UNIVERSIDAD DE GRANADA

GRADO EN INFORMACIÓN & DOCUMENTACIÓN

TRABAJO FIN DE GRADO

Nociones básicas sobre Opinion Mining y Machine Learning. La polaridad en twitter de Cristiano y Messi durante los octavos de final del Mundial de Rusia.

Presentado por:

D. EDUARDO MORALES PÉREZ

Tutor:

Prof. Dr. o D. ENRIQUE HERRERA VIEDMA

Curso académico 2017 / 2018

D./Dña.: **Enrique Herrera Viedma**, tutor/a del trabajo titulado Nociones básicas sobre **Opinion Mining y Machine Learning. La polaridad de Cristiano y Messi durante los octavos de final de Rusia** realizado por el alumno/a **Eduardo Morales Pérez** INFORMA que dicho trabajo cumple con los requisitos exigidos por el Reglamento sobre Trabajos Fin del Grado en Información y Documentación para su defensa.

Granada, _____ de _____ de _____

Fdo.: _____

Por la presente dejo constancia de ser el/la autor/a del trabajo titulado **Opinion Mining y Machine Learning. La polaridad de Cristiano y Messi durante los octavos de final de Rusia** que presento para la materia Trabajo Fin de Grado del Grado en **Información y Documentación**, tutorizado por el/la profesor/a **D. Enrique Herrera Viedma** durante el curso académico 2017-2018.

Asumo la originalidad del trabajo y declaro que no he utilizado fuentes (tablas, textos, imágenes, medios audiovisuales, datos y software) sin citar debidamente, quedando la Facultad de Comunicación y Documentación de la Universidad de Granada exenta de toda obligación al respecto.

Autorizo a la Facultad de Comunicación y Documentación a utilizar este material para ser consultado con fines docentes dado que constituyen ejercicios académicos de uso interno.

____ / ____ / ____

Fecha Firma

AGRADECIMIENTOS

A mi familia, que siempre me ha apoyado en mis decisiones y que supo darme los valores que me han permitido llegar hasta aquí. Gracias por el increíble esfuerzo.

A mis amigos, los cuales saben quiénes son, y que los cuento con una mano. Gracias por escucharme cuando os necesitaba y pido perdón si alguna vez no supe ayudaros.

A mis compañeros de clase, por ser tan auténticos y únicos. Gracias por 4 de los mejores años de mi vida. Os deseo sinceramente mucho ánimo y suerte en la vida.

A la gente del bloque de pisos en el que viví mi último año universitario en Granada. Allí donde viví os dejaré mi corazón y mil experiencias vividas. Pronto volveremos a vernos

A todas esas personas, compañeros de clase, ex-amigos/as, profesores, gente de la calle que alguna vez me lo hizo pasar mal. También os dedico este trabajo y os doy las gracias. Sin vosotros tampoco habría llegado al lugar que ocupo hoy. Gracias por hacerme más fuerte y más preparado para las adversidades.

ÍNDICE

RESUMEN 12

ABSTRACT 12

1. INTRODUCCIÓN 14
1.2.1 Twitter 16
1.2.2 Facebook 16
1.2.3 Instagram 17
1.2.4 Los blogs 17

2. OBJETIVOS 19

3. MATERIALES Y MÉTODOS 19

4. TÉCNICAS DE ANÁLISIS DE SENTIMIENTOS 21

4.1 Niveles de clasificación 22

4.2 Selección de características 23

4.3 Tareas 24

4.4 ¿Cómo funciona? 25
El pre-procesamiento de los datos 26
Minería de datos 27
El post-procesamiento 28

4.5 Técnicas de clasificación 29
4.5.1 Aprendizaje automático (Machine Learning Approach) 29
4.5.1.1 Aprendizaje supervisado 30
4.5.1.1.1 Clasificaciones basadas en reglas. 32
4.5.1.1.2 Clasificaciones basadas en árbol de decisión 34
4.5.1.1.3 Clasificaciones lineales 35
4.5.1.1.4 Clasificaciones probabilísticas 38
4.5.1.2 Aprendizaje no Supervisado 39

5. HERRAMIENTAS 42

5.1 Algunas de las herramientas 43
Google Cloud Natural language API 43
Microsoft Azure 43
Standford CoreNLP 43
TheySay Preceive REST API 43
MonkeyLearn 44
IBM Natural language Understanding 44

5.2 Herramientas analizadas 46

6. RESULTADOS 54

7. CONCLUSIONES 56

BIBLIOGRAFÍA 59

RESUMEN

Las tecnologías de la información aportan facilidades en nuestras vidas desde hace ya algún tiempo. Bien es sabido que cada vez pasamos más tiempo conectados, lo que nos ha permitido ser más resolutivos con los problemas a los que nos enfrentamos en el día a día. Esto también ha permitido un cambio en el paradigma social y la forma en que nos comunicamos. Esto se debe en gran medida al uso de las RRSS que logran la conexión entre personas, eliminando las barreras del tiempo y espacio. De esta nueva forma en la que nos comunicamos se deriva gran cantidad de contenido subjetivo que expresa opinión, vivencias, sentimientos... Para las empresas ésta información puede ser de gran utilidad, ya que conociendo los gustos de sus clientes respecto a un producto podrán ofrecer servicios cada vez más adaptados al usuario final.

Del procesamiento de los textos con subjetividad que existe en Internet surge el análisis de sentimientos. Este trabajo presenta una serie de cuestiones básicas (como algunos algoritmos empleados en el aprendizaje automático) junto con el análisis de dos herramientas de forma empírica, con la intención de que el lector se familiarice con éste ámbito y así pueda seguir sacando provecho de las TIC.

Palabras clave: Análisis de sentimientos, aprendizaje automático, clasificación.

Abstract

Information technologies make our lives easier than before. It is well known that we spend much longer on the Internet, which make us being more resolutives with the problems we deal everyday. This have allowed a change in social paradigm and the way we communicate. This is such in a big way thanks to social media that allows people to be more connected, avoiding barriers of time and distance. This new way of communication generates a great amount of subjective content that expresses opinion, experiences, feelings… For business this information could be very useful because knowing the pleasures of their customers related to a product may offer services increasingly adapted to the final user.

From text processing with subjectivity we found on the Internet, it arise sentiment analysis. This Project presents some basic points (like some algorithms used in machine learning) along with the analysis of two tools on a empirical way with the intention that the reader become familiar in this field so he can continue to take advantage on ICT.

Keywords: Sentiment analysis, machine learning, classification.

1. INTRODUCCIÓN

Desde hace ya algunos años las tecnologías de la información y comunicación han inundado nuestras vidas de una forma rápida y sutil. Es tal el calado en ella, que los usos que le damos a estas tecnologías van desde satisfacer una simple necesidad de información cualquiera (como conocer cuánto pesa un elefante) hasta realizar la compra del mes. Podemos hablar incluso de un cambio en el modelo en el que consumimos información, puesto que Internet nos ha abierto las puertas a la sociedad de la Información (San Juan 2009).

Uno de los usos más importantes de Internet es el de crear comunidades virtuales sobre ciertos temas en los que una comunidad interactúa entre sí o con otras comunidades. Esto permite expresar lo que sus usuarios sienten, lo que les gusta o no…. Por esta razón el análisis del uso que hacemos en estas tecnologías puede ser de gran importancia para conocer más sobre nuestra propia naturaleza y apetencias, especialmente para las marcas.

Sobre el contexto del 'Big data' numerosas empresas se encargan de recopilar cantidades masivas de datos, a los que aplican técnicas del procesamiento del lenguaje natural (PLN) para dotarles de una estructura que permita su análisis. Aquello que se publica en las redes sociales por tanto puede ser utilizado con el fin de conocer más profundamente qué opinan los internautas sobre un tema o producto, lo que puede permitir a una empresa conseguir mayor ventaja en el mercado (Gómez-Torres et al. 2018). De esta idea surgen las técnicas de análisis de sentimientos que se presentan en este trabajo. Siguiendo una serie de pasos para extraer y procesar el contenido de los textos (sin estructura) se puede conocer si contiene connotaciones positivas o negativas.

1.1 Características del análisis de sentimientos

Las técnicas de análisis de sentimientos están ligadas al procesamiento del lenguaje natural (PLN) que es la disciplina encargada de producir sistemas informáticos que faciliten la comunicación entre las personas y el ordenador (ya sea voz o texto). Su finalidad es la de buscar, procesar y extraer información relevante sobre diferentes elementos expresados por un sujeto, entidad, etc.(Correa & Paula Andrea Benavides Cañón 2007). Es decir, 'traducir' el lenguaje natural para poder ser expresado en un lenguaje estructurado que una máquina sea capaz de procesar. En este caso los elementos que se busca encontrar son las entidades y los conceptos contenidos en las oraciones para detectar los hechos subjetivos expresados como opiniones, emociones, sentimientos..., por ello el éxito del análisis de sentimientos radica en que la extracción de características realizado por el PLN sea adecuado (Vallez & Pedraza-Jiménez

Rafael 2007). Sin embargo existen ciertas limitaciones a la hora de interpretar el lenguaje, ya que el PLN cuenta aún con algunos problemas que impiden el reconocimiento total de ciertos textos. Entre ellos encontramos (Alberich 2007):

- Ambigüedad: Ciertas construcciones lingüísticas parecidas difieren en su significado en función del contexto, el tono o la intención del emisor. Esto dificulta el correcto procesamiento. También las palabras con múltiples significados son una traba al respecto. Identificar patrones lingüísticos es importante para superar el problema en la medida de lo posible. La ambigüedad radica en la mayoría de los casos en problemas sintácticos, léxicos o semánticos.
- Imprecisión: El lenguaje natural en ocasiones no designa de forma clara lo que trata de expresar por el uso de palabras inadecuadas, lo que puede confundir bastante.

También es importante mencionar los sistemas de recomendación en este punto (Huecas & Salvachúa n.d.) ya que pueden estar relacionados con el análisis de sentimientos. Un sistema de recomendación es capaz de sugerir a un usuario información que le pueda ser de utilidad en base al análisis de sus búsquedas anteriores(Huecas & Salvachúa n.d.). Por esto, si se le aplica el análisis de sentimientos un sistema de recomendación podría ser capaz de ofrecer ítems basándose en lo que detecta de las emociones de sus usuarios. He aquí otra razón por la que el estudio de la subjetividad puede parecer interesante.

El análisis de sentimientos es por tanto tremendamente útil, ya que permite hacer estudios para conocer sentimientos globales de una población alrededor de una temática que pueda tratar sobre campañas de cualquier tipo, cuestiones sociales, deportes, economía, etc.

Entre las distintas y numerosas plataformas que encontramos en Internet, en este trabajo se trata principalmente la red social Twitter, porque los usuarios expresan abiertamente sus gustos y opiniones. Cabe mencionar sin embargo que el análisis de sentimientos no es exclusivo a esta plataforma y que existen numerosos lugares en Internet donde puede realizarse.

1.2 Ámbito de aplicación

El marco muestral sobre el que se mueve este trabajo es el de las redes sociales, ya que contienen numerosa información de sus usuarios.

En primera instancia una red social es una plataforma que se aloja en varios servidores de internet. Está compuesta por una estructura de usuarios con cuenta que pueden acceder a ella para publicar o compartir contenido, de manera que se puede interaccionar socialmente de forma que así se forma una comunidad.

Es necesario aclarar que cada una de estas redes sociales posee sus propias características que

las diferencian en su uso y su finalidad. Algunas permiten a los usuarios ponderar las publicaciones de otros usuarios cualitativamente dependiendo de la emoción que les ha hecho sentir. A continuación se comentan brevemente las más conocidas:

1.2.1 Twitter

Twitter permite enviar mensajes públicos de corta longitud (280 caracteres) que llegan sólo a los seguidores propios. Existe la posibilidad de hacer Retweet (RT) en el tweet de un usuario para éste llegue a los seguidores propios (aunque no sigan a la otra persona). Si mucha gente los comparte cabe la posibilidad de que se vuelvan virales y lleguen a mucha más gente. También existen los Hashtags (#) que son términos clave de lenguaje libre que permiten a los usuarios vincular su comentario de forma pública a un tipo de contenido. No nos equivocamos si llamamos a esta plataforma la 'Red del impacto' porque permite la comunicación entre empresas y sus clientes, entre periodistas y lectores, entre un usuario y el mundo. Cualquier persona con un dispositivo con Twitter puede ser un 'pequeño periodista' de su localidad. Por estas razones Twitter permite analizar a la sociedad por su contenido. Si a un usuario le ha gustado una publicación, puede darle a 'me gusta', de manera que quedará guardado en su perfil para verlo todas las veces que desee.

1.2.2 Facebook

Facebook se ha encuadrado como la red social por excelencia, ya que gran cantidad de personas crearon sus cuentas de manera masiva al poco tiempo de lanzarse. En un primer momento sólo estaba abierta a estudiantes, además permitía subir contenido como información personal, fotos y comentarios en tu muro, actualmente cuenta con numerosas aplicaciones conectadas, noticias... es una red bastante completa de información que permite conocer con bastante exactitud el círculo de personas del que se rodean los usuarios. A día de hoy ha cobrado un gran papel en Internet, ya que en casi todos los lugares de confianza en los que es necesario crear una cuenta Facebook ofrece la posibilidad de hacerlo en un 'clic' (se importan los datos necesarios desde la red social hasta la plataforma) y no es necesario introducir manualmente la información personal, lo que es muy cómodo.

Los usuarios reciben las publicaciones más recientes en la página principal y pueden ponderarlas en función de uno de los 6 sentimientos que se ofrecen. Es conocida como la red del enganche, ya que genera publicaciones (scroll infinito) que atraen la atención del

usuario impidiéndole abandonar la red, puesto que continuamente se recibe información de su interés (ésta información aparece gracias a que Facebook cuenta con un sistema de recomendación que nos conoce 'mejor que nuestras madres').

1.2.3 Instagram

Instagram aparece en octubre de 2010. Es la red del estilo, pues es una red dedicada a la fotografía e imágenes. Las publicaciones están orientadas a llegar a la parte emocional de cada usuario. Hay gran tipología de usuarios en esta cuenta, ya que no es obligatorio formar una cuenta personal, es decir, cada cuenta tiene libertad en la temática a publicar (ornitología, humor, interés general, etc.). También se permite subir vídeos de corta duración (máximo 1 minuto).

Todas las publicaciones dan la posibilidad de indicar si te ha gustado y/o comentarla. También hay un apartado que muestra las publicaciones que han gustado a los usuarios que sigues.

Desde no hace mucho Instagram implementó un sistema de recomendación en la aplicación de manera que un usuario que sigue a distintas cuentas sobre un mismo asunto puede agrupar publicaciones de un mismo tema con el fin de que éste algoritmo le presente una serie de publicaciones que quizás le puedan gustar.

1.2.4 Los blogs

Los blogs hoy día se presentan como sitios web en los que el autor publica entradas de forma periódica (ésta es la unidad fundamental de información de esta plataforma) que se ordenan cronológicamente (lo más reciente aparece lo primero). Suelen girar en torno a uno o varios temas de interés. El autor posee el control total del sitio, pudiendo así publicar, moderar, borrar y editar contenido. Los lectores de un blog tienen capacidad (en la mayoría de los casos) de escribir comentarios y responder a otros usuarios, fomentando así el diálogo y la discusión.

Este trabajo centra la atención en la unidad fundamental de información de Twitter, es decir, en los tweets. Éstos son candidatos perfectos por la cantidad de elementos subjetivos que permiten expresar a los usuarios en tiempo real y la sencilla recuperación de los mismos gracias a los hashtags. Es por ello que si son correctamente procesados se podrían extraer opiniones, emociones… de ellos.

1.3 Estructura de la memoria

Una vez que se conoce el tema que se va a tratar, sus características más generales así cómo el

ámbito sobre el que se puede trabajar, se procede a continuación a definir la estructura que seguirá de aquí en adelante este TFG.

