TRABAJO FIN DE MÁSTER

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

Análisis de tendencias en 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, Septiembre de 2019

Análisis de tendencias con Big Data

Palabras clave: Reglas de Asociación, Big Data, Minería de textos, Twitter

Resumen:

La minería de medios sociales es uno de los ámbitos de aplicación de la minería de datos más estudiados en los últimos años. Tanto en el ámbito de empresarial como en el de investigación, estas técnicas suscitan un gran interés debido a que con el correcto procesamiento pueden obtenerse una gran cantidad de información y valor de datos que apriori parecen desectruturados. En este trabajo, se propone un sistema basado en minería de textos para análisis de medios sociales mediante en cual se dará un flujo de análisis de datos en Big Data en Twitter. Esto se conseguirá mediante el análisis de patrones, proporcionados por reglas de asociación, cuya utilidad en este ámbito de aplicación quedará constatada en el exhaustivo estudio del estado del arte llevado a cabo. Se discuten y compran diversas técnicas de extracción de reglas así como se evidencian las limitaciones de los algoritmos habituales, los cuales queda demostrada su poca utilidad en problemas enmarcados en el paradigma Big Data. Para poder constatar que los resultados son aceptables o se ajustan a la realidad, el sistema será probado con un caso de uso real de Big Data sobre las elecciones generales del Gobierno de España del 28 de abril, constando el buen funcionamiento del sistema, a pesar de tener más de 1.5 millones de transacciones.

Trend Analysis with Big Data

Keywords: Association rules, Big Data, Text mining, Twitter

Abstract:

Social media mining is one of the most studied areas of data mining in recent years. In both business and research, these techniques arouse a great deal of interest because with proper processing can be obtained a large amount of information and value of data that apriori seem desecruturados. In this work, we propose a system based on text mining for social media analysis through which a flow of data analysis will be given in Big Data on Twitter. This will be achieved through the analysis of patterns, provided by association rules, whose usefulness in this field of application will be verified in the exhaustive study of the state of the art carried out. Various techniques for extracting rules are discussed and purchased, as well as the limitations of the usual algorithms, which have been shown to be of little use in problems framed in the Big Data paradigm. In order to be able to verify that the results are acceptable or conform to reality, the system will be tested with a case of real use of Big Data on the general elections of the Government of Spain on the 28th of April, showing the good functioning of the system, despite having more than 1.5 million transactions.

Agradecimientos

Este último año de estudio del Máster de Ciencia de Datos, no ha sido fácil pero también ha estado lleno de buenos momentos sobre todo en lo profesional. Ha sido en este año 2019, y en el pasado 2018, cuando he decidido focalizar mi carrera al ámbito de la investigación y por consiguiente todos los esfuerzos han ido en vías a la obtención de una beca FPU, para estudios de doctorado que al fin puedo decir que he conseguido.

Durante esta época, he podido realizar las primeras publicaciones en congresos nacionales e internacionales así como la colaboración en proyectos de investigación en los que el grupo IdBIS participa. Me gustaría por tanto agradecer a los miembros de este grupo y concretametne a Carlos Fernandez-Basso y Karel Gutiérrez, su ayuda prestada tanto en este proyecto como en otros.

Por último agradecer a las tutoras del proyecto, María José Martín Bautista y María Dolores Ruiz Jiménez el haberme dado la oportunidad de comenzar los estudios de doctorado bajo su tutela así como, permitirme trabajar y desarrollarme intelectualmente con grandes profesionales en el sector de informática y la investigación como ellas y los demás miembros del equipo de investigación.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El mundo que nos rodea está en constante cambio y ha sido en las últimas décadas cuanto este cambio ha sido notado y asimilado por la totalidad de los sectores económicos y sociales con una mayor fuerza. Este proceso ha sido influenciado sin duda alguna por la revolución de las *tecnologías de la información y la comunicación* y más recientemente de la inteligencia artificial.

Esta tecnología en conjunción con otras, ha propiciado el establecimiento de la conocida como *sociedad de la información*, denominación que se le da a una sociedad cambiante donde la manipulación de datos e información ha tomado un papel crucial en las actividades económicas, sociales y culturales. El tratamiento de estos datos para su posterior puesta en valor puede suponer una ardua labor, sobre todo cuando el volumen, naturaleza desestructurada de los mismos o necesidades de trabajo en tiempo real hacen que sea imposible su almacenamiento y procesado acorde a las técnicas habituales. Es aquí donde como una escisión o rama de la inteligencia artificial, nace el Big Data. El Big Data ha propiciado por tanto el nacimiento de aplicaciones basadas en el análisis de datos que antes eran simplemente imposibles de llevar a cabo por las limitaciones de la tecnología. Entre sus usos más habituales están aquellos sectores o ámbitos en los que el volumen de los datos son el principal escollo a salvar. Algunos ejemplos pueden encontrarse en el ámbito del internet de las cosas [29] donde se genera una gran cantidad de datos provenientes de sensores en apenas segundos, o las redes sociales [14] donde la generación de contenido por parte de los usuarios alcanza tales dimensiones

que las técnicas habituales de minería de datos no pueden abarcar. Como hemos visto anteriormente, las redes sociales son una de las grandes factorías de datos actuales. La cantidad de datos generada actualmente por los usuarios de las mismas hacen que un procesado eficiente y útil de las mismas tenga que enmarcarse en soluciones de Big Data. Estas soluciones 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 de obtener valor de datos desestructurados y de gran volumen provenientes de redes sociales se denomina *social media mining*. Esta técnica está enmarcada, dependiendo de las necesidades de información, dentro de las técnicas de minería de textos, procesamiento del lenguaje natural o análisis de grafos y es una de las áreas de máximo apogeo actual entre los investigadores del ámbito de la inteligencia artificial.

En el presente proyecto, tratamos de aportar al ámbito de la minería de medios sociales, una metodología y un sistema final de análisis de tendencias u opiniones en un entorno de Big Data. El sistema será capaz de obtener de manera eficiente valor e información relevante de una gran cantidad de datos proveniente de plataformas de microblogging, como por ejemplo Twitter. En la siguiente sección veremos una breve motivación del proyecto para continuar con el enumerado de los objetivos principales que el proyecto cubre. El capitulo finaliza detallando la organización de la presente memoria.

1.1. Motivación

La totalidad de las actividades económicas y sociales del mundo actual se basan en la posibilidad de acceder a golpe de click a ingentes bases de datos de información. Esto hace que los sistemas automáticos de procesado de datos usados para obtener, procesar y mostrar esta información tomen cada vez un papel más relevante en nuestro día a día. Como vimos en la introducción estas soluciones se enmarcan en multitud de sectores y herramientas como por ejemplo, las enfocadas al marketing [37] en pequeñas y grandes compañías, a la elaboración de modelos predictivos en ámbitos financieros o de seguros [30], y dada la reciente incursión de las redes sociales en los paradigmas sociales,

se ha propiciado la aparición de un gran número de soluciones enfocadas a la minería de redes sociales [21] [36].

El ámbito de la minería de medios sociales, se ha convertido recientemente en de una las vertientes más estudiadas, aplicadas e investigadas dentro de la tradicional minería de datos. Esto es así, debido a que tanto el uso en alza que estas tienen, como el elevado nivel de implantación en la sociedad de las mismas, ha acrecentado la necesidad de sistemas y soluciones, capaces de extraer valor de estas cantidades ingentes de contenido. Es en esta necesidad de obtener valor de los datos desestructurados provenientes de redes sociales, donde surge el análisis de tendencias o minería de opiniones. La minería de opiniones, trata en última instancia de comprender o analizar comportamientos, actividades y opiniones, por ejemplo, de consumidores de cierto producto o usuarios de cierta red social. Si atendemos a su finalidad, esta radica en la obtención de conocimiento útil proveniente de las redes sociales y que pueda traducirse en ventajas competitivas en el proceso de toma de decisiones de una pequeña o gran compañía.

