TRABAJO FIN DE MASTER

MASTER PROFESIONAL EN INGENIERÍA EN INFORMÁTICA

Análisis de tendencias con Big Data

Autor
José Ángel Díaz García

Directoras
María José Martín Bautista
María Dolores Ruiz Jiménez

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Diciembre de 2016

Índice general

Índice de tablas

Capítulo 1

Introducción

Actualmente nadie debería sorprenderse cuando escuche que vivimos en la *sociedad de la información*, concepto acuñado para referenciar a una sociedad cambiante y donde la manipulación de datos e información juega un papel más que relevante en las actividades sociales, culturales y sobre todo, económicas. El tratamiento de estos datos puede suponer una ardua labor, más aún cuando el volumen de estos es tan grande que los paradigmas para su procesado deben migrar hacia nuevas vertientes y aún más cuando estos datos provienen de fuentes tan dispares como nuestras tendencias en la compra diaria, el uso que le damos a una tarjeta de crédito o a una red social... Es por ello, que fruto de la necesidad del análisis y la obtención de información de estos datos en especie desestructurados y aparentemente carentes de significado surgen técnicas y herramientas capaces de procesar y obtener información útil y relevante.

Como hemos mencionado anteriormente, las redes sociales son grandes factorías de datos. Datos que una vez procesados pueden servir de ayuda para comprender temas relevantes de la sociedad actual o incluso desvelar patrones aparentemente ocultos en los hábitos de comportamiento de usuarios que pueden ser de ayuda en procesos de toma de decisiones o para diversos estudios posteriores. A este proceso se le denomina minería de redes sociales o *social media mining* y es una de las vertientes de estudio sobre la que más se investiga actualmente dentro del ámbito de la minería de datos.

Continuación, haremos una breve introducción al problema a resolver, para continuar con la puntualización de los objetivos principales del proyecto de

fin de máster y concluir la sección dando al lector una idea de la organización final de la memoria.

1.1. Problema a resolver

Vivimos en un mundo aparentemente obsesionado por etiquetar, clasificar y buscar relaciones entre todo lo que nos rodea. El buscar relaciones entre distintos factores como por ejemplo asociar la existencia de un tipo determinado de nubes con una probabilidad más alta de lluvias, es algo innato del ser humano desde tiempos inmemoriales e inherente a la totalidad de los ámbitos de estudio habidos y por haber a lo largo de la historia.

Las técnicas de minería de datos y extracción de conocimiento, tales como reglas de asociación, clustering o modelos de clasificación entre otras, no son muy distintas al menos en el concepto general de la búsqueda de relaciones en cualquier ámbito o problema. Pese a que estas técnicas están presentes en casi todas las vertientes de estudio y desarrollo con las que los seres humanos actualmente trabajan, hay ciertos problemas o enfoques en los que destacan notablemente y en los cuales son herramientas esenciales. Estos problemas, son tales como la detección de comunidades [1], la realización de diversos estudios y herramientas enfocados al marketing [4] en pequeñas y grandes compañías, la elaboración de modelos predictivos en ámbitos financieros o de seguros [5] y por supuesto la minería de redes sociales o el análisis de sentimientos [2] [3] que actualmente se ha convertido en una de las vertientes más estudiadas, dado su interés para comprender los hábitos de los usuarios desde una perspectiva de análisis más fiable comparable a preguntar de forma particular sujetas al estudio en cuestión.

Es en este último punto, la minería de redes sociales, junto con las técnicas anteriormente introducidas y que veremos con más detalle en los puntos siguientes, donde surge lo que conocemos como análisis de tendencias o **minería de opiniones**. Objeto de estudio en el que se trata de comprender o analizar comportamientos, actividades y opiniones, por ejemplo, de consumidores de cierto producto o usuarios de cierta red social. El fin de estas técnicas es por tanto la extracción de conocimiento útil que pueda traducirse en ventajas competitivas en el proceso de toma de decisiones de una pequeña o gran compañía, sin olvidar claro está las connotaciones científicas y áreas de estudio que se pueden desarrollar en el proceso.

Como hemos mencionado en el punto anterior, todas las técnicas mencionadas buscan extraer conocimiento y valor sobre los datos de sus procesos de negocio o servicios, pero, ¿qué pasa cuando el volumen de estos datos es tal que no podemos procesarlo con las técnicas tradicionales que han venido usándose en las últimas décadas? La respuesta es que de poco valdría el poder y la potencia de las técnicas de minería de datos en procesos de extracción de conocimiento, donde el volumen de datos es tan elevado, sin la fusión y utilización en conjunto de estos métodos con las nuevas técnicas de proceso de datos basados en Big Data que permitan el manejo y procesado de estos datos de una manera eficiente, útil y replicable.

Es por esto anteriormente mencionado, que el presente trabajo de fin de máster presenta una solución que podría ser enmarcada dentro del campo del análisis de tendencias o minería de opiniones y que aúna el uso de reglas de asociación acorde a los paradigmas estudiados dentro del tan sonado Big Data, para el minado de la red social Twitter de manera que nos permitan obtener valor y conocimiento sobre los datos (tweets) a lo largo de un periodo de tiempo de varios meses con la idea y finalidad de obtener patrones de opinión dentro de esta red social, centrando los esfuerzos en las opiniones sobre personajes conocidos en el ámbito mundial tales que puedan ser considerados como *influencers* dentro de esta red social.