- A lo largo del punto dos se exponen los objetivos perseguidos, es decir la razón/razones que tratan de justificar el trabajo.
- Los materiales y métodos están claramente definidos en el punto tres. Es necesario establecer una división, ya que para la comprensión ha sido necesario articular una primera parte teórica y una parte práctica a continuación.
- El punto cuatro contiene el verdadero corpus teórico de ésta memoria. Es la consecución de los objetivos teóricos expuestos en el apartado dos. Se han tratado de plasmar las nociones básicas aplicadas al análisis de sentimientos, así como los pasos que se han de seguir para procesar cualquier texto del que se desean obtener connotaciones subjetivas. Se han abordado diferentes algoritmos de aprendizaje que se pueden utilizar para ello. Las áreas del conocimiento que intervienen, así como las relaciones que se establecen entre ellas para procesar texto se abordan a lo largo de los cinco sub-apartados en los que se divide este punto. Algunos de los puntos a los que se trata de dar respuesta son: ¿dónde se aplica el análisis de sentimientos?, ¿qué es necesario extraer? y ¿cómo se trabaja con todo esto? El 'cómo' es quizás el apartado más extendido, ya que aborda los métodos que se emplean para la identificación, extracción y procesamiento de los datos con el fin de obtener una estructura textual que permita trabajar y extraer conclusiones.
- El quinto punto de este trabajo se divide. Comienza con una lista de herramientas que están actualmente disponibles en la red para el uso y disfrute de sus usuarios. El segundo punto se trata de una muestra empírica de la experiencia de uso propia que presentan dos herramientas.
- Resultados. A través de este punto se comentan los resultados que se han obtenido de la extracción y el análisis de una serie de tweets durante los octavos de final del mundial de fútbol de Rusia en 2018. Su finalidad: Conocer qué usuarios publican mejores comentarios en función de su lengua, así como de descifrar qué personaje público recibe más mensajes positivos y cuál más negativos de los usuarios de twitter.
- Conclusiones. De este punto se derivan las deducciones a las que se ha llegado a través la resolución de los resultados obtenidos, especialmente en la parte práctica. Se abordan algunos de los puntos a mejorar del trabajo así como posibles mejoras.

2. Objetivos

La intención principal a la hora de realizar este trabajo es la de recopilar una serie de conceptos teóricos que sean capaces de responder a las siguientes preguntas:

- ¿Qué es el análisis de sentimientos?
- ¿Por qué debería de ser estudiado?
- ¿Dónde se puede aplicar?
- ¿Cómo funciona?
- ¿Cuáles son algunas de las herramientas que se pueden utilizar?

El propósito de dar respuestas a éstas preguntas es muy simple: Abordar los principios básicos que conforman los distintos modelos de aprendizaje que se aplican al análisis de sentimientos, así como tratar algunos de los algoritmos empleados por medio de una explicación clara y ordenada. Por ello, el objetivo de este trabajo es el de explicar cómo funciona esta tecnología con el fin de que cualquier persona que nunca ha oído hablar de ella sea capaz de comprender a grandes rasgos sus mecanismos. Mi objetivo es que la persona que lea este trabajo sea capaz de desarrollar un mapa mental de los conceptos más importantes y que se anime a probar las herramientas mencionadas. El enfoque del trabajo no se centra en las operaciones que siguen los sistemas de aprendizaje, sino en facilitar la comprensión global de su funcionamiento (es decir, cómo interactúan éstos los sistemas con los datos de entrada y de salida).

Puesto que no existe una clara diferenciación entre tareas y conceptos en esta disciplina (ya que cada año se publican numerosos trabajos) con la redacción de esta memoria lo que se pretende es realizar una agrupación distintiva y clara de los conceptos en común que tienen varios autores. Existe una necesidad real de crear un sumario o recopilación de lo que se publicado recientemente, especialmente para la comunidad hispanoparlante El presente trabajo trata de cubrir esta necesidad. Para ello se disponen dos partes con el fin de facilitar la comprensión, una teórica y otra práctica. Los objetivos de éstas se detallan más profundamente en los materiales y métodos.

3. Materiales y métodos

Para la primera parte (teórica) se procedió a hacer una revisión bibliográfica de (34) trabajos relacionados de alguna forma con la minería de opinión, cuyo propósito era el de dar a entender los pasos y el funcionamiento que sigue una herramienta de análisis de sentimientos siguiendo un método deductivo. Se ponen de manifiesto los conceptos, modelos y pasos a

seguir más importantes que se deben de tener en cuenta para extraer los sentimientos de un texto.

Para comenzar se procedió con la lectura completa de algunos trabajos como (Serrano-Guerrero et al. 2015;Giraldo-Luque et al. 2018; Lin et al. 2017) que permitieron la comprensión de los puntos sobre los que se tenía que trabajar. Para las búsqueda de los primeros trabajos se emplearon los términos 'Opinion mining'|'Sentiment analysis' tras lo cual se refinaron los resultados por categorías y por veces citado. Posteriormente la lectura de algunos puntos de (Liu 2011) junto con otros de (Agrawal & Shafer 1996) permitió profundizar enormemente en el tema, ya que contenían las cuestiones teórico-prácticas relacionadas con el aprendizaje automático, especialmente los sistemas de aprendizaje supervisados.

En relación con el aprendizaje no supervisado, autores como (Huang 1998)(Macqueen 1967) han sido de gran ayuda. Otros trabajos como (Cortes & Vapnik 1995) requirieron de hacer una búsqueda más exacta incluyendo 'unsupervised learning' en la búsqueda. También se encontró en la red algunos trabajos como (Justicia de la Torre 2017) o (Rios Alcobendas 2017) que al parecer pertinentes respecto al área se incluyeron. Así ocurrió también con webs como (Vallez & Pedraza-Jiménez Rafael 2007).

A excepción del primer artículo, los demás fueron buscados en la colección principal de la WoS a lo largo de los meses de mayo, junio y julio. También se emplearon trabajos encontrados en las bibliografías.

Para la segunda parte (práctica), se realizó un estudio sobre los usuarios de twitter que publicaron tweets durante el mundial de fútbol de Rusia de 2018. Se extrajeron un total de 6241 tweets recopilados desde una herramienta online. Las sintaxis empleadas para extraer estos comentarios son las siguientes:

- (Portugal|#Worldcup)+(Cristiano|Cristiano Ronaldo)-(RT @)*. De los que se obtuvieron 3356 tweets para Cristiano Ronaldo
- (Argentina|#Worldcup)+(Lionel Messi|Messi)-(RT @)*. De los que se obtuvieron 2885 tweets para Messi.

Se decidió eliminar de la recuperación aquellos comentarios que fueran Retweets, ya que esto podría dar lugar a la repetición de información (ruido) así como limitar el uso de la aplicación empleada. Para seleccionar la muestra se procedió a limpiar todos aquellos mensajes que no fueran aptos para realizar análisis de sentimientos en ellos (en este caso fueron aquellos que

contenían enlaces, ya que gran parte del contenido contaminaba los resultados). De esta limpieza quedaron 604 tweets con los que trabajar.

El punto 5, que aborda las aplicaciones puede subdividirse por:

- Una breve recopilación de herramientas encontradas en Internet con sus diferentes características con el objetivo de que el lector se anime a probarlas en función de sus intereses, ya que numerosas son de uso gratuito (aunque algunas cuenten con limitaciones de peticiones al servidor).
- El análisis empírico de la polaridad de los mensajes extraídos de twitter. Para ello se han analizado dos herramientas: TAGS V6 y la herramienta integrada en Excel que ofrece MeaningCloud.

  De la primera herramienta se han extraído 6241 tweets para conocer cuál de los dos personajes públicos elegidos despierta más opiniones positivas y negativas.

  Tras la limpieza de los datos se emplea la herramienta integrada en Excel con la intención de construir un modelo capaz de clasificar las polaridades de los tweets.

4. TÉCNICAS DE ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos o minería de opinión comprende el estudio de las opiniones, actitudes o emociones de las personas en relación a una entidad. El objetivo del análisis es el de identificar lo que se expresa en un texto (Liu 2011), independientemente de la estructura que guarde. Sin embargo cuanto menos estructurado sea un documento mayor tratamiento de los datos requiere. Encontrar las opiniones encerradas en el texto requiere del conocimiento de un gran paquete de conceptos, tales como clasificación de sentimientos, clasificación de subjetividad, sumario de opinión, detección de spam, etc. (Serrano-Guerrero et al. 2015). A continuación se definen algunos de los más relevantes. Es importante diferenciarlos ya que todos son conceptos que expresan subjetividad y que es fácil confundir ya que sus diferencias son pequeñas:

1. Opinión:
   a) Según la RAE: ‘Juicio o valoración que se forma una persona respecto de algo o alguien. (Española 2014)
   b) Según Herrera-Viedma (Serrano-Guerrero et al. 2015): ‘Es un sentimiento positivo o negativo, punto de vista, actitud, emoción o valoración sobre una entidad (producto, persona, evento, tema u organización).

Siguiendo la segunda definición para la explicación, podemos establecer una pequeña

diferenciación entre las opiniones. Podemos distinguir dos tipos según su intención:

a) Regulares: Cuyo propósito es el de expresar un juicio acerca de una entidad (de forma directa o indirecta)
b) Comparativas: Expresan similitud o diferencia entre entidades.

Las opiniones también pueden clasificarse por su causa:

a) Directas: Expresan un aspecto de una entidad.
b) Indirectas: La causa de que expresen un aspecto de una entidad es que se hayan basado en el efecto de otra entidad (son el resultado de algo)

La importancia de las opiniones radica en que está formada por 5 aspectos, que bien identificados posibilitan el análisis de sentimientos:

- Entidad: Producto, servicio, persona, evento, organización, etc.
- Aspectos de la entidad: Componentes o atributos de la entidad
- Orientación de la opinión respecto al aspecto de la entidad (también conocido como polaridad)
- Contenedor de opinión: Sujeto que expresa la opinión
- Tiempo: momento en que se expresa

2. Subjetividad: Permite expresar sentimientos personales, visiones o creencias, aunque no tiene por qué incluir ningún sentimiento. La diferencia es que una oración objetiva contiene información de hecho sobre el mundo, mientras que la subjetiva expresa vivencias o juicios. (Leskovec et al. 2010)
3. Emoción (Liu 2011): Sentimiento subjetivo y/o pensamiento. Para medir la fuerza de una opinión antes deberíamos de conocer la intensidad de las emociones como la alegría o la tristeza.
4. Estado anímico: Es la mezcla de emociones y vivencias que mueven a una persona a escribir un comentario o una crítica.

4.1 Niveles de clasificación

El análisis de sentimientos pues, considera un proceso de clasificación que define las relaciones que se producen entre los elementos contenidos en uno o varios textos dados. La clasificación resultante (que puede ser una ontología, una taxonomía, tesauro…) se obtiene del análisis que puede producirse en tres niveles (Serrano-Guerrero et al. 2015; Liu 2011):

1. Nivel de documento: El documento se considera como una unidad informativa que representa una opinión o un sentimiento (positivo o negativo) y se trata de identificar.
2. Nivel de oración: La unidad informativa son las oraciones. En primer lugar es preciso

identificar si la oración denota objetividad o subjetividad (ya que las oraciones objetivas no expresan sentimientos). No todas las oraciones son objetivas o subjetivas en su totalidad, sino que hay un porcentaje de objetividad/subjetividad en cada una de ellas. Una vez que se conocen las oraciones que contienen subjetividad, el siguiente paso es el de determinar las opiniones que expresan (positivas o negativas) a través de la identificación de patrones.

El análisis es similar al nivel de documento, la diferencia radica en su tamaño, ya que en ambos casos se tratan de establecer las relaciones que se dan entre los elementos de la oración a través de análisis morfosintáctico y léxico-semántico. No ocurre así en el siguiente nivel.

3. Nivel de aspecto: El objetivo es el de clasificar el sentimiento en función de las características propias de una entidad, lo que requiere en primer lugar identificar las entidades y sus atributos. Los aspectos representan las características del objeto que recibe el sentimiento, es decir que puede tratarse de un nombre, una frase nominal, verbos, adjetivos o adverbios. La intención es la de conocer si se ha expresado opinión y si ésta es positiva, negativa o neutral. (Zafra n.d.). Numerosos trabajos realizan el análisis a nivel de documento (Pang et al. 2002) o de oración (Wilson et al. 2005).

Así pues, entre estos y algunos elementos más la minería de opinión persigue encontrar patrones desde los recursos digitales. (Liu 2011). Aquí intervienen numerosos disciplinas, tales como el aprendizaje automático, estadística, bases de datos, la inteligencia artificial y la recuperación de información.

4.2 Selección de características

La minería de textos es una disciplina de la lingüística computacional que se encarga de extraer conocimiento inherente a los documentos. Emplea el descubrimiento de conocimiento en las bases de datos (KDD knowledge discovery in database) con la intención de buscar información en conjuntos de datos que permitan hallar patrones útiles. Se puede definir como:

«Data mining (also called knowledge discovery in databases) IS the efficient discovery of previously unknown patterns in large databases.»(Agrawal & Shafer 1996)

La diferencia principal entre minería de textos y minería de datos radica en sus naturalezas. Los datos poseen una estructura que permiten tratarlos y extraer conocimiento, mientras que los textos además de reflejar conocimiento cuentan con diferentes tipologías de información y orígenes diversos, así como relaciones internas de varios tipos (Justicia de la Torre 2017). Aunque estas relaciones le dan mayor expresión, dificulta el procesamiento para manejar su

información. Es por esto que si se quiere hallar patrones útiles que nos permitan realizar un análisis de sentimientos sobre un texto no estructurado es preciso definir una organización con su contenido previamente. La estructura o clasificación que sigue se forma a partir de partes específicas del texto como oraciones, palabras, expresiones….tras su procesamiento. Algunas de las características que hemos de conocer para extraer y procesar son (Liu 2011):

a) Frecuencia de los términos o de n-gramas: Los términos son las palabras relevantes (estadísticamente hablando) que aparecen en un texto y lo identifican. Los N-gramas es la consecución de términos de n elementos, en este caso de palabras, aunque pueden tratarse de fonemas, palabras, etc. Mediante los n-gramas podemos construir cadenas de texto basándonos en la probabilidad condicional. Permite encontrar el siguiente elemento dada una secuencia. En ocasiones se consideran las posiciones de las palabras. (Rodríguez 2012)
b) Partes del discurso: Los adjetivos son claros indicadores de opinión, aunque también hay expresiones que indican expresión de forma menos clara (por ejemplo algunas expresiones compuestas o localismos)
c) Negaciones: Normalmente las palabras negativas suelen cambiar la orientación de la oración.
d) Dependencias sintácticas: se pueden generar mediante árboles de dependencia.
e) Palabras/frases de opinión: Las cuales manifiestan sentimientos positivos o negativos

4.3 Tareas

Muchas son las tareas que se vinculan al análisis de sentimientos. Son numerosas y comparten muchos aspectos en común, lo que requiere hacer una diferenciación entre ellas. Consisten en muchos casos en establecer una serie de clasificaciones que nos permitan trabajar con los datos, de manera que aplicando posteriormente el KDD podamos extraer patrones de comportamiento comunes. Estas tareas permiten estructurar el contenido de un texto por medio de los aspectos que componen una opinión. Se diferencia (Serrano-Guerrero et al. 2015):

1- Extracción de entidades y de las características. Este paso es básico para conocer sobre qué o quién giran las opiniones que se desean conocer. Tras su identificación y agrupación (estructuración) se pueden extraer patrones que reflejen la polaridad de las opiniones hacia ellas.
2- Clasificación de subjetividad: Trata de detectar si una oración/frase es subjetiva o no, puesto que las objetivas totales no reflejan opiniones ni sentimientos. Una oración que

expresa sentimientos sí que es subjetiva, pero una frase que no exprese sentimientos también puede serlo.

3- Clasificación de sentimientos: La clasificación de sentimientos trata de medir el sentimiento expresado en un texto en relación a una clasificación. Ésta debe de permitir medir la polaridad (id est si es positivo, negativo o neutral). Hay numerosas formas de definir ésta clasificación. Puede realizarse mediante los valores de un intervalo (por ejemplo [0,5] o lo que es lo mismo una valoración de 0 a 5 estrellas…

   Hay que dejar claro que no todo es siempre blanco o negro, porque los humanos tendemos a combinar objetividad y subjetividad al mismo tiempo cuando nos expresamos, así que lo importante de la clasificación es que sea capaz de encontrar la subjetividad de las oraciones y clasificarlas en un valor positivo, negativo o neutral. Una buena clasificación de subjetividad permite asegurar una mejor clasificación de sentimientos.

