Se determinará la importancia de cada palabra clave dependiendo de si la encontramos en el documento o si está ausente. La gran ventaja de este modelo es su simplicidad y rapidez. Por otro lado, la desventaja de este modelo es que solo toma por palabras relevantes o importantes aquellas que el autor de dicho estudio ha seleccionado como palabras clave. Si existe otra palabra clave que el autor no ha determinado como tal, o simplemente se ha olvidado de tal palabra, los resultados pueden no ser del todo acertados.

### 3.3. Latent Dirichlet Allocation (LDA)

El modelo de asignación de Dirichlet latente es la técnica no supervisada más moderna para extraer información temática (temas o tópicos) de una colección de documentos (Allahyari et al., 2017). Esta técnica se basa en que los documentos son un conjunto de temas distribuidos aleatoriamente y mediante LDA podemos identificar los diferentes tópicos que forman dicho documento. La proporción de los tópicos es distinta en cada documento pero los tópicos son los mismos para todos los documentos (Elkan, 2014).

**¿Cómo funciona?**

LDA es una forma de distribuir las palabras de un documento en tópicos para así hacer más fácil su estudio. Esta técnica elige que palabras son más probables en distintos tópicos y que proporción de tópicos existe en cada documento. Posteriormente elige la palabra mas probable de cada tópico tras haber analizado el corpus en cuestión.

Es muy parecido al modelo de LSA que vimos anteriormente, pero en vez de analizar las palabras y su posible relación, se centra en los tópicos de cada documento para así hacer más sencilla la interpretación de esta información.

### 3.4. Modelo de Espacio Vectorial Semántico

“La forma más común de representar documentos es convertirlos en vectores numéricos. Esta representación se llama "Modelo de espacio vectorial" (*Vector Space Model; VSM*). Aunque su estructura es simple y originalmente se introdujo para la indexación y recuperación de información, VSM se usa ampliamente en varios algoritmos de minería de textos y permite el análisis eficiente de una gran colección de documentos (Allahyari et al., 2017, p. 4). Esta técnica analiza las expresiones lingüísticas de un texto mediante la noción de distancia. Esto quiere decir, por ejemplo, que la palabra “pistola” está mas cerca de la palabra “guerra” que de “sofá”.

Según Valero Moreno (2017), este modelo considera los textos como si fuesen bolsas de palabras, pero no es capaz de analizar las relaciones semánticas entre las distintas palabras del corpus. Mediante esta técnica, el enfoque no es en la relación semántica entre las palabras, sino en semántica distribucional y composicional. La semántica distribucional se encarga de analizar el significado de cada palabra en sí

mientras que la composicional hace lo mismo con frases y párrafos. A continuación, vamos a proceder a hacer un análisis mas profundo de estos dos términos.

### 3.4.1. Semántica Distribucional

El significado de una palabra está determinado por el contexto en el que se usa (Martí Antonín, 2019) Las diferencias de significado de las palabras de un texto están en correlación con las diferencias de distribución. "Se trata de una aproximación extensional y relacional al significado que sortea el problema de la representación formal del mismo" (Martí Antonín, 2019). El significado como tal de una palabra entonces es representado por medio de la suma de los contextos en los que aparece.

A continuación, voy a analizar los distintos modelos que existen dentro de la semántica distribucional.

#### 3.4.1.1. Modelos distribucionales semánticos basados en vectores de conteo.

Los modelos distribucionales semánticos basados en vectores de conteo representan la frecuencia de aparición de palabras en un documento mediante el uso de matrices. Existen tres tipos de matrices distintas que se diferencian según sean las similitudes que analicen. Son las siguientes:

- Matrices palabra-documento: esta matriz se ha utilizado con frecuencia a la hora de recuperar información para descubrir que documentos son relevantes dependiendo de lo que se busca averiguar. Cada palabra se encuentra en las filas de la matriz mientras que los distintos documentos se encuentran en las columnas. Mediante esta matriz podemos averiguar el número de veces que aparece un término en dicho documento. La única desventaja es que no detecta el orden de dichos términos por lo que la información no es tan específica como podría ser.

