



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

1. Introducción	6
1.1. Problema a resolver	7
1.2. Objetivos del proyecto	8
1.3. Organización de la memoria	9
2. Planificación del proyecto	10
2.1. Gestión de recursos	10
2.1.1. Personal	11
2.1.2. Hardware	11
2.1.3. Software	12
2.2. Planificación temporal	12
3. Marco de trabajo	15
3.1. Minería de medios sociales digitales	15
3.2. Minería de opiniones	16
3.3. Técnicas	18
3.3.1. Clustering	18
3.3.2. Reglas de asociación	19
3.4. Big Data	21
3.4.1. Historia	21
3.4.2. Las V's del Big Data	21
Análisis de tendencias con Big Data	1

3.4.3. Aplicaciones	22
4. Estado del arte	23
4.1. Motivación	23
4.2. Trabajos previos y relacionados	24
4.2.1. Reglas de asociación y <i>frequent itemset mining</i>	25
4.2.2. <i>Frequent itemset mining</i> , reglas de asociación y Big Data	25
4.2.3. Reglas de asociación y minería de medios sociales	26
4.2.4. <i>Opinion mining</i> y aprendizaje no supervisado	27
4.2.5. Reglas de asociación y análisis de sentimientos	27
4.2.6. Reglas de asociación, Big Data y minería de medios sociales	28
4.3. Trabajo a realizar	29
5. Dataset	31
5.1. Twitter	31
5.1.1. Funcionamiento	32
5.1.2. Anatomía de un tuit	32
5.1.3. Twitter API	33
5.2. Persistencia de los datos	35
5.3. Obtención de datos	35
5.3.1. Tweepy	35
5.3.2. Scrappy	36
5.4. Especificaciones del dataset	36
6. Carga y pre-procesado de datos	38
6.1. Carga de datos	38
6.2. Pre-procesado	39
6.2.1. Integración	40

6.2.2.	Limpieza	41
6.2.3.	Valores perdidos	41
6.2.4.	Selección de instancias	41
7.	Análisis exploratorio de datos	44
7.1.	Proceso exploratorio	44
7.1.1.	<i>Term Document Matrix</i>	45
7.1.2.	Reducción del problema basada en EDA	46
7.1.3.	Visualización	48
7.1.4.	N-gramas	50
7.2.	Análisis de sentimientos	50
7.2.1.	Distribución de sentimientos en Twitter	52
7.2.2.	Palabras asociadas a los sentimientos	53
8.	Minería de datos: Reglas de asociación	55
8.1.	Algoritmos usados	55
8.1.1.	Apriori	56
8.1.2.	FP-Growth	57
8.2.	Proceso experimental	58
8.2.1.	Comparativa de rendimiento de ambos algoritmos	58
8.3.	Visualización de reglas	58
8.4.	Interpretación de las reglas	58
9.	Conclusiones	59
9.1.	Resumen del trabajo realizado	59
9.2.	Conclusiones finales	59
9.3.	Líneas futuras	59

Índice de figuras

4.1. Flujo del proceso de obtención del modelo y tecnologías usadas.	30
5.1. Ejemplo de un tuit.	32
5.2. Ejemplo de un tuit que no sería enmarcado como opinión. . .	34
6.1. Arquitectura de Spark R.	39
7.1. Nube de palabras raras.	46
7.2. Frecuencia de palabras en función del tamaño.	47
7.3. Nube de palabras con 300 ocurrencias mínimo.	48
7.4. Palabras más frecuentes con al menos 1700 ocurrencias. . . .	49
7.5. 2-gramas más frecuentes.	51
7.6. Histograma de los sentimientos.	52
7.7. Palabras asociadas a los sentimientos.	53

Índice de tablas

2.1. Especificaciones técnicas de la máquina personal usada.	11
2.2. Especificaciones técnicas del cluster.	11
5.1. Especificaciones del dataset	37
6.1. Tiempos de ejecución del proceso NER con distintos métodos .	43
7.1. Ejemplo de matriz de frecuencias.	45

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 procesamiento 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 procesamiento de estos datos de una manera eficiente, útil y replicable.

Es por esto anteriormente mencionado, que el presente trabajo de fin de máster presenta, un exhaustivo estudio del campo del análisis de tendencias y minería de opiniones, así como el desarrollo de una solución que podría ser enmarcada dentro de estos campos 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 por medio de técnicas de minería de textos un conjunto de datos que hable de personas relevantes dentro de ésta, como por ejemplo políticos o artistas y obtener sobre estos patrones y tendencias, momentáneas o 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.
- Almacenamiento y procesamiento de los datos usando técnicas de minería de textos y Big Data.
- Aplicación de técnicas basadas en reglas de asociación para obtener patrones interesantes en los datos.
- 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, así como en el estudio y la documentación de las técnicas usadas para la limpieza, integración y visualización de los datos. 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 ciencia de datos, esta parte no es menos importante, ya que una correcta planificación temporal que ayude al científico de datos, a distribuir su tiempo y esfuerzos entre las distintas partes que integran un proyecto de análisis de datos (integración, preprocesado, minería de datos) puede ser de vital importancia de cara al triunfo o el fracaso del proyecto.

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 mismo.

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.

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 personal usada.

Además de la máquina personal se ha utilizado un cluster de procesamiento de datos para los procesos más costosos formado por cuatro máquinas con las especificaciones técnicas que podemos ver en la tabla 3.3.1.

Elemento	Características
Procesador	Intel Xeon E5-2665
Memoria Ram	32 GB
Núcleos	8

Tabla 2.2: Especificaciones técnicas del cluster.

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 la minería de opiniones y de las reglas de asociación así como de 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 realizó 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, 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 octubre de 2017.
6. **Limpieza de datos:** Dado que partimos de un problema no supervisado, donde los datos carecían de filtrado alguno, esta fue una de las etapas que más tiempo tomó. Tras la aplicación de técnicas básicas de limpieza en minería de textos, se aplicaron técnicas experimentales de procesamiento del lenguaje natural para filtrar los datos y poder poner el foco del problema en aquel subconjunto de datos que hace referencia a personas. Dado el volumen de datos esta etapa necesito el uso de un cluster de procesado así como de técnicas de programación paralela y concurrente que podríamos enmarcar como Big Data. Esta tarea fue llevada a cabo entre octubre y diciembre de 2017.
7. **Análisis exploratorio de datos:** Sobre el dataset final, se han realizado gráficos y estudios estadísticos básicos con el fin de conocer y

entender mejor la naturaleza de los mismos. Esta tarea fue llevada a cabo durante el mes de diciembre de 2017.

8. **Análisis de sentimientos:** Sobre los datos, se aplicaron técnicas de análisis de sentimientos para poder realizar gráficos que nos ayudaran a discernir que palabras o expresiones estaban relacionadas en nuestro dataset con sentimientos para en el paso de obtención de reglas de asociación poder polarizar en cierta medida las mismas, o tener al menos, otro enfoque subjetivo de éstas, pudiendo así desambiguar en cierta medida las mismas. Esta tarea fue llevada a cabo durante el mes de diciembre de 2017.
9. **Reglas de asociación y experimentación:** Con los datos limpios y estudiados, se obtienen un conjunto de reglas de asociación sobre la temática y se experimenta sobre el mismo obteniendo distintos conjuntos en función de itemsets frecuentes, así como de la variación de los parámetros de confianza y soporte en las reglas. Esta tarea comprendió los meses de diciembre de 2017 y enero de 2018.
10. **Elaboración de la memoria:** La memoria ha constado de una elaboración continua, ya que continuamente han ido añadiendo y refinando capítulos en función de como se avanzaba en el proceso de desarrollo y experimentación. Los meses que ha comprendido su elaboración, han sido por tanto desde primeros de febrero de 2017 hasta enero de 2018.

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. Gundeche [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 critica 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 quintupla 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 el punto 5.1.2 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 inclu-ya 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

Las 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]. Su utilidad es la de obtener conocimiento relevante de grandes bases de datos y se representan según la forma $X \rightarrow Y$ donde X , es un conjunto de ítems que representa el antecedente e Y un ítem 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.

Medidas

La forma clásica de medir 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 $supp(X \rightarrow Y)$, y representa la fracción de las transacciones que contiene tanto a X como a Y .
- Confianza: Se representa como $conf(X \rightarrow Y)$, y representa la fracción de transacciones en las que aparece el ítem Y , junto en las que aparece el ítem X .

Pese a que estas medidas son las más comunes y extendidas, hay innumerables propuestas de medidas complementarias en la literatura, tales como el **lift**, **convicción**, **factor de certeza**, **diferencia absoluta de confianza** entre otras muchas.