4- Sumario de opinión: Es la tarea de extraer expresiones:
   a) De texto en documentos aislados
   b) De datos estructurado donde se agrupan las oraciones para identificar los sentimientos que se expresan a las entidades

5- Recuperación de opinión: Mediante una consulta a la base de datos que recoge la información estructurada tras su procesamiento, trata de recuperar documentos que expresan una opinión concreta. Existen dos ponderaciones de los documentos en estos sistemas.
   a) Relevancia contra la consulta
   b) Opinión sobre la consulta

6- Sarcasmos e ironía: Es una de las tareas más complicadas en éste área puesto que no existe forma clara para detectar declaraciones de sarcasmo o ironía.

4.4 ¿Cómo funciona?

Hoy día contamos con numerosas aplicaciones en la red que permiten ejecutar varias de las tareas anteriormente explicadas, entre de otras. Éstas, se ejecutan en la nube y en muchos casos son gratuitas, aunque a menudo puede ser necesario estar afiliado a una organización para tener acceso, otras limitan su uso mediante un número concreto de peticiones al servidor. Hay muchas aplicaciones y puesto que los caminos para realizar el análisis de sentimientos son variados cada una de ellas ofrecen distintos modelos de análisis.

La forma de trabajo de estas herramientas es simple: los datos se introducen tras su extracción

y filtrado para generar un modelo clasificador. Emplean métodos de aprendizaje que construyen clasificaciones con datos de entrenamiento (que reflejan hechos del mundo real) o sin ellos. En algunos casos, son capaces de producir un modelo capaz, mientras que en otros es necesario segmentar el trabajo. Esto es muy normal ya que, construir un modelo que pueda encontrar, extraer y estructurar el conocimiento de textos, realizar todas las tareas de clasificación y además sacar conclusiones es tremendamente complejo (incluyendo además que cada caso es muy específico) por lo que es mejor hacer una diferenciación en los pasos a seguir. El problema del análisis de sentimientos es que aunque intervienen diferentes ramas del saber, de las tareas anteriormente nombradas, ninguna es un problema resuelto en su totalidad. Entre las disciplinas que intervienen en el análisis de sentimientos encontramos (Justicia de la Torre 2017):

- Recuperación de información: Filtrar y recuperar los documentos adecuados
- Procesamiento del lenguaje natural (pre-procesamiento y etiquetado del texto)
- Extracción de información. Como el reconocimiento de entidades.
- Minería de datos. Permite descubrir patrones y asociaciones en textos que nos capaciten para comprender el conocimiento que poseen con la finalidad de trabajar con ellos

Combinando todas éstas ramas del saber se consigue extraer conocimiento de los textos (‘Knowledge Discovery in Text’ en inglés). De esta forma se derivan patrones que permiten que se produzca el análisis de sentimientos. Los pasos a seguir para descubrir conocimiento en el texto (KDT) se divide en pre-procesamiento, minería de datos y post-procesamiento:

El pre-procesamiento de los datos

Este paso es muy importante ya que fijará el tipo de información obtenida tras él. La estructura debe conservar la riqueza documental que poseía previamente al análisis, el buen resultado depende en gran medida de cómo se aplica el correcto PLN en ésta fase. Una correcta estructuración de la información previa al procesamiento por parte de un algoritmo facilitará en gran medida la construcción de una clasificación.

Varios autores difieren en la tipología de documentos con los que trabajar. (Feldman et al. 1998) hablan de documentos categorizados con términos que los identifican, mientras que (Hearst 1999) somete el corpus mediante la lingüística computacional (donde trabaja el PLN) para la extracción de términos.

Para superar el problema que representa el procesamiento de la información en los documentos, se pueden aplicar técnicas de extracción de información. Éstas técnicas permiten proporcionar información contenida en el documento (entidades, relaciones, etc.). Este proceso se realiza por

medio de Ontologías o lexicones. Algunas de las técnicas que se emplean aquí son (Justicia de la Torre 2017):

- Análisis léxico: Que nos permita identificar el idioma
- Tokenización o división de texto. Limitar los conceptos del texto para que representan términos. De esta forma se acota la información y se provee de una estructura común
- Consulta a diccionarios
- Procesadores de conceptos específicos (fechas, siglas, etc.)
- Lexicones de nombres propios
- Categorizar cada palabra o token en su categoría gramatical mediante la definición de relaciones. También conocido como post-tagging.

Mediante las técnicas de extracción de información se consigue definir una plantilla (estructura) con la información del texto para ser almacenada en una base de datos, con la que, de en adelante se podrá consultar y definir reglas. El proceso de extracción de información que se ha comentado, requiere el seguimiento de una serie de pasos que se articulan de la siguiente forma:

- Extraer los hechos: Vinculado a algún tipo de área del conocimiento que es necesario dominar.
- Integrar los hechos: Para encontrar relaciones de co-ocurrencia, como por ejemplo el uso de referencias catafóricas (mecanismo por el que se establece una relación entre palabras, debido a que una palabra remite a otra) (Cervantes n.d.).
- Representar el conocimiento en un soporte útil para trabajar con él. Puede tratarse una plantilla para introducir en una base de datos relacional.

Cuando ya se cuenta con el texto analizado, procesado y estandarizado la siguiente fase trata de extraer cada uno de los sentimientos que están expresados en él. Esto se consigue por medio de un algoritmo de aprendizaje capaz de clasificar o predecir la clase a la que pertenecen los nuevos datos en base a una clasificación previa (modelo predictivo) o a las relaciones que existen entre conceptos (modelo descriptivo).

Minería de datos

Es la fase más importante puesto que integra modelos de aprendizaje y estadísticos para la obtención de patrones (Justicia de la Torre 2017).

En minería de datos existen dos tipos de modelos según las relaciones entre las variables:

- Modelo descriptivo: Forman agrupaciones de datos cuyas conexiones pueden ser de cualquier tipo. Suelen coincidir con los sistemas de aprendizaje no supervisado.

- Modelo predictivo: Para clasificar un conjunto nuevo de datos se emplean las reglas que se basan en un conjunto de datos ya observados (esto se verá más profundamente en el apartado 4.5.1.1, Máquinas de aprendizaje supervisado). Mediante datos de entrenamiento se alimenta un algoritmo que se encargará de procesarlos para así extraer los patrones que ayudarán a formar la clasificación. El siguiente apartado (4.5) presenta más profundamente todos estos procesos se llevan a cabo

El post-procesamiento

Aquí se trata de identificar los patrones que pueden sernos útiles. No siempre los resultados son los esperados y por ello en ocasiones es necesario evaluar el trabajo en busca de mejoras o errores. Antes de comenzar esta fase de hecho, es recomendable evaluar los patrones que se extraigan. Estos patrones lingüísticos deberían de ser útiles, nuevos e interesantes en función de la dirección que tenga el estudio que pretenda realizarse.

En ocasiones se emplean técnicas de visualización para ello (Justicia de la Torre 2017), ya que permiten completar el proceso de descubrir conocimiento. Este apartado está dedicado a algunas de las herramientas que permiten visualizar la información y así encontrar patrones que permitan evaluar el trabajo realizado o construir conocimiento nuevo.

El entorno debe de ser lo más amigable posible para facilitarle la comprensión al usuario final. Para lograr la visualización se aplican algunas transformaciones al conjunto de datos final para la representación, como gráficas o nubes de palabras. Esto dependerá en gran medida del algoritmo utilizado para clasificar los datos, ya que las técnicas pueden diferir en la tipología de las salidas que producen. A continuación se muestra una lista de software empleado en algunos trabajos:

- VOSviewer: Mapeo de bibliografías (van Eck & Waltman 2011)
- SOCNETV (social network visualization)(V. Kalamaras n.d.)
- HTML (hyper text markup Language). Que en realidad es un lenguaje utilizado para la estructuración de documentos.
- OntoGen (Fortuna et al. n.d.)
- Pajek (Batagelj & Mrvar 2004)

Antes de pasar a explicar algunos de los algoritmos que emplea la minería de datos para crear clasificaciones, se presenta a modo de resumen el siguiente esquema con el propósito de que ayude a facilitar la comprensión relativa a los pasos que se tienen que seguir:

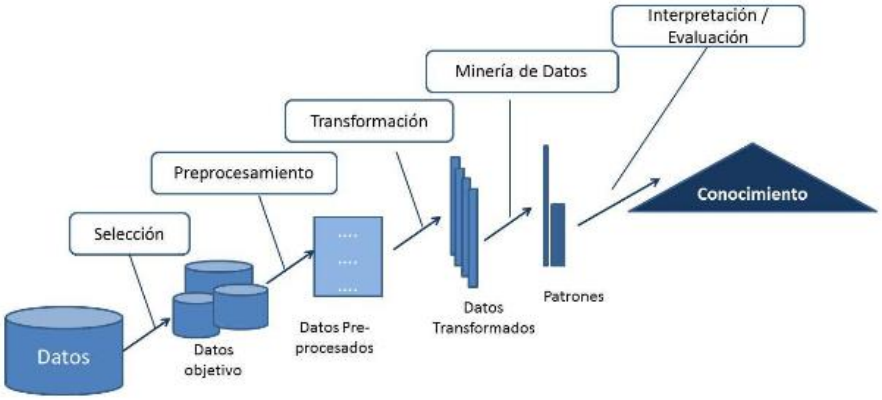

Ilustración 1- Proceso extracción de conocimiento de los textos

(Justicia de la Torre, 2017)

4.5 Técnicas de clasificación

Las técnicas que se emplean para clasificar sentimientos pueden dividirse principalmente en tres bloques que se reflejan en la Ilustración 2 (Liu 2011):

a) El aprendizaje automático. Mediante la aplicación de algoritmos se espera que la máquina comprenda el lenguaje humano.
b) El aprendizaje basado en el léxico: Su nombre se debe a que se basa en un léxico de sentimientos, es decir, una lista con términos que expresan sentimientos.
c) Aprendizaje híbrido. Combina los enfoques anteriores para encontrar un mejor resultado.

4.5.1 Aprendizaje automático (Machine Learning Approach)

El aprendizaje automático pertenece a una rama de las ciencias de la computación y la inteligencia artificial y se emplea en gran cantidad de dominios web. Su objetivo es proveer de técnicas a las máquinas para que sean capaces de aprender mediante la inducción de ejemplos.

Trata el análisis de sentimientos como si se tratara de un problema de clasificación de texto. (Liu 2011). Se crean programas que generalizan comportamientos gracias a una serie de ejemplos. Los métodos se dividen en aprendizaje supervisado y aprendizaje no supervisado:

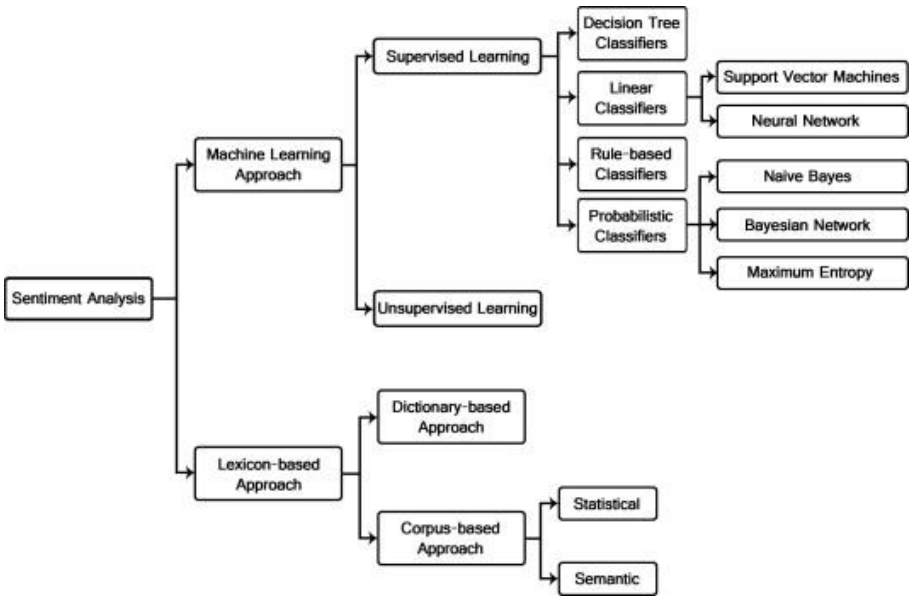

Ilustración 2. Técnicas de clasificación de sentimientos

(Liu, 2011)

4.5.1.1 Aprendizaje supervisado

El objetivo del aprendizaje supervisado es que la máquina sea capaz de predecir valores o etiquetas tras enseñarle cómo hacerlo (Medhat et al. 2014). También se le conoce como clasificación. Trata de enseñar a una máquina a aprender mediante datos que representan ejemplos de una realidad pasada. Esto se consigue aportándole las condiciones, es decir, aportándole los valores que debe obtener en las salidas en función de los valores de entrada. En el aprendizaje supervisado se cuenta con dos variables: La variable predictora y la variable clase que refleja la categoría a la que pertenece un elemento del mundo. A los conjuntos de datos de ambas variables se les denomina atributos.

Asignando a cada valor de la variable predictora los valores de la variable clase (atributos) se genera una función (predicción) a la que se conoce como modelo de clasificación, modelo predictivo o clasificador (José Solano Rojas n.d.). Este modelo, puede seguir cualquier forma, ya sea árbol de decisión, mediante un conjunto de reglas, un modelo bayesiano, etc.

La diferencia principal con el aprendizaje no supervisado es que en este tipo de aprendizaje se conocen las etiquetas o clases que se pretenden predecir. El aprendizaje es supervisado, porque al algoritmo se le proporcionan en principio los resultados que deseamos.

La máquina aprende a través de un conjunto de datos al que se le conoce como *'datos de entrenamiento'(X)*, gracias al cual se desarrolla un modelo mediante un algoritmo de aprendizaje. Para comprobar su funcionamiento se emplea un conjunto de *'datos de prueba'* (Y) (resultados que deseamos) para asegurar la confiabilidad del modelo.

La relación que guardan los *datos de práctica* con los *datos de prueba* es que se asume que ambos tienen una distribución de los ejemplos parecida (lo que nos permite predecir) aunque

esto no siempre se cumple.(Liu 2011) Cuando no se cumple en gran medida se aprecia que los datos de entrenamiento y los de prueba son muy dispares. Esto suele ocurrir cuando la muestra no es representativa.

La finalidad de todo esto es encontrar patrones en los datos para predecir los valores de la clase en futuros casos. Para extraer patrones se deduce la función generada por un vector con los *datos de prueba y de entrenamiento.* La salida va en función del tipo de algoritmo empleado, si es de regresión la salida es un valor numérico, si es de clasificación se obtiene una etiqueta de clase. La etiqueta de clase refleja la categoría a la que pertenece un elemento del mundo (es decir al lugar que ocupan los atributos) (Venegas 2007).

Para medir la exactitud de la función empleamos la siguiente fórmula:

$$Exactitud(M) = \frac{Nº\ de\ clasificaciones\ correctas}{Nº\ total\ de\ casos\ de\ prueba}$$

El número de clasificaciones correctas corresponde al número de las predicciones exitosas del modelo (Liu 2011).

Si no es satisfactorio el resultado que obtenemos de la exactitud, necesitamos elegir otro algoritmo (mediante nuevos conjuntos de datos). Las tareas de aprendizaje en las máquinas requieren muchas pruebas hasta alcanzar una buena definición del modelo. Puede deberse a la aleatoriedad de los datos, a una muestra no representativa de la población (como se menciona antes) o limitaciones de los algoritmos actuales. Es cuestión de probar y probar (entrenar y procesar).