- Matrices palabra-contexto: estas son las mas adecuadas para calcular la similitud de las palabras de un corpus. En este caso volvemos a encontrar las palabras en las filas de la matriz mientras que en las columnas se encuentran los diferentes contextos del documento. También existe la posibilidad de que los contextos sean reducidos a una palabra objetivo muy cerca a la palabra de la que queremos obtener términos similares. En este instante estamos hablando de matrices palabra-palabra (Valero Moreno, 2017).

- Matrices palabra-patrón: estas matrices se utilizan para medir la similitud semántica de pares de palabras respecto de un determinado patrón (Martí Antonín, 2019). En las filas encontraremos los pares de palabras mientras que en las columnas estarán los distintos patrones.

Para calcular la importancia relativa de una palabra dentro de un texto frente a otra se realiza lo que se conoce como **esquema de pesos**. Según Ana Isabel Valero (2017) el más utilizado es la frecuencia de aparición de las palabras en el texto. Esto se hace mediante una ecuación que tiene en cuenta la cantidad de veces que aparece dicha palabra en el documento en cuestión (tf (w,d)) y la forma inversa también, es decir, lo poco que aparece una palabra en el documento (idf (w)). Esto es debido a que una palabra que aparece poco en un texto, como ya explicamos anteriormente, suele indicar una mayor importancia con relación a aquellas que aparecen más. Aparte de esto, también se incluye la cantidad de documentos que existen en el corpus. La ecuación sería la siguiente:

$$\textbf{tf idf (w, d) = tf (w, d) · idf (w)}$$

Una vez establecidas los diferentes tipos de matrices que nos podemos encontrar, debemos comentar los diferentes métodos que se llevan a cabo para reducir el tamaño de estas matrices. Algunos de estos métodos ya hemos analizado anteriormente:

- **Análisis Semántico Latente** (LSA; *Latent Semantic Analysis*): descomposición de valores singulares es la técnica que se utiliza mediante LSA. Esta técnica descompone una matriz de palabra-documento en diferentes factores a los cuales la matriz original puede aproximarse mediante una combinación lineal (Valero Moreno, 2017).

- **El Hiperespacio Análogo al Lenguaje** (HAL; *Hyperspace Analogue to Language*): se trata de una matriz de coocurrencia palabra-palabra, es decir, este modelo considera como contexto de una palabra las palabras que rodean inmediatamente a ese término en concreto. Se suele tomar un parámetro de 10 palabras. "Las coocurrencias son medidas según la distancia entre las palabras. Las palabras que ocurren cercanas en la ventana de contexto tienen mayor peso y las que ocurren en lados opuestos tienen menor peso" (Valero Moreno, 2017). Cuando dos palabras aparecen juntas dentro del marco de 10 términos, se incrementará el grado de asociación entre ambas. Cuanto más cerca estén estas palabras entre sí, mayor será el grado de asociación. Es por esto por lo que podemos asumir que el grado de asociación de dos palabras es inversamente proporcional a la distancia que las separa. Se podrá confirmar entonces que dos palabras están relacionadas semánticamente si tienden a aparecer juntas.

- **El Análisis Semántico Latente Probabilístico** (PLSA; *Probabilistic Latent Semantic Analysis*): este modelo se puede considerar como una versión mas fiable o avanzada del *Latent Semantic Analysis*. "En comparación con el análisis semántico latente estándar que se deriva de álgebra lineal y *downsizes* las tablas de ocurrencia (por lo general a través de una descomposición de valor singular), el análisis semántico latente probabilístico se basa en una mezcla de derivados de la descomposición de un modelo de clases latentes" (Benítez Andrades, 2011).

- **El modelo de Asignación Latente de Dirichlet** (LDA; *Latent Dirichlet Allocation*): el objetivo de este modelo es analizar como aparecen los distintos tópicos en el corpus. Los pasos a seguir serían los siguientes:
    - Primero se seleccionan unos tópicos ya definidos anteriormente y se eligen las palabras asociadas a dichos tópicos con mayor probabilidad de aparecer.
    - Posteriormente se escoge un conjunto de tópicos para cada palabra y una palabra para cada tópico.
    - Finalmente se agrupan los tópicos relacionados con las palabras que aparecen en el texto con mayor frecuencia. Esto nos haría saber los tópicos que mejor explican la información contenida en el documento.