Obtención de reglas

Si nos centramos en la manera de obtener las reglas, estas 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. La primera etapa de obtención de itemsets frecuentes puede conllevar problemas de memoria ya que en una base de datos con muchos items o transacciones el número de estos será muy elevado, es por ello que surgen aproximaciones en el proceso de representación de itemsets frecuentes que nos permitirán obtener estos en bases de datos de gran tamaño. Estas aproximaciones son:

- Itemsets maximales: Son aquellos itemsets frecuentes para los que ninguno de los superconjuntos inmediatos al itemsets en cuestión, son frecuentes. A partir de estos podremos recuperar todos los itemsets frecuentes de manera sencilla sin tener que mantenerlos todos en memoria.
- Itemsets cerrados: Son aquellos itemsets frecuentes para los que ninguno de los superconjuntos inmediatos al itemsets en cuestión, tienen un soporte igual. Con esta aproximación, tendremos soportes e itemsets frecuentes que podremos recuperar fácilmente, aunque al ser más numerosos que los maximales mantenerlos en memoria puede llegar a ser complicado.

Aplicaciones

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 pérdidas 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 auténtica 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 última 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 aún, 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 [41], 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 minería 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 y análisis de sentimientos

Dentro de esta categorización encontramos bastantes trabajos relacionados en la literatura. Estos se centran en la obtención de reglas de asociación para

caracterizar mejor las características de los productos o de la entidad sobre la que se opina en un *review*, como por ejemplo el trabajo [35] donde los autores proponen un enfoque basado en reglas de asociación, co-ocurrencias de palabras y clustering, para obtener las características más comunes respecto a determinados grupos de palabras que representan que puedan representar una opinión, el fin del estudio es ofrecer una vuelta de tuerca al proceso de análisis de sentimientos que simplemente polariza una opinión para poder refinarlo de manera que no solo se polarice ésta sino que se pueda ver acorde a que palabras o características de opinión se ha llevado a cabo esta polarización.

También es interesante el trabajo [36], donde los autores proponen una nueva medida para la discriminación de términos frecuentes sin orientación aparente de las opiniones, lo que favorece el proceso de análisis de sentimientos posterior.

4.2.6. 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 [37]. 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 [38] 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

[39] 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 [40] 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.

Por último, cabe destacar el estudio [41], donde se estudia la evolución de los comportamientos de votantes en Twitter, antes, durante y tras las elecciones obteniendo por medio de SentiWordNet [42] la polarización de las palabras más usadas dentro de las opiniones de las personas, tras lo cual, se llevan a cabo representaciones gráficas usando nubes de palabras.

4.3. Trabajo a realizar

El presente trabajo propone por tanto, un enfoque novedoso ante el problema, donde por medio del uso de técnicas de minería de textos, minería de datos basada en reglas de asociación y programación de alto rendimiento realizaremos el minado de la red social Twitter para obtener patrones de opinión y el estudio de las opiniones y palabras relacionadas con estas opiniones acerca de personajes con cierta relevancia en el ámbito social. Por último se polarizarán las reglas por medio de técnicas de análisis de sentimientos, con el fin enriquecer más las mismas y dotarlas de mayor comprensión.

Pese a que hemos visto trabajos similares a lo largo de la sección 4.2, ninguno de ellos se centra en la aplicación de técnicas de minería de datos sobre un volumen tan bruto de datos como es nuestro caso, sino que realizan filtrados sobre temáticas concretas, como pueden ser tuits políticos, o de opiniones sobre productos. La ausencia de filtrado previo, permitirá a nuestro modelo, afrontar el problema de la minería de opiniones desde un enfoque no dirigido en absoluto, en el que trataremos de detectar tendencias y obtener información relevante acerca de las opiniones de los usuarios de Twitter,

acerca de las acciones o ideas de personajes concretos, que durante el proceso seguro aparecerán.

Para ayudar al lector a comprender las diferentes partes de las que se compondrá el proceso y los capítulos siguientes donde se detalla cada uno de los mismos, se ha elaborado un esquema del modelo. Este puede verse en la figura 4.1.

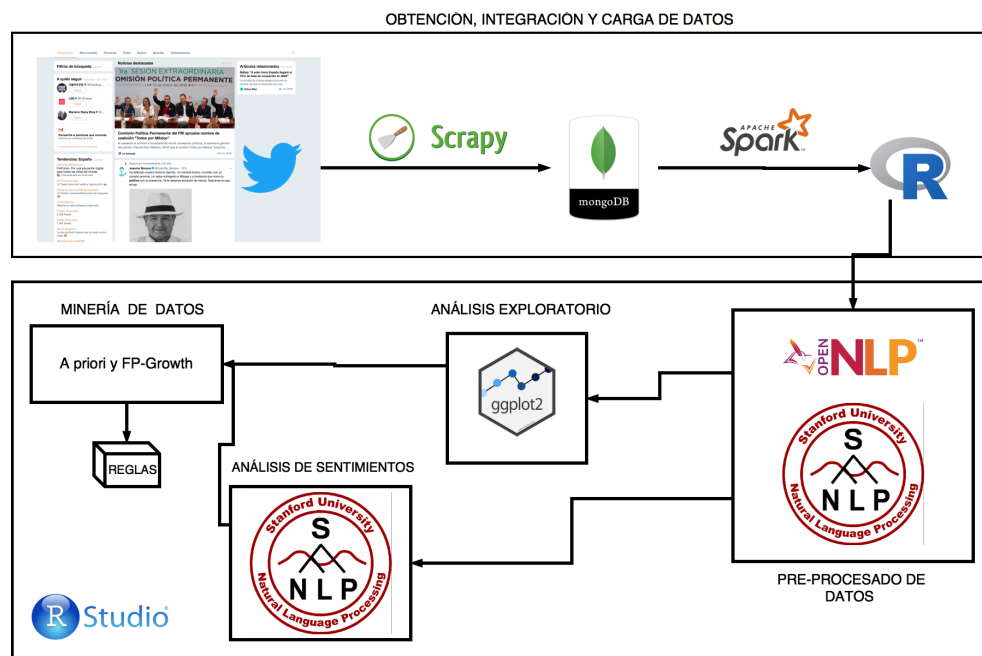


Figura 4.1: Flujo del proceso de obtención del modelo y tecnologías usadas.

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

Twitter nace en Estados Unidos en el año 2006, partiendo de la idea de los antiguos mensajes de texto (SMS) limitó el número de caracteres en cada tuit a 140 favoreciendo que el intercambio de información fuera rápido, conciso y fluido, dando comienzo a una nueva vertiente en la web 2.0 que posteriormente se conocería como *microblogging*.

El crecimiento de la red social en los últimos años ha sido exponencial, hecho que no la ha alejado de tener serios problemas de rentabilidad, pero que confirman su éxito y aceptación por parte del gran público. A principios de 2010 el número de usuarios activos al mes de la misma se fijaba en torno a los 30 millones, número claramente superado en la actualidad donde se estima en torno a los 313 millones de usuarios activos mensuales (Dreamgrow Marketing, 2017).

Recientemente, la red social ha aumentado el número de palabras en cada tuit a 280, lo que en conjunción con nuevas medidas como la facilidad para in-

cluir videos o demás contenido multimedia y lo vistoso de estas publicaciones con un diseño intachable, constatan la salud de la red social.

5.1.1. Funcionamiento

El funcionamiento de la red social en sí es muy sencillo, en esta podemos acotar un rol muy sencillo, el de **seguidor** que serán aquellas personas que quieren seguir nuestras publicaciones y de las cuales podremos ser seguidores o no, es decir, puede no es de obligada existencia el carácter bidireccional en una relación de ‘amistad’ dentro de esta red social al igual que existe en otras redes sociales como por ejemplo Facebook.

Este tipos de relaciones, son muy interesantes y han sido estudiados en la literatura como parte de la teoría de grafos y extendidos a la minería de redes sociales, para la detección de comunidades o de personas influyentes dentro de la red social [43]. Dado que nuestro trabajo versa sobre la minería de opiniones, no es menester entrar en más detalle entre las relaciones entre usuarios dentro de twitter (*retweets*, *follows*). Por otro lado, si que es necesario dado nuestro problema, diseccionar las partes que componen un tuit para ver su estrecha relación con los trabajo de minería de opiniones y la importancia de estos datos en este ámbito de la minería de datos.

5.1.2. Anatomía de un tuit

En la figura 5.1, podemos ver el ejemplo de un tuit real. Este puede tener variantes como enlaces o imágenes, pero esencialmente es texto y algunos *hashtags* o etiquetas que sirven para acotar u opinar sobre temas en concreto.