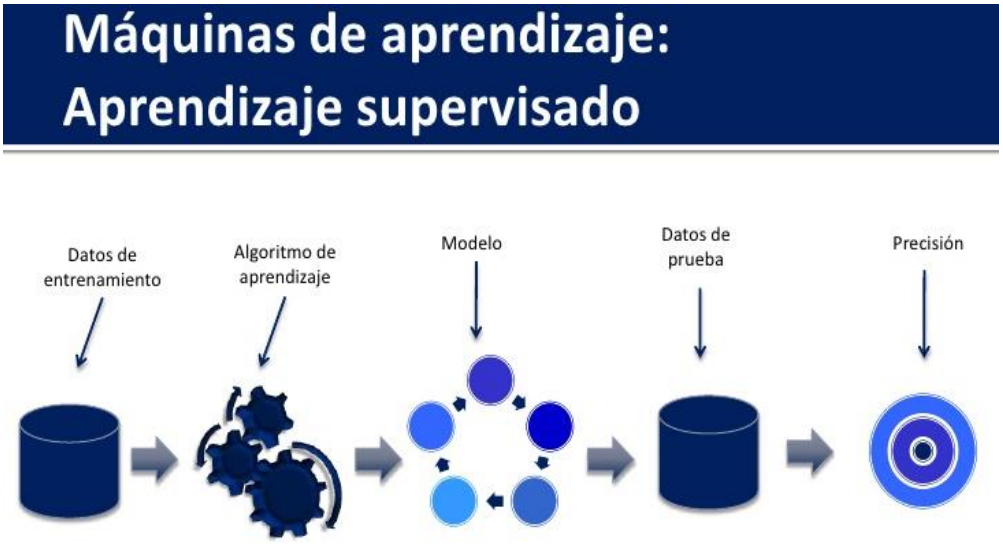

Ilustración 3-Esquema del funcionamiento del aprendizaje supervisado

https://es.slideshare.net/jorgaco/presentacin-titulo

Los pasos a seguir para alcanzar el aprendizaje son los siguientes:

1) Usar un algoritmo de aprendizaje sobre los datos de entrenamiento para construir un modelo de clasificación.
2) Se calcula la exactitud (M) utilizando los datos de prueba. Si la exactitud es buena el modelo se puede usar.
3) Si la exactitud no es satisfactoria es necesario volver atrás para cambiar algo (algoritmo de aprendizaje o el procesamiento de los datos)

El algoritmo puede aplicarse siguiendo distintos modelos:

4.5.1.1.1 Clasificaciones basadas en reglas.

Las reglas de clasificación son un problema típico del aprendizaje supervisado que sigue una estrategia de aprendizaje secuencial. En este apartado se detalla el funcionamiento de las reglas de asociación como modelo clasificador.

Las reglas de asociación son unas reglas de medida de regularización de los datos. Se diferencian de las reglas de clasificación en que pueden predecir el valor/etiqueta de cualquier atributo. El objetivo de las reglas de asociación es el de encontrar relaciones de co-ocurrencia entre una serie de ítems (Agrawal & Shafer 1996) (productos, entidades, organizaciones, etc.) con las que se obtienen asociaciones. De esta manera se puede predecir el valor o las clases a las que pertenecen los atributos de un conjunto de datos. Las reglas de asociación son útiles en el PLN para encontrar patrones lingüísticos en textos ya que permiten descubrir hechos comunes en un conjunto de datos.

La predicción de las etiquetas o del valor de los atributos se consigue mediante los valores de otros atributos ya conocidos. El clasificador se construye a partir de los *datos de prueba (X) y de entrenamiento (Y),* que ya se conocen. De esta manera asociando los atributos de estos conjuntos de manera correspondiente entre ellos, se forma un conjunto de patrones que relacionan valores de atributos con las clases tal y como se hace en el aprendizaje supervisado.

El clasificador utilizado para predecir atributos necesita medir la fuerza de las reglas, por lo que se emplean una serie de restricciones, las más conocidas son el apoyo y la confianza (Hu & Chen 2006).

- El soporte o apoyo: Dado un conjunto de ítems, el soporte es la proporción de los documentos de la BD que contiene ese conjunto ítems. Es una estimación de la probabilidad, es decir, determina cómo de frecuente se puede aplicar una regla. Para comprenderlo mejor se presenta la fórmula:

$$\text{Soporte (Conjunto de ítems X)} = \frac{|X|}{|Documentos\ de\ la\ BD|}$$

Y un ejemplo:

Si contamos con 5 documentos (en la base de datos), en los que los términos yogur y nueces (conjunto de ítems X) aparecen en dos de ellos, el soporte sería del 40%, ya que:

$$\text{Soporte (yogur, nueces)} = \frac{2}{5} = \mathbf{0.4}$$

Es decir, se da en 2 de cada 5 documentos.

El soporte es útil como medida, ya que si es demasiado bajo puede ser que se aplique por pura casualidad. En algunos entornos, como el empresarial, donde hay grandes volúmenes de información una regla que cubra pocos casos no es útil para comerciar con ella.

- La confianza: Es el porcentaje de documentos totales que contienen un conjunto Y si se da el conjunto X. Es una estimación de la probabilidad condicional para determinar la previsibilidad de una regla. Si la confianza es baja, no se pueden predecir X e Y de forma confiable (de ahí su nombre). Veamos la fórmula:

$$\text{Confianza (Conj. de ítems X} \rightarrow \text{Conj. de ítems Y)} = \frac{sop(X \cap Y)}{sop(X)}$$

Y un ejemplo:

Contamos con 5 documentos en la base de datos, en los que aparece yogur y nueces juntos en dos ocasiones, y queremos conocer con cuánta confianza aparece zumo junto a ellos (en este caso aparece sólo en 1 de esas ocasiones):

$$\text{Confianza (Yogur, nueces} \rightarrow \text{zumo)} = \frac{|\frac{1}{5}|}{|\frac{2}{5}|} = \mathbf{0.5}$$

Estas reglas de asociación deben de satisfacer unos umbrales mínimos para que sean verdaderamente útiles. El umbral mínimo de soporte (al que llamaremos 'sopmin' a partir de ahora) limita a frecuencia del conjunto de datos (encontrar conjuntos más frecuentes), así como las reglas generadas. Sin embargo es necesario fijar también un umbral de confianza mínimo ('confmin') ya que de no hacerlo estaríamos considerando que todos los datos tienen la misma

naturaleza (la misma frecuencia en la BD). Ya que es difícil encontrar los conjuntos frecuentes en la base de datos (se necesita considerar todos los subconjuntos de ítems en la BD) existen ciertos algoritmos de minería de reglas de asociación. Debido a su eficiencia y requisitos de memoria el algoritmo más empleado es el algoritmo A priori (Liu 2011).

Las reglas de asociación pueden ser empleadas para construir una clasificación. A estos clasificadores construidos por reglas a menudo se las conoce como clasificadores asociativos, los cuales pueden ser presentados desde tres enfoques principalmente:

1. Usar las reglas de asociación de clases para crear la clasificación.
2. Usar las reglas de asociación de clases como atributos.
3. Usar las reglas de asociación normales.

4.5.1.1.2 Clasificaciones basadas en árbol de decisión

Mediante una serie de construcciones lógicas se construye un árbol que categoriza una serie de condiciones que ocurren sucesivamente (José Solano Rojas n.d.). Se trata de un modelo jerárquico de predicción. Calcula el valor de los atributos de una variable dependiente gracias a los valores de los atributos en la variable independiente (Liu 2011). Se puede emplear para categorizar texto desde un enfoque estadístico. Es necesario cumplir unas normas:

a) Debe estar formado por nodos, vectores de números, flechas y etiquetas.
b) Debe contar con un nodo raíz que no es apuntado por ninguna flecha.
c) Todos los nodos son apuntados por una flecha salvo el nodo raíz.
d) Hay un camino único para llegar al nodo raíz desde cualquier otro nodo.

Los árboles de decisión son un método muy sofisticado por su eficiencia y facilidad de comprensión en humanos. Están formados por los nodos de decisión (que reflejan una prueba o pregunta relativa al valor de un atributo) y los nodos de las ramas (que indican la clase) (Biot & Academy 1977). Los nodos se conectan por medio de aristas que reflejan la condición que se ha de cumplir para seguir ‘el camino’. En la teoría de grafos se le conoce con ese nombre (o

path en inglés) por ésta razón.

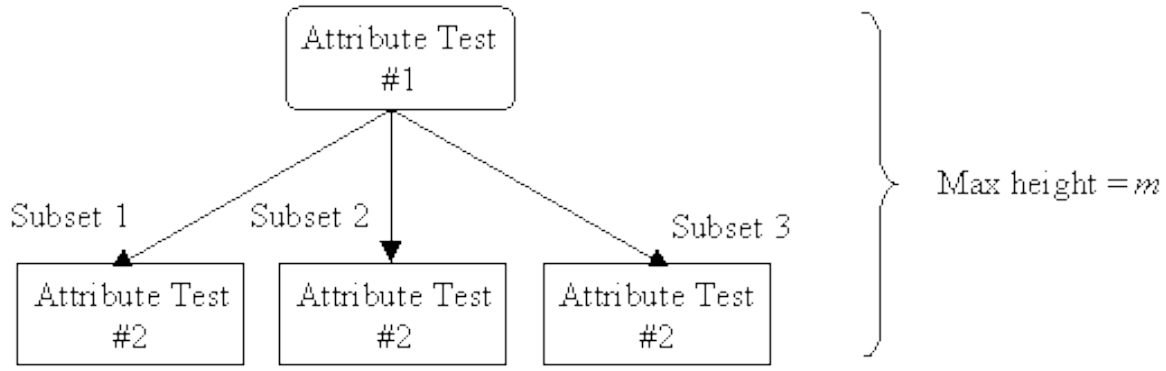

Ilustración 3- Ejemplo de árbol de decisión

http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/dtexample.gif

Los árboles de decisión (por su estructura) pueden ser convertidos en un conjunto de reglas, que son muy semejantes a las de las reglas de asociación. Estas reglas, que permiten formar la clasificación, siguen el mismo formato, sin embargo el árbol de decisión no permite encontrar todas las reglas con la confianza y el apoyo mínimo porque construye las reglas con una sola conclusión (a diferencia de las reglas de asociación).

En la práctica es más favorable realizar árboles de decisiones pequeños (con pocos nodos de las ramas), ya que éstos cuentan con resultados más exactos. En ciertas ocasiones por tanto es necesario considerar si es necesario realizar una 'poda' sobre el árbol para excluir contenido irrelevante.

4.5.1.1.3 Clasificaciones lineales

Los clasificadores lineales emplean técnicas de asociación de características con las entidades. Es decir, utilizan las características (atributos) de los objetos para agrupar (encontrar la clase a la que pertenece). Autores como (Liu 2011) lo consideran como aprendizaje supervisado, ya que se le pueden incluir las etiquetas con las que el algoritmo aprende, sin embargo, otros autores lo acercan más al aprendizaje no supervisado (Kim et al. 2017), ya que al definir los atributos sobre un plano se requiere una función de distancia.

Partiendo de que se cuentan con los datos de entrenamiento, el clasificador representa las características de los objetos como un vector sobre un plano mediante los pesos obtenidos de las frecuencias. A continuación el clasificador elige el valor de la clase a la que pertenece un nuevo objeto por una combinación lineal de las características. Suelen ser los clasificadores más rápidos junto con los árboles de decisión. Su funcionamiento se acerca a las clasificaciones probabilísticas basadas en la probabilidad condicional.

Encontramos dos modelos empleados para determinar parámetros en un clasificador lineal:

- Modelo generativo o de densidad condicional: Emplea modelos gaussianos y binomiales de densidad.
- Modelo discriminativo: persiguen maximizar la salida en los conjuntos de entrenamiento.

Uno de los clasificadores más conocidos es la máquina vectorial de soporte (SVM). Éste sigue un modelo generativo y pueden seguir distribuciones gaussianas, multinomial o de Bernouilli. Construye un hiperplano entre conjuntos para construir dos clasificadores.

Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial son un sistema de aprendizaje lineal formado por un conjunto de algoritmos. Fueron desarrollados por Vladimir Vapnik junto con su equipo en los laboratorios AT&T en los años 90. Estas máquinas permiten construir un modelo capaz de predecir la clase a la que pertenece un nuevo punto (dato). Para predecir la clase es necesario (como el todos los sistemas de aprendizaje supervisado) aportarle al sistema los datos de entrenamiento con los que definir el modelo(Cortes & Vapnik 1995). Se emplean mucho en clasificaciones de páginas web o en bioinformática. Es uno de los métodos de aprendizaje más conocidos, lo cual es comprensible porque permite crear clasificaciones (y regresiones) más exactas que muchos algoritmos de clasificación de texto.

Es un sistema de aprendizaje lineal, en el que se construyen clasificadores de dos clases (positivo y negativo, por ejemplo) y un hiperplano (línea) trata de separar los conjuntos.

Sobre un espacio de muy alta dimensión (Z) se mapean los vectores de entrada de manera que los datos quedan representados. El hiperplano trata de diferenciar las dos clases para construir la clasificación, lo que nos presenta dos problemas (Cortes & Vapnik 1995):

- Seleccionar la línea de decisión más adecuada que separe a los conjuntos.
- Cómo definir el hiperplano si los datos de distintas clases son demasiado

homogéneos cómo para establecer una función lineal.

Para resolver la primera incógnita es necesario imaginar los puntos más cercanos entre dos de las clases (o más alejados del núcleo de su conjunto). Gracias a estos puntos se definen los vectores de soporte, que ayudan a diferenciar los conjuntos de clases distintas. Puesto que existen casi infinitos lugares entre los vectores de soporte, en los que definir el hiperplano, queda preguntar: ¿cuál será de todos ellos el que debamos de elegir? La mejor solución será siempre aquella que maximice la distancia entre ambos conjuntos (ilustración 4):

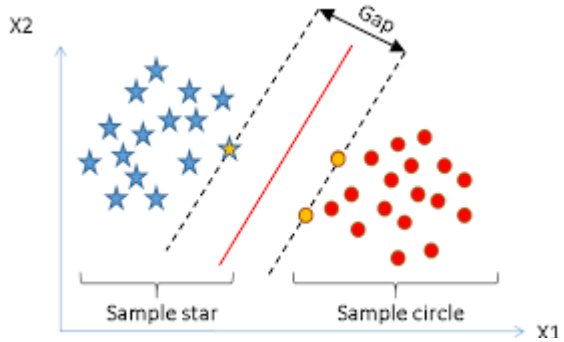

Ilustración 4- Distancia entre conjuntos

https://www.researchgate.net/figure/Main-ideas-of-Support-Vector-Machine-SVM-Source-Simon-and-Melton-2014_fig4_297236887

Es decir, si consideramos los vectores de soporte de las dos clases, la distancia que separa a ambos vectores puede ayudar a construir el hiperplano adecuado. El hiperplano se define entre un espacio igual a ambos vectores (las distancias entre cada conjunto se queda así delimitado de la misma forma). De esta manera se puede establecer una buena diferenciación entre las clases (con el mayor margen posible).

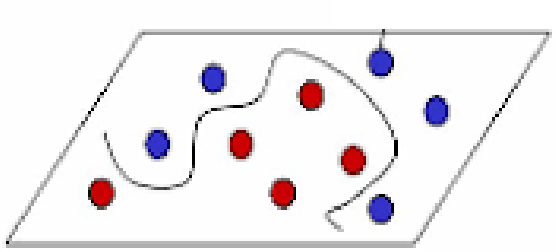

Ilustración 5- Separación débil

http://compuinteligencia.blogspot.com/2009/06/maquinas-de-soporte-vectorial.html

La homogeneidad que presentan los datos influye directamente a la hora de establecer un hiperplano. Un hiperplano de separación es débil cuando hay una ligera mezcla entre los conjuntos, mientras que la separación es nula cuando la diferenciación entre los datos es clara. En ocasiones la diferenciación no es nada clara, como la que podemos ver en la ilustración 5.