Todos estos modelos o técnicas tienen en algo en común. Esto es su objetivo de encontrar la similitud entre frases, párrafos y documentos. Para hallar la similitud entre documentos es fundamental hallar la similitud entre palabras (Valero Moreno, 2017). Existen dos formas en las que distintas palabras se pueden asemejar. Estas son la similitud léxica y la similitud semántica. La primera se presenta cuando dos palabras se parecen en cuanto a su estructura (morfema, raíz, etc.) mientras que la segunda hace alusión al parecido en significado o contexto de estas dos palabras. La similitud léxica se presenta mediante algoritmos basados en cadenas mientras que la similitud semántica a través algoritmos basados en el corpus y conocimiento (Torres López, Arco García, 2016):

- Medidas basadas en cadenas: se encargan de medir la similitud entre dos cadenas de textos para compararlas y analizar su disposición.

- Medidas basadas en corpus: determinan la semejanza entre palabras de acuerdo con la información obtenida de grandes textos.

- Medidas basadas en conocimiento: especifican el grado de similitud entre palabras usando información obtenida de redes semánticas.

#### 3.4.1.2. Modelos basados en la predicción de contextos

Recientemente se han realizado diversos estudios para la construcción de vectores de significado que tratan de aprender representaciones de vectores para palabras (Grefenstette, Sadrzadeh, Clarl, Coecke, 2014). El objetivo principal de estos modelos es representar vectores de aprendizaje de palabras usando redes neuronales. Se trata de que cada palabra se representa mediante un vector el cual es promediado por medio de vectores palabra en un contexto, y el vector resultante es el que usamos para predecir otras palabras en el mismo contexto (Torres López, Arco García, 2016).

Un buen ejemplo de esto sería escoger un vector con la palabra “rápido” y posteriormente realizar una concatenación de los vectores palabra “veloz”, “pronto”, etc.

Para aprender el modelo de red neuronal, existen 3 pasos a seguir (Valero Moreno, 2017):

- Asignar a cada palabra que encontremos en el texto un vector de rasgos de palabra.

- Expresar la *joint probability function (*función de probabilidad unificada) de secuencias de palabras dependiendo de los vectores de rasgos de estas palabras en la secuencia.

- Aprender los vectores de rasgos de palabra y los parámetros de la función de probabilidad unificada.

El vector de rasgos de una palabra simboliza diferentes aspectos de dicha palabra y cada palabra está asociada a un punto en el espacio vectorial. La función de probabilidad se expresa como el producto de probabilidades condicionales de la próxima palabra teniendo en cuenta las anteriores. Este modelo generaliza las combinaciones de palabras ya que se asume que palabras similares van a tener vectores de rasgos similares. Un ejemplo de esto es "coche" y "moto". Estas palabras poseen roles semánticos y sintácticos parecidos en las siguientes oraciones: "El coche iba muy rápido en la autopista" y "La moto estaba derrapando en la rotonda". A partir de estas dos frases, podemos generalizar distintas combinaciones como "La moto iba muy rápido por la autopista" y "El coche estaba derrapando en la rotonda" (Bengio, Ducharme, 2003).

Durante muchos años se han comparado los modelos de predicción con los de conteo que ya explicamos anteriormente. Varios autores obtuvieron mejores resultados mediante los modelos de predicción. Estos autores exponen que los pesos de los vectores son establecidos para predecir el contexto de las palabras que tienden a aparecer a posteriori. Debido a que las palabras similares aparecen en contextos similares, el programa aprende automáticamente a asignar vectores similares a términos similares. "Declaran que esta nueva forma de entrenar los modelos semánticos distribucionales es atractiva porque reemplaza el cálculo heurístico de las transformaciones de vectores de los modelos iniciales, con un paso de aprendizaje supervisado. La supervisión no tiene un costo de anotación manual, dado que la ventana de contexto usada para entrenar puede ser extraída automáticamente de un corpus no anotado. Sin embargo, este enfoque es dependiente de la calidad del corpus original y del dominio" (Torres López, Arco García, 2016).

### 3.4.2. Semántica Composicional

Anteriormente hemos analizado, mediante la semántica distribucional, como elaborar modelos para obtener representaciones vectoriales de palabras. La semántica

composicional, por otro lado, se encarga de obtener representaciones de vectores para frases, oraciones y documentos.