1.2. Objetivos del proyecto

En el presente proyecto de fin de máster podemos encontrar un objetivo principal del cual posteriormente se desgranarán objetivos secundarios. El objetivo principal del proyecto es elaborar un sistema basado en reglas de asociación y Big Data aplicado a la red social Twitter, con el fin de obtener, acotar y limpiar un conjunto de datos que hable de personas relevantes dentro de ésta, como por ejemplo políticos y obtener sobre estos patrones y tendencias a lo largo del tiempo.

Este objetivo a su vez podemos desgranarlo en objetivos con menos granularidad, que serían los siguientes:

- Obtención de información sobre la minería de redes sociales, el análisis de tendencias y las técnicas de Big Data aplicadas en estos campos anteriormente.

- Estudio del estado del arte en el campo del análisis de tendencias y los paradigmas del Big Data sobre el mismo.
- Extracción de los datos provenientes de la red social Twitter para analizar y aplicar las técnicas desarrolladas.
- Desarrollo de la solución basada en reglas de asociación y Big Data.
- Pruebas y experimentación.
- Análisis de resultados y comparación de los mismos con posibles eventos políticos y sociales.

1.3. Organización de la memoria

Tras el estudio del problema e introducción al tema visto en este punto, los siguientes capítulos se centran en el estudio del estado del arte de la materia y finalmente en el desarrollo de la solución aportada. En el siguiente capítulo podemos encontrar detalladamente la planificación seguida durante la elaboración del proyecto, así como el estudio de los recursos empleados; tras este capítulo encontramos una serie de capítulos donde estudiamos el estado del arte del uso de reglas de asociación en minería de redes sociales y por supuesto, en conjunción con los paradigmas y arquitectura propuestos usando el BigData. En la parte central del proyecto se entrará en detalle en la solución aportada. Se concluirá con un estudio de los resultados y las vías futuras que la elaboración de este proyecto abre.

Capítulo 2

Planificación del proyecto

Una correcta planificación puede suponer el éxito o rotundo fracaso del proyecto en cualquier ámbito o disciplina aplicable. Si esta disciplina es a su vez la ingeniería en cualquiera de sus vertientes, la necesidad de una correcta planificación se acentúa aún más llegando a convertirse en una de las partes cruciales y más importantes del proyecto en sí.

En este capítulo haremos un resumen de la planificación del proyecto versando este en los recursos software, humano y hardware empleados así como de la la planificación temporal seguida por el proyecto.

2.1. Gestión de recursos

En esta sección se hará un repaso por los recursos utilizados, siendo estos como vimos en la introducción del capítulo tres categorías bien diferenciadas, recursos de personal, hardware y software los cuales son a su vez los tres pilares clave de un proyecto de ingeniería informática, a pesar de que en este caso que nos compete esté más enfocado al ámbito de investigación que al de la diseño y desarrollo de un producto final, como sería el caso de un proyecto íntegro de ingeniería del software.

2.1.1. Personal

El personal del proyecto radica exclusivamente en el autor José Ángel Díaz García, encargado de todas las partes del mismo, bajo la supervisión de los tutores del mismo.

2.1.2. Hardware

Elemento	Características
Procesador	2,6 GHz Intel Core i5
Memoria Ram	8 GB 1600 MHz DDR3
Disco duro	SATA SSD de 120 GB

Tabla 2.1: Especificaciones técnicas de la máquina usada.

2.1.3. Software

El software utilizado es en su práctica totalidad software libre, siendo el restante software propietario cuyas licencias vienen incluidas en el sistema operativo de la máquina usada siendo este OS X . El software usado es:

- **TeXShop**: procesador de textos basado en Latex usado para elaborar la documentación del presente proyecto. Web de TeXShop
- **Scrapy**: Librería de Python que ofrece un *framework* para la creación de *web crawlers*.
- **Twitter**: Red social de microblogging.
- **MongoDB**: Base datos noSQL usada como almacén persistente de los datos.
- **RStudio**: Entorno de Desarrollo en R donde se ha realizado la mayor parte del proceso del proyecto.
- **RSpark**: Librería para R que ofrece grandes ventajas a la hora de procesar grandes cantidades de datos bajo este lenguaje de programación.

2.2. Planificación temporal

La parte más importante de esta sección radica en la planificación temporal seguida en los meses de trabajo que el proyecto ha ocupado, siendo este elaborado continuamente etapa a etapa.

1. Obtención de información y estudio del tema: La primera parte del proyecto consistió en la obtención de información acerca de las reglas de asociación y la aplicación de estas en el ámbito de la minería de redes sociales y más concretamente en Twitter. En este primer proceso de recopilación de información también se estudiaron temas más genéricos dentro del Big Data y la minería de datos con el fin de tener una visión global de las herramientas y técnicas a estudiar y usar en el problema. Esta etapa aunque ha sido continua, tuvo especial importancia desde mediados de noviembre de 2016 a finales de diciembre de ese mismo año.

2. Estudio del estado del arte: Tras obtener buena cantidad de información y comprender el problema a resolver se dio comienzo a desarrollar un estudio exhaustivo del estado del arte de la materia así como a comenzar a desarrollar los primeros capítulos de la memoria en cuestión. Esta etapa tuvo lugar desde finales de diciembre de 2016 hasta finalizar el proyecto debido a que se ha realizado un estudio continuo de los nuevos estudios que iban apareciendo sobre la temática.

3. Selección de herramientas: Una vez fijado Twitter como medio objetivo, se llevó a cabo una investigación sobre las herramientas más oportunas para la obtención de los tuits de la red social. Esta etapa tomo lugar entre final de junio y principio de julio de 2017.