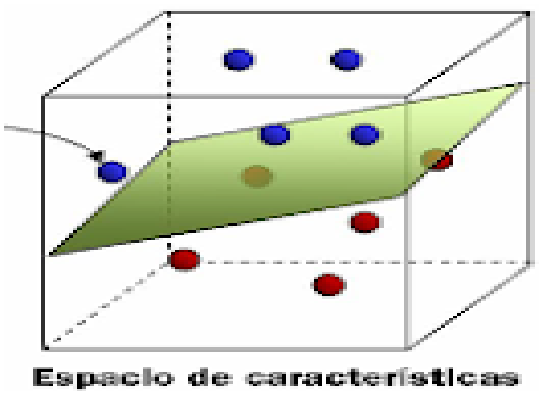

Ilustración 6- Separación en un espacio de tres dimensiones

http://compuinteligencia.blogspot.com/2009/06/maquinas-de-soporte-vectorial.html

Esto se debe a que la dimensión del espacio es poco intuitiva. Para resolver, por tanto la segunda incógnita y ver si son separables necesitamos tratar de ver el espacio desde otra perspectiva, empleando una función denominada kernel. El kernel transforma el espacio situando los puntos en otro espacio de mayor dimensionalidad (Biot & Academy 1977).

En la ilustración 5 se puede comprobar que no se puede emplear una función lineal para separar los datos, sin embargo, en la ilustración 6 se comprueba cómo generando otro plano de mayores dimensiones se puede definir una línea que separa claramente los conjuntos.

Debido a la alta necesidad de abstracción que se requiere en ciertas ocasiones, los sistemas de soporte vectorial (los espacios de altas dimensiones) no suelen emplearse en aplicaciones que requieran el entendimiento humano.

4.5.1.1.4 Clasificaciones probabilísticas

Mediante el desarrollo de un modelo gráfico basado en la probabilidad de ocurrencia de las palabras intenta obtener conclusiones por medio de las variables.

Basados en modelar la distribución que siguen los documentos de cada clase, las variables son consideradas como supuestos independientes. (Biot & Academy 1977). A continuación se explica el clasificador Bayesiano ingenuo (Naive Bayes) empleado para establecer una clasificación basada en las probabilidades de los términos del corpus de un documento:

Clasificador de Bayesiano ingenuo

Este clasificador está basado en el teorema de Bayes. Dicho teorema fue desarrollado por el párroco Thomas Bayes en 1773. Puesto que es una extensión de la probabilidad condicional el teorema es útil para establecer una clasificación calculando la probabilidad posterior. El nombre de ingenuo radica en que las variables utilizadas para redecir la clase son consideradas independientes. Relaciona las probabilidades condicionadas de varios sucesos

independientes con su intersección. Se emplea para encontrar la probabilidad condicional contraria (Rios Alcobendas 2017). Para esto es necesario contar con una serie de datos calculados con anterioridad de manera que así podamos predecir una clase.
La siguiente fórmula define la teoría:

$$P(B|A) = \frac{P(B|A) \cdot P(A)}{\mathrm{P(B)}}$$

O lo que es lo mismo: La probabilidad de que un suceso B ocurra cuando A se da, es igual a la probabilidad de la intersección de los sucesos por la probabilidad del suceso A y dividido entre la probabilidad del suceso B. Este teorema puede aplicarse de forma acumulativa, es decir considerando nuevas variables que se relacionan.
Aunque un documento consiste en una secuencia de oraciones formadas por palabras secuenciales, el clasificador de Bayes considera a cada documento como una bolsa de palabras independientes (Liu 2011). La probabilidad de una palabra por tanto es independiente de su posición en el documento o del tamaño del texto. La probabilidad de observar una palabra de una clase C se estima computando la frecuencia relativa de ésta respecto a todos los términos derivados de la colección del conjunto de datos de entrenamiento. Mediante un clasificador de Bayes ingenuo se calcula la probabilidad posterior de que un atributo pertenezca a una clase y se selecciona el resultado más probable. Cada documento puede considerarse pues, como una distribución multinomial de las palabras que contiene.
Esto es un ejemplo de distribución probabilística multivariante que posee varias respuestas aleatorias (esto es una generalización de una binomial, donde cada suceso cuenta con un número finito de soluciones).
También puede formularse mediante una distribución de Bernoulli, solo que aquí cada palabra del vocabulario puede estar representada por 0 o 1 (ausencia o presencia del mismo).
Dadas una serie de características independientes podemos considerar la probabilidad de que describan un objeto. Así pues, a modo de ejemplo la clasificación establecida por Naive Bayes es del tipo:
Un atributo X puede encuadrarse dentro de la clase manzana cuando sea una fruta, sea roja, y tenga un diámetro de 10 cm.

4.5.1.2 Aprendizaje no Supervisado

En el aprendizaje no supervisado en ningún momento se aportan los datos observados para que la máquina aprenda, es por ello que el sistema debe de ser capaz de generar conocimiento

únicamente con los datos de entrada.

Por esta razón el aprendizaje no supervisado tiene que predecir el lugar que ocupará un nuevo punto dentro de una clasificación basándose en lo que ya conoce de las entradas. Aunque esto suena complicado no lo es en realidad, ya que se basa en la búsqueda de comportamientos o patrones comunes en los datos. Las agrupaciones pueden producirse por cualquier patrón que permita asociar los datos. Puesto que en el aprendizaje no supervisado el valor de las etiquetas de clase no viene dado, es la máquina la que debe buscar estos patrones de asociación entre los datos para generarlas. Esto hace las características de estos datos más importantes, ya que gracias a ellas se pueden encontrar comportamientos comunes entre ellos y mediante un algoritmo generar las etiquetas de la clase que se quiera predecir. La clase debe de obtenerse de la información intrínseca que contienen los datos (es decir de los atributos o características de los objetos que se analizan).

También conocido como clustering, o clusterización, el aprendizaje no supervisado trata de agrupar los datos de entrada en base a la similaridad o disimilaridad de la estructura que presentan. La clusterización es el proceso mediante el cual se encuentran agrupaciones entre los datos. Así pues los datos agrupados conjuntamente (i.e clúster) presentan mayor similaridad entre ellos que aquellos datos que se encuentran en otro clúster, o lo que es lo mismo, hay mayor disimilaridad entre los datos de distintos clústers…

Aunque parece una tarea sencilla para una persona, cuanta mayor dimensionalidad presenta el espacio mayor dificultad presenta encontrar dichas agrupaciones. Esta es la razón por la que se necesita emplear técnicas automáticas que permitan establecer grupos. Sin embargo el proceso que conlleva difiere en la morfología. Dentro de la literatura podemos encontrar dos formas de clusterizar los datos (Pedro Larranaga, Inaki Inza 2008).

- Clustering particional o no jerárquico: Las agrupaciones se realizan asignando una serie de k grupos.
- Clustering hierático o jerárquico: Agrupan o separan clusters para formar otros. Sigue forma de árbol, al que se le conoce como dendrograma donde los datos se agrupan por distancias (medidas de similitud).

La meta de clusterizar es descubrir agrupaciones instrínsecas de los datos de entrada a través de un algoritmo de agrupamiento y una función de distancia.

Clustering particional (K-means)

El algoritmo más conocido en el clustering particional es el algoritmo k-means (Macqueen 1967). Éste sigue dos modelos. Uno establece en primera instancia una serie de particiones

(agrupaciones de datos) mientras que el otro modelo utiliza exclusivamente unas semillas. Se procede a explicar a continuación el segundo método, ya que es el más empleado (Huang 1998).

Su objetivo es el de agrupar los elementos de tal forma que todos pertenezcan a algún clúster. Sobre un espacio d-dimensional se representa un conjunto de datos y a continuación un algoritmo se encarga de agruparlos en base a sus parecidos en k clusters. La meta de éste método está en encontrar el número de clúster K adecuado para cada conjunto de datos. El número de agrupaciones K adecuado es aquel que minimice la distancia entre todos los elementos de los conjuntos (Biot & Academy 1977).

Para comenzar se elige la cantidad de grupos que se desea establecer (k). Una vez fijado, se elige un k número de centroides, que se pueden elegir aleatoriamente de entre los datos. A éstos centroides se les denomina 'semillas' ya que van a permitir encontrar la distancia a la que se encuentran el resto de datos. El algoritmo establece una clasificación de las distancias a las que se encuentran los elementos de los centroides, lo que permite agrupar los más cercanos. Con estas proto-agrupaciones, el algoritmo reasigna el centroide entre los elementos de cada conjunto, de manera que éste sea el que se encuentre a una distancia lo más cercana posible del resto de puntos de los conjuntos. Estos dos últimos procesos de reasignación y clasificación de distancias se repiten hasta que no se pueda escoger un centroide mejor dentro de cada clúster. Los pasos de este proceso se detallan en la *illustración 7*.

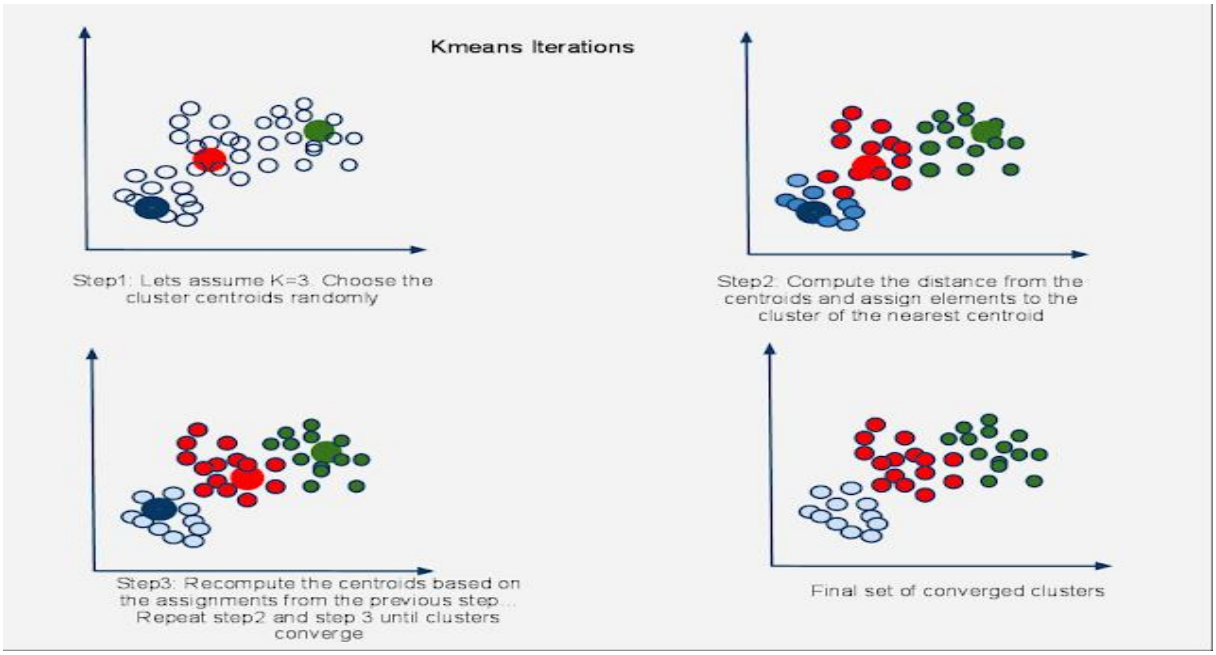

Ilustración 7- Pasos para asignar centroides en k-means

http://humble-developer.blogspot.com/2011/01/kmeans-clustering-algorithm-part-1.html

Para llevar a cabo este algoritmo es necesario procesar los datos de entrada antes de alimentarlo ya que no se conocen las clases. Los objetos que entran se tratan como un conjunto de variables aleatorias. En el análisis de sentimientos de este tipo de aprendizaje la importancia radica en las palabras de opinión, ya que indica la positividad, negatividad o neutralidad de éstas (bueno, lo peor, increíble, horrible, etc.)

Clustering hierático

Consiste en una secuencia de agrupaciones (clústeres) cuya construcción tiene forma de árbol al cual se le conoce como dendrograma. Por esta razón reciben el nombre de jerárquicos. A continuación se explican los tipos de clustering hierático (Biot & Academy 1977):

a) Métodos aglomerativos o ascendentes. Se realiza un análisis para obtener el total de elementos que forman los datos. A continuación se comienzan a agrupar con los elementos que tenga más similitud. El proceso continúa hasta que sólo queda un grupo con todos los elementos.

   En el dendrograma se agrupan desde abajo los más parecidos. Se parte de los puntos distribuidos individualmente a los pies del árbol. Éstos se agrupan en niveles hasta que se obtiene un cluster que contiene todos los puntos en la copa del árbol.

b) Métodos disociativos: Es el proceso inverso, por el que se constituye un grupo inicial con todos los datos y se procede a formar grupos más pequeños (realizando divisiones). En el dendrograma por lo tanto se comenzaría desde el conjunto total de datos en la copa hasta llegar a los pies del árbol con cada uno de los elementos.

El clustering hierático denota ciertas ventajas sobre el método k-means puesto que no requiere de una función de distancia y permite explorar los clústeres al detalle (Gallardo 2014). Sin embargo consume muchos recursos computacionalmente hablando, ya que es muy ineficiente para las grandes cantidades de información.

5. Herramientas

A continuación se trata una serie de herramientas recopiladas de internet, de las que se han extraídos las características principales para conocer sus términos generales. Actualmente todas pueden utilizarse online en la red y son gratuitas (aunque limitadas, cada una de una forma).

En el segundo apartado se explica el uso de dos herramientas de forma empírica.

5.1 Algunas de las herramientas

Google Cloud Natural language API

Google también se suma al conjunto de empresas que no quieren dejar de beneficiarse de las ventajas que aporta el análisis de sentimientos, es por ello que ha desarrollado este servicio en la nube. Explica detalladamente los pasos a seguir para realizar análisis de sentimientos en cualquier texto, aunque es necesario conocer ciertas nociones básicas de programación sobre Python. Emplea algoritmos de aprendizaje automático que estableen la polaridad de un texto en distintos idiomas. Da la posibilidad de utilizar esta herramienta de manera gratuita hasta que se gaste el saldo que te da la compañía.

Microsoft Azure

Además de Google, Microsoft también ofrece servicios en la nube para el análisis de textos. Permite realizar el análisis sobre distintos idiomas sin necesidad de especificarlo, ya que posee una detección automática del lenguaje. Para utilizar todas las herramientas de forma ilimitada es necesario pagar, aunque también permite crear una cuenta de prueba para conocer los servicios que ofrece

Standford CoreNLP

Elaborado por la universidad de Standford, este software consiste en una biblioteca de código abierto, con múltiples aplicaciones en el procesamiento del lenguaje natural. Presenta varios idiomas, aunque para realizar análisis de sentimientos sólo se dispone actualmente del inglés. Ésta herramienta de clasificación de sentimientos permite determinar la polaridad del texto que se introduzca (neutral, positivo o negativo).

TheySay Preceive REST API

Ésta herramienta es bastante sencilla de utilizar. Dispone de un cuadro de texto que permite el análisis de textos cortos. Acepta el español y alemán aunque no funcionan demasiado bien, por ello lo mejor es emplearlo con mensajes en inglés. La ventaja principal radica en que no necesita conocer nada de programación para establecer un análisis de la polaridad en los mensajes que se introducen. Permite realizar 30 peticiones por minuto, con un límite de 500 al día.

Cuenta con más utilidades que otras aplicaciones, ya que cuenta con distintos niveles de análisis tales como:

- Sentimientos: La polaridad entre las entidades, entre las oraciones y en todo el corpus de texto.

- Emociones: Expresa en porcentaje (y con una imagen que lo representa) cuánto de las siguientes emociones están plasmadas en el texto: Enfado, calma, miedo, felicidad, simpatía, vergüenza, seguridad, sorpresa.
- Tema (o tópicos): Sobre los que trata lo que se introduce.
- Análisis sintáctico
- Tipos de entidades: Encuentra las entidades de las oraciones y establece una clasificación (diferencia entre personas, organizaciones, lugares, etc.)
- Especulación: El lenguaje especulativo describe eventos que están (o no por suceder). Aquí la oración se clasifica entre especulativa, de riesgo o de intención

MonkeyLearn

Ésta herramienta es capaz de clasificar tweets en Inglés con la finalidad de establecer su polaridad, también cuenta con un extractor de palabras clave. Monkeylearn presenta un clasificador de NPS (net promoter Score) que mide la lealtad de los clientes (única herramienta encontrada con ésta característica). Está muy enfocada al uso para empresas que quieren conocer en profundidad los que opinan sus clientes de la marca (como hoteles).