Mediante la semántica distribucional, podíamos elaborar una representación vectorial que analizaba la similitud entre palabras. Sin embargo, no podemos realizar el mismo método cuando analizamos frases u oraciones ya que no se pueden aprender rasgos distribucionales a ese nivel. La semántica composicional permite aprender una jerarquía de rasgos donde niveles más altos de abstracción se obtienen mediante niveles más bajos (Grefenstette, Hermann, Dinu, Blunsom, 2013). Una función general de composición semántica se puede manifestar mediante u y v que son las representaciones más pequeñas, R es la información relacional y K el conocimiento histórico (Torres López, Arco García, 2016).

La similitud semántica entre oraciones es más difícil de percibir que la de palabras independientes. El contenido semántico de una frase está relacionado con la de sus constituyentes y la habilidad de recombinarlo mediante una serie de reglas específicas (Torres López, Arco García, 2016). Las redes neuronales son capaces de analizar objetos individuales distintos como palabras, pero, a la hora de analizar múltiples objetos como ocurre con oraciones o documentos, es más difícil determinar que rasgos están relacionados o no (Mitchell, Lapata, 2010).

## 4. Software

En este apartado compararemos los distintos software o herramientas que existen actualmente en el mercado para la minería de textos. Definiremos algunos de los softwares que considero de mayor utilidad e importancia y que llevan a cabo las distintas técnicas que hemos descrito anteriormente.

### 4.1. SemanticVectors

Se trata de una biblioteca de código abierto cuya función es crear modelos de espacio palabra a partir de texto en lenguaje natural (Torres López, Arco García, 2016). Estos modelos son diseñados para representar palabras o documentos a partir de distintos conceptos y se crean al aplicar algoritmos de conceptos a matrices palabra-documento creadas con el programa *Apache Lucene*. Los distintos algoritmos que podemos utilizar para crear un modelo son *Random Projection, Latent Semantic Analysis y Reflective Random Indexing (RRI).* Existen tres etapas básicas para la creación de un modelo mediante *SemanticVectors:*

- Crear valores aleatorios básicos para cada documento.
- Sumar los vectores de documentos básicos donde el término ocurre para crear los vectores de palabras.
- Sumar los vectores de términos para así crear vectores de documentos.

El algoritmo RRI es capaz de analizar las conexiones entre palabras que no coocurren juntos dentro del texto. Para hacer esto existen dos formas. La primera se centra en las palabras o términos por lo que crea vectores aleatorios para cada término mientras que la segunda forma se centra en los documentos de tal forma que crea vectores aleatorios para cada documento introducido en el corpus (Widdows, Cohen, 2010).

### 4.2. Word2Vector

Este software necesita de un texto de entrenamiento para posteriormente poder construir representaciones vectoriales de palabras del texto que deseamos analizar. Este software implementa las arquitecturas conocidas como CBOW y Skip-Gram. Skip-Gram es conocida por ser mas lenta, aunque es buena para capturar palabras poco frecuentes (Torres López, Arco García, 2016). La fuente de entrada para esta herramienta es el corpus que deseamos analizar y en cambio produce vectores palabra como salida. Es indispensable entrenar el modelo con hasta billones de palabras para garantizar resultados fiables en cuanto al espacio vectorial de palabras.

Dos de los algoritmos más usados para el entrenamiento son *hierarchical softmax* y *negative sampling*. El primero es preferible para palabras poco frecuentes mientras que *negative sampling* es mejor para palabras frecuentes en vectores de baja dimensión (Mikolov, Chen, Corrado, Dean, 2013).

### 4.3. Glove

*Global Vectors* es una propuesta de código abierto que captura estadísticas del corpus de manera directa para representar un espacio vectorial de palabra con subestructuras (Torres López, Arco García, 2016). Glove realiza un entrenamiento sobre una matriz de coocurrencia palabra-palabra y tiene un enfoque de aprendizaje no supervisado. Esto significa que el modelo aprende por sí solo tras recibir el corpus de entrenamiento. Este software solo necesita una pasada sobre el corpus para analizar la coocurrencia de las palabras. Con corpus de grandes dimensiones puede ser muy costoso.

### 4.4. R

R es un programa cuya función es la de la manipulación de datos, cálculo y representación gráfica. Existen una serie de paquetes dentro de este software. El usuario puede elegir el que mejor le convenga dependiendo del estudio que va a realizar (Valero Moreno, 2017):

- *Tm*: es específico para la minería de textos.