4. Obtención del dataset: Para poder comenzar a hacer pruebas y desarrollar el sistema basado en reglas, una vez elegida la herramienta, (Scrapy), se comenzó a obtener datos de la red social durante unos días ininterrumpidamente para tener un conjunto de entrenamiento suficiente. Esta tarea tomo lugar a mediados de julio de 2017.

5. Carga y preprocesado de los datos: Una vez obtenidos los datos y almacenados en MongoDB se hizo necesaria su carga y limpieza, esta tarea

no es trivial ya que necesitó de técnicas de procesado del lenguaje natural y aplicaciones de Big Data para poder trabajar con un volumen de datos muy elevado en una máquina estándar como es el caso. Esta tarea fue llevada a cabo entre los meses de julio y agosto de 2017.

2.3. Costes

Capítulo 3

Marco de trabajo

Antes de comenzar a abordar el estado del arte del análisis de tendencias o *minería de opiniones*, y más concretamente su aplicación en el ámbito de la web 2.0, es necesario introducir algunos conceptos teóricos que nos permitan comprender mejor los conceptos de los trabajos de los que discutiremos capítulo 4.2. Sobre ello, hablaremos en este capítulo.

3.1. Minería de medios sociales digitales

La reciente incursión de las redes sociales digitales en nuestro mundo han cambiado el paradigma de trabajo, económico y social de la sociedad. Dada su importancia, diversos sectores y ámbitos de estudio han puesto el punto de mira en el estudio de estos nuevos paradigmas sociales. La minería de datos es uno de los campos que estudia los medios sociales digitales originando una nueva vertiente de la misma denominada como **minería de medios sociales**.

La **minería de medios sociales**, acorde a P. Gundecha [6], comprende el proceso de representar, analizar y extraer de datos provenientes de medios sociales patrones con significado y valor. La minería de medios sociales es por tanto un campo multidisciplinar y su alcance puede ser dividido en los siguientes ámbitos de aplicación:

- **Análisis de comunidades**: Por medio de teoría de grafos, se obtienen comunidades dentro de nuestra población objetivo. Estos pueden ser usuarios con similares intereses, gustos o preferencias.
- **Sistemas de Recomendaciones colaborativos** : Se basa en la hipótesis en que usuarios similares tendrán gustos similares por lo que se pueden afinar los sistemas de recomendación teniendo estos factores en cuenta.
- **Estudios de Influencia**: Se basan en la obtención de la influencia de marcas o personas en determinados sectores.
- **Difusión de la información**: En un mundo saturado de información como el actual, saber de qué manera tendremos que difundirla para llegar a un mayor número de personas es un factor decisivo. Esto es lo que estudia este área dentro de la minería de medios sociales.
- **Privacidad, seguridad y veracidad**: Este punto se centra en la verificación automática de cuentas falsas, identificación de fuentes de spam así como de la identificación de la veracidad de información o identificación de problemas de violación de privacidad.
- **Opinion mining**: Este punto, es uno de los más estudiados en **minería de medios sociales**, podemos encontrarlo junto al análisis de sentimientos aunque como veremos en el punto siguiente hay ligeras diferencias. Dada la relevancia de cara al presente trabajo ampliaremos este concepto en la sección siguiente.

3.2. Minería de opiniones

La minería de opiniones, conocida en el ámbito internacional como *opinion mining*, es una vertiente al alza dentro de la famosa minería de textos y tiene su raíz por tanto en las técnicas de procesamiento de lenguaje natural. Si analizamos la web o las publicaciones en redes sociales, encontraremos cientos de miles de *reviews* o posts de personas acerca de un producto o marca, el potencial de analizar la finalidad de esta opinión, ver si es una crítica constructiva, si se promueve el producto o si simplemente lo crítica puede suponer una gran ventaja competitiva para las empresas y marcas,

por ello, son más las que cada vez usan estas técnicas en sus procesos de vigilancia tecnológica u obtención del *feedback* del consumidor.

Como todas las especializaciones o vertientes dentro del área de la minería de textos, en *opinion mining* tratamos por tanto de obtener información relevante y valor a partir de textos, como los que hemos mencionado anteriormente, blogs, tweets o diversas redes sociales, de ahí que sea estudiada dentro del proceso de *social media mining* descrito anteriormente, ya que podríamos decir que una técnica complementa a la otra. Pero, ¿qué es una opinión? Acorde a la definición dada por Liu en [7], una opinión es una quíntupla compuesta de los siguientes elementos:

1. **Entidad**: Puede ser un objeto, persona, servicio, lugar sobre el que se emite la opinión.
2. **Emisor**: Entidad que emite la opinión.
3. **Aspecto**: Es un aspecto que se valora sobre la **entidad** en cuestión.
4. **Orientación**: Puede ser positiva, negativa o neutra.
5. **Momento temporal**: Corresponde al momento en que la opinión se emite, ya que mismos **emisores, entidades y aspectos** podrán cambiar de **orientación** en momentos distintos, por lo que es un registro importante a tener en cuenta.

Pese a que aún no hemos entrado en el estudio de las redes sociales ni de la **anatomía** de un tweet, estos serán la fuente y la unidad mínima de información en nuestro proyecto. En puntos siguientes trazaremos un claro paralelismo entre esta definición y los tweets en concreto.

La minería de opiniones, se centrará por tanto en obtener de textos que podrán provenir de diferentes fuentes, *aspectos* de opinión, esto difiere en cierta medida del proceso de *análisis de sentimientos* [8] [9] que se centra desde un enfoque mayormente supervisado en la clasificación de estas entidades textuales acorde a sentimientos u orientación. Analizando estos *aspectos* y sus implicaciones sobre su *entidad* relacionada, podremos obtener por tanto ventajas muy relevantes como por ejemplo saber qué opinan los consumidores de una marca en concreto, posicionar productos u obtener análisis de confianza entre otras muchas aplicaciones.