Es de las herramientas que más limita el consumo de peticiones, ya que son 300 al mes de manera gratuita y el empleo de un solo modelo. Además presenta gran cantidad de spam, ya que cada vez que se procesa una petición se envía un correo electrónico a tu cuenta, lo que llena el buzón en muy poco tiempo. He de incidir en esto, ya que durante la prueba de uso de la misma llegaron más de 300 correos en cuestión de 5 minutos.

IBM Natural language Understanding

Ésta es la herramienta de análisis de sentimientos de IBM. Analiza texto y extrae conceptos, entidades, palabras clave, aspectos, relaciones y conexiones semánticas gracias a la comprensión del lenguaje natural (ésta es una rama del NLP centrado en la comprensión de la lectura).

Tabla comparativa (Rubio Cortés 2016)

Herramienta	**Idiomas**	**Funciones**	**Límite**	**Licencia**	**tipo**
Google Cloud	Chino Inglés Español Francés Alemán	Polaridad	300$ (dura 12 meses)	Propietario	Biblioteca

	Italiano Japonés Coreano Portugués				
Microsoft Azure	Inglés Español	Polaridad Palabras clave	200$ (dura 12 meses)	Propietario	API REST
Standfordcore	Inglés Árabe Chino Francés Alemán Español (Actualmente se están desarrollando más)	Polaridad	-	Open Source	Biblioteca
TheySay Preceive	Inglés Español Alemán	Polaridad Emociones Temas Análisis sintáctico Entidades Especulaciones	20000 caracteres por cuerpo de texto	Propietario	API REST
Monkey learn	13 idiomas (entre los que encontramos Inglés y Español)	Polaridad Palabras clave clasificador de NPS	300 peticiones	Propietario	API REST
IBM Natural language Understanding	13 idiomas (entre los que encontramos Inglés y	Polaridad	50 columnas de texto, no mayores de	Propietario	API REST

	Español)		1 MB		

5.2 Herramientas analizadas

Puesto que para comenzar a trabajar necesitamos contar con la materia prima, se plantea la cuestión: ¿Cómo puedo extraer una serie de tweets que reflejen el contenido que trato de analizar? Tras la atenta lectura de distintos trabajos como el de (Giraldo-Luque, et al., 2018) se eligieron dos tipos de software, uno de recolección y otro de análisis.

El primero es TAGS V6. Esta herramienta se trata de una interfaz de programación de aplicaciones (API) de búsqueda. La segunda es una herramienta que permite realizar análisis de sentimientos en Excel ofrecida por la plataforma MeaningCloud.

TAGS V6

Permite la búsqueda y extracción de información de los tweets en tiempo real.(Giraldo-Luque et al. 2018) Se puede programar para que las tablas se actualicen con tweets nuevos cada hora. Presenta la desventaja (tal y como indica en la FAQ de la página) de que es un índice de los tweets más recientes, así que la API no puede mostrar todos los tweets que se han publicado (desde que twitter empezó). Está enfocada a la recolección de los tweets más relevantes (Gonzalez-Bailon et al. 2012), es por eso que es una buena herramienta para extracción de datos, ya que interesa conocer la opinión de los usuarios con más seguidores que siguen el mundial.

El mecanismo para extraer los tweets se explica a continuación:

1. Descargar la hoja maestra de Excel que se guarda en drive (lo que requiere tener una cuenta de google)
2. Creamos una copia de la hoja maestra (lo que guardará el archivo en google spreadsheets que es similar a un Excel en la nube, solo que permite ejecutar ciertas funcionalidades).
3. Para poder utilizar ésta herramienta es necesario crear una cuenta en la plataforma de desarrollador de twitter (ya que es necesario introducir el código propio para conectarnos).

4. Vincular nuestra cuenta de Twitter (de hecho, ésta es una de las utilidades más importante de la aplicación, ya que sólo es necesario iniciar sesión una vez, aunque realicemos varias consultas y tratemos de extraer múltiples archivos)
5. Establecer los patrones de la consulta (permite el uso de operadores booleanos). La ilustración 4 se muestra un ejemplo.
6. En Advanced settings se estipulan las condiciones que deben de cumplir los tweets para ser recolectados:

    a) Fecha: se puede especificar hasta un máximo de 7 días anteriores al presente
    b) Contador de seguidores de usuarios que publican los tweets. Esto evita tweets con Spam
    c) Número máximo de tweets que se pretenden recolectar
    d) Podemos realizar la búsqueda sobre tweets, biografías o listas de favoritos.

7. En la pestaña 'Stat' podemos ver un resumen de la recolección (el número total de tweets, la fecha y hora del primer y último tweet)
8. Si clicamos en la pestaña TAGS> RUN NOW! Ejecutaremos la API, lo cual creará otra pestaña con los resultados de la búsqueda (ARCHIVE)

Se han utilizado dos hojas de Excel para la extracción de tweets. A continuación se explican las condiciones empleadas para buscar tweets en ambos archivos:

	Cristiano Ronaldo	Lionel Messi
Términos empleados para la búsqueda:	(Portugal|#Worldcup)+ (Cristiano|Cristiano Ronaldo)- (RT @)*	(Argentina|#Worldcup)+ (Messi|Lionel Messi)- (RT @)
Seguidores mínimos por usuario	500	500
Número máximo de tweets a extraer:	7500	7500

Tweets extraídos:	3356	2885
Período de búsqueda:	12/06/2018-27/06/2018	12/06/2018-27/06/2018
Enlace de descarga	https://drive.google.com/file/d/1nWDc4NkqnlYLSNGbe0dQqbIa6-_PRpU0/view?usp=sharing	https://drive.google.com/file/d/1Fd0lB7o-6RoDw_WGkbL4u-N1GfN6qYz5/view?usp=sharing

*El método de búsqueda empleado no recupera Retweets, lo que facilita la limpieza de datos.

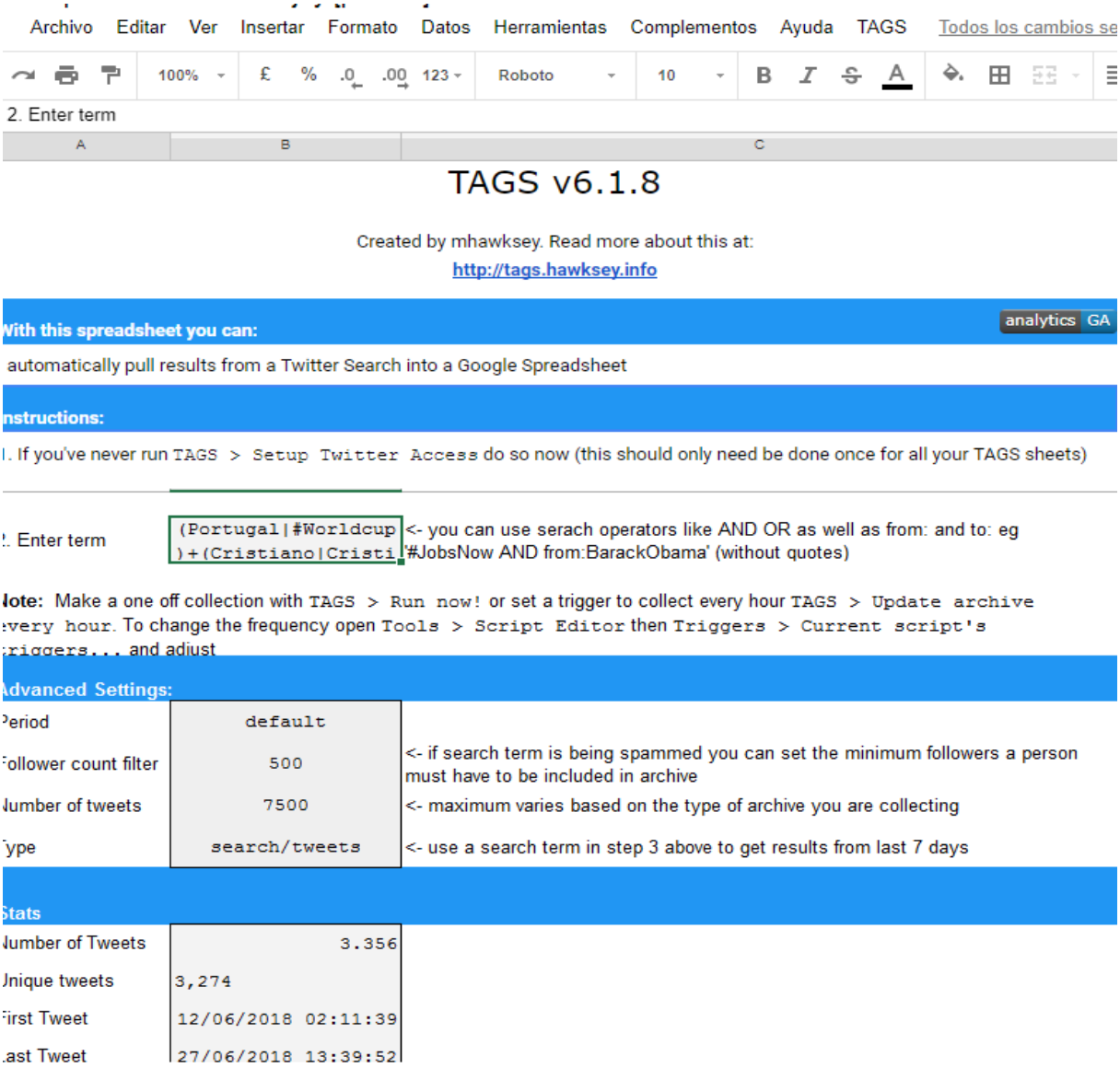

Ilustración 4 – Ejemplo de página de búsqueda en TAGS V6 (Settings)

(Las imágenes relativas a TAGS V.6 fueron recolectadas mediante la experiencia propia de uso y pueden visualizarse en los enlaces de descarga de la tabla anterior)

TAGS V6 ofrece múltiples funcionalidades, entre ellas se incluyen todas las que ofrece Excel (como por ejemplo una herramienta de filtrado por columnas que permite limpiar los datos).

Entre sus funcionalidades destaca la creación automática de hojas con gráficos, clasificaciones… Se procede a continuación a explicar brevemente cada una de las hojas con sus características:

a) Hoja Archive. Muestra los tweets que se han descargado y los datos relativos a él. Éstos datos son el nombre del usuario que ha publicado el tweet, personas a las que sigue y número de seguidores, el contenido del tweet, la fecha y la hora de publicación, ubicación, enlace a la foto de perfil, idioma del tweet, etc. En la ilustración 5 se muestran algunos de los datos (los más relevantes). Es en esta hoja sobre la que se trabaja, ya que los cambios en ella afectarán al resto de hojas.

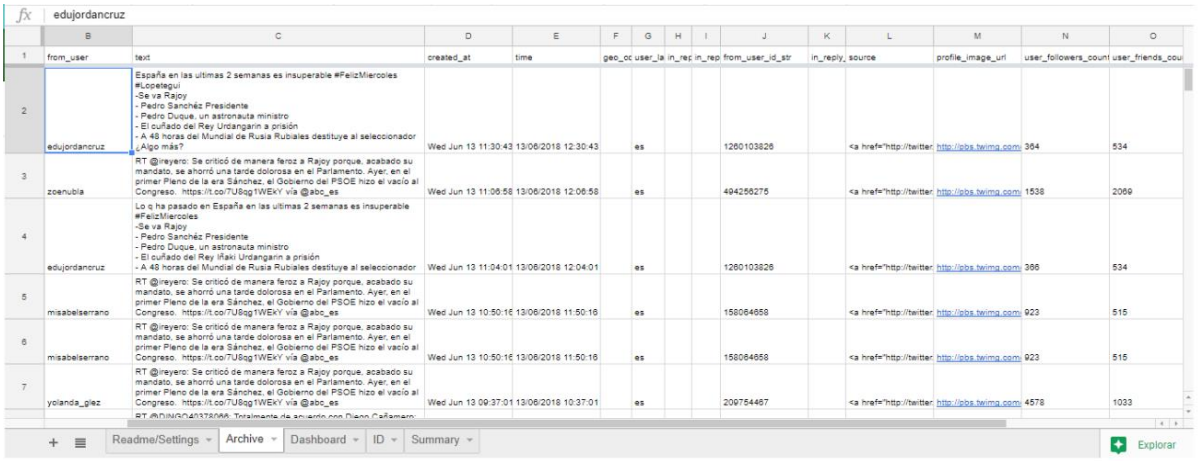

Ilustración 5- Hoja Archive.

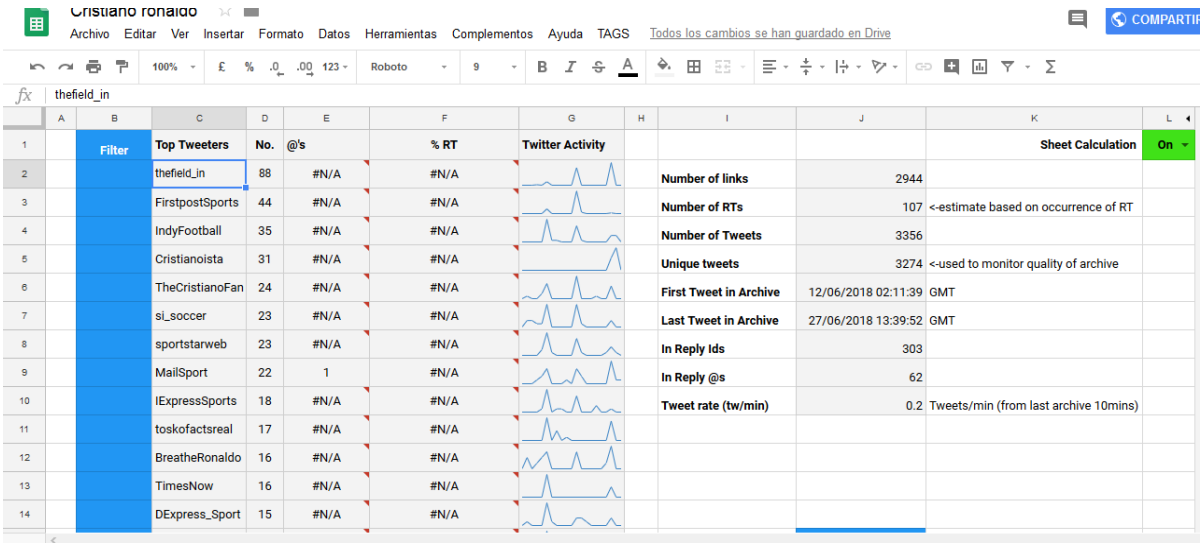

Ilustración 4 - Hoja Summary.

b) Hoja Summary: Mediantes tablas, ésta hoja nos aporta métricas interesantes para su estudio. Éstas son:

La primera tabla contiene aquellos usuarios qué mayor cantidad de tweets han enviado a la red, así como la cantidad de menciones, el impacto del tweet (%RT) y una pequeña línea temporal. La segunda tabla muestra el total de links, retweets, así como la hora y fecha del primer y último tweet recolectado. En este caso, se eliminaron los resultados que contenían retweets, por lo que no es de mucha utilidad, sin embargo podemos conocer los usuarios que más han publicado.

c) Hoja Dashboard: Ésta hoja muestra gráficos.

El primer gráfico de columnas muestra los usuarios que más tweets han publicado.

El segundo gráfico es interactivo, permite interactuar con él. Enseña la línea temporal de los tweets. Es decir muestra cronológicamente cuando comenzaron a publicarse tweets sobre este tema y se puede establecer una franja temporal

El tercer gráfico es un gráfico cronológico como al anterior, pero es global y estático.

Por último existe una tabla que muestra aquellos tweets con más RT's.