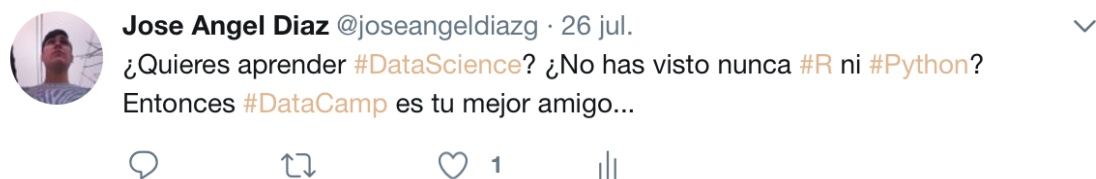


Figura 5.1: Ejemplo de un tuit.

Atendiendo al ejemplo anterior (figura 5.1), estaríamos hablando sobre *data science*, R, Python y el portal Data Camp. Cabe destacar, que aunque en este caso todos van precedidos del carácter # esto no tendría porque ser así, y alguna de estas palabras podría aparecer sin éste.

Tracemos ahora por tanto un pequeño paralelismo entre el tuit del ejemplo anterior y la definición dada por Liu de qué es una opinión que hemos podido ver en en la sección 3.2.

1. **Entidad:** En el caso de un tuit, la entidad sería sobre lo que se opina. En el caso del ejemplo anterior, sería #DataCamp.
2. **Emisor:** El paralelismo es obvio, el emisor es en el caso de un tuit la persona que lo tuitea en este caso, el usuario joseangeldiazg.
3. **Aspecto:** Recoge lo que se valora, y aunque puede parecer abstracto del anterior tuit podemos deducir que se valora la capacidad de un portal en internet para formar a las personas sobre conceptos tales como *data science*, R o Python.
4. **Orientación:** En el caso que nos ocupa, es positiva.
5. **Momento temporal:** Este elemento de la quintupla definida por Liu, al igual que el emisor siempre está presente dentro de un tuit y en este caso corresponde con el 26 de julio de 2017.

Es por tanto evidente, la relación entre un tuit y una opinión, poniendo al descubierto la importancia de este tipo de datos en el proceso y estudio de la minería de opiniones. Por otro lado, también es menester remarcar que no todos los tuits podrían enmarcarse dentro de la definición de opinión, como por ejemplo el que podemos ver en al figura 5.2 que simplemente es informativo.

5.1.3. Twitter API

Twitter abrió sus datos al mundo al hacer disponible una serie de APIs mediante las cuales se permite a terceros tanto la obtención de estos datos para su estudio como la implementación de software que trabaje sobre estos datos. Estas APIs necesitan del protocolo de seguridad y autenticación



Figura 5.2: Ejemplo de un tuit que no sería enmarcado como opinión.

OAuth, además ofrecen ciertas limitaciones a la hora de obtener los datos por lo que solo se permiten entre 150 y 300 solicitudes por hora y además hay una ventana temporal cerrará el flujo de información cada 15 minutos. Las APIs disponibles son:

- **Search API:** Obtiene los tuits de hasta 7 días, es similar a lo que nos ofrecería la búsqueda básica de Twitter en la interfaz web al buscar por un termino.
- **Streaming API:** Obtiene información en tiempo real.
- **REST API:** Obtiene los datos mediante HTTP, el formato puede venir en XML, HTML, o JSON, la limitación aquí viene definida por el número de resultados devueltos por página que no puede ser superior a 3200 tweets.

Como podemos observar, estas limitaciones pueden suponer un gran problema a la hora de obtener una gran cantidad de datos para que nuestro

trabajo tenga un cierto rigor y peso, por ello, en los puntos siguiente ahondaremos en el proceso seguido y tecnologías usadas para la obtención y almacenamiento de un gran volumen de datos a pesar de estas restricciones.

5.2. Persistencia de los datos

Tras analizar los requisitos de los datos a almacenar y las operaciones que realizaríamos sobre ellos, la opción por la que el proyecto se ha decantado ha sido MongoDB [44]. Esta base de datos es de tipo NoSQL y es la más extendida en procesos que van a trabajar con una gran cantidad de datos (Big Data).

Dado que en nuestro problema no necesitamos una gran consistencia, sino una versatilidad y facilidad a la hora de trabajar con grandes volúmenes de datos, así como una gran facilidad para conectarse a las herramientas que veremos en el punto anterior, nos hemos decantado por este sistema de base de datos.

5.3. Obtención de datos

Como hemos visto en el punto 5.1.3, la obtención de los datos mediante la API de Twitter tiene serias restricciones a la hora de permitir peticiones de datos a la misma. Es por esto, que para la obtención de los datos, se usaron y probaron distintas herramientas y librerías disponibles de manera que esta tarea fuera lo más sencilla y eficiente posible.

5.3.1. Tweepy

Tweepy [45] es una de las librerías de código abierto más extendidas entre la comunidad a la hora de conectar el lenguaje de programación Python con la API de Twitter. Esta librería ofrece distintos métodos y funciones útiles por ejemplo, para el proceso de conexión y autenticación de nuestras aplicaciones con la propia red social, así como también facilita la creación de métodos tanto para obtener datos en streaming (Streaming API) como por búsqueda (Search API).

Si nos centramos en la relevancia de esta librería respecto a nuestro proyecto, podríamos categorizarla como la primera herramienta que barajamos ya que se había visto a lo largo de los estudios de máster, pero rápidamente fue desechada ya que es imposible abolir las restricciones de peticiones a las API de Twitter, lo que hacia muy difícil, sino imposible, obtener una gran cantidad de datos en un tiempo aceptable.

5.3.2. Scrappy

Scrappy [46], al igual que en el caso anterior, es una librería *open source* para Python que nos permite mediante una *framework* de desarrollo la creación de *web crawlers*, conocidos como *spiders*. Estos *spiders*, sirven para recorrer la web, acorde a patrones previamente programados, y obtienen datos que pueden ser relevantes para múltiples funciones.

Utilizando por tanto, esta herramienta y códigos de ejemplo disponibles en las especificaciones de la misma en internet, se modificó un crawler para recorrer la web de Twitter, en un rango de fechas y lugar especificados. Estos datos, se obtenían de la página de búsqueda de Twitter por lo que permite evitar las restricciones de las API vistas anteriormente. Los pasos realizados por el crawler serían:

- Definición de parámetros, en nuestro caso fecha y lugar.
- El crawler comenzaría a buscar en la web de twitter tuits acorde a nuestros parámetros.
- Dado que el html que ofrece esta web es muy fácil de parsear, se construye un objeto con los principales datos del tuit.
- Se almacena este objeto en la base de datos MongoDB con la que el programa ha conectado.

5.4. Especificaciones del dataset

Una vez vista la naturaleza de los datos, el método de almacenamiento y el proceso de obtención de los mismos, es turno de hablar de las especificaciones técnicas del mismo.

Variable	Tipo	Uso
ID	Entero	Identifica cada tuit en la red social.
datetime	String	Contiene la fecha y la hora de emisión del tuit.
has_media	Booleano	Indica si el tuit tiene elementos multimedia.
is_reply	Booleano	Indica si el tuit es una respuesta o no.
is_retweet	Booleano	Indica si el tuit es un RT o no.
nbr_retweet	Entero	Indica el número de RTs que tiene el tuit.
nbr_favourite	Entero	Indica el número de favoritos que tiene el tuit.
nba_reply	Entero	Indica el número de respuestas del tuit.
text	String	Es el cuerpo del texto del tuit.
url	String	Urls que pueda haber en el tuit.
userid	Entero	Es el id del usuario emisor del tuit.
usernameTweer	Srting	Es el nombre del usuario emisor del tuit

Tabla 5.1: Especificaciones del dataset

El dataset está formado por un total de 1.697.229 tuits, obtenidos en EEUU, entre los meses de enero y junio de 2016 y que son de habla inglesa. Estos tuits se organizan en un dataframe de R cuyos datos y especificaciones podemos ver en la tabla 5.1

Queda constatada la potencia del proceso de obtención de los datos ya que se ha generado un dataset muy rico y que se presta a la utilización del mismo en múltiples problemas que estudiaremos en el capítulo 9 . Aún así, nuestro dataset se reducirá a un objeto de tipo *corpus* donde nos quedaremos con el texto en cuestión, ya que es la parte más interesante para aplicar técnicas de minería de opiniones.