En el presente proyecto, obviaremos la rama supervisada, para centrarnos en el enfoque no supervisado dentro del campo de la *minería de opiniones*, en el que no conocemos las clases o *etiquetas* a priori, aunque si que si incluya una cierta polarización básica sobre las opiniones, ya que es información relevante en el conjunto del proceso. Las técnicas de aprendizaje no supervisado que estudiaremos, son el clustering y las reglas de asociación, ambas las trataremos en el siguiente punto.

3.3. Técnicas

Pese a la gran relación que existe entre las técnicas mencionadas en 1.1, estas difieren en factores tan dispares como su utilización, su aplicación o la información que aportan sobre un problema. El análisis de estos factores nos ayudará a elegir la técnica o el conjunto de técnicas adecuadas para cada problema concreto, es decir, podemos partir de enfoques diferentes que se apoyen y retroalimenten mutuamente. Cabe destacar y diferenciar las técnicas más relevantes aplicadas a los problemas inherentes al estudio del análisis de tendencias, para así poder comprender mejor los siguientes capítulos y nuestro problema en cuestión.

3.3.1. Clustering

Las técnicas de clustering, se basan en la obtención de grupos o clases en función de un determinado conjunto de muestras o población, sin conocer a priori estas clases. Las técnicas de clustering, están enmarcadas dentro del aprendizaje no supervisado y basan la obtención de estos grupos y clases en dos factores como pueden ser la distancia o la similitud. De todos los problemas mencionados anteriormente estas técnicas son muy utilizadas en estudios relativos al marketing y estudios sociales, donde es relevante obtener agrupaciones. Un ejemplo concreto sería discernir entre los distintos tipos de cliente que compran regularmente en un supermercado, para poder ofrecer ofertas concretas en función del grupo de manera que estas sean personalizadas en función de cada cliente, permitiendo así que los beneficios se incrementen [10].

3.3.2. Reglas de asociación

La reglas de asociación dentro del ámbito de la informática no son muy distintas, al menos en el concepto general, de la búsqueda de relaciones en cualquier ámbito. Las reglas de asociación se enmarcan dentro del aprendizaje automático o minería de datos y no es algo nuevo sino que llevan siendo usadas y estudiadas desde mucho tiempo atrás, datando una de las primeras referencias a estas, del año 1993 [11]. Estas se representan según la forma **X->Y** donde **X**, es un conjunto de items que representa el antecedente e **Y** un item consecuente, por ende, podemos concluir que los ítems **consecuentes** guardan una relación de co-ocurrencia con los ítems **antecedentes**. Esta relación puede ser obvia en algunos casos, pero en otros necesitará del uso de algoritmos de extracción de reglas de asociación que podrán desvelar relaciones no triviales y que puedan ser de mucho valor. Podremos presentar por tanto a las reglas de asociación, como un método de extracción de relaciones aparentemente ocultas entre ítems o elementos dentro de bases de datos transaccionales, *datawarehouses* u otros tipos de almacenes de datos de los que es interesante extraer información de ayuda en el proceso de toma de decisiones de las organizaciones.

La bondad o ajuste de las reglas de asociación a un determinado problema, vendrá dada por las medidas del **soporte** y la **confianza**, que podremos definir de la siguiente manera:

- Soporte: Se representa como *supp (X->Y)*, y representa la fracción de las transacciones que contiene tanto a X como a Y.
- Confianza: Se representa como *conf (X->Y)*, y representa la fracción de transacciones en las que aparece el ítem Y, junto en las que aparece el ítem X.

Las reglas de asociación pueden abordarse desde dos perspectivas, solución por fuerza bruta (prohibitivo) o desde un enfoque basado en dos etapas. La primera de estas etapas es la generación de itemsets frecuentes, a partir de los cuales, en la segunda etapa se obtienen las reglas de asociación, que tendrán si todo ha ido correctamente un valor de confianza aceptable o elevado.

Su uso ha sido extendido en campos como las telecomunicaciones, gestión de riesgos, control de inventarios [12] [13] o almacenes y recientemente en el

minado de redes sociales representando en este ámbito una de las vertientes más estudiadas actualmente en campos de estudio como por ejemplo el análisis de sentimientos [14]. Dada la importancia de estas técnicas en nuestro trabajo las estudiaremos con detalle en el capítulo 4.2, donde veremos su aplicación en diversos trabajos relacionados en menor o mayor medida con el nuestro.

3.4. Big Data

Como hemos visto en puntos anteriores, la 'explosión' y expansión de la era digital ha hecho que el volumen de datos de los que disponemos, así como de las fuentes que generan estos datos se hayan multiplicado exponencialmente. Erraríamos por tanto, si pensáramos que las técnicas tradicionales de carga, procesado y análisis de datos tradicionales pudieran ser aplicadas a estos grandes volúmenes de datos, por lo que ha sido necesario la implantación y creación de nuevas técnicas capaces de lidiar con estos grandes volúmenes de datos, y a esto es lo que conocemos como *Big Data Analytics*.