- *Qordcloud*: sirve para realizar nubes de palabras.

- *Ggplot2*: este paquete se trata de una gramática de gráficas que se expande mas allá de las funciones básicas de R.

- *Readr*: permite leer y escribir documentos.

- *Cluster*: contiene funciones específicas para realizar análisis de grupos de palabras.

- *Dplyr*: sirve para manipular y transformar datos.

**Ejemplo práctico**

R es utilizado por muchas compañías para obtener datos de Twitter cuando realizan estudios con el objetivo de acercarse a sus clientes y conocer mejor su público objetivo. Los datos de entrada los proporciona twitter a través de sus APIs. Mediante este modelo podemos obtener los tweets de posibles clientes, sus seguidores a quién sigue, etc. Además de esto podemos obtener la localización de estas personas o descubrir las tendencias de uso de una o varias palabras (Valero Moreno, 2017). Esta información puede ser muy valiosa para cualquier empresa, especialmente hoy en día donde las redes sociales son un aspecto muy importante de cada negocio.

### 4.5. SAS Text Miner

Esta herramienta es capaz de procesar los distintos documentos de un corpus en varios formatos como pdf y html, extraer palabras o conjuntos de palabras, eliminar palabras que no aportan valor o no tienen importancia (palabras vacías) y reducir palabras a sus lexemas (Eíto Brun & Senso, 2004). Posee la capacidad también de reconocer las palabras individualmente en un documento para así evitar ambigüedades. Otras de sus funciones son:

- Como las otras herramientas ya mencionadas, *SAS Text Miner* también es capaz de representar los textos mediante vectores de palabras los cuales miden su frecuencia.

- *Feature extraction* o la identificación de nombres propios.

- *Clustering* o la agrupación de documentos

- Categorización automática de documentos.

### 4.6. Google Cloud Platform

Parecido al ejemplo mencionado anteriormente con R en Twitter, esto también se trata de un API de Google que posee bases de datos y almacenamiento a parte de servicios de *big data* y *machine learning* entre otros (Valero Moreno, 2017). Se puede usar para extraer información de blogs, artículos y opiniones de usuarios respecto a un producto o servicio.

### 4.7. SPSS LixiQuest

*SPSS Lixiquest* es otro software muy conocido por su capacidad de analizar textos con el fin de conocer su contenido. Puede procesar una gran cantidad de documentos a la vez, a parte de poder identificar nombres propios y analizar la relación entre términos (Valero Moreno, 2017).

## 5. Aplicación práctica

A lo largo de este trabajo hemos explicado las diferentes técnicas que encontramos en *text mining* y las herramientas que podemos utilizar para llevar a cabo estas técnicas. A continuación, voy a describir las cuatro aplicaciones que tiene la minería de textos.

### 5.1. Clustering

*Clustering* o clasificación de documentos es uno de los usos mas populares que tiene la minería de textos (Allahyari et al., 2017). Se puede usar para la clasificación, visualización, organización de documentos. *Clustering* trata de encontrar y agrupar grupos similares de documentos en una extensa colección de documentos. Aunque se utilice mayoritariamente para la agrupación de documentos, también se emplea para la agrupación de párrafos, oraciones o hasta palabras.

### 5.2. Feature extraction

*Feature extraction* o extracción de información se centra en obtener nombres propios de personas u organizaciones a parte de localizar fechas en un documento y analizar la relación entre estas (Valero Moreno, 2017).

### 5.3. Sentiment Analysis

También conocido como el análisis de sentimientos o minería de opiniones, esta es la aplicación de la minería de textos que trata de analizar el vocabulario de un corpus con el objetivo de determinar sus cargas emocionales (Allahyari et al., 2013). Es muy útil para averiguar las opiniones del público sobre un tema en concreto.

### 5.4. Creación de resúmenes

Esta es una de las funciones más comunes del *text mining* y consiste en obtener una descripción general de un conjunto de documentos predeterminados. Es muy útil para el estudio o investigación de un tema en concreto ya que nos permite obtener una idea general de una serie de documentos sin tener que analizar cada documento uno por uno.