Capítulo 6

Carga y pre-procesado de datos

En este capítulo veremos las técnicas de tratamiento de datos utilizadas para la limpieza y refinamiento de un dataset que pueda ser usado para procesos posteriores como la visualización, el análisis de sentimientos y la obtención de reglas de asociación. Dado el volumen de los datos, y la naturaleza de los mismos, donde prácticamente cada uno de los 1.7M de tuits contiene algún elemento que hace que sea totalmente distinto de los demás, esta etapa fue una de las que más tiempo requirió. En la primera parte del capítulo veremos por tanto el proceso y técnicas usados para obtener e integrar los datos en el sistema RStudio mientras que en la parte final del mismo estudiaremos las distintas técnicas de pre-procesado de datos usadas para obtener un conjunto de datos de calidad de cara a los procesos posteriores.

6.1. Carga de datos

Una vez obtenidos y almacenados los datos en MongoDB, el siguiente paso lógico del problema es pasarlos a RStudio, donde por medio de nuestros scripts comenzaríamos con el tratamiento de los mismos. Es en este punto, donde topamos con el primer problema que nos lleva a un enfoque basado en Big Data del problema dado que ninguna de las herramientas nativas de R ni las conexiones directas de R con MongoDB de paquetes como *rmongodb* pueden manejar el dataset completo para obtener 1.7M de documentos almacenados en MongoDB y pasar su contenido a un tipo de dato *data-frame* de R.

La solución, la encontramos en el paquete ***SparkR***. Este paquete [48], crea una sesión distribuida por medio de virtualización (figura 6.1) en Spark que ofrece funciones de filtrado, agregación y selección entre otras muchas, de manera similar a como se podría hacer con los dataframes nativos de R, con la diferencia de que al ser desde un enfoque distribuido, hace uso de unos objetos denominados, *SparkDataFrames*. Estos objetos, pueden manejar grandes colecciones de datos ya que los distribuyen en columnas, que pueden ser construidas como en nuestro caso con una base de datos noSQL externa, aunque hay otras técnicas viables.

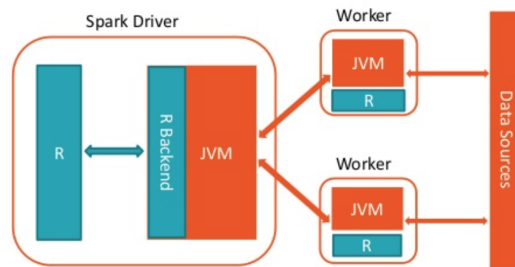


Figura 6.1: Arquitectura de Spark R.

Una vez obtenidos los datos, en nuestra sesión de Spark en RStudio, debemos pasarlos a la sesión básica de R. Esto es así debido a que Rstudio es el anfitrión de Spark, pero las sesiones difieren, por lo que debemos operar entre ambas por medio de funciones básicas de Big Data como *collect* para el caso que nos compete de fusionar nuestro *SparkDataFrame* a un *DataFrame* de R.

6.2. Pre-procesado

El pre-procesado de datos es una de las tareas más importantes en un proyecto de minería de datos. Podríamos definirlo como aquellas técnicas enmarcadas en ciencia de datos cuya finalidad es la de obtener datos de mayor calidad, más comprensibles, de menor dimensión y que puedan ser tratados apropiadamente por aquellas técnicas de minería de datos o *machine learning* que habría que aplicar después. La importancia de este proceso viene dada por motivos tales como:

1. Los datos en origen pueden ser impuros o de mala calidad, lo que conducirá a malos modelos una vez apliquemos minería de datos.
2. El pre-procesado, puede generar un conjunto de datos de dimensiones inferiores al original, con la consiguiente mejora que esto ofrece.
3. Al final del proceso de pre-procesado de datos, obtendremos datos de calidad, o al menos, mejores que si no fueran pre-procesados. Por consiguiente, los modelos basados en minería de datos, funcionarán mejor y podrán ser en muchos casos más interpretables.

Los procesos de pre-procesado de datos abarcan métodos que van desde la integración de los datos, hasta la reducción de variables u observaciones, pasando por distintos métodos de limpieza como filtrados de ruido, eliminación de palabras vacías o imputación de valores perdidos entre otros. Dado que nuestro problema, podría enmarcarse en minería de textos, las técnicas usadas vienen marcadas distintas técnicas de tratamiento de textos y procesamiento del lenguaje natural que veremos a continuación.

6.2.1. Integración

Al finalizar la etapa vista en la sección 6.1 tendremos datos en forma de `DataFrame`, o lo que es lo mismo, similares a una tabla. De cara a aplicar técnicas de minería de textos, estamos manteniendo mucha más información (tabla 5.1) que la necesaria para nuestro análisis del campo *text* de nuestros tuits. Lo ideal sería una estructura de datos que cumpliera las siguientes premisas:

- Pudiera trabajar de manera eficiente con grandes conjuntos de textos.
- Mantuviera para cada texto (tuit) metadatos de manejos de cadenas, como longitud, ids.

La solución a estas necesidades reside por tanto en los objetos de tipo **Corpus** del paquete **tm** [49] de minería de textos para R. Con simples instrucciones, tendremos integrados todos los tuits en nuestro *corpus* de tal manera, en la que cada tuit es considerado un documento independiente en nuestra colección.

6.2.2. Limpieza

El proceso de limpieza, ha sido sencillo ya que son pasos bastante estandarizados en el campo de la minería de textos. En resumen, las técnicas aplicadas han sido:

1. Eliminación de palabras vacías en inglés. A estas se le ha añadido la palabra *via*, que podemos considerar vacía en el ámbito que nos incumbe.
2. Eliminación de enlaces. Dado el alcance del problema, esta tarea ha conllevado la localización de las principales redes sociales que se usan para compartir enlaces en Twittter, tales como Facebook, Youtube, SmartURL, Vine, OwLy o BitLy entre otras varias. El motivo de esta localización has sido la elaboración de expresiones regulares que por medio de funciones se han usado para su eliminación.
3. Eliminación de signos de puntuación y caracteres no alfanuméricos.

Cabe remarcar que hemos obviado el proceso de *steaming* o lo que es lo mismo, guardar solo las raíces léxicas de cada palabra, debido a que consideramos que se podría perder interpretabilidad de cara a los procesos posteriores de obtención de reglas de asociación.

6.2.3. Valores perdidos

Tras el proceso de limpieza anterior en un volumen tan grande de datos cabe esperar que algún documento (tuit) estuviera formado por tan solo palabras vacías, enlaces o combinaciones de estos, es por ello, que por medio de filtrado básico de R se obtienen aquellos que no contienen ninguna palabra y se elimina del conjunto del dataset para evitar problemas en los procesos posteriores.

6.2.4. Selección de instancias

La selección de instancias trata de obtener un conjunto de datos de dimensión inferior al original de manera que los procesos posteriores de minería

puedan manejar mejor estos datos, u obtener información con valor de una manera menos influenciada por el ruido de observaciones no relevantes para el problema en cuestión.

Dado el gran volumen de datos que manejamos, y la variedad infinita de temáticas posibles que se pudieran estar hablando en Twitter durante los meses de obtención de datos, cabe esperar en la etapa de minería de datos una gran explosión de itemsets frecuentes y reglas de asociación, que en la mayoría de los casos no serían relevantes para el estudio de tendencias u opiniones respecto a personas, lo cual es nuestro objetivo. Por ello, parece un paso claro que antes de llegar a etapas superiores del modelo, filtremos y eliminemos aquellos tuits que no hacen referencia a personas.