Dada la naturaleza, volátil y masiva de las redes sociales, las técnicas y paradigmas habituales de extracción de conocimiento se ven superadas. Es en esta encrucijada donde la minería de medios sociales se encuentra con el Big Data, siendo las soluciones enmarcadas dentro del mismo las únicas aplicables en innumerables ocasiones, sobre todo en aquellas en las que el volumen de datos o la necesidad de un procesado en tiempo real suponen un impedimento para los algoritmos y tecnologías de minería de datos tradicionales. La aplicación del Big Data a las redes sociales es bastante novedosa y es que no debemos perder de vista que las redes sociales mas famosas apenas tienen unos 13 años de vida, por lo que estaríamos hablando de una de las vías de investigación más novedosas del momento. Esto es así, tanto si ponemos el foco en la novedad que aportan las redes sociales, como si lo ponemos sobre la novedad de los aspectos puramente informáticos, siendo el Big Data uno de los más recientes avances de la computación a gran escala, haciendo que nos encontremos aún en los albores de la explotación de esta tecnología.

Debido a esta novedad y a la necesidad de soluciones de minería de opinión aplicadas a grandes cantidades de datos de manera eficiente, donde surge este proyecto, cuya finalidad y objetivos veremos en la próxima sección.

1.2. Objetivos del proyecto

El presente proyecto fin de máster podría encuadrarse en un objetivo principal que a su vez quedaría definido por un conjunto de objetivos secundarios. El objetivo principal del proyecto sería por tanto, el estudio, desarrollo y validación de un sistema capaz de minar opiniones sobre plataformas de microblogging en un entorno de Big Data, permitiendo visualizar los resultados de una manera amigable y útil para el usuario final (investigador), aportando valor a las posibles preguntas de investigación que este pudiera formular. Este objetivo final, puede definirse como hemos visto antes en función de los siguientes objetivos secundarios:

1. Estudio del estado del arte en el campo de la minería de opinión basada en Big Data en plataformas de microblogging.
2. Desarrollo y aplicación de una metodología de preprocesado de datos eficaz para conjuntos de textos provenientes de plataformas de microblogging.
3. Aplicación de técnicas de minería de datos descriptiva para obtener patrones interesantes en los datos.
4. Obtención y salvado de un corpus de datos de gran tamaño que permitan elaborar la experimentación y validación del sistema final.
5. Experimentación, análisis de resultados y comparación de los mismos con posibles eventos políticos y sociales.
6. Puesta en valor del sistema mediante técnicas de visualización dinámicas.

1.3. Organización de la memoria

La memoria del proyecto está organizada de la siguiente manera. En el capítulo 2, se abordan los conceptos técnicos del trabajo de fin de máster, tales como los costes, personal, equipos utilizados. Tras este capitulo, en los capítulos 3 y 4 abordaremos el primer objetivo del estudio del estado del arte de la minería de opinión basada en Big Data. El peso técnico del proyecto

recae sobre los capítulos 5 y 6 donde abordaremos los objetivos 2, 3, 4, 5 y 6 del presente proyecto. Finalizaremos la memoria con el capitulo 7, donde veremos las conclusiones y vías futuras de investigación que el proyecto abre, así como con una pequeña valoración personal del mismo.

Capítulo 2

Planificación del proyecto

Planificar un proyecto informático de manera correcta ser factor crucial que determine el éxito, o en su defecto el fracaso del mismo. La necesidad de una correcta planificación se acentúa más en proyectos enmarcados dentro del ámbito del Big Data. Estos proyectos son verdaderas obras de ingeniería teniendo en cuenta la necesidad de equipos, recursos humanos, software, variables (tiempo de computo, memoria, costes) que pueden implicar. Es por ello, que en este capítulo haremos un pequeño resumen de la planificación del proyecto, aportando una visión general de los recursos implicados.

2.1. Gestión de recursos

En la primera sección de este capítulo, se hará un repaso por los principales recursos implicados pudiendo estos ser categorizados como personal, hardware y software, los cuales son a su vez los tres pilares clave de un proyecto de tecnologías de la información y la comunicación.

2.1.1. Personal

El personal a cargo del proyecto, consta principalmente del autor José Ángel Díaz García, encargado de desarrollar todas las partes del mismo mediante la supervisión de los tutores. Por otro lado se ha contado con cierta

asesoría y ayuda de miembros del equipo de investigación de bases de datos y sistemas de información inteligentes.

2.1.2. Hardware

Para la elaboración de memorias, notas, artículos así como para los procesos de datos menos complejos se ha utilizado el sistema descrito en la tabla 2.1. Por otro lado, todo el proceso basado en Big Data se ha llevado a cabo en el clúster de procesado de datos del grupo de investigación, este clúster está formado por 4 máquinas cuyas especificaciones pueden verse en la tabla 2.2.

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.

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

Tabla 2.2: Especificaciones técnicas del clúster.

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.

- **Twitter**: Red social de microblogging.
- **MongoDB**: Base datos noSQL usada como almacén persistente de los datos.
- **RStudio**: Entorno de Desarrollo en R donde se han realizado test estadísticos o la visualización.
- **PyCharm**: Entorno de desarrollo en Python donde se ha llevado a cabo la mayor parte del código.
- **PySpark**: Pasarela entre Python y Spark que permite usar funciones de Spark de manera nativa en Python.
- **Spark**: Entorno de computación en clúster usado para elaborar los procesos de Big Data más complejos.
- **Git**: Sistema utilizado para el control de versiones.

2.2. Planificación temporal

En este punto estudiaremos la planificación temporal seguida, así como los pequeños hitos que en cada una de las etapas se fueron consiguiendo y la duración de las mismas.

1. **Estudio del estado del arte**: El primer paso del proyecto, versa en el estudio del estado del arte de la minería de medios sociales y más concretamente en plataformas de microblogging, así como aquellos trabajos que usen reglas de asociación en el ámbito de las redes sociales. Esta etapa tuvo lugar durante el mes mayo de 2019.
2. **Obtención del dataset**: Para testear el sistema se recogieron datos durante el mes anterior a las elecciones generales del 28 de abril de 2019.
3. **Carga de los datos**: Tras la obtención de los datos se hizo necesaria la carga y almacenamiento de los mismos en bases de datos de tipo MongoDB de manera que estos puedan ser cargados en el sistema posteriormente de manera eficiente. Esta etapa tuvo lugar en los primeros días de mayo de 2019.

4. **Limpieza de datos**: Aplicación de técnicas de minería de textos para limpiar los datos así como enriquecer los mismos. Esta tarea fue llevada a cabo entre mayo y junio de 2019.

5. **Reglas de asociación y experimentación**: Sobre los datos limpios, se ejecutaron algoritmos de reglas de asociación difusos y crisp para obtener patrones de opinión o tendencias sobre los datos de las elecciones. Esta tarea tuvo lugar durante el mes de julio y agosto de 2019.

6. **Elaboración de la memoria**: La memoria ha ido elaborándose de manera paulatina a la obtención de resultados aunque el grueso de la elaboración ha recaído sobre el ultimo mes de trabajo agosto de 2019.

2.3. Costes

Tras el análisis de los recursos empleados y la planificación temporal seguida, es menester estimar los costes del proyecto en el supuesto caso de su implantación en una empresa o grupo de investigación. Esta estimación de costes está realizada en función de dos verticales, los gastos de personal y los gastos de ejecución.

Personal

Como hemos descrito en la sección 2.1.1 el personal radica en un solo investigador y la dedicación total atendiendo a la carga lectiva del proyecto final de máster, estaría en torno a los 3 meses a jornada completa. Teniendo en cuenta una estimación de unos 2000 euros brutos al mes tendríamos un total de 6000 euros en gastos de personal.