3.4.1. Historia

Pese a que es un término que llevamos pocos años escuchando, su acuñamiento data del año 1998, donde el libro *Predictive data mining: a practical guide.* [15], ya hacía referencia a los grandes volúmenes de datos y sus problemas relacionados, bajo el término de BigData, pero no fue hasta entrado el año 2000 cuando empezaron a aparecer los primeros artículos académicos, que podrían enmarcarse dentro del BigData. Pocos años después, con la aparición y expansión de las redes sociales, estas empresas necesitaron nuevos paradigmas y algoritmos para procesar esta gran cantidad de información que venía de las mismas. Fue en este punto, y tras otros estudios como el llevado a cabo por Alex 'Sandy' Pentland en el MIT [16], cuando se comenzó a hablar de las **3 V's del Big Data** [17], tomando por tanto este nuevo concepto su forma actual y comenzando la expansión que le llevaría a ser hoy en día una de las 'tecnologías' más punteras.

3.4.2. Las V's del Big Data

En este punto, entraremos a hablar de las conocidas **V's del Big Data**, adjetivos que en su conjunción lo definen como tal y que en sus orígenes, fueron 3, aunque pronto se fueron complementando y extendiendo, hasta nuestros días donde el BigData quedaría caracterizado por 5 V's:

1. **Volumen**: La relación de esta palabra con el concepto Big Data es clara. Y es que el tamaño de los datos continua aumentando, hasta volúmenes de los mismos nunca antes vistos.
2. **Variedad**: Los tipos de los datos son muy distintos y provienen de fuentes muy dispares.
3. **Velocidad**: Los datos son muy volubles y deben ser recogidos y analizados rápidamente, véase por ejemplo en el concepto de una aplicación de Big Data en bolsa, donde tan solo un segundo puede suponer perdidas o beneficios muy importantes.
4. **Variabilidad**: Los datos pueden cambiar de estructura o interpretación.
5. **Valor**: En última instancia, sin valor, no hay Big Data y es que estos datos una vez procesados deben aportar conocimiento y valor a la empresa y organización.

3.4.3. Aplicaciones

Las aplicaciones del BigData, dado su interés, están presentes en numerosas áreas, algunas de las cuales pueden ser las siguientes:

- Negocios y marketing: Análisis de comportamientos en el comprador, detección de comunidades.
- TIC: En este sector los beneficios son muy relevantes y evidentes, como por ejemplo reducir el tiempo de procesamiento de horas e incluso días a unos pocos segundos.

- Salud y ciencia: En este área el BigData ha supuesto una autentica revolución. Disponer de nuevos algoritmos y formas de procesar datos más eficientes y potentes han supuesto la posibilidad de obtener el mapa genético de una persona en concreto a velocidades y costes antes nunca pensados, esto tiene grandes beneficios para la ciencia y la salud de esta persona que podrá incluso prevenir enfermedades futuras.

Todas estas aplicaciones, tienen como último fin mejorar los procesos de negocio y en ultima medida la vida diaria de las personas de a pie, lo que hace que aún cuando el concepto del Big Data esté aun en sus albores de lo que podrá ser en un futuro sus beneficios pueden notarse desde ya en el día a día de la sociedad.

Capítulo 4

Estado del arte

En este capítulo se realiza un estudio exhaustivo de los trabajos realizados en el campo del análisis de tendencias y minería de opiniones en medios sociales, así como una pequeña introducción y explicación sobre el creciente interés en la materia. Se concluye con la motivación del trabajo y la diferenciación que este aporta respecto a los trabajos relacionados que veremos a lo largo del presente capítulo.

4.1. Motivación

El ámbito que nos incumbe y que aúna, como hemos visto a lo largo de los capítulos de la introducción, técnicas de minería de datos, redes sociales y BigData, es relativamente nuevo, debido sin duda alguna a la novedad que las redes sociales ofrecen. Por poner algún ejemplo, Twitter fue fundada en el año 2006 y Facebook en el 2005, lo que nos da una media de unos 11 años de vida en las redes sociales más famosas, antiguas y usadas. Por otro lado, debemos tener en cuenta que su implantación y comercialización en la sociedad no tuvo lugar el mismo día de fundación por lo que su 'edad' sería aún menor.

Si dejamos apartado el 'problema' de la reciente novedad de las redes sociales, y nos centramos en los aspectos puramente informáticos del proyecto (BigData y minería de opiniones), también tienen un notable carácter de novedad. El BigData por su parte, es uno de los más recientes avances de la

computación a gran escala y la minería de datos, haciendo que nos encontremos aún en los albores de la explotación de esta tecnología. Por su parte, la minería de opiniones, íntimamente ligada a la minería de redes sociales y la aparición de estas por tanto, promueve un gran interés tanto en los aspectos empresariales y comerciales de la sociedad como en ámbitos relacionados con la investigación.

La novedad por tanto del estudio de estas técnicas, hace que haya pocos trabajos previos completamente relacionados con el ámbito de estudio, pero también hace que actualmente sea una de las áreas de investigación que más interés suscita entre la comunidad científica, dada la creciente importancia que las redes sociales digitales están tomando en casi la totalidad de las acciones y tareas de nuestro día a día.

A lo largo de esta sección, veremos algunos estudios concretos que nos ayuden a situarnos donde estamos y hacia donde vamos en el campo de la minería de datos y la minería de opiniones en concreto, haciendo una pequeña mención a trabajos relacionados cuyo principal objeto de estudio han sido algunas de las técnicas vistas a lo largo de la sección anterior y que son de obligado estudio para comprender y poder llevar a cabo nuestro trabajo. Concluiremos ahondado en los trabajos previos que tratan sobre las técnicas de minería de datos, redes sociales y BigData en conjunción, siendo este el ámbito de estudio final del presente trabajo.