Necesitamos por tanto reconocer las entidades presentes en un tuit y esto puede hacerse usando la técnica de *Name Entity Recognition* [50], de ahora en adelante NER por sus siglas en inglés. Propuesto por la Universidad de Stanford, el método está implementado en Java, aunque viene integrado en diversos paquetes para R, por lo que permite de una manera sencilla la obtención de ‘*entidades nombradas*’ en un texto en función del tipo que deseemos buscar, como por ejemplo, personas, lugares, empresas e incluso monedas. El algoritmo usado para este proceso ha sido el siguiente:

Algorithm 1 Obtiene los tuits que hacen referencia a personas

```

for all tuit in tuits do
  string  $\leftarrow$  entidad(tuit)
  if String distinto de null then
    listaNombres  $\leftarrow$  string
  else
    listaNombres  $\leftarrow$  idTuit
  end if
end for
for all elemento en listaNombres do
  if elemento.id = tuit.id then borramos el tuit
  else
    datasetFinal  $\leftarrow$  tuit
  end if
end for

```

METODO	10 TUIITS	100 TUIITS	1000 TUIITS	10000 TUIITS	1.7M TUIITS
SECUENCIAL	47s 148ms	8min 23s 756ms	1h 21min 29s 29ms	-	-
PARALELO (3 CORES)	7s 737ms	1min 15s 212ms	11min 21s 953ms	1h 47min 26s 858ms	-
LAPPLY	47s 712ms	8min	15min 54s 123ms	-	-
PARALELO (3 CORES) CON IFELSE	18s 483ms	1min 5s 682ms	8min 49s 789ms	1h 29min 39s 516ms	-
CLUSTER (PARALELO CON IFELSE)	1min 55s 682ms	1min 54s 453ms	2min 57s 810ms	20min 15s 324ms	1d 5h 0min 31s 836ms

Tabla 6.1: Tiempos de ejecución del proceso NER con distintos métodos

Este proceso, es un proceso lento, pero fácilmente paralelizable con las herramientas para programación paralela que R ofrece. Concretamente, en este punto se investigó bastante las opciones más rápidas de ejecución y como optimizar el código en R al máximo para obtener buenos resultados. Las configuraciones y pruebas realizadas, cuyos tiempos de ejecución pueden verse la tabla 6.1, fueron las siguientes.

- Ejecución secuencial.
- Ejecución en paralelo con 3 núcleos.
- Utilización de la función `lapply` en lugar de bucles.
- Sustitución de los bloques *if-else* por la función *ifelse()* de R.
- Sustitución de los bloques *if-else* por la función *ifelse()* de R + Ejecución en paralelo con 3 núcleos.
- Utilización de la función *ifelse()* de R, junto con ejecución paralela en un cluster con 31 núcleos.

Tras la ejecución del proceso *NER*, obtenemos resultados bastante aceptables donde se localizan 140.718 tuits que hacen referencia a personas. Sobre este conjunto de tuits se vuelven a aplicar un nuevo refinamiento de las técnicas de limpieza vistas en la sección 6.2.2, a las que se añade el **paso a minúsculas** de todo el contenido. Este paso había sido obviado anteriormente para favorecer el proceso de *NER*.

Como resumen final una vez finalizado el pre-procesado de datos, hemos conseguido reducir nuestro dataset de 1.7M de instancias a 140.718 limpias de caracteres raros o palabras vacías, donde además, todas y cada una de ellas hablan o hacen referencia a personas, por lo que muy seguramente con los procesos posteriores de análisis exploratorio de datos y minería de datos podremos obtener información relevante.

Capítulo 7

Análisis exploratorio de datos

El análisis exploratorio de datos o EDA por sus siglas en inglés, se centra en gran medida en indagar en los datos que recibimos de la etapa de pre-procesado para comenzar a obtener información relevante o que pueda ser de utilidad para afinar en procesos posteriores o incluso anteriores. El análisis exploratorio, tiene su mayor aliado en los gráficos, ya que una representación gráfica de los mismos será de gran ayuda para comprender el problema y poder dirigir mejor las demás acciones.

Este capítulo comienza por tanto en el estudio y visualización de los datos obtenidos en el capítulo 6, y finalizará con un análisis de sentimiento básico, necesario para posteriores etapas, donde trataremos de obtener los sentimientos asociados a las reglas de asociación que obtendremos en el capítulo 8.

7.1. Proceso exploratorio

Llegados a este punto, sabemos que nuestro dataset está formado por 140.718 tuits que hacen referencia a alguna persona, no tienen caracteres raros o palabras vacías, contienen pocos enlaces y su contenido está en minúsculas. Pero, ¿qué hay del contenido de estos tuits? ¿cómo sabremos si nuestro posterior proceso de minería de datos obtendrá opiniones relevantes sobre personas? La respuesta a estas preguntas reside en los gráficos y su interpretación.

TERM \ DOC	DOC1	DOC2	DOC3	DOC4	DOC5
<i>TERMINO1</i>	0	1	0	0	1
<i>TERMINO2</i>	0	0	0	1	0
<i>TERMINO3</i>	0	1	0	0	0
<i>TERMINO4</i>	1	1	0	0	1
<i>TERMINO5</i>	0	1	1	0	0

Tabla 7.1: Ejemplo de matriz de frecuencias.

Hasta el momento nos hemos dedicado a limpiar y procesar los datos usando técnicas aprobadas por la comunidad en cuanto a minería de textos respecta, pero es a partir de ahora cuando entra en juego el criterio y la técnica del analista de datos para comenzar a obtener información relevante y refinar los procesos de limpieza sobre los datos del problema en cuestión.

7.1.1. *Term Document Matrix*

En el ámbito de un problema de minería de textos, si queremos comenzar a trabajar y representar los datos, dependemos de una estructura de datos denominada *Term Document Matrix*. La idea es poder representar el número de documentos donde un término aparece, tal y como podemos ver en el ejemplo de la tabla 7.1, donde tendríamos 5 documentos y 5 términos.

Notar la facilidad de procesamiento de esta estructura para recuperar por ejemplo, el número de ocurrencias del término 1 (2 ocurrencias) y los documentos en los que aparece (doc2 y doc5). Cabe mencionar la posibilidad de encontrar como variante a esta estructura, otra denominada como *Document Term Matrix*, donde en la columna inicial tendríamos los documentos y en la fila los términos.

En nuestro problema, tendremos tantos documentos como tuits tenemos, es decir 140.718 y tantas filas como palabras distintas puedan haberse usado en Twitter, la variabilidad es tanta y la matriz es tan dispersa que sin procesamiento previo se hace imposible mantener una matriz tan grande en memoria por lo que es aquí donde el análisis exploratorio y la visualización toman un papel relevante.

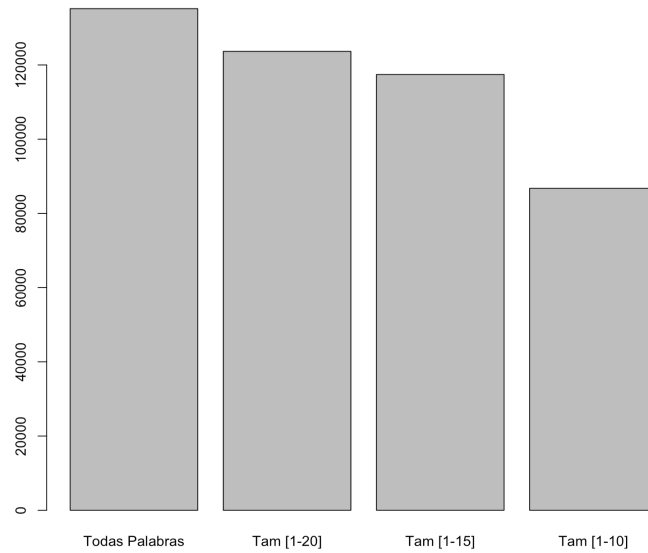


Figura 7.2: Frecuencia de palabras en función del tamaño.

nos quedaremos con aquellas en el rango de tamaño [1-13] eliminando las de tamaño 14 y superiores donde encontramos entre otras las que vimos en el gráfico 7.1.

Una vez acotadas estas palabras, nuestro problema se ha reducido a una matriz de 109790x140718, aún imposible de mantener en memoria, pero tal y como hemos mencionado antes, al estar en un problema donde el resultado final es buscar tendencias de opinión, un paso obvio sería cortar este dataset en función de la frecuencia de las palabras ya que una palabra que aparece en Twitter solo 5 o 6 veces nunca podrá ser considerada parte de una tendencia por lo que nuestra *term document matrix* final estará solo compuesta por aquellas palabras que aparezcan más de 20 veces en conjunto de tuits, reduciendo el dominio del problema a una matriz de 7323x140718, fácilmente manejable en memoria y donde por medio de suma y conteos en sus columnas y filas nos permitirán la representación de gráficos para una mejor comprensión del problema.

Al reducir el problema, podemos comenzar a realizar gráficos para obtener información relevante. Para ello, nos basaremos en dos de los gráficos más extendidos en minería de textos, las nubes de palabras y los gráficos de frecuencia o histogramas.

come work paul status shit white rd wasnt atwar tript
book head even happy 1st said simmons play
name yet well free heart want obama tonight
dr tell know youtube john kim says
bad mike never E get u live run back youre
look way best video one ed one yang oh hes jay
call start see love just trump people
world mac b think fan
hilarity good will im dont cant give
st sale hit us via day im now cant bill
jobbig today man vsal like now cant bill
chris man ben buddy can clinton
tho winim that nick kind dear last sure talk got right
make tmda luna coo de new field still miss
please drake greatdonald liked thank
tatus air watch how much thanks idol game really uk feat say
watch loimusicask last ever game going story

En este gráfico comenzamos a ver patrones interesantes, como por ejemplo los nombres propios de **Trump**, **Clinton**, **Obama**, **Drake**, **Simmons** o **Hillary**, entre otros, por lo que parece que el proceso NER descrito en la sección 6.2.4 ha funcionado correctamente. Por otro lado, la premisa de los objetivos del proyecto de obtener tendencias y opiniones sobre personas desde un enfoque no dirigido, está comenzando a cumplirse.