## 6. Ejemplos

En este apartado pretendo exponer una serie de situaciones en los que la minería de textos aplica al área de marketing de empresas de la actualidad. Estos ejemplos están basados en empresas o negocios los cuales, mediante *text mining*, son capaces de aumentar su competitividad en el mercado. Debo mencionar que los siguientes ejemplos son situaciones hipotéticas de empresas que han sido creadas con el fin de mostrar la necesidad de *text mining* en empresas de la actualidad.

### 6.1. Heartland Consumer

*Heartland consumer* es una gran multinacional que ofrece ayuda a personas que tienen dificultad con las nuevas tecnologías, especialmente ordenadores. Este servicio ayuda a clientes a instalar sus ordenadores u otros aparatos informáticos en casa a parte de también resolver cualquier duda o consulta que puedan tener.

Para poder estar al tanto de las opiniones de sus clientes, *Heartland Consumer* tiene un departamento específico cuya función es la de analizar cada mención de la compañía en redes sociales. Esto no solo requiere muchas horas de trabajo, sino que también es muy poco eficaz. Es por esto por lo que la empresa ha decidido adoptar técnicas de *text mining*. Mediante esta nueva iniciativa *Heartland Consumers* podrá lograr lo siguiente:

- Detectar los consumidores que realmente están descontentos con los servicios prestados mediante *sentiment analysis*.
- Determinar realmente cual es el verdadero problema para poder abordarlo cuánto antes.
- Agrupar todas las quejas (*clustering*) de los clientes para mejorar el servicio prestado.
- Detectar clientes que han mencionado nuestra empresa más de una vez para poder dirigirnos a ellos directamente (*feature extraction*). A los clientes les gusta ser mencionados directamente por grandes empresas ya que les hace sentir importantes y también les proporciona confianza el saber que sus quejas y opiniones realmente están siendo escuchadas y analizadas por el negocio en cuestión.

### 6.2. Albión Shoes

*Albión Shoes* es una nueva start-up que ofrece zapatillas modernas y baratas para jóvenes. Están preparados para desvelar al público su primera colección de zapatillas para el verano, pero antes quieren realizar una campaña publicitaria destacando su nuevo lanzamiento. Para ello, *Albión Shoes* ha decidido hacer uso de técnicas de *text mining* para, no solo analizar lo que sus consumidores y el mercado necesita, sino también examinar los resultados posteriores a la campaña y considerar si ha sido un éxito o no.

Esta fue la preparación de Albión Shoes antes de lanzar su campaña publicitaria:

- ¿Qué está *trending* ahora mismo en la industria del zapato? Es decir, ¿cuál es un tema del que se está hablando mucho hoy en día entre los entusiastas de las zapatillas? *Albión Shoes* podría centrar su estrategia alrededor de este topic para así captar la atención de posibles clientes.
- Detectar temas *taboo* de su público objetivo. Existen ciertos temas o palabras que tienen un efecto negativo en consumidores. Es necesario detectar cuales son para así poder evitarlos.
- Analizar el movimiento de nuestros competidores para poder percibir sus puntos fuertes y débiles. Al ser una empresa emergente, es muy importante tener en cuenta a nuestros competidores para poder aprender de sus errores o examinar sus campañas más exitosas.

Tras la campaña publicitaria, *Albión Shoes* utilizó técnicas de minería textual para analizar lo siguiente:

- Reacción general del público a la campaña. Queremos saber si la reacción fue positiva o no.
- ¿Qué fue lo que más se destacó de la campaña?
- ¿Quiénes fueron los *influencers* que mas participaron o opinaron sobre nuestro cliente?

## 7. Conclusiones

Como consecuencia al gran desarrollo tecnológico que hemos vivido en las últimas dos décadas, cada vez existe más información y conocimiento. La capacidad para analizarla y almacenarla ya no puede depender solamente de nosotros y es por esto por lo que la minería de textos resulta tan útil.

En este trabajo hemos expuesto lo importante que pueden ser las distintas técnicas de la minería de textos ya que nos permite agrupar miles de documentos, clasificarlos y reducir sus dimensiones para hacer más factible su estudio.

Cabe destacar la cantidad de softwares que están a disponibilidad de cualquier usuario hoy en día. En este proyecto mencioné los 7 que yo considero como los más útiles e importantes, pero existen decenas de ellos al alcance de cualquier negocio o persona.