4.2. Trabajos previos y relacionados

En esta sección, veremos estudios que guardan relación con nuestro trabajo, para facilitar la comprensión del mismo al lector, se ha dividido la sección en diferentes técnicas. Comenzaremos en *frequent itemset mining* y reglas de asociación para posteriormente ir 'agregando' técnicas que desemboquen en el estudio final de los trabajos íntimamente ligados con el del presente proyecto, que aúnan técnicas de minería de medios sociales, aprendizaje no supervisado y Big Data.

4.2.1. Reglas de asociación y *frequent itemset mining*

La minería de datos basada en reglas de asociación ha sido ampliamente estudiada como podemos ver en las referencias [18] y [19] donde se realiza una introducción a este campo de estudio. Si entramos en ámbitos de aplicación más concretos y de estudios relativamente recientes, encontramos trabajos como el que podemos ver en [20], donde se usan reglas de asociación y minería de textos para analizar las oportunidades que un determinado producto de una marca concreta tendrá en el mercado, con el fin de poder tomar decisiones sobre el mismo antes de su lanzamiento, momento el que estas muy probablemente llegarían tarde de cara a una mala aceptación por parte de los consumidores. En esta línea encontramos también el trabajo [21] donde se lleva un paso más allá el problema típico afrontado por las reglas de asociación de análisis de cestas de la compra. En el artículo, los autores tratan de identificar cambios en las tendencias de compra antes incluso de que estas lleguen a ser tendencia como tal, con el fin de anticipar las acciones de venta de una compañía.

4.2.2. *Frequent itemset mining*, reglas de asociación y Big Data

El crecimiento exponencial de los datos desde los inicios del siglo XXI hasta nuestros días y la aparición de nuevas técnicas enmarcadas dentro del tan sonado BigData que permitan el procesado de estos datos desde un enfoque no privativo en cuanto a tiempo y eficiencia respecta, han hecho que surjan multitud de estudios relacionados como los que podemos encontrar en las referencias [22], [23] y [24], cuyos estudios ofrecen nuevos algoritmos de minería de reglas de asociación basados en modelos de programación basados en BigData. Concretamente, los estudios realizados en [22] y [24] se centran en el marco de trabajo de MapReduce para obtener de manera eficiente y favoreciendo el paralelismo, *itemsets* frecuentes en el caso del primero, y reglas de asociación en el caso del segundo, por otro lado, el estudio [23] hace uso también del paradigma de programación MapReduce y la plataforma Hadoop para dar una nueva versión del algoritmo Apriori, el cual comparan con tres versiones diferentes. En esta misma línea de desarrollo, encontramos también el estudio [25] que afronta una nueva forma de obtener reglas de asociación eficientemente en grandes conjuntos de datos basada en algoritmos evoluti-

vos, este algoritmo es comparado con las versiones estándar de los algoritmos FP-Growth y Apriori, obteniendo este mejores resultados en las evaluaciones de los modelos resultantes. Por último, cabe mencionar el artículo [26], que supone un gran apoyo para el presente trabajo ya que versa íntegramente sobre minería de reglas de asociación sobre BigData, concretamente usando MapReduce sobre un cluster Hadoop, donde se analizan los retos y beneficios que la conjunción de ambas técnicas conlleva.

4.2.3. Reglas de asociación y minería de medios sociales

La reciente aparición de las redes sociales y el cambio que éstas han producido en nuestro mundo, han conseguido que un gran número de vías de investigación de campos muy diferenciados se hayan centrado en el análisis y estudio de los datos producidos por las mismas redes sociales. Por tanto, los trabajos en relación con ellas son numerosos y de diversos campos de estudio como pueden ser las ciencias sociales, los estudios financieros o las humanidades, si bien, nosotros nos centraremos en los estrictamente relacionados con las ciencias de la computación y más concretamente con la minería de datos. Algunos trabajos interesantes podrían ser por tanto el de la referencia [27] donde se realiza un estudio de diversas técnicas de minería de datos aplicadas a redes sociales en el primer caso y por otro lado la referencia [28] que presenta un enfoque más preciso sobre el análisis del uso de las drogas ilegales a través del minado de las redes sociales, para ello obtiene las etiquetas usadas por usuarios en publicaciones de Instagram y los compara con un gran diccionario de términos relacionados con el uso de drogas ilegales, una vez acotado los primeros post, identifican comportamientos comunes en los mismos. Si centramos aún más nuestro objeto de estudio, las referencias [29] y [30] presentan diversas técnicas basadas en reglas de asociación para el minado de las redes sociales online, concretamente la red social Twitter. El primer estudio usa reglas de asociación dinámicas con el fin de obtener datos sobre los hábitos y comportamientos de los usuarios, y por otro lado, de obtener datos relacionados con emociones y sentimientos alrededor de ciertos *trending topics* en el segundo.

Por último, encontramos una serie de trabajos que abordan el problema de la minería de medios sociales y las reglas de asociación y que guardan una es-

trecha relación con el problema que en este proyecto resolveremos. Debemos mencionar por tanto el trabajo [31], donde los autores proponen un minado de reglas de asociación sobre las redes sociales para obtener reglas sobre los hobbies de los usuarios, o el trabajo [32], donde los autores proponen un análisis basado en reglas de asociación para encontrar *influencers* en Twitter, estudio que aborda el punto de **estudios de influencia** dentro de las distintas vertientes de la mineria de medios sociales y que está íntimamente ligado al nuestro, donde haremos algo similar pero capaz de procesar mayor cantidad de datos y enfocado a un personaje en concreto.