Análisis de tendencias con Big Data

partir de ahora cabrá esperar temas políticos y opiniones sobre los mismos en nuestro proceso de minería de datos.

Para finalizar este punto, se realizó un histograma para ver cuales eran las palabras más frecuentes. El fin de este es obtener y entender posibles tendencias en Twitter en esta época, que como hemos averiguado corresponde a las elecciones presidenciales. El punto de corte para el gráfico lo situaremos en palabras con al menos 1700 ocurrencias, es decir, haremos un estudio de las tendencias más acentuadas en esta época.

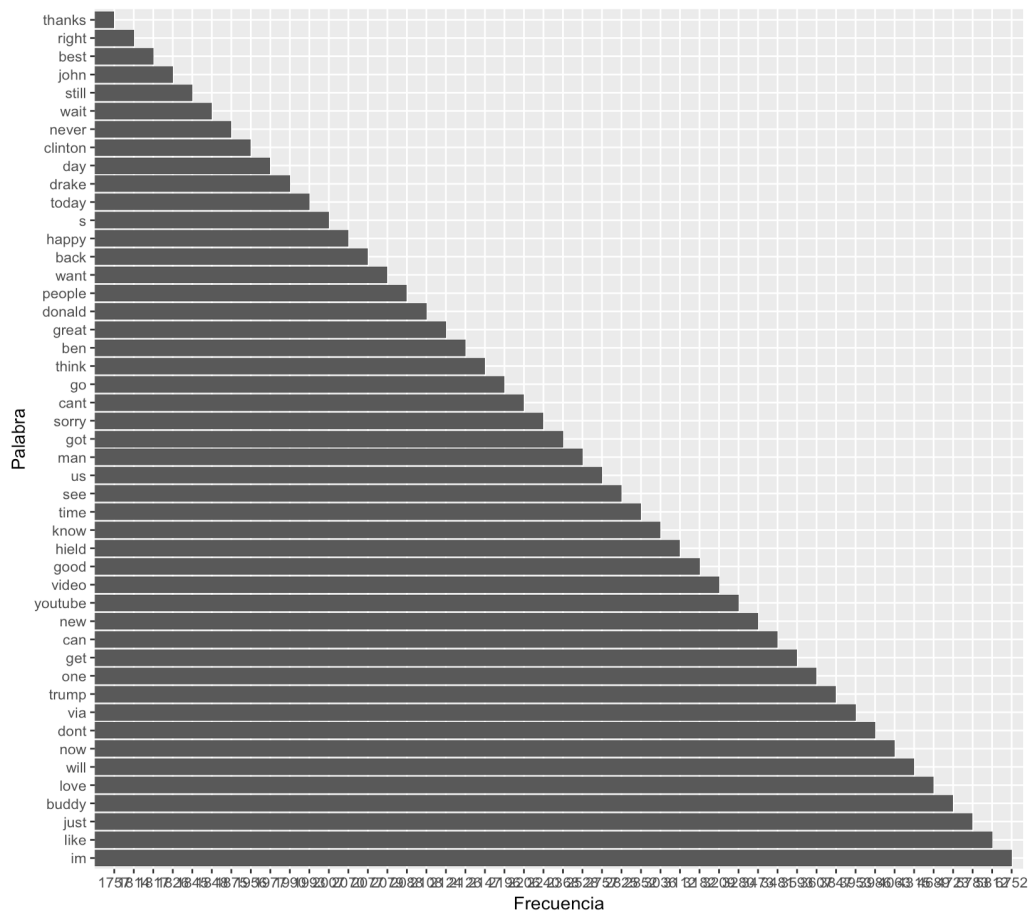


Figura 7.4: Palabras más frecuentes con al menos 1700 ocurrencias.

Si estudiamos en detalle el gráfico 7.4 vemos como la tendencia en Twitter era hablar de **Trump**, **Clinton** u **Obama**, por lo que podremos centrar

nuestros esfuerzos en el área política donde seguro obtendremos resultados e información relevante. Cabe también destacar, la presencia de nombres propios como, **donald**, que casi con probabilidad representará el nombre de pila del actual presidente de EEUU. Este factor es clave, ya que si minamos reglas de asociación, estas tendrán más fuerza y funcionarán mejor y eliminarán ruido, si pudiéramos relacionar o fusionar términos como **donald**, en un solo nombre propio del tipo **donald-trump** o **hillary-clinton**. Para realizar esto, nos basaremos en un estudio basado en *bi-gramas*.

7.1.4. N-gramas

Los n-gramas, son una técnica extendida en minería de textos y recuperación de información, que se basa en la probabilidad de co-ocurrencia. Es decir, dato un término a estudiar los **n** términos siguientes al mismo para descubrir patrones como en nuestro caso, nombres propios formados por dos palabras. Realizaremos un estudio sobre nuestros tuits basado en *2-gramas*, donde trataremos de comprobar si la premisa de que habrá una gran representación de nombres compuestos que podremos unir para obtener mejores resultados en la etapa posterior de minería de datos.

Para obtener los bi-gramas, primero usamos un *tokenizer* del paquete **RWeka** [52], tras lo cual, podremos usar normalmente la generación de gráficos de barras para ver cuales son las palabras que aparecen juntas con más frecuencia.

Acorde al gráfico 7.5, podemos corroborar lo esperado de que los bi-gramas más comunes corresponderían a nombres propios, al menos, en gran medida. Debido a esto y para favorecer el siguiente proceso de obtención de reglas de asociación, fusionaremos los nombres propios más comunes de nuestro problema, para que sean tenidos en cuenta como una sola palabra en lugar de dos.

7.2. Análisis de sentimientos

Como vimos al comienzo de este capítulo, la motivación para realizar análisis de sentimientos es el tener la posibilidad de polarizar las reglas de asociación que obtendremos en la siguiente etapa en función de los términos

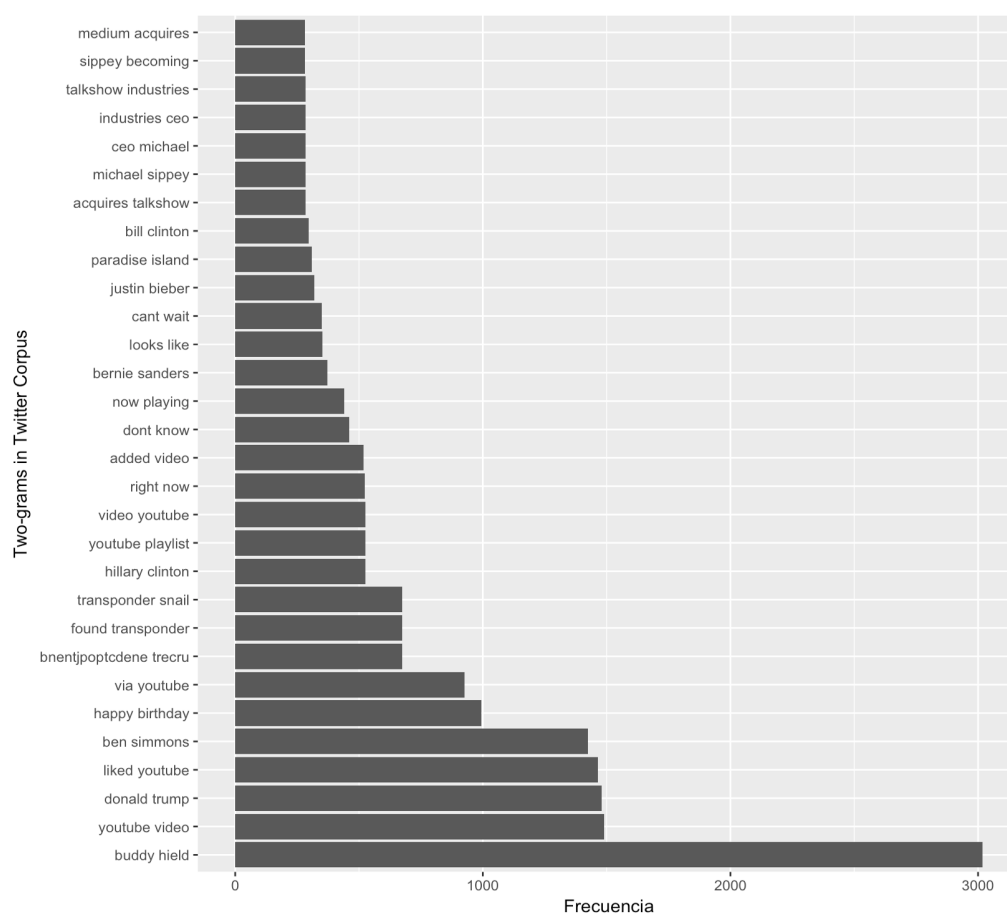


Figura 7.5: 2-gramas más frecuentes.

que aparezcan en el antecedente o consecuente de las mismas. R ofrece una gran cantidad de paquetes para el análisis de sentimientos, en este trabajo se ha usado concretamente el paquete *syuzhet*, que usa en su capa inferior el coreNLP de la Universidad de Stanford [51].