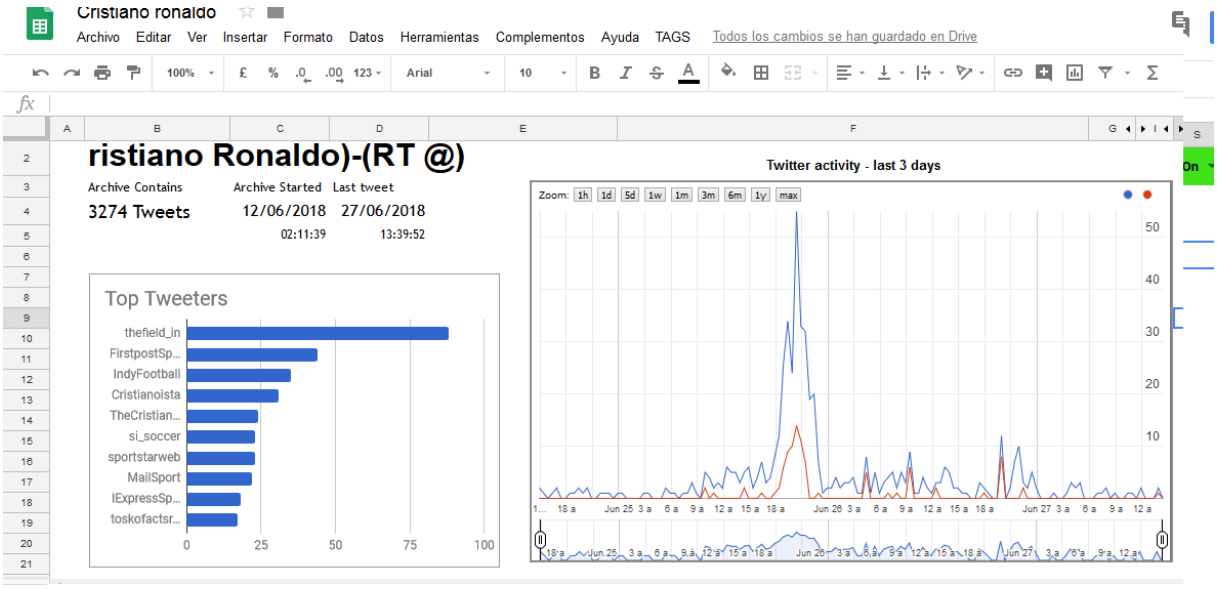

Ilustración 6- Hoja Dashboard

d) Hoja ID actualmente no está disponible

e) TAGS V6 incluye además (si compartimos la hoja, es decir, se pone visible para todo Internet) una funcionalidad extra, que permite visualizar los datos que hemos extraído mediante un mapa de grafos. Los usuarios son los nodos, y las relaciones las menciones entre ellos. También podemos ver los hashtags más empleados en los tweets así como los que más publican y los que más mencionan. Si se pincha en uno de los nodos podemos ver cómo se conecta con el resto de usuarios. Gracias al filtrado de la hoja de Excel se ha podido extraer un mapeo de los usuarios de habla inglesa que coincidían con los términos de búsqueda. Un punto en contra de esta herramienta es que no permite acotar el mapa de forma que se muestren sólo los usuarios con conexiones.

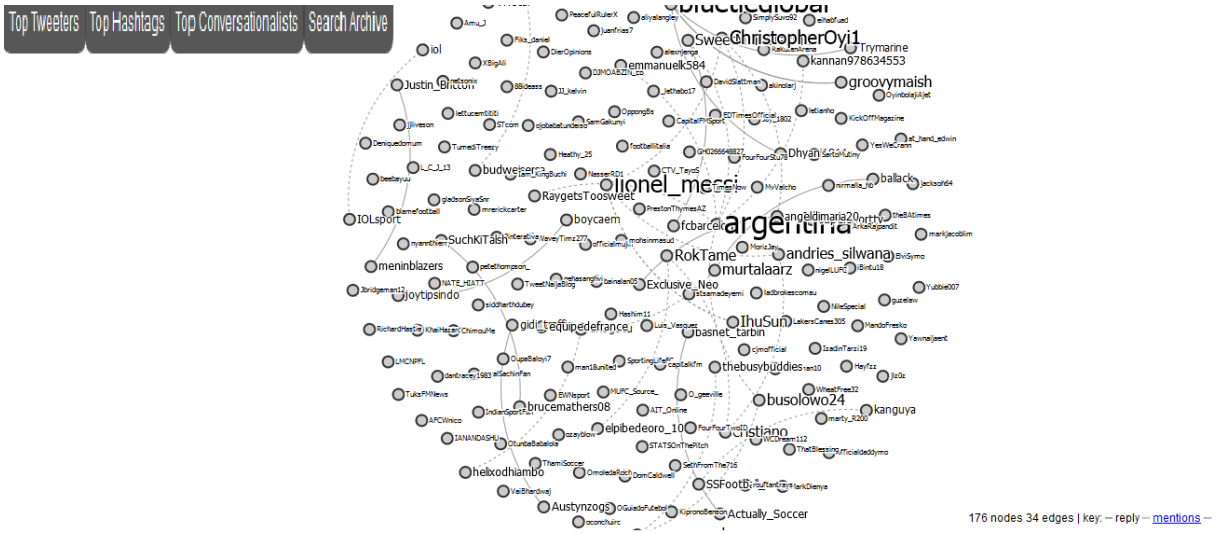

Ilustración 7-Visualización de los twitteros de habla inglesa que mencionaron 'Messi' y Argentina en sus tweets

https://tags.hawksey.info/tagsexplorer/

Antes de trabajar con los datos es necesario establecer una serie de filtros que permitan trabajar con ellos. Para limpiar los datos de aquella información que no nos sirve (ruido) vamos a aplicarle un filtro a la columna 'text' que contiene la información que se quiere analizar.

En este caso lo que se ha eliminado ha sido:

a) Los tweets que contenían enlaces a otras páginas, ya que contaminarían el contenido (http), tras lo que quedaron 604 tweets que contenían el contenido que se quería procesar.

b) Se han eliminado símbolos tales como: #,@,!,?,|, *,",'. Además de las palabras vacías. Ésta medida se ha tomado sólo para los términos en inglés con la finalidad de poder crear una nube de los términos del documento. Para el análisis de la polaridad el contenido de los tweets ha sido íntegro.

Ya que no hay retweets (contenido duplicado) tampoco se ha considerado eliminar las menciones ('@') ya que esto limitaría la cohesión del texto). A continuación se ha procedido a

separar los tweets por idioma, de manera que se puedan procesar individualmente para trabajar con ellos en una aplicación de procesamiento de datos.

Para realizar el análisis de sentimientos sobre los tweets que se han recolectado se necesita alguna herramienta que permita utilizar un modelo de clasificación de texto y de clasificación de sentimientos. La herramienta empleada para ello es Add-in for Excel de MeaningCoud.

Meaningcloud

MeaningCloud Add-in for excel' es un complemento de Excel desarrollado por la plataforma

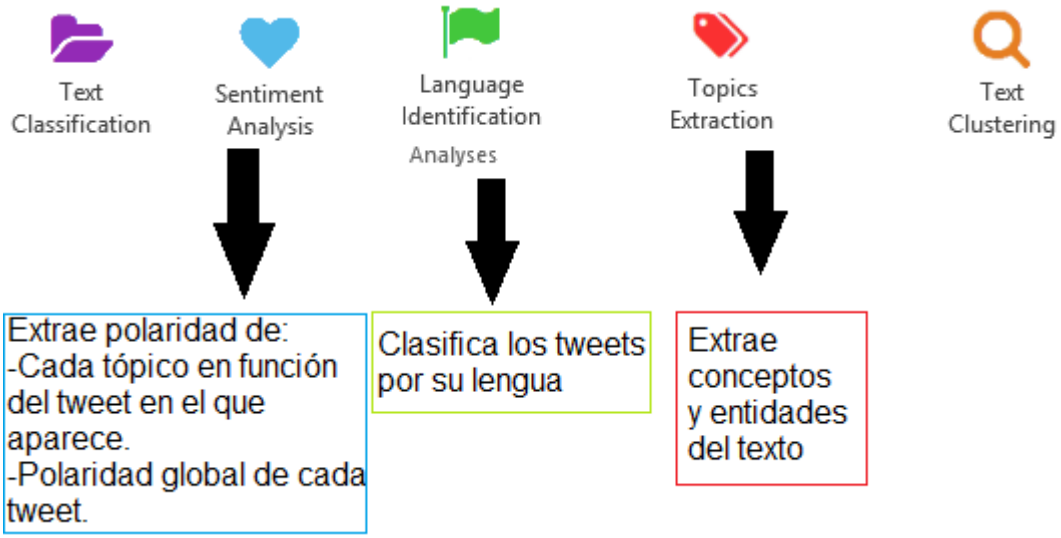

Ilustración 8-Funcionalidades de Add-in for Excel

(Ilustración extraída de la aplicación)

MeaningCloud, especializada en minería de datos. Ofrece herramientas capaces de extraer información de cualquier texto no estructurado, aunque es necesario crear una cuenta para poder utilizarlas. En la ilustración 8 se muestran las herramientas que ofrece, y se explican las utilizadas.

La facilidad principal es que permite realizar el trabajo sobre Excel, lo que la hace una herramienta para casi todos los públicos. Ésta plataforma no exige de pago por uso, sin embargo la cantidad de texto a analizar no es ilimitada, ya que la versión gratuita permite el análisis de 20000 peticiones mensuales, que se renuevan cada cinco meses.

Ofrece modelos de clasificación basados en estadísticas y en reglas, también ofrece la posibilidad de utilizar ambos a la vez, es decir, un modelo híbrido. Éste modelo presenta la ventaja de que el etiquetado no debe de ser muy exhausto, ya que sólo se empleará cuando el modelo estadístico no sea exacto. Sin embargo el modelo híbrido es mejor emplearlo cuando se hayan construido las clases ya que la evaluación es más complicada aquí (la web recomienda comenzar con el modelo estadístico).

Se planteó la idea de crear una clasificación acorde con el trabajo, ya que esto permitiría crear una clasificación de contenido mucho más específica. Sin embargo la plataforma ofrece numerosas clasificaciones que funcionan bien, además de la exhaustividad que requeriría el trabajo, (la ilustración 9 resume los pasos a seguir para ello) por no hablar de que esto obligaría a realizar más peticiones al servidor (limitar la prueba de uso). Finalmente que se emplearon las taxonomías/tesauros en distintos idiomas que proporciona Meaningcloud. Se empleó la clasificación de 'The International Press Telecommunications Council (IPTC)' para los idiomas inglés, español y portugués ya que eran los que más clases contenían.

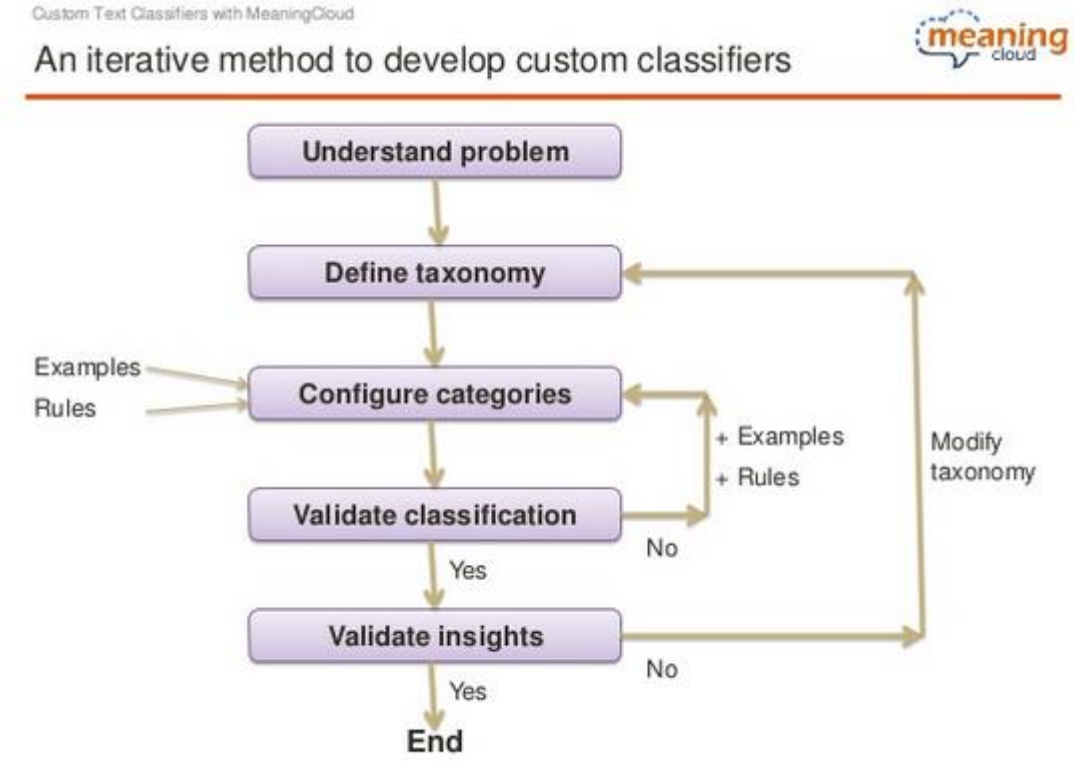

Ilustración 9- Pasos para desarrollar una clasificación

https://www.meaningcloud.com/blog/learn-to-develop-custom-text-classifiers-recorded-webinar

Para realizar el análisis de la polaridad fue necesaria la construcción de un diccionario basado en el contenido de los tweets, de manera que se definieran las entidades y conceptos más relevantes de los que se deseaba obtener información (también se define el lugar que ocupa en la ontología, que puede ser definido como el usuario desee). La extracción de los tópicos fácil gracias a la funcionalidad que ofrece la aplicación, lo único que fue necesario fue definirlos en un Excel siguiendo una estructura para importarlos en la web. Se incluyeron 27 tópicos de los 30 máximos que permite la aplicación. En él se incluyeron las entidades y conceptos más

importantes entre todos los tweets (los más repetidos).

A continuación se procedió al análisis de sentimientos de los documentos que contenían los tweets de Messi y Cristiano. Para asegurar la ausencia de fallos se volvió a separar el contenido por su lengua, ésta vez con la herramienta de identificación del lenguaje que ofrece la aplicación. Puesto que ambos contenían tweets en varios idiomas (de los cuales lo más importantes estaban en inglés, español y portugués) se usaron clasificaciones de distintos idiomas. Una vez llevado a cabo todo este procesamiento se contaron las celdas que contenían las diferentes polaridades globales de los mensajes. Entre los tweets encontramos las siguientes tipologías:

- Muy positivo: P+
- Positivo: P
- Neutral: NEU. No es positivo ni negativo, aunque indica que el mensaje posee contenido subjetivo
- Ninguno: NONE. No contiene contenido subjetivo, es decir, que no expresa ningún sentimiento
- Negativo: N
- Muy negativo: N+

6. Resultados

Antes de interpretar los resultados es interesante conocer los términos que componen al documento (ilustración 10). Como podemos comprobar worldcup es el término de mayor tamaño, ya que es el evento principal por el cual se relaciona a Cristiano y a Messi en los tweets. Si observamos los términos que los identifican, ya se augura un ligero mayor tamaño de Cristiano Ronaldo sobre Messi. El resto de términos más empleados indican en su mayoría países o continentes, ya que los usuarios de esta forma muestran el apoyo a su región.