4.2.4. Opinion mining y aprendizaje no supervisado

Como ya hemos introducido en puntos anteriores, sobre minería de opiniones hay un gran número de trabajos publicados. Uno interesante debido a su ámbito de aplicación podría ser el estudio realizado en [33], donde se usa Twitter como elemento principal para obtención de datos que servirán para construir un clasificador que pueda ser usado *opinion mining* aunque este punto, lo abordan desde un punto supervisado de opiniones buenas o malas, por lo que está más ligado al análisis de sentimientos. Si nos centráramos en el enfoque supervisado de la minería de datos, podríamos dedicar un capítulo completo a los estudios realizados, pero, dado el enfoque de nuestro trabajo en el aprendizaje no supervisado, nos centraremos en estudiar algún estudio interesante que versen sobre estas técnicas.

Como por ejemplo el estudio visto en [34] donde los autores, proponen un sistema no supervisado, basado en diez tipos diferentes de reglas que se usan para identificar las características de los productos, determinar las opiniones acorde a estas características, determinar la polaridad y por último obtener un ranking de opiniones basados en la confianza y soporte de estos con las reglas de asociación. Podemos destacar, la similitud de este trabajo con el nuestro, en el que deberemos identificar características sobre un personaje y tras ello las opiniones sobre esas características.

4.2.5. Reglas de asociación, Big Data y minería de medios sociales

El creciente volumen de datos proveniente de las redes sociales, ha hecho que en los últimos años los estudios relacionados con la minería de las redes sociales haya mutado hacia la inclusión de nuevos paradigmas de programación basados en BigData. Uno de los primeros estudios que enfoca el minado de redes sociales desde un punto de vista no supervisado y mediante el uso de reglas de asociación en conjunción de BigData lo encontramos en [35]. El estudio data del año 2013 y en él, los autores detallan los problemas asociados al gran aumento de los datos que provienen de las redes sociales y del problema que los algoritmos tradicionales tienen al enfrentarse con estos grandes volúmenes de datos. Proponen por tanto una versión del algoritmo Apriori para la obtención de *frequent itemsets* basada en MapReduce y que funciona sobre clusters Hadoop en la nube, que solventa estos problemas. Tan solo unos años después, en el año 2015, Sheela Gole y Bharat Tidke en [36] afrontan de nuevo el problema de mantener en memoria grandes conjuntos de datos relacionados con las redes sociales y poder obtener de ellos reglas de asociación. La solución que aportan al problema de nuevo radica en el algoritmo MapReduce y en un nuevo algoritmo denominado como *ClustBigFIM* que aporta una nueva versión del algoritmo BigFIM, propuesto en 2013 en [37] para favorecer la extracción de *frequent itemsets* de grandes almacenes de datos, de manera que se extiende al marco de trabajo MapReduce de una manera más escalable y extensible que su predecesor.

Atendiendo a estudios más recientes si cabe, y en parte muy relacionados con nuestro problema a resolver, encontramos el estudio [38] donde Jie Yang y Brian Yecies, aplican técnicas de minería de datos basadas en reglas de asociación para obtener información de la red social china de reviews de películas Douban. Para ello, siendo conscientes del problema de la gran cantidad de datos, realizan una aproximación con BigData usando minado paralelo de reglas de asociación.

4.3. Trabajo a realizar

El presente trabajo propone por tanto, un enfoque novedoso ante el problema donde por medio de reglas de asociación y programación de alto ren-

dimiento basada en los paradigmas del BigData realizaremos el minado de la red social Twitter para obtener patrones de opinión y el estudio de la tendencia de estas opiniones acerca personajes de un determinado sector, como por ejemplo políticos, que será el caso que nos compete. Pese a que hemos visto trabajos relacionados a lo largo de la sección 4.2, ninguno de ellos se centra en el estudio de la evolución de las opiniones sobre un determinado sector que pueda agrupar a varios *influencer* en concreto, sino que se centran en la detección de los mismos acorde a ciertos patrones y tendencias de los usuarios. Nuestro estudio por tanto, propone un enfoque novedoso en el que partiendo de los datos en bruto obtenidos de la red social a lo largo de un período de tiempo, se procesaran usando minería de textos y se aplicaran técnicas basadas en aprendizaje no supervisado mediante las cuales podremos categorizar la tendencia de opiniones acerca de un particular grupo de *iinfluencers*, dentro de un determinado ámbito, esta información podrá ser visualizada en forma de nube de etiquetas pudiendo catalogar en un solo recurso gráfico a uno de los miembros de este grupo en concreto.

Capítulo 5

Dataset

En este capítulo se detalla el conjunto de datos empleado para la elaboración del sistema de minería de opiniones basado en reglas que veremos en capítulos siguientes. Como estos datos provienen de la red social Twitter, se lleva a cabo una pequeña introducción a la misma para continuar con una explicación del proceso llevado a cabo para la obtención del *data set*.

5.1. Twitter

5.1.1. Anatomía de un tuit

5.1.2. Twitter API

5.2. Obtención de datos

5.3. Especificaciones del dataset

Bibliografía

[1] Moosavi, S.A. and Jalali, M. Community detection in online social networks using actions of users. 2014 *Iranian Conference on Intelligent Systems, ICIS*.

[2] K. Kwon, Y. Jeon, C. Cho, J. Seo, In-Jeong Chung, H. Park: Sentiment trend analysis in social web environments. *BigComp 2017*, 261-268

[3] M. Pilar Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Ángel Rodríguez-García, R. Valencia-García: Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Comp. Math. Methods in Medicine* 2017.

[4] Serrano-Cobos, Jorge. Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos. *El profesional de la información*, 2014, vol. 23, n. 6, pp. 561-565.