Este paquete, contiene diccionarios muy potentes para la identificación de sentimientos y permite obtener estos por distintas técnicas y algoritmos en función de lo que se pase como argumento a la función *get_sentiments()*. Otro factor muy interesante del paquete es la posibilidad de obtener gráficos de como evolucionan los sentimientos en un texto, por ejemplo, los personajes de un libro empiezan tristes y acaban felices o viceversa. Dado nuestro problema

donde en ningún caso podríamos considerar un tuit, como la continuación del otro, o que exista relación alguna entre ellos, esta funcionalidad carece de uso, y tendremos que fijarnos solo en los conteos de sentimientos y las palabras asociadas a cada uno de ellos.

7.2.1. Distribución de sentimientos en Twitter

Al igual que hicimos para obtener las frecuencias de las palabras más usadas, obtendremos las frecuencias de los sentimientos para hacernos una idea del sentimiento de los americanos en Twitter en la primera mitad del año 2016.

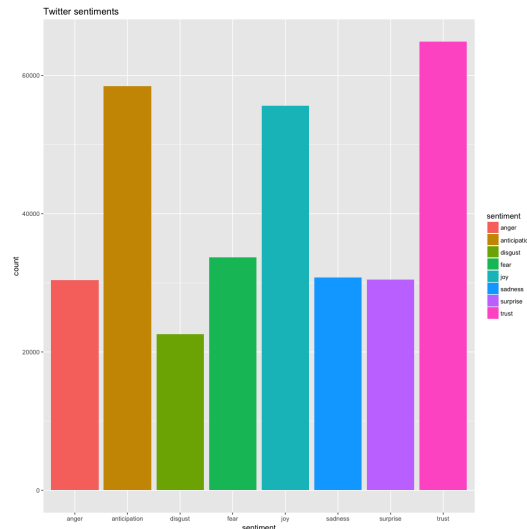


Figura 7.6: Histograma de los sentimientos.

El histograma realizado puede verse en el gráfico 7.6, donde parece que un número muy elevado de tuits hablan de **veracidad o afirmaciones**. Esto es normal ya que en twitter mucha gente afirma hechos o noticias por lo que era de esperar que este sentimiento fuera el mayoritario. Por otro lado, vemos que la **diversión** o el agrado tiene también una gran representación. Del mismo modo ocurre con la **anticipación**, este último debe ser analizado para ver que términos se asocian con anticipación, ya que es ambiguo. El resto de los sentimientos están más o menos equilibrados y frente a lo que

7.2.2. Palabras asociadas a los sentimientos

El punto anterior nos informa del número de sentimientos de cada tipo, muy útil para ver la polaridad o el ‘estado de ánimo’ el pueblo americano durante aquellas fechas, pero de cara a la obtención de que palabras representan que sentimientos, usaremos una nube de palabras, similar a las vistas anteriormente, pero categorizadas por colores en función de los sentimientos. El resultado de este gráfico podemos verlo en la figura 7.7.

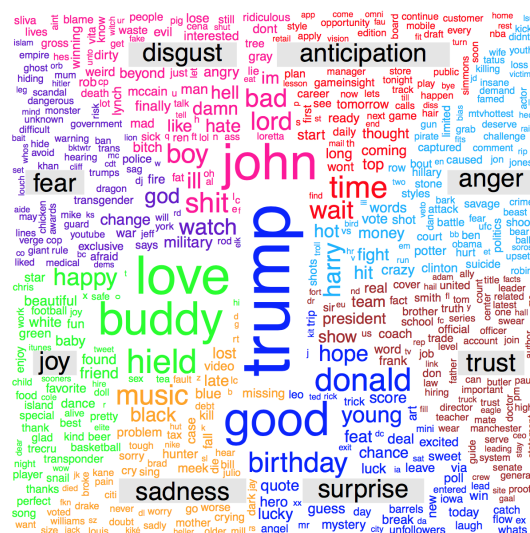


Figura 7.7: Palabras asociadas a los sentimientos.

Si entramos en el análisis del mismo, queda bastante más claro los datos con los que estamos trabajando y como se han polarizado los sentimientos. Vemos que trump, suscita sorpresa y que es la palabra más usada dentro de esta categoría. El sentimiento anticipación parece focalizarse en tiempos, y momentos temporales. Además hay cosas interesantes como relacionar **enfa-**
do con los políticos, los asesinatos o la piratería entre otros casos.

Un sentimiento muy interesante es el miedo, donde vemos un claro ejemplo de la sociedad americana. Vemos que la policía o el ejercito suscitan miedo.

pero también aparece la palabra **transgenero**, bien sabido es la homofobia del país que tenemos entre manos, por lo que parece que nuestro proceso ha funcionado bastante bien.

Si paráramos en este punto, en base al conocimiento adquirido, podríamos obtener los temas tratados en Twitter durante aquel período de tiempo, así como categorizar sin ninguna duda cuales eran las personas más influyentes e incluso que políticas o sentimientos se podrían asociar con ellos. Queda constatada la potencia del análisis exploratorio de datos y la importancia de esta etapa en un proyecto de análisis de datos, pero, también cabe preguntarse si aún hay conocimiento de valor implícito en estos datos. El tratar de resolver esta pregunta será lo que abordaremos en el siguiente capítulo, donde desarrollaremos la última etapa de nuestro modelo, la etapa de **minería de datos**.

Capítulo 8

Minería de datos: Reglas de asociación

En este capítulo nos centramos en la última etapa del modelo que vimos en el gráfico 4.1, donde por medio de técnicas de minería de datos basadas en el uso de las reglas de asociación trataremos de obtener información implícita en los datos y que sirva bien de apoyo o bien para corroborar, las ideas o premisas que hemos obtenido de nuestro proceso de análisis exploratorio de datos.

Para la obtención de reglas se usarán dos algoritmos, el **Apriori** y el **FP-Growth**, con el fin de poder contrastar información y realizar una comparativa entre los resultados ofrecidos por ambos. Se estudiarán por tanto al inicio de capítulo los detalles de cada uno de los algoritmos. Al finalizar el capítulo se verán distintas técnicas de visualización de las reglas obtenidas en el proceso experimental, así como la interpretación de las mismas.

8.1. Algoritmos usados

En esta sección veremos una introducción teórica a los algoritmos empleados en el proceso experimental. Dado que el objetivo del trabajo no está ligado al rendimiento o mejora de los mismos sino a los resultados sobre el dominio del problema, no entraremos en detalle en los mismos sino que se

mencionará la idea subyacente de su funcionamiento para facilitar la comprensión de los puntos siguientes.

8.1.1. Apriori

El algoritmo **Apriori**, fue propuesto por Agrawal y Srikant en 1994 [53] y desde entonces sigue siendo el algoritmo más extendido para la obtención de itemsets frecuentes, con los que construiremos en una segunda etapa las reglas de asociación. Se basa en el principio de que si un itemset es frecuente, entonces todos sus subconjuntos también lo son por lo que al encontrar uno de estos, podremos podar el árbol de búsqueda evitando hacer comprobaciones y aumentando la eficiencia. Para obtener los itemsets frecuentes, el algoritmo en base a un valor mínimo de soporte fijado por el experto en la materia, generará todas las posibles combinaciones de itemsets y comprobará si son o no frecuentes. En cada iteración, se generan todos los posibles itemsets distintos que se pueden formar combinando los de la anterior, por lo que los itemsets irán creciendo de tamaño.