Ilustración 10- Tag Cloud

(Elaborada con Voyant Tools - https://voyant-tools.org/)

Los resultados obtenidos de los 604 tweets se expresan mediante las tablas que se muestran a continuación:

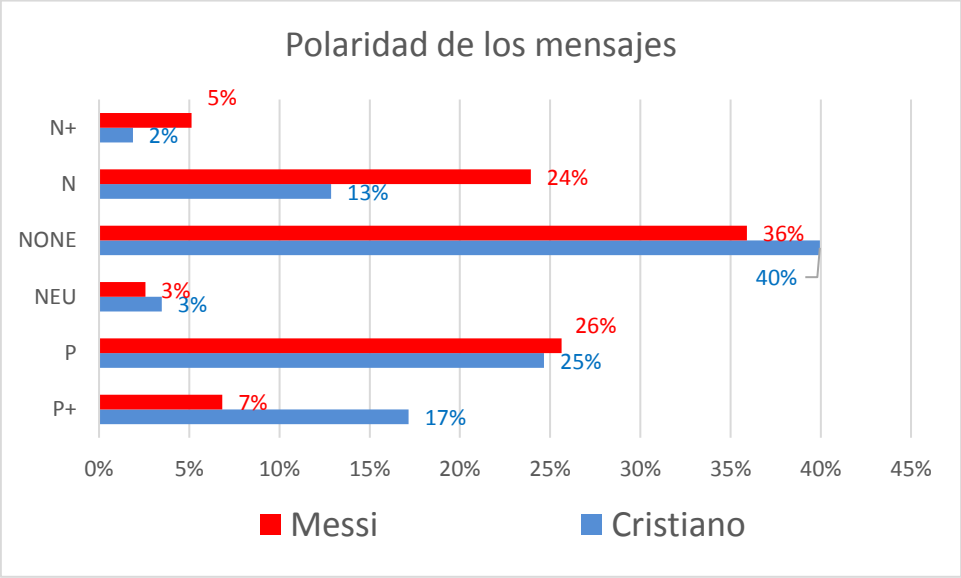

Ilustración 11- Polaridad de los mensajes

Los resultados se normalizaron sobre 100 para su representación. De los 604 tweets, 373 contienen el término Cristiano|Ronaldo y los restantes Leo Messi|Messi. La ilustración 11 contiene la gráfica que muestra la polaridad que contienen los mensajes hacia estos futbolistas.

Del total de los mensajes, se puede comprobar que la mayor parte corresponde en ambos casos a aquellos tweets que no expresan subjetividad (es el caso de aquellos mensajes que expresan los resultados de fútbol, o carecen de importancia léxica en la mayoría de los casos).

Como se puede comprobar Cristiano Ronaldo cuenta con un 17% de tweets muy positivos, mientras que Messi acumula sólo el 7%. Aunque Messi le supera en votos positivos (P), no se puede obviar el hecho de que cuenta con mayor cantidad de votos negativos y muy negativos. Éstos hechos demuestran que Cristiano ha despertado más opiniones positivas, mientras que Messi ha despertado más opiniones negativas a lo largo de los octavos del mundial de fútbol de Rusia de 2018.

Respecto al origen de los mensajes, la ilustración 12 muestra la polaridad de los mensajes por idiomas. Así pues los usuarios de habla inglesa concentran la mayor parte de los comentarios, mostrando actitudes tanto positivas como negativas (aunque bien es cierto que las positivas concentran mayor número).

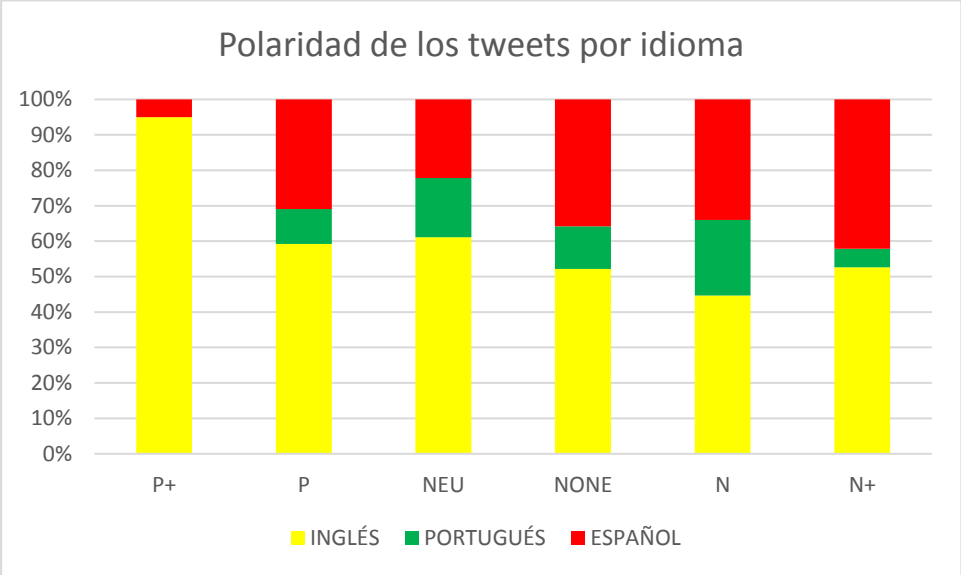

Ilustración 12- Polaridad de los tweets según el idioma

7. Conclusiones

Respecto a los resultados del ejercicio práctico:

- Tanto Messi como Cristiano son dos jugadores que en la actualidad juegan en equipos de fútbol de España, sin embargo una gran parte de los tweets que contenían a cualquiera de estos dos personajes públicos procede de usuarios de habla hispana. Sería necesario conocer las ubicaciones de las personas que manifestaron sus opiniones, en caso de querer conocer más profundamente acerca de la procedencia de estos usuarios y así aclarar de dónde se reciben peores comentarios.
- Tras el análisis de datos se observó que muchas de las entidades se trataban de otros

jugadores que los usuarios reflejaban en sus tweets. Este punto es interesante, ya que eran comparados entre ellos y especialmente con Cristiano y Messi. De aquí surge la hipótesis de que si se extraen tweets que mencionen a los compañeros de su equipo tal vez el procesamiento de los datos permita producir una polaridad más exacta. Realizar esta búsqueda además permitiría establecer un ranking entre los jugadores para conocer por ejemplo, el más querido entre ellos.

- Posiblemente una mejor y mayor extracción de tweets habrían posibilitado obtener unos resultados más apurados (más exactos).

Respecto al uso de las aplicaciones:

- Ambas presentan una interfaz fácil para el usuario, así como de una experiencia de uso cómoda y rápida. El método que emplea la recolecta y su posterior filtrado combinan bastante bien (en buena parte gracias a Google y sus hojas de cálculo).
- Tanto la extracción de tweets como el procesamiento de los datos en ambas herramientas me parece algo pobre, ya que en la actualidad existen muchas otras aplicaciones con resultados mucho mejores sin aumentar la complejidad en el proceso. Theysay Api, con un interfaz sencillo y claro permite encontrar el sentimiento que expresa cualquier texto que se le introduzca.
- Se podrían haber obtenido unos resultados más precisos si se hubiera construido una taxonomía antes del análisis.
- Aunque el análisis no permite extraer los sentimientos de facto, el análisis de polaridad es a grandes rasgos una forma de medir lo que un público siente. Y esto es interesante aun así para las organizaciones. Todas aquellas que cuenten con pocos recursos pueden permitirse realizar este tipo de análisis, así como estudiantes, casas de apuestas o emprendedores que deseen conocer los gustos de una población.
- Como puntos a mejorar creo que el uso de la interfaz de Add-In for Excel está bien actualmente, sin embargo trataría de intentar ofrecer a los usuarios más facilidades a la hora de construir las clasificaciones. Además añadir alguna funcionalidad que permita identificar los sentimientos a través de la polaridad y el contenido del texto favorecería enormemente la aplicación.

Respecto a la parte teórica:

- Este trabajo recoge a grandes rasgos los puntos principales necesarios para comprender el funcionamiento de los sistemas de aprendizaje, sin embargo algunos puntos interesantes no han sido tratados, como por ejemplo el uso de las taxonomías y

sus tipos.

- De los trabajos leídos uno de los que más me ha interesado ha sido el de (Lin et al. 2017) ya que trata de refinar a los usuarios de un blog mediante una clasificación de sentimientos, lo que en mi opinión guarda muchas futuras posibles aplicaciones. La posibilidad de refinar a los usuarios por sus gustos tendrá muchas utilidades en un futuro no muy lejano, ya sea para que algunas empresas desarrollen campañas de mercado basándose en nuestros perfiles comerciales así como en la lucha contra el terrorismo.
- El análisis de sentimientos es sin embargo aún, un conglomerado de técnicas de múltiples disciplinas que no asegura un modelo confiable al completo. Es necesario seguir perfeccionando los algoritmos. Algunas redes sociales ya se han dado cuenta de esto, como Facebook que cuenta con diferentes estados de ánimo que se emplean para ponderar cualquier publicación.

BIBLIOGRAFÍA

Agrawal, R. & Shafer, J.C., 1996. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp.962–969.

Alberich, M., 2007. Procesamiento del Lenguaje Natural - Guía Introductoria. *Guía Introductoria*, p.27. Available at: https://es.slideshare.net/menamigue/procesamiento-del-lenguaje-natural [Accessed June 3, 2018].

Batagelj, V. & Mrvar, A., 2004. Pajek — Analysis and Visualization of Large Networks. In Springer, Berlin, Heidelberg, pp. 77–103. Available at: http://link.springer.com/10.1007/978-3-642-18638-7_4 [Accessed June 23, 2018].

Biot, M. a & Academy, R., 1977. Data Mining and Analysis: Fundamental concepts and Algorithms. , 22, pp.183–198.

Cervantes, I., CVC. Diccionario de términos clave de ELE. Falsos amigos. Available at: https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/referenciacatafor ica.htm [Accessed July 3, 2018].

Correa, S.R. & Paula Andrea Benavides Cañón, 2007. Procesamiento Del Lenguaje Natural En La Recuperación De Información. *PROCESAMIENTO DEL LENGUAJE NATURAL EN LA RECUPERACIÓN DE INFORMACIÓN*. Available at: http://eprints.rclis.org/9598/1/PROCESAMIENTO_DEL_LENGUAJE_NATURAL_EN_LA_RECUPERACI_N_DE_INFORMACI_N.pdf [Accessed May 27, 2018].

Cortes, C. & Vapnik, V., 1995. Support-Vector Networks. *Machine Learning*, 20(3), pp.273–297. Available at: https://link.springer.com/content/pdf/10.1007%2FBF00994018.pdf [Accessed June 24, 2018].

van Eck, N.J. & Waltman, L., 2011. Text mining and visualization using VOSviewer. Available at: http://arxiv.org/abs/1109.2058 [Accessed July 3, 2018].

Española, R.A.E. y A. de A. de la L., 2014. Diccionario de la Lengua Española (23° Edición). *Empresario*. Available at: http://dle.rae.es/?id=R6gqDaZ [Accessed June 4, 2018].

Feldman, R. et al., 1998. Text mining at the term level. In Springer, Berlin, Heidelberg , pp. 65–73. Available at: http://link.springer.com/10.1007/BFb0094806 [Accessed July 2, 2018].

Fortuna, B., Grobelnik, M. & Mladenic, D., OntoGen » About. Available at: http://ontogen.ijs.si/ [Accessed July 3, 2018].

Gallardo, J.Á., 2014. Métodos jeráquicos de análisis de cluster. *Métodos Jerárquicos de Análisis Multivariante*, pp.1–26. Available at: http://www.ugr.es/~gallardo/pdf/cluster-

3.pdf.

Giraldo-Luque, S., Fernández-García, N. & Pérez-Arce, J.-C., 2018. *El profesional de la información information world en español.*, Swets & Zeitlinger. Available at: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2018.ene.09/38562 [Accessed June 1, 2018].

Gómez-Torres, E. et al., 2018. Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en Ecuador (Influence of social networks on the analysis of sentiment applied to the political situation in Ecuador). , 1, pp.67–78.

Gonzalez-Bailon, S. et al., 2012. Assessing the Bias in Communication Networks Sampled from Twitter. *SSRN Electronic Journal*. Available at: http://www.ssrn.com/abstract=2185134 [Accessed July 3, 2018].

Hearst, M.A., 1999. Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -, pp.3–10. Available at: http://portal.acm.org/citation.cfm?doid=1034678.1034679.

Hu, Y.H. & Chen, Y.L., 2006. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision Support Systems*, 42(1), pp.1–24. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167923604002052 [Accessed July 3, 2018].

Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3), pp.283–304. Available at: http://link.springer.com/article/10.1023/A:1009769707641.

Huecas, G. & Salvachúa, J., Filtros Colaborativos y Sistemas de Recomendación. Available at: https://es.slideshare.net/ghuecas/filtros-colaborativos-y-sistemas-de-recomendacin [Accessed May 29, 2018].

José Solano Rojas, B., Tareas de la minería de datos: clasificación CI-2352 Intr. a la minería de datos.

Justicia de la Torre, M. del C., 2017. *Nuevas técnicas de minería de textos: Aplicaciones*, Available at: https://dialnet.unirioja.es/servlet/tesis?codigo=110045.

V. Kalamaras, D., SocNetV - Social Network Analysis and Visualization Software. Available at: http://socnetv.org/ [Accessed July 3, 2018].

Kim, J. et al., 2017. *Advances in knowledge discovery and data mining: 21st Pacific-Asia conference, PAKDD 2017 Jeju, South Korea, may 23–26, 2017 proceedings, part II*,

Leskovec, J., Huttenlocher, D. & Kleinberg, J., 2010. Predicting Positive and Negative Links

in Online Social Networks. *Handbook of Natural Language Processing*, (1), pp.1–38.

Lin, J., Mao, W. & Zeng, D.D., 2017. Personality-based refinement for sentiment classification in microblog. *Knowledge-Based Systems*, 132, pp.204–214. Available at: http://dx.doi.org/10.1016/j.knosys.2017.06.031.

Liu, B., 2011. *Web Data Mining*, Available at: http://books.google.com/books?id=jnCi0Cq1YVkC&printsec=frontcover&dq=web+data+mining&hl=&cd=1&source=gbs_api%5Cnpapers2://publication/uuid/95AAEC46-AF3D-41FF-8050-319896FFFDE0.

Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(233), pp.281–297.

Medhat, W., Hassan, A. & Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp.1093–1113. Available at: https://www.sciencedirect.com/science/article/pii/S2090447914000550 [Accessed June 11, 2018].

Pang, B., Lee, L. & Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. *Empirical Methods in Natural Language Processing (EMNLP)*, 10(July), pp.79–86.

Pedro Larranaga, Inaki Inza, A.M., 2008. *Clustering*, Available at: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t14clustering.pdf.

Rios Alcobendas, G., 2017. *Técnicas estadísticas en análisis de redes sociales*. Universidad de Sevilla.

Rodríguez, J.M., 2012. Procesamiento de Lenguaje Natural: Modelos de Lenguaje: Introducción a N-Gramas. Available at: http://pdln.blogspot.com/2012/10/modelos-de-lenguaje-n-gramas.html [Accessed June 18, 2018].

Rubio Cortés, D., 2016. Herramienta para el análisis de opiniones y sentimientos sobre Twitter.

San Juan, U. de, 2009. Una nueva Generación. *Concepto de Sociedad de la Información*, pp.16–30. Available at: http://www.unsj.edu.ar/unsjVirtual/comunicacion/seminarionuevastecnologias/wp-content/uploads/2015/05/concepto.pdf [Accessed May 27, 2018].

Serrano-Guerrero, J. et al., 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, pp.18–38. Available at:

http://dx.doi.org/10.1016/j.ins.2015.03.040.

Vallez, M. & Pedraza-Jiménez Rafael, 2007. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. *Hipertext: Anuario Académico sobre Documentación Digital y Comunicación Interactiva.*, 5(5). Available at: https://www.upf.edu/hipertextnet/numero-5/pln.html#problematica-procesamiento-lenguaje-natural [Accessed July 2, 2018].

Venegas, R., 2007. Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista signos*, 40(63), pp.239–271. Available at: https://media.utp.edu.co/referencias-bibliograficas/uploads/referencias/articulo/1248-clasificacion-de-textos-academicos-en-funcion-de-su-contenido-lexico-semanticopdf-z8M2F-articulo.pdf [Accessed June 21, 2018].

Wilson, T., Wiebe, J. & Hoffman, P., 2005. Recognizing contextual polarity in phrase level sentiment analysis. *Acl*, 7(5), pp.12–21.

Zafra, S.M.J., Análisis de Sentimientos a nivel de aspecto y estudio de la negación en opiniones escritas en español * Sentiment Analysis at aspect level and study of negation in Spanish reviews. Available at: https://gplsi.dlsi.ua.es/sepln15/sites/gplsi.dlsi.ua.es.sepln15/files/attachments/paper19.pdf [Accessed July 1, 2018].