Por último, quiero destacar es lo útil y valioso que es la minería de textos para las empresas de la actualidad. Hoy en día existen diversas plataformas como Reddit o Twitter donde cualquiera puede exponer sus opiniones y comentarios acerca de un tema en concreto. La minería textual no solo proporciona a las empresas la capacidad de conocer las preferencias de sus clientes sino también analizar su reacción ante lanzamientos de productos o campañas de marketing.

## 8. Bibliografía


Allahyari, M., Pouriyeh, S., Assefi, M., Safei, S., Trippe, E. D., Gutiérrez, J. B., Kochut, K. (2017, Julio). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.* Recuperado de https://arxiv.org/pdf/1707.02919.pdf

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003, Febrero). *A Neural Probabilistic Language Model.* Recuperado de http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Benítez Andrades, J. A. (2011, Octubre). *Estado del arte en Probabilistic Latent Semantic Analysis aplicado a problemas de acceso a la información en la Web.* Recuperado de https://www.jabenitez.com/personal/MASTER/MOTORES_DE_BUSQUEDA_WEB/TAREAS/MBW-PLSA-JoseAlbertoBenitezAndrades-71454586A.pdf

Botana, G. J. (2010, Septiembre). *La técnica del Análisis de la Semántica Latente (LSA/LSI) como modelo informático de la comprensión del texto y el discurso.* Recuperado de https://repositorio.uam.es/bitstream/handle/10486/6181/37597_jorge_de_guillermo_botana.pdf?sequence=1&isAllowed=y

Deloitte. (2013). *Text Analysis: the three-minute guide*. Recuperado de https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-us-ba-textanalytics3minguide.pdf

Dr. Sheakh, T. (2017, Agosto). *Text Mining and its Applications*. Recuperado de https://www.researchgate.net/publication/323254549_Text_Mining_and_its_Applications

Eito-Brun, R., Senso, J. (2004, Febrero). *Minería Textual*. Recuperado de: https://www.researchgate.net/publication/28157651_Mineria_textual

Elkan, C. (2014, Febrero). *Text Mining and Topic Models*. Recuperado de http://cseweb.ucsd.edu/~elkan/250B/topicmodels.pdf

González, A. (2014, Julio). *Conceptos Básicos de Machine Learning.* Recuperado de https://cleverdata.io/conceptos-basicos-machine-learning/

Grefenstette, E., Hermann, K. M., Dinu, G., Blunsom, P. (2013, Diciembre). *New Directions in Vector Space Models of Meaning*. (no publicado)

Grefenstette, E., Sadrzadeh, M., Clarl, S., Coecke, B., Pulman, S. (2011). *Concrete Sentence Spaces for Compositional Distributional Models of Meaning.* Recuperado de https://www.aclweb.org/anthology/W11-0114

Historia de la Minería de Textos. (2010) Recuperado de http://mineriadetextos.tripod.com/historia.html

Infochannel. (2019). *¿Cuántos datos genera el mundo cada minuto?* Recuperado de https://www.infochannel.info/cuantos-datos-genera-el-mundo-cada-minuto

JISC. (2012). *Value and benefits of text mining*. Recuperado de https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

Joung, J., Jung, K., Ko, S., Kim, K. (2018, Diciembre). *Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development.* Recuperado de https://www.researchgate.net/publication/329858374_Customer_Complaints_Analysis_Using_Text_Mining_and_Outcome-Driven_Innovation_Method_for_Market-Oriented_Product_Development

Kalra, P. (2013, Marzo). *Text Mining: Concepts, Process and Applications.* Recuperado de https://www.researchgate.net/publication/277160258_TEXT_MINING_CONCEPTS_PROCESS_AND_APPLICATIONS

Kloptchenko, A., Eklund, T., Barbro, B., Karlsson, J., Vanharanta, H. (2002, Febrero). *Combining Data and Text Mining Techniques For Analyzing Financial Reports.* Recuperado de https://pdfs.semanticscholar.org/fff3/51143937f3b6eb3d92a0731d40ddf8a94a78.pdf

Kroeze, J. H., Matthee, M. C., Bothma, T. J. D. (2007, Julio). *Differentiating between data-mining and text-mining terminology.* Recuperado de https://www.researchgate.net/publication/272644253_Differentiating_between_data-mining_and_text-mining_terminology