Apriori tiene bastantes factores o limitaciones relacionados con la eficiencia del algoritmo y que pueden afectar en gran medida al proceso de minería de datos que en algunos problemas específicos podría incluso resultar prohibitivo por tiempos o espacio. Algunas de estas limitaciones serían:

1. Soporte: Umbrales demasiado bajos conllevarán a una explosión del número de itemsets frecuentes lo que está directamente relacionado con una mayor necesidad de memoria y tiempo.
2. Número de ítems distintos: Esta limitación, está ligada a la necesidad del algoritmo apriori de almacenar el soporte de cada uno de éstos, lo que puede conllevar problemas de memoria.
3. Tamaño de la base de datos: Este punto está ligado, al anterior, pero en lugar de tener en cuenta los ítems individuales se tienen en cuenta el número de transacciones. Apriori al ser exhaustivo realiza múltiples pasadas por toda la base de datos por lo que el tiempo de ejecución puede ser muy elevado o incluso no llegar a acabar en varios días o semanas.

4. Longitud de las transacciones: Ligado al problema anterior, si las transacciones a su vez están formadas por muchos ítems, almacenar esto en memoria puede llegar a ser privativo e incluso imposible.

Estas limitaciones, nos han llevado a el estudio de otro método menos sensible a los requisitos temporales o de espacio, de cara a las posibles ampliaciones del problema a mayores cantidades de datos aún. Este método es el algoritmo FP-Growth y lo estudiaremos en el siguiente punto.

8.1.2. FP-Growth

El algoritmo **FP-Growth** [54] fue propuesto en el año 2000, como una solución a los problemas de memoria generados por los métodos típicos como el Apriori, visto anteriormente. Es un algoritmo muy eficiente y ampliamente extendido en problemas y soluciones que podrían ser enmarcados bajo el nombre de Big Data.

FP-Growth, crea un modelo comprimido de la base datos original utilizando una estructura de datos que denomina como ***FP-tree*** que está formada por dos elementos esenciales:

- Grafo de transacciones: Gracias a este grafo la base de datos completa puede abreviarse. En cada nodo, se describe un itemsets y su soporte que se calcula siguiendo el camino que va desde la raíz hasta el nodo en cuestión.
- Tabla cabecera: Es una tabla de listas de ítems. Es decir, para cada ítem, se crea una lista que enlaza nodos del grafo donde aparece.

Una vez se construye el árbol, utilizando un enfoque recursivo basado en divide y vencerás, se extraen los itemsets frecuentes. Para ello primero se obtienen el soporte de cada uno de los ítems que aparecen en la tabla de cabecera, tras lo cual, para cada uno de los ítems que superan el soporte mínimo se realizan los siguientes pasos:

1. Se extrae la sección del árbol donde aparece el ítem reajustando los valores de soporte de los ítems que aparecen en esa sección.

2. Considerando esa sección extraída, se crea un nuevo ***FP-tree***.
3. Se extraen los itemsets que superen el mínimo soporte de este último ***FP-tree*** creado.

En función a lo estudiado, es obvio ver que la memoria que ocupa es mucho menor que la generada por Apriori, así como al generar itemsets por medio del principio divide y vencerás, **FP-Growth** se presta a ser usado en entornos distribuidos como por ejemplo el entorno de Big Data, Apache Spark, aumentando sus prestaciones de manera notable.

8.2. Proceso experimental

8.2.1. Comparativa de rendimiento de ambos algoritmos

8.3. Visualización de reglas

8.4. Interpretación de las reglas

Capítulo 9

Conclusiones

Este capítulo cierra el trabajo de varios meses en la materia y se da por finalizada la memoria de este proyecto. Como el propio título indica, se comentarán las conclusiones obtenidas en el transcurso del trabajo desde el punto de vista de un pequeño resumen del trabajo realizado y el análisis de las vías futuras de investigación que este trabajo de fin de grado abre y amplía.

9.1. Resumen del trabajo realizado

9.2. Conclusiones finales

9.3. Líneas futuras

Tal y como se habló en la sección 5.3, sería interesante la extensión del proyecto a un enfoque en la nube, de manera que se pudiera mantener este en ejecución en máquinas virtuales de algún proveedor de servicios cloud, como puede ser Amazon. Esto, eliminaría las restricciones de las máquinas personales, ya que prescindir de la misma y dejarla ejecutando grandes franjas de tiempo es inviable y permitirían un análisis más detallado, por ejemplo manteniendo la búsqueda de Tweets en streaming y estos siendo almacenados directamente en otra máquina virtual en la nube con MongoDB a la

que podríamos acceder en remoto y analizar los datos eliminando el proceso de carga y sincronización y restringiendo este la configuración inicial de la arquitectura.

Por otro lado, el dataset generado es muy rico en información y se presta a la ampliación del problema y su enriquecimiento, como por ejemplo, en lugar de eliminar los tuits cuyo valor de *is_rt* sea igual a *true* estos pueden ser considerados como pesos y darle más valor, ya que en prácticamente la totalidad de los casos un RT implicará que se está de acuerdo con una opinión. Teniendo también los id de tuit y usuario, podría incluso montarse una red de grafos en el que se pudiera ver en el espacio la red de usuarios formada en torno a una opinión o tendencia. En relación también a otras técnicas que pudieran aplicarse sobre los datos, cabría destacar la posibilidad del uso de *hierarchical clustering* para obtener relaciones entre las palabras con una visualización fácilmente interpretable.

Una última ampliación del trabajo que se ha abierto recientemente y que ofrece grandes posibilidades es la comparación de resultados obtenidos en este trabajo, con el mismo proceso si realizáramos la obtención de los datos ahora mismo, donde los tuits en lugar de 140 caracteres pueden llegar a ocupar 280. Cave de esperar por tanto un aumento de palabras de opinión y sentimientos, que al menos sería más elaboradas y podrían aportar mucha más información de cara a un nuevo estudio.

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. Indurkha. *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 Publishers, 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 NGEF. *Cluster Computin*, 2015, 18:2, 577-585.
- [26] Dr. R Nedunchezian 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] Hai Z., Chang K., Kim J. (2011) Implicit Feature Identification via Co-occurrence Association Rule Mining. In: Gelbukh A.F. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2011*. Lecture Notes in Computer Science, vol 6608. Springer, Berlin, Heidelberg
 - [36] Yuan M., Ouyang Y., Xiong Z., Sheng H. (2013) Sentiment Classification of Web Review Using Association Rules. In: Ozok A.A., Zaphiris P. (eds) *Online Communities and Social Computing. OCSC 2013*. Lecture Notes in Computer Science, vol 8029. Springer, Berlin, Heidelberg
 - [37] Z Farzanyar, N Cercone: Efficient mining of frequent itemsets in social network data based on MapReduce framework. *ASONAM 2013*: 1183-1188.
 - [38] 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.
 - [39] S. Moens, E. Aksehirli, B. Goethals: Frequent Itemset Mining for Big Data. *BigData Conference 2013*: 111-118.

- [40] 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.
- [41] Abascal-Mena R., López-Ornelas E., Zepeda-Hernández J.S. (2013) User Generated Content: An Analysis of User Behavior by Mining Political Tweets. In: Ozok A.A., Zaphiris P. (eds) Online Communities and Social Computing. OCSC 2013. Lecture Notes in Computer Science, vol 8029. Springer, Berlin, Heidelberg
- [42] Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC 2006, Genova, Italy, pp. 417-422 (2006)
- [43] D. Ediger, K. Jiang, J. Riedy, D. A. Bader and C. Corley, "Massive Social Network Analysis: Mining Twitter for Social Good,"2010 39th International Conference on Parallel Processing, San Diego, CA, 2010, pp. 583-593.
- [44] Web del proyecto MongoDB. <https://www.mongodb.com>
- [45] Web del proyecto Tweepy. <http://www.tweepy.org>
- [46] Web de Scrappy. <https://scrappy.org>
- [47] Web de Scrapinghub. <https://scrapinghub.com>
- [48] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: a unified engine for big data processing. *Commun. ACM* 59, 11 (October 2016), 56-65.
- [49] I. Feinerer ,K. Hornik. (2017) Text Mining Package (Versión 0.7-3) [Software] Recuperado de <https://cran.r-project.org/>
- [50] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

- [51] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.
- [52] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, A. Zeileis (2018) R/Weka Interface (Versión 3.9.2) [Software] Recuperado de <https://cran.r-project.org/>
- [53] R. Agrawal and R. Srikant Fast algorithms for mining association rules in large databases. 1994. *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB, pp. 487-499.
- [54] Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns without Candidate Generation. 2000. *Proc. Conf. on the Management of Data (SIGMOD 2000)*, Dallas, TX, pp. 1?12.