Ejecución

En esta categoría encontramos los gastos de adquisición del material inventariable así como los gastos del material fungible. Como inventariable, tenemos los equipos descritos en la tabla 2.1.2, es decir, el equipo personal y el cluster de procesado cuyo coste en función del período de amortización

(precio por uso dividido entre tiempo de amortización) puede verse en la tabla 2.3.

Unidad	Precio	Periodo Amortización	Duración proyecto	Total
Mac pro 2,6GHz Intel Core i5	1300	2 años	1,5 meses	81,25
Clúster	6000	6 años	1 mes	120

Tabla 2.3: Detalle de costes inventariables.

Si atendemos a los costes fungibles, aquellos cuyo inventariado es menos relevante como por ejemplo el material de oficina podríamos estimarlo en unos 100 euros.

Resumen de gastos

En función a los cálculos estimados de costes realizados en esta sección, en la tabla 2.4 podemos ver el total de los gastos que podrían derivarse del proyecto.

Gastos elegibles	Total
Personal	6000 euros
Costes inventariables	201,25 euros
Costes fungibles	100 euros
TOTAL	6301,25 euros

Tabla 2.4: Resumen final de costes.

Capítulo 3

Marco de trabajo

Antes de comenzar a estudiar el estado del arte del análisis de tendencias o minería de medios sociales como Twitter es necesario presentar los conceptos teóricos relacionados y que permitirán comprender de una mejor manera los conceptos de los trabajos que serán discutidos en el capítulo 4.

3.1. Minería de medios sociales digitales

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

La **minería de medios sociales**, acorde a P. Gundecha [15], comprende el proceso de representar, analizar y extraer de datos provenientes de medios sociales patrones con significado y valor, de manera que puedan utilizarse en el proceso de toma de decisiones de una pequeña o gran compañía. 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 o en inglés, NLP. 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 critica constructiva, si se promueve el producto o si simplemente lo

crítica puede suponer una gran ventaja competitiva para las empresas y marcas, por ello, son más las que cada vez usan estas técnicas en sus procesos de vigilancia tecnológica, obtención del *feedback* del consumidor o simplemente para la creación de *data lakes* de opiniones que puedan ser analizadas, por ejemplo para afinar el futuro lanzamiento de un nuevo producto.

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 de 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 [23], una opinión es una quíntupla compuesta de los siguientes elementos:

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

Pese a que aún no hemos entrado en el estudio de las redes sociales ni de la **anatomía** de un tweet, estos serán la fuente y la unidad mínima de información en nuestro proyecto. En el punto 3.5.2 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* [31] [8] 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 las aplicaciones basadas en clasificación, para intentar indagar en el uso de las reglas de asociación en el ámbito de la obtención de patrones sobre textos u opiniones en Twitter. Dado que los datos provenientes de redes sociales carecen de clases o *etiquetas* a priori, análisis descriptivos basados en reglas de asociación como el propuesto en este trabajo, pueden ser muy útiles para comprender los datos o extraer conocimiento de esas grandes factorías de datos que son las redes sociales.

3.3. 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 [3]. Su utilidad es la de obtener conocimiento relevante de grandes bases de datos y se representan según la forma $\mathbf{X} => \mathbf{Y}$ $\mathbf{Y}$ donde $\mathbf{X}$, es un conjunto de ítems que representa el antecedente e $\mathbf{Y}$ un ítem o conjunto de ítems 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**, la **confianza** y el **lift**, que podremos definir de la siguiente manera:

- Soporte: El soporte de un ítem se representa como *supp (X)*, y representa la fracción de las transacciones que contienen al ítem X entre el total de transacciones (t) de la base de datos (D). Su fórmula sería:

$$supp(X) = \frac{|t \in D : X \subseteq t|}{|D|} \tag{3.1}$$

- El soporte de una regla de asociación estaría representado como *supp(X → Y)* y por consiguiente correspondería al total de transacciones que contiene tanto al ítem X como al ítem Y. Podríamos definirlo matemáticamente como sigue:

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{3.2}$$

- Confianza: Se representa como *conf (X→ Y)*, y representa la fracción de transacciones en las que aparece el ítem Y, de entre aquellas transacciones donde aparece el ítem X. Su definición sería:

$$conf(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X)} \tag{3.3}$$

- Lift: El *lift*, es una medida útil para evaluar la independencia entre los ítems de una determinada regla de asociación. En una regla del tipo *lift (X→ Y)*, esta medida representa el grado en que X tiende a ser frecuente cuando A está presente en la regla, o viceversa. El lift, quedará definido matemáticamente de la siguiente manera:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} \tag{3.4}$$

Pese a que estas medidas son las más comunes y extendidas, hay muchas más propuestas de medidas complementarias en la literatura, tales como la **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 ítems 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 itemset 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 itemset 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.

3.3.1. Reglas Generalizadas

Las reglas de asociación pueden ser estudiadas e interpretadas desde un punto de vista jerárquico [39], por ejemplo, en el problema de la cesta de la compra, la regla *{Manzanas, Platanos}* $\rightarrow$ *{Yogurt}* podría ser reemplazada por *{Fruta}* $\rightarrow$ *{Yogurt}*. Esto nos permite lograr un mayor grado de abstracción, lo que es interesante para obtener información relevante. Esta abstracción también nos permite resumir enormemente el conjunto de reglas, lo que resultará en un análisis más sencillo de los problemas en cuestión sin perder información relevante por el camino. Las reglas de asociación generalizadas también son interesantes para los entornos Big Data, ya que el

tamaño de los resultados obtenidos puede ser altamente resumido, mejorando consecuentemente el tiempo y los recursos de procesamiento.

Aplicaciones

Su uso ha sido extendido en campos como las telecomunicaciones, gestión de riesgos, control de inventarios [25] [41] 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 [10]. Dada la importancia de estas técnicas en nuestro trabajo las estudiaremos con detalle en el capítulo 4, 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.* [40], 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 [34], cuando se comenzó a hablar de las **3 V's del Big Data** [22], 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 continúa 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. Spark

Una de las tecnologías más extendidas dentro del Big Data es Apache Spark. Este es un framework opensource de procesamiento distribuido cuya funcionalidad se centra en la orquestación de trabajos en una red de computación o clúster. Spark es un framework muy potente que permite al usuario

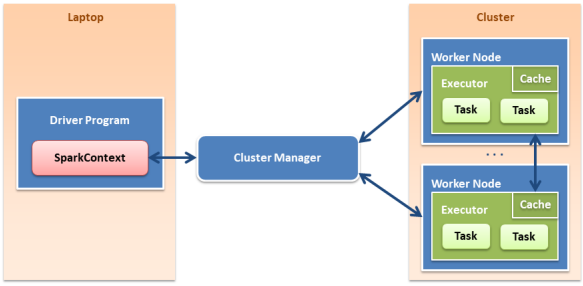

Figura 3.1: Arquitectura spark.

realizar prácticamente todas las tareas de un pipeline de análisis de datos habitual de manera distribuida y por supuesto, mucho más eficiente que si fuera secuencial. En cuanto a versatilidad y posibilidad de adaptación también tiene grandes ventajas, pues ofrece APIs para trabajar en distribuido desde los lenguajes más usados en análisis de datos como R, Python, Sacala o Java entre otros. En esta línea de versatilidad se encuentran también los datos con los que este puede trabajar yendo estos de más a menos estructurados sin causar problemas mayores al procesado.

La arquitectura de Spark, quedaría definida por un nodo maestro, desde el cual se ejecutan los programas de usuario, un nodo que se encarga de distribuir los datos y una serie de nodos denominados *workers*, que serían los encargados de ejecutar su parte del trabajo enviando de nuevo al nodo de orquestación los resultados, que serían unidos y ofrecidos al nodo maestro como una misma entidad. Podemos ver una esquema de esta arquitectura en la figura 3.1.

3.4.4. Spark vs Hadoop

El mundo del Big Data, como en todos los ámbitos de las TIC no recae sobre una sola tecnología sino que hay varias donde elegir, y que a menudo, son competencia entre si. Este es el caso de Spark y Hadoop. Hadoop fue el primer framework de Big Data, está basado en la tecnología Map-Reduce (figura 3.2) y ha sido y es ampliamente utilizado para problemas de Big Data.

Una de las características principales de Hadoop recae en el hdfs, un sistema de archivos distribuido que permite trabajar con una gran cantidad de

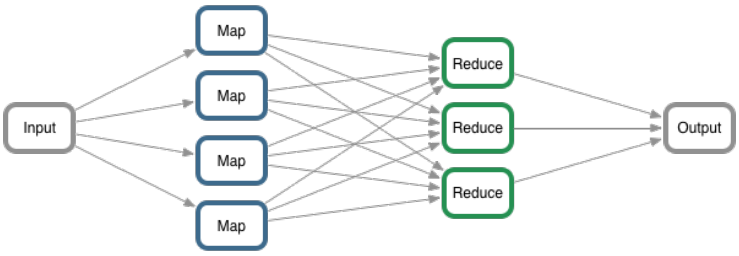

Figura 3.2: Arquitectura Map Reduce.

datos así como escalar de manera sencilla en el caso de necesitar más recursos o más capacidad de procesado. El hdfs, también aporta redundancia y tolerancia a fallos, pero por contra, al tener los datos en disco los procesos son más lentos en comparación a Spark.

En cuanto a la versatilidad y usabilidad, Spark también sería más recomendable dado que como hemos mencionado anteriormente tiene APIs muy sencillas para los lenguajes más usados. Estas ventajas, han hecho que recientemente el uso de Spark se haya visto incrementado en comparación con Hadoop.

3.4.5. Aplicaciones

El Big Data está presente en muchas áreas socio-económicas actuales, algunas de las cuales pueden ser:

- 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 aunque el concepto del Big Data esté aún 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.

3.5. 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).

Aunque en sus inicios el número máximo de caracteres que podíamos encontrar en un tweet era 180, en 2017 la red social aumentó el número de palabras en cada tuit a 280, lo que en conjunción con nuevas medidas como la facilidad para incluir videos o demás contenido multimedia y lo vistoso de estas publicaciones con un diseño intachable, constatan la salud de la red social.

3.5.1. Funcionamiento

El funcionamiento de la red social es en sí trivial, 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, 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 estudiadas en la literatura como parte de la teoría de grafos y ha sido extendido a la minería de redes sociales, para la detección de comunidades o de personas influyentes dentro de la red social [12]. Dado que nuestro trabajo versa sobre la minería de opiniones, no es menester entrar en más detalle en las relaciones entre usuarios dentro de twitter (*retweets*, *follows*). Por otro lado, sí 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.

3.5.2. Anatomía de un tuit

En la figura 3.3, 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.

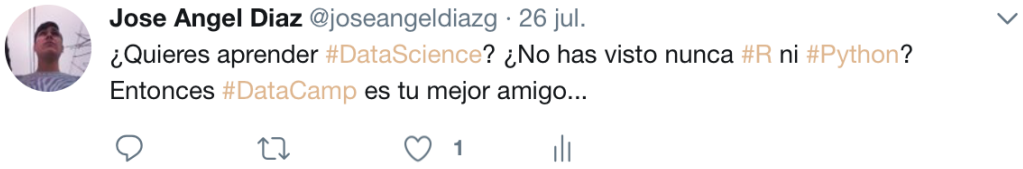

Figura 3.3: Ejemplo de un tuit.

Atendiendo al ejemplo anterior (figura 3.3), estaríamos hablando sobre *data science*, R, Pythton 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.

Figura 3.4: Ejemplo de un tuit que no sería enmarcado como opinión.

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 3.4 que simplemente es informativo.

3.5.3. 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 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 que 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 término.
- **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.

En los puntos siguientes ahondaremos en el proceso seguido y tecnologías usadas para la obtención y almacenamiento de un gran volumen de datos sobre los cuales realizar el proceso de experimentación.

Capítulo 4

Estado del arte

El campo de la minería de opinión en redes sociales es relativamente nuevo, debido, sin duda, a la novedad de las redes sociales. Twitter fue fundado en 2006 y Facebook en 2005, lo que nos da un promedio de 14 años de vida para las redes sociales más famosas. Por otro lado, hay que tener en cuenta que su implantación en la sociedad no se produjo desde su fundación, por lo que su 'edad' es aún menor. Si nos centramos solo en los aspectos informáticos del estudio, también son notablemente nuevos, porque a pesar de haber sido ampliamente estudiados, todavía estamos en los albores de lo que la minería de datos podría ofrecer en el futuro. La novedad de estas técnicas y el predominio de técnicas supervisadas que abordan problemas similares, hace que existan pocos enfoques previos completamente relacionados con el campo de estudio, que aúna reglas de asociación y minería de medios sociales. Sin embargo, también es uno de los campos de investigación que genera más interés entre la comunidad científica.

4.1. Trabajos relacionados

En esta sección veremos trabajos relacionados con el nuestro. Para facilitar la comprensión se han dividido en distintas secciones en función de las tecnologías utilizadas y el ámbito de aplicación. Por último, finalizaremos el capítulo con una sección dedicada a nuestra propuesta y nuestra aportación al estado del arte.

4.1.1. Reglas de asociación y microblogging

Uno de los principales estudios en el campo de las reglas de asociación es el propuesto en 2000 por Silverstein et al. [38], donde se utilizan técnicas de reglas de asociación para el conocido problema de la cesta de la compra, que relaciona la compra de un determinado producto con la posibilidad de comprar uno diferente. Después de esto, el tema ha sido ampliamente estudiado y aplicado en multitud de artículos de investigación, aunque la aplicación de reglas de asociación en redes sociales no se abordaría por primera vez hasta 2010. El artículo es propuesto por Oktay et al. [32] y estudia la relación entre la aparición de ciertos términos en las preguntas del sitio web de Stack Overflow con la aparición de ciertos términos en las respuestas a estas preguntas. En cierto modo está relacionado con nuestro estudio, en el que intentamos obtener e interpretar la relación entre los términos, aunque las técnicas y el dominio difieren completamente. El uso de las reglas de asociación en las redes sociales ha sido mostrado en documentos como el propuesto por Erlandsson et al. [13], en el que se presenta un análisis basado en las reglas de asociación para encontrar influencers en Facebook.

Si nos centramos en nuestro dominio, Twitter, el trabajo de Pak y Paroubek [33], ha sido un punto de partida para destacar a Twitter como una fuente importante de análisis de opinión y sentimiento. Atendiendo al uso de las reglas de asociación en Twitter se diversifica el dominio de los estudios y aplicaciones. Las propuestas en [43] y [26] proponen la detección de patrones de palabras asociados al ciberbulling para detectar estos comportamientos a través de la red social Twitter, aunque en estos experimentos el dominio es muy pequeño y el número de tweets es muy limitado. Esta especificidad del dominio se encuentra también en el trabajo [18] donde Hamed et al. propusieron un sistema basado en reglas de asociación para determinar la coocurrencia de hashtags en el campo del tabaquismo, con el objetivo de crear un sistema experto para dejar de fumar. También en el campo de la minería de patrones, pero en el campo de los seguros y con un mayor volumen de tweets de entrada para el proceso de minería, encontramos el trabajo propuesto por Mosley y Roosevelt [28]. En este trabajo los autores utilizan las reglas de la asociación y el clustering para obtener patrones interesantes relacionados con las personas y los seguros. Nuestra propuesta estaría ligada a estos artículos, con la diferencia de que no usamos prefiltrado de tweets y la cantidad de

estos es mucho mayor para que los patrones que se obtengan puedan ser considerados más fuertes, porque aparecen con más representación.

Otro campo de aplicación en el que encontramos reglas de asociación en Twitter es el resumen de información, ya sea de los usuarios por su influencia en la red de microblogging [1] o de los elementos más relevantes (tweets, post) [35]. En estas áreas, la teoría de grafos y otros tipos de reglas de asociación, como las reglas maximales, también entran en la ecuación. Una vez más, en ambas propuestas encontramos un gran problema en el número de tweets cuyo volumen es muy bajo. Esto hace muy difícil transferir los resultados a un problema real, donde el volumen de datos y variables sería mayor. Por otro lado, ambas propuestas destacan el poder de las reglas de la asociación para resumir la información y obtener patrones en la red social, algo que nuestro trabajo también hace pero a mayor escala (es decir, con un mayor número de tweets).

Todos estos enfoques muestran que las reglas de asociación se utilizan con suficiente regularidad en el dominio de las redes de microblogging como Twitter, aunque, como se ha comprobado, con una seria limitación con el tamaño de la entrada. Esta limitación es salvada por algunos trabajos que podrían ser enmarcados dentro del alcance de Big Data. Aquí encontramos el trabajo propuesto por Adedoyin-Olowe et al. [2] y el trabajo propuesto por Fernandez-Basso et al [14]. En el primero se extraen reglas de asociación sobre un stream de datos, por el contrario en el segundo trabajo se extraen itemsets frecuentes, con el fin de usarlos para detección de eventos en el campo de los deportes y de la política. En el primer trabajo se usa un gran número de tweets de alrededor de 3,8 millones, aunque se particionan para simular posteriormente el streaming. También enmarcado dentro del paradigma de Big Data, pero solo para la metodología porque el volumen de tweets es muy pequeño, encontramos el trabajo [20] que hace un sistema de recomendaciones de películas que se nutre de Imdb[1] y Twitter. Además del volumen de datos que el sistema puede procesar, nuestra propuesta difiere de estas dos visiones, ya que nuestro sistema puede obtener reglas de asociación mientras que en el presentado por Fernández-Basso sólo se obtienen conjuntos de ítems frecuentes. Además, los otros dos sistemas propuestos por [2] y [20], utilizan reglas de asociación para la detección de tópicos en streaming ignorando el análisis de sentimientos sobre, por ejemplo, los temas detectados, algo que sí

[1]Internet Movie Database

hace nuestra propuesta. De acuerdo a los patrones revelados sobre los políticos, partidos o usuarios, se realiza una etapa posterior de identificación de sentimientos que ofrece una gran cantidad de información al usuario final.

4.1.2. Reglas de asociación y análisis de sentimientos

En cuanto a los campos del análisis de sentimientos y reglas de asociación, hay pocos estudios relacionados debido al predominio de los métodos de clasificación [44] en estas áreas, pero, vale la pena mencionar el reciente interés en este tema. Encontramos estudios como el de Hai et al. [17] donde se aplica un enfoque basado en reglas de asociación, co-ocurrencia de palabras y clustering para obtener las características más comunes respecto a ciertos grupos de palabras que pueden representar una opinión. El objetivo del estudio es dar un paso adelante en el análisis de sentimientos, que normalmente solo clasifica una opinión. El método propuesto no solo clasifica, sino que también ofrece al usuario las palabras o características de opinión que se han empleado en la clasificación. El trabajo de Yuan et al. [42] propone una nueva medida para la discriminación de términos frecuentes sin orientación aparente de las opiniones, lo que favorece el posterior proceso de análisis de sentimientos. Vinculado a este punto está el estudio realizado por Dehkharghani et al. [11] donde se propone el uso de reglas de asociación para vincular la co-ocurrencia de términos en tweets, los cuales son posteriormente clasificados de acuerdo a los sentimientos de estos términos vinculados en las reglas obtenidas. En términos generales, el vínculo entre estos estudios es el uso de reglas de asociación y conjuntos de elementos frecuentes para mejorar el proceso de análisis de sentimientos. Esto difiere de nuestro estudio en que una vez que se extraen las reglas, utilizamos un enfoque jerárquico de las reglas para mejorar la interpretación de las reglas de asociación usando para ello análisis de sentimientos.

En esta vía de investigación, hemos encontrado dos trabajos que proponen un enfoque mixto de reglas de asociación y análisis de sentimientos para obtener patrones en Twitter. El propuesto por Mamgain et al. [24] y el propuesto por Bing et al. [5]. Ambos proponen una etapa previa de análisis de sentimientos, asociando sentimientos a cada elemento y, posteriormente, obtienen patrones utilizando el algoritmo Apriori. El primer trabajo crea un modelo que puede ayudar a los estudiantes a elegir la mejor universidad de la India y el segundo lo aplica para la predicción del mercado de valores. La fuerza

de usar ambas herramientas desde un enfoque mixto es por lo tanto contrastada en la literatura, aunque en ambas propuestas el número de tweets que emplearon es muy limitado. Nuestra propuesta difiere de estas en el ámbito de aplicación así como el volumen de tweets utilizados es mucho mayor y puede considerarse un problema de Big Data debido a su gran volumen. Nuestra propuesta también difiere de las anteriores ya que utilizamos reglas de asociación generalizadas para el análisis de sentimientos, obteniendo así patrones de sentimientos más fuertes que los que se han visto anteriormente.

4.1.3. Reglas de asociación generalizadas

Los enfoques jerárquicos en el proceso de extracción de reglas de asociación han sido últimamente bastante estudiados, debido en gran parte a la necesidad de condensar la información que representan, por ejemplo, para mejorar los procesos de visualización. Un ejemplo reciente de este uso lo presentan Hahsler y Karpienko [16] donde se propone una visualización basada en matrices, que hace uso de una simplificación jerárquica de los ítems que forman las reglas de la asociación. En el presente estudio, el enfoque jerárquico también se utiliza para simplificar o generalizar las reglas, pero en lugar de hacerlo por categorías de elementos, lo hacemos por sentimientos. Otros enfoques que utilizan reglas de asociación generalizadas en Twitter son el análisis de [6] y [7] ambos propuestos por Cagliero y Fiori. En el primero, los autores utilizan reglas de asociación dinámicas, es decir, reglas en las que la confianza y soporte cambian con el tiempo, con el fin de obtener datos sobre los hábitos y comportamientos de los usuarios en Twitter, y en el segundo, estas reglas se generalizan para obtener reglas más fuertes. En este último, los autores proponen generalizar las reglas obtenidas de los tweets según taxonomías como lugares, tiempo o contexto, de manera que puedan ser utilizadas para analizar la propagación o evolución de los contenidos en el tiempo. El presente trabajo utiliza reglas de asociación generalizadas utilizando los sentimientos obtenidos en el proceso anterior de análisis de sentimientos, en lugar de utilizar lugares o contextos como las otras propuestas descritas en esta sección. Con este uso, el sistema puede partir de un conjunto de datos proveniente de redes sociales y obtener patrones que muestren la distribución de los sentimientos en los datos, basados en un tema específico sobre el que se desea obtener información, todo ello sin elevados tiempos de cómputo lo que lo hace muy interesante para problemas de Big Data.

Reference	N tweets	Purpose	AS	PM	RAG
[43]	8275	Detección de patrones de cyberbulling.	No	Si	No
[26]	14000	Detección de patrones de cyberbulling.	No	Si	No
[18]	35000	Co-ocurrencia de hashtags para sistema experto contra el tabaco	No	Si	No
[28]	68370	Patrones en el ámbito de los seguros.	No	Si	No
[1]	24026	Identificación de usuarios activos durante ataques terroristas.	No	Si	No
[35]	500	Resumen de tweets sobre Obama.	No	Si	No
[20]	20000	Recomendación de películas.	No	Si	No
[2]	224291-3837291	Detección de eventos.	No	Si	No
[44]	57000	Análisis de las elecciones de Australia.	Si	No	No
[9]	80563	Obtención de patrones para promover el ciclismo.	Si	Si	No
[24]	8.772	Patrones sobre las mejores universidades de India.	Si	Si	No
[5]	150000	Prediccion de stock de productos.	Si	Si	No
[11]	3000	Resumir conversaciones de Twitter.	Si	Si	No
[6]	450000	Detección de tópicos.	No	Si	Si
[7]	450000	Estudios de propagación de información.	No	Si	Si
Proyecto	**1517477**	**Obtener patrones y sentimientos sobre las elecciones del 28A.**	**Si**	**Si**	**Si**

Tabla 4.1: Comparación de propuestas acorde al número de tweets, el uso de análisis de sentimientos (AS), pattern mining (PM) y reglas de asociación generalizadas (RAG).

Para concluir esta sección se ha recopilado en la Tabla 4.1.3 los trabajos relacionados y revisados que utilizan Twitter como corpus para el posterior proceso de minería de datos.

4.1.4. Nuestra propuesta

Nuestra propuesta presenta un enfoque de Big Data al ámbito de la minería de opinión y extracción de patrones. Mediante el uso de Spark y técnicas de minería de datos, trataremos de obtener patrones interesantes sobre Twitter con una tokenización especial para análisis de sentimientos de manera que en posteriores trabajos puedan usarse las reglas generalizadas por sentimientos ofreciendo una nueva capa de análisis. La propuesta actual, ofrece por tanto la primera capa del sistema que permite trabajar con grandes conjuntos de datos. Por tanto la contribución de este estudio a los campos del pattern mining y minería de medios sociales es:

- La propuesta de una metodología capaz de trabajar con un número muy alto de tweets que puede ser considerado Big Data de una manera eficiente. Este punto difiere de otros estudios porque el volumen de datos estudiado en la mayoría de ellos es muy limitado y está muy lejos de los problemas reales.

- Se ha llevado a cabo una revisión detallada de los estudios publicados que aplican las reglas de la asociación en el campo de la minería de datos (incluyendo el análisis de Twitter) y la minería de medios sociales.

Capítulo 5

Metodología

En este capítulo veremos la metodología seguida para la creación del sistema. Entraremos en detalle en las tecnologías y técnicas usadas en el modulo de obtención de datos y del preprocesado, dejando la experimentación y los detalles del proceso de minería de datos con reglas de asociación para el capítulo 6. Para una mejor comprensión se ha plasmado el flujo de información en la figura 5.1.

5.1. Dataset

En esta sección veremos, el proceso de obtención de los datos así como la persistencia de los mismos y una breve explicación de la composición del dataset que será utilizado en el proceso experimental.

5.1.1. Obtención de datos

Los datos han sido obtenidos mediante la API de streaming de Twitter. Esta API, permite obtener datos sobre un determinado hashtag, usuario o tópico indefinidamente. Para obtener los datos, se ha llevado a cabo un script en python que realiza las siguientes tareas:

1. Identifica la app con las credenciales de twitter.

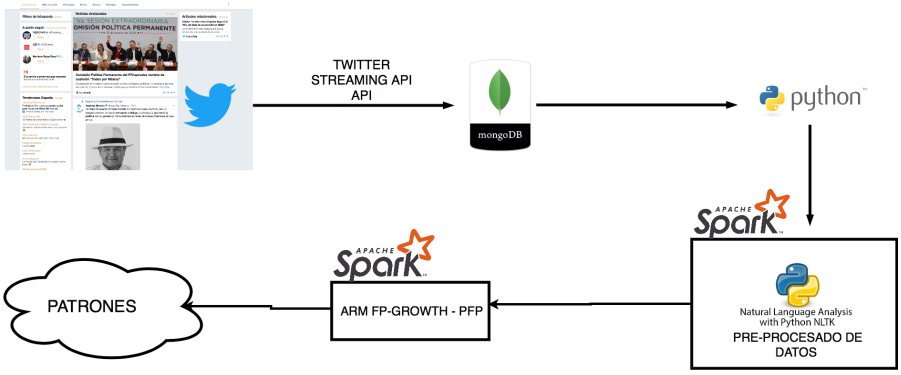

Figura 5.1: Arquitectura del sistema.

2. Obtiene parámetros de entrada, en nuestro caso, queremos tuits relativos a las elecciones del 28A , nuestro parámetro de búsqueda será 28A, 28Abril y Elecciones2019.

3. Obtiene datos y los guarda en la base de datos de manera indefinida hasta que no se pare el script.

La API de streaming de Twitter es muy útil para analizar datos en tiempo real, pero en nuestro caso también es útil para almacenar datos a lo largo del tiempo de manera que se pueda obtener un dataset de gran tamaño. Dado que la obtención de datos se realizó durante el mes de abril, mes en el que continuamente se generaban tuits relativos al 28A, se consideró obtener tuits relativos a este hashtag de manera que la experimentación y los patrones, podrían ser contrastados con eventos de la vida social y política de España durante el mes de abril de 2019.

5.1.2. Persistencia

La persistencia de los datos, se ha llevado a cabo con una de las bases de datos más usadas dentro del paradigma Big Data, MongoDB. Esta base de datos, es noSQL[1]. Dado que en nuestro sistema no necesitamos una gran consistencia, sino que buscamos la versatilidad y facilidad de conexión con las APIs de Python, hemos considerado MongoDB como la mejor base de datos para nuestro proceso.

[1]Not Only SQL

5.1.3. Especificaciones de los datos

El conjunto final de datos, se compone de 1517476 registros (tuits), con un lenguaje de 140727 palabras o símbolos distintos. El tamaño máximo del texto de un tweet es de 280 caracteres, y aunque la obtención de datos incluye información relativa al usuario que emite el tweet, en este análisis solo hemos mantenido el texto pues los demás datos pueden ser útiles en trabajos futuros pero no en este.

5.2. Preprocesado

El preprocesado es una de las tareas más relevantes e importantes dentro del flujo de un proyecto de ciencia de datos. Podríamos definirlo como el conjunto de técnicas enmarcadas en ciencia de datos cuya finalidad es obtener datos de mayor calidad de forma que los algoritmos de minería de datos, bien sean descriptivos o predictivos, puedan aplicarse de una manera más eficiente y con mejores resultados.

En minería de textos, el preprocesado de datos es ligeramente diferente a otros problemas de minería de datos, dada la naturaleza de los textos como datos no estructurados. En esta sección, veremos las técnicas de preprocesado de textos llevado a cabo sobre el dataset descrito en el punto 5.1.3.

5.2.1. Tokenización y limpieza

En este punto, estudiaremos las técnicas llevadas a cabo para limpiar los datos provenientes de Twitter. Para conseguir estas tareas en python, el primer paso es crear dataframes de Pandas, sobre los cuales podremos ir aplicando técnica de minería de textos mediante applys y herramientas de procesamiento de lenguaje natural. Las técnicas usadas, han sido:

1. El primer paso pasa por eliminar los enlaces. Para ello, creamos una expresión regular que elimina este contenido, teniendo en cuenta que los enlaces no son necesarios en el proceso de extracción de reglas de asociación, pues para nuestras transacciones textuales solo necesitamos los ítems o palabras mencionados en un tweet.

2. Tras esto hemos tokenizado cada palabra haciendo uso de un tokenizador especial para Twitter implementado dentro del paquete *nltk*. Este paquete tokenizador mantiene caracteres especiales como los emoticonos y procesa de una manera distinta los signos de puntuación. Un ejemplo de uso podría ser, el tweet '*Esto es un ejemplo!!!! Con emoticonos :) y símbolos)*' seria tokenizado como, '*Esto, es, un, ejemplo, !!!, Con, emoticonos, :), y, símbolos,)*'. Esto es así debido a que se mantienen símbolos para permitir analizar un tipo de lenguaje coloquial, algo muy útil para dominios como el de Twitter.

3. El tercer paso es eliminar las palabras vacías en Español.

4. En nuestro análisis no nos interesan los signos de puntuación como *!*, pero sí que nos interesa mantener otros como *:)*, por ello, una solución es eliminar aquellos ítems de tamaño inferior a 2 símbolos, con lo que mantendremos emoticonos y eliminaremos signos de puntuación y posibles tokens erróneos.

5. Se ha creado una lista especial de palabras vacías del dominio de Twitter, tales como *lol, RT, via*. Tras su creación, se han eliminado del dominio del problema pues son palabras que meten ruido en el dataset y no aportan nada.

6. El sexto paso para la limpieza de los datos ha sido el paso de todos los tokens a utf-8, ya que había caracteres especiales que no sirven en nuestro procesado.

7. El siguiente paso, se ha basado en el paso a minúsculas de todos los tokens.

8. Tras el paso a minúsculas, se ha utilizado un proceso de corrección del lenguaje para evitar palabras mal escritas o con faltas de ortografía. Para ello para cada token, se ha entrenado un *spellcheker* y en caso de que el resultado sea una palabra errónea con un valor de confianza máximo, se cambia por la correcta.

9. Por último se han eliminado los tokens compuestos íntegramente por números.

Es necesario mencionar antes de finalizar esta sección que dos de las labores, que tradicionalmente se realizan en procesado de tuits, como son la eliminación de usuarios y hashtags, en nuestro caso no se han llevado a cabo pues queremos ver qué ítems se relacionan con qué hashtags y con qué personajes y partidos políticos, por lo que para nuestro análisis ítems como @sanchezcastejon o @vox_es entre otros son muy interesantes.

5.2.2. Big Data

Aunque el proceso de limpieza en el cluster de computación no es privativo en cuanto al tiempo de ejecución, se ha trasladado al paradigma BigData. Para realizar esta labor, se ha usado la API de Spark para Python, *pyspark*. Esta API pone a nuestro servicio múltiples funciones que permiten distribuir el trabajo de una manera sencilla, en este caso dado que tenemos un conjunto de independiente de datos muy grande sobre el que realizar un flujo de procesos, la mejor opción para distribuir los datos ha sido mediante *mapPartitions*, de manera que el sistema está preparado para aumentar el número de tuits aún más si cabe y obteniendo un mejor rendimiento. En el siguiente capítulo, continuaremos ahondando en el proceso de distribución de los datos y así como un estudio comparativo de la obtención de reglas de forma secuencia y de forma distribuida mediante Spark.

Capítulo 6

Minería de datos

En este capítulo abordaremos los objetivos finales del proyecto, de aplicación de técnicas de minería de datos, obtención de patrones y puesta en valor del sistema final mediante técnicas de visualización. También, se abordará la experimentación y se analizará un caso de uso relativo a las elecciones generales del 28 de abril de 2019.

6.1. Transacciones textuales

Como vimos en el capítulo 4 las reglas de asociación pueden aplicarse sobre texto para obtener relaciones de co-ocurrencia entre ítems (palabras) dentro de una base de datos textual, esto nos permite resumir la información y obtener patrones interesantes que relacionan una palabra con personas, marcas... en nuestro caso políticos y partidos políticos.

El primer paso para poder aplicar algoritmos de extracción de reglas sobre texto, pasa por crear transacciones de texto. Estas fueron definidas en el paper [27], y teniendo una colección de documentos cada uno de estos documentos sería una transacción, y cada una de las palabras presentes en el vocabulario constituirían un ítem, que podría aparecer o no en cada una de las transacciones, la representación mas eficiente, al menos en tamaño, pasaría por usar booleanos dado que tenemos matrices muy dispersas y de gran tamaño.

6.2. Reglas de asociación

En esta sección veremos todo lo relativo al proceso de extracción de reglas de asociación, desde el estudio teórico de los algoritmos utilizados a las complicaciones encontradas dado el volumen de los datos y la solución Big Data aportada.

6.2.1. 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 a la mejora o estudio matemático de los algoritmos 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.

Apriori

El algoritmo **Apriori**, fue propuesto por Agrawal y Srikant en 1994 [4] 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 al 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.

FP-Growth

El algoritmo **FP-Growth** [19] 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 itemset 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.

6.2.2. Reglas de asociacion en Big Data

En función a lo estudiado sobre los dos algoritmos anteriores, es obvio que la memoria que ocupa FP-Growth 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.

Debido a esto nos hemos decantado por usar el algoritmo FP-Growth presente en PySpark para el entorno distribuido, aunque hemos usado Apriori para intentar comparar el procesado de ambos algoritmos. Como veremos en la próxima sección Apriori solo ha conseguido funcionar con un conjunto de datos muy pequeño y niveles de soporte muy altos, constatando su nula funcionalidad para problemas de Big Data.

Para la extracción de reglas, el algoritmo FP-Growth ha sido ejecutado con umbrales mínimos de soporte y confianza de 0.001 y 0.6 respectivamente.

Modo	Rango	Media
Secuencial	[12,3min-14,1min]	12,67min
Distribuido	[9,10min-10,9min]	10,1min

Tabla 6.1: Comparación de preprocesado en secuencial y Big Data.

6.3. Big Data vs secuencial

Para poder tener un punto de vista crítico sobre las mejoras que el Big Data, ofrece frente a soluciones secuenciales se llevo a cabo un estudio comparativo entre la solución secuencial y Big Data.

6.3.1. Comparativa preprocesado

Acorde a la parte de preprocesado de datos de nuestro sistema, en la tabla 6.1 podemos ver una comparativa de los resultados y tiempos de ejecucion del preprocesado acorde al paradigma Big Data o el tradicional procesado secuencial.

Acorde a los resultados de la anterior tabla, podemos concluir como la diferencia entre preprocesado en secuencial y Spark no es muy pronunciada, debido sin duda al tiempo de creación de contexto de Spark.

6.3.2. Comparativa extracción de reglas

Sobre la extracción de reglas en modo secuencial, solo pudieron obtenerse resultados con valores de soporte muy altos y lejanos de la realidad de 0.1, por lo que solo se extraen reglas muy obvias. Apriori, no ha sido capaz de funcionar ni siquiera con un 10 % de los datos, ni en secuencial ni en distribuido, por lo que el estudio comparativo y experimentación ha sido imposible debido a las limitaciones técnicas del algoritmo. Estas limitaciones radican en que el algoritmo Apriori es capaz de trabajar con un gran numero de transacciones pero con un reducido número de ítems, del orden de 10000 tipos distintos, en nuestro caso tenemos mas de 1500000 transacciones, algo que aunque complicado no sería problema, pero con un lenguaje de 120000 ítems distintos, lo que imposibilitan la ejecución de Apriori que en estuvo

más de 4 días en ejecucion sin apenas obtener resultados. En cuanto al algoritmo FP-Growth en formato distribuido, el algoritmo tarda el orden en 1 a 2 horas en terminar su ejecucion dependiendo de si tenemos valores de soporte de 0.01 o 0.001.

6.4. Caso de uso: 28A

El pasado 28 de abril de 2019 tuvieron lugar las elecciones al generales para elegir el Gobierno de España. El PSOE [1] ganó las elecciones, que además pasarán a la historia como una de las elecciones generales donde la participación ciudadana fue más alta, situándose esta en el 71,76 %. Esta alta participación, junto con el 'clima' político de la época en el país, hacen de un análisis de patrones sobre las conversaciones generadas sobre el 28 de abril en Twitter, una aplicación muy interesante. Con el que intentaremos dar respuestas a preguntas como:

- ¿Hay relación entre patrones que asocian términos negativos o positivos con determinados políticos y los resultados posteriores?
- ¿Qué preocupaba a la sociedad española durante los días previos a las elecciones?

En las siguientes secciones, veremos patrones interesantes hallados así como métodos de visualización que permitan una mejor interpretación de las reglas obtenidas, con las que trataremos de dar respuesta a estas preguntas.

6.4.1. Patrones interesantes

En este capítulo entramos en detalle sobre algunos patrones interesantes de las reglas. Para una mejor comprensión de esta sección, iremos detallando patrones que puedan ayudar a responder las preguntas de investigación propuestas al inicio de esta sección.

[1] Partido Socialista Obrero Español

Antecedente	Consecuente	Confianza
mejor, candidato	pablo_casado	0.999
porespaña	vox	1.0
presagio, fantastico	santi_abascal	0,993
amenazado, agredido, simpatizantes	vox_es	1.0
blas, lezo, madrid, victoria	colon	0.999

Tabla 6.2: Reglas de apoyo y en contra.

¿Hay relación entre patrones que asocian términos negativos o positivos con determinados políticos y los resultados posteriores?

A priori, parece que no se puede trazar una relación entre los términos negativos y positivos y los resultados posteriores, debido a que se generan casi por igual número contenido que trata de apoyar como contenido dedicado a descalificar y atacar a los otros partidos. Por tanto, este tipo de análisis debería ser realizado de manera más exhaustiva y por circunscripciones electorales. En la tabla 6.2 podemos ver algunos de estos patrones.

¿Qué preocupaba a la sociedad española durante los días previos a las elecciones?

Podemos ver como en la sociedad española, había casi por igual preocupación por la irrupción de la extrema derecha así como por que el PSOE se mantuviera en el poder y el trato de favor que este partido pudiera tener con los partidos nacionalistas catalanes o vascos. Esto denota, como en España a pesar de que el bi-partidismo es cosa del pasado, aún los bloques son derecha e izquierda y las redes sociales en momentos de elecciones tienen un flujo de generación de contenido en contra y a favor muy similar. Algo que se vio remarcado con unos resultados donde el bloque de la derecha y la izquierda están casi igualados. Algunas reglas que constatan pueden verse en la tabla 6.3

6.4.2. Visualización

En esta última sección del capítulo de minería de datos, abordaremos el último de los objetivos visto en la sección 1.2. Trataremos de poner en valor

Antecedente	Consecuente	Confianza
complicidad, socios, pnv, proetarras	sánchez	1.0
pnv, socios, sánchez	hostigando	1.0
junto, independencia	cataluña	0,993
miedo, nadie, 28a	vox_es	1.0

Tabla 6.3: Reglas interesantes sobre los temas tratados en el discurso político.

Figura 6.1: Reglas sobre Sánchez en formato tag cloud.

el sistema con técnicas de visualización que permitan de una manera intuitiva para el usuario final obtener información sobre el gran conjunto de datos de entrada.

Dado que el número de reglas y términos es muy elevado, una manera interesante de visualizar los datos es crear *tag clouds* sobre los términos que aparecen en las reglas, de esta manera podremos en un mismo análisis gráfico ver todas las palabras asociadas con un determinado término, por ejemplo para visualizar algunas de las reglas relativas a Sánchez, podemos tener el gráfico 6.1.

6.4.3. Discusión sobre las reglas

En un conjunto de datos tan grande como el utilizado el número de reglas obtenidas es muy grande, lo que hace a ojos de una persona inexperta en política que sea una tarea muy complicada el obtener patrones relevantes, igualmente ha quedado constatado que el uso de reglas de asociación en Big Data textual es un método potente y que resume el contenido de de varios millones de tuits en sets de reglas, mas cómodos de trabajar y sobre todo, con relaciones sólidas entre los términos. En cuanto a la visualización, es menester crear nuevas técnicas de visualización de reglas para Big Data pues las técnicas actuales tienen muchas limitaciones cuando el número de reglas es muy elevado.

Capítulo 7

Conclusiones

El proyecto concluye con este capítulo donde se realiza una pequeña retrospectiva sobre los objetivos marcados, la valoración personal del proyecto y por último, concluiremos con las vías futuras que el trabajo abre y que sin duda serán exploradas en la próxima etapa de formación predoctoral que estoy a punto de comenzar.

7.1. Conclusiones y valoración

En el presente proyecto se ha llevado a cabo un exhaustivo trabajo de investigación en el que se ha realizado un estudio del estado del arte que deja en evidencia la gran mejora de capacidad de procesamiento en cuanto al volumen de datos que el sistema propuesto aporta. Se ha realizado un sistema que es, por tanto, capaz de operar con Big Data proveniente de Twitter, procesarlo y posteriormente obtener patrones interesantes en este caso, en el mundo de la política pero que podría llevarse a cabo con cualquier otro dominio.

En cuanto a las complicaciones encontradas, denotar cómo el volumen de los datos hace que aún con paradigmas distribuidos su procesado sea arduo, aunque los resultados baten tanto en tiempo como en resultados a los procesos secuenciales. También, es necesario mencionar la vital importancia de la fase de limpieza de datos cuando estos vienen de redes sociales, pues la cantidad

de ruido y problemas que presentan es muy elevada y pueden entorpecer a los algoritmos de minería de datos.

Para finalizar, en cuanto a retrospectiva del proyecto se han conseguido cada uno de los objetivos marcados al inicio del proyecto, así como se han mejorado e investigado aquellas vías futuras de investigación que se marcaron en el anterior proyecto final de máster en ingeniería y del cual este es una continuación y ampliación. En lo personal, ampliar el proyecto a un enfoque de Big Data, me ha ayudado a comprender mejor el paradigma del Big Data, así como el lenguaje Python y Spark dotándome de una serie de herramientas que sin duda me serán muy útiles en la próxima fase de estudio del doctorado.

7.2. Investigación futura

Las vías de investigación futuras que un trabajo de este tipo, que aúna redes sociales y Big Data, abre son muchas pues se trata de dos de los ámbitos de aplicación que mas interés suscitan hoy en día entre la comunidad científica. Nos centraremos por tanto, en aquellas vías de investigación que serán exploradas en la futura investigación predoctoral que me encuentro en vías de comenzar, sirviendo por tanto estas vías futuras como antesala de los objetivos de la futura tesis.

Primeramente, sería muy interesante trasladar el problema a una versión en streaming, de manera que podamos obtener información en tiempo real sobre conversaciones en Twitter. Por otro lado, sería interesante dado que el número de caracteres de un Tweet da bastante identidad a cada ítem, anteriormente con solo 140 caracteres era bastante pobre, sería muy interesante realizar una aproximación usando reglas de asociación difusas en las que cada ítem en una transacción no sería representado con 0 o 1, sino con el valor de frecuencia o por ejemplo el tf-idf nuevamente.

Una vía futura centrada en aplicación y visualización, podrían realizarse mapas de calor en función de las reglas y la ubicación geográfica de los tweets, algo muy útil en dominios de la política como es el caso que nos ocupa, pero que sería extrapolable a otros dominios de manera igualmente útil.

Para finalizar, otra vía futura relativa a visualización sería el estudio de nuevas técnicas de visualización para reglas de asociación, que permitieran de una manera dinámica seleccionar las reglas a visualizar, o permitir una

mejor visualización