[5] E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis. Support Syst. 50, 3 (February 2011), 559-569.

[6] Pritam Gundecha, Huan Liu. Mining Social Media: A Brief Introduction. Arizona State University, Tempe, Arizona.

[7] B Liu, L Zhang . A survey of opinion mining and sentiment analysis. *Mining text data*, 2012. Springer.

[8] S. Noferesti, and M. Shamsfard. Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews. *PLOS ONE*, 2015.

[9] Cambria E, Speer R, Havasi C, Hussain A. SenticNet: A publicly available semantic resource for opinion mining. *AAAI CSK*. 2010, 14-8.

[10] Baier D., Daniel I. Image Clustering for Marketing Purposes. In: Gaul W., Geyer-Schulz A., Schmidt-Thieme L., Kunze J. Challenges at the Interface of Data Analysis, Computer Science, and Optimization. *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Heidelberg. 2012.

[11] Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD* Rec. 22, 1993, 207-216.

[12] P. Mandave, M. Mane, S. Patil. Data mining using Association rule based on APRIORI algorithm and improved approach with illustration. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Vol. 3 Issue2 November 2013.

[13] Yong Yin, Ikou Kaku, Jiafu Tang, JianMing Zhu. Data Mining. Chapter 2, Association Rules Mining in Inventory Database (pp 9-23). Springer, 2011.

[14] R. Dehkharghani, H. Mercan, A. Javeed, Y. Saygin: Sentimental causal rule discovery from Twitter. *Expert Syst. Appl.* 41(10): 4950-4958 (2014).

[15] S. M. Weiss and N. Indurkhya. *Predictive data mining: a practical guide.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[16] A. Petland. Reinventing society in the wake of big data. Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012.

[17] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note, February 6*, 2001.

[18] Han, J.W. and Kamber, M. (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann Pulishers, Inc., San Francisco.

[19] Tan, P.N., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. Pearson Education, Inc., London, 30-336.

[20] W. Seo, J. Yoon, H. Park, B. Coh, J. Lee, O. Kwon. Product opportunity identification based on internal capabilities using text mining and association rule mining. *Technological Forecasting & Social Change* 105 (2016) 94-104.

[21] M. Kaura, S. Kanga. Market Basket Analysis: Identify the changing trends of market data using association rule mining. International Conference on Computational Modeling and Security (CMS 2016). *Procedia Computer Science* 85 (2016) 78 - 85.

[22] K. Jayabal, Dr. P. Marikkannu. An Efficient Big Data processing for frequent itemset mining based on MapReduce Framework. International Journal of Novel Research in Computer Science and Software Engineering Vol. 3, Issue 1, pp: (130-134).

[23] Lin, Ming-Yen and Lee, Pei-Yu and Hsueh, Sue-Chen. Apriori-based Frequent Itemset Mining Algorithms on MapReduce. *ICUIMC* , 2012. pp(76:1-76:8).

[24] X. Zhou and Y. Huang. An improved parallel association rules algorithm based on MapReduce framework for big data. 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, 2014, pp. 284-288.

[25] Y. Chen, F. Li, J. Fan. Mining association rules in big data with NGEP. *Cluster Computin*, 2015, 18:2, 577-585.

[26] Dr. R Nedunchezhian and K Geethanandhini. Association Rule Mining on Big Data. International Journal of Engineering Research & Technology (IJERT). Volume. 5 - Issue. 05. (2015).

[27] M Adedoyin-Olowe, M Medhat Gaber, Frederic T. Stahl: A Survey of Data Mining Techniques for Social Media Analysis. *JDMDH* 2014.

[28] YZhou, N Sani, Chia-Kuei Lee, J Luo: Understanding Illicit Drug Use Behaviors by Mining Social Media. *CoRR* abs/1604.07096 (2016).

[29] L Cagliero and A Fiori. Analyzing Twitter User Behaviors and Topic Trends by Exploiting Dynamic Rules. Behavior Computing: Modeling, Analysis, Mining and Decision. Springer, 2012 pp. 267-287.

[30] L. Maria Aiello, G Petkos, Carlos J. Martín, D Corney, S Papadopoulos, R Skraba, A Göker, I Kompatsiaris, A Jaimes: Sensing Trending Topics in Twitter. *IEEE Trans.* Multimedia 15(6): 1268-1282 (2013).

[31] X Yu, S Miao, H Liu, Jenq-Neng Hwang, W Wan, J Lu: Association Rule Mining of Personal Hobbies in Social Networks. *Int. J. Web Service Res.* 14: 13-28 (2017).

[32] F Erlandsson, P Bródka, A Borg, H Johnson: Finding Influential Users in Social Media Using Association Rule Learning. Entropy 18: 164 (2016).

[33] A Pak, P Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Lrec.* 2010.

[34] Ana M. Popescu and O Etzioni. Extracting product features and opinions from reviews. *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* Pages 339-346. 2005.

[35] Z Farzanyar, N Cercone: Efficient mining of frequent itemsets in social network data based on MapReduce framework. *ASONAM* 2013: 1183-1188.

[36] S. Gole and B. Tidke, Frequent itemset mining for Big Data in social media using ClustBigFIM algorithm. International Conference on Pervasive Computing (ICPC), Pune, 2015, pp. 1-6.

[37] S. Moens, E. Aksehirli, B. Goethals: Frequent Itemset Mining for Big Data. BigData Conference 2013: 111-118.

[38] J Yang and B Yecies. Open AccessMining Chinese social media UGC: a bigdata framework for analyzing Douban movie reviews, 2016, *Journal of Big Data*, vol 1.