Landauer, T. K., Foltz, P. W., Laham, D. (1998). *An Introduction to Latent Semantic Analysis*. Recuperado de http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Liang, P., Potts, C. (2015). *Bringing Machine Learning and Compositional Semantics Together.* Recuperado de https://www.annualreviews.org/doi/pdf/10.1146/annurev-linguist-030514-125312

Martí Antonín, M. A. (2019). *Modelos de Semántica Distribucional*. Recuperado de http://cilx2018.uvigo.gal/actas/pdf/plen03.pdf

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013, Septiembre). *Efficient Estimation of Word Representations in Vector Space.* Recuperado de https://arxiv.org/pdf/1301.3781.pdf

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., Nisbet, R. (2012, Enero) Extraído de: *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.*

Mitchell, J., Lapata, M. (2010, Marzo). *Composition in Distributional Models of Semantics.* Recuperado de https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2010.01106.x

Morgan, E. L. (2013, Septiembre). *Text Mining in a Nutshell.* Recuperado de https://cds.library.nd.edu/expertise/documents/text-mining.pdf

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., Ngo, D. C. L. (2014, Junio). *Text Mining for Market Prediction: A Systematic Review.* Recuperado de https://www.researchgate.net/publication/274037232_Text_mining_for_market_prediction_A_systematic_review

Ordenes, F. V., Burton, J., Theodoulidis, B., Gruber, T., Zaki, M. (2014). *Analyzing Customer Experience Feedback using Text Mining: a Linguistics-Based Aproach*. Recuperado de https://journals.sagepub.com/doi/abs/10.1177/1094670514524625?journalCode=jsra

Pejic, M. B., Krstic, Z., Seljan, S., Turulja, L. (2019, Enero). *Text Mining for Big Data Analysis in Financial Sector: A Literature Review.* Recuperado de https://www.mdpi.com/2071-1050/11/5/1277

Preuss, B. (2017, Abril). *The Application of Text Mining in Business Research*. Recuperado de https://www.researchgate.net/publication/316829582_The_Application_of_Text_Mining_in_Business_Research

SAS. *What is SAS Text Miner 12.1?* Recuperado de http://support.sas.com/documentation/cdl/en/tmgs/65668/HTML/default/viewer.htm#p0ddepwag5tv2tn17ibibha0a0cd.htm

Shah, T. (2017, Diciembre). *About Train, Validation and Test Sets in Machine Learning*. Recuperado de https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Shinde, P., (2015, Agosto). *A Systematic Study of Text Mining Techniques*. Recuperado de https://www.researchgate.net/publication/282836265_A_Systematic_study_of_Text_Mining_Techniques?enrichId=rgreq-8c2256f1bd88b22237c168eee8ce7c9b-XXX&enrichSource=Y292ZXJQYWdlOzI4MjgzNjI2NTtBUzozNjYyMzU5NTE4NzgxNDRAMTQ2NDMyODg2MzA5NQ%3D%3D&el=1_x_3&_esc=publicationCoverPdf

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., Lichtendahl, K. C., Jr. (2017). *DATA MINING FOR BUSINESS ANALYTICS: Concepts, techniques and applications in R*. New York, NY: John Wiley & Sons.

Talib, R., Hanif, M. K., Ayesha, S., Fatima, F. (2016). *Text Mining: Techniques, Applications and Issues.* Recuperado de https://thesai.org/Downloads/Volume7No11/Paper_53-Text_Mining_Techniques_Applications_and_Issues.pdf

Torres López, C., Arco García, L. (2016, Junio). *Representación Textual en Espacios Vectoriales Semánticos.* Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992016000200011

Turegun, O. (2019, Enero). *Text Mining in Financial Information*. Recuperado de https://www.researchgate.net/publication/330213489_Text_Mining_in_Financial_Information

Valero Moreno, A. I. (2017). *Técnicas Estadísticas en Minería de Textos*. Recuperado de https://idus.us.es/xmlui/bitstream/handle/11441/63197/Valero%20Moreno%20Ana%20Isabel%20TFG.pdf?sequence=1&isAllowed=y

Widdows, D., Cohen, T. (2010). *The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics.* (no publicado)

Witten, I. H. *Text Mining.* Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=29B0C8BADEBF28ED4141031FD9E58890?doi=10.1.1.74.3588&rep=rep1&type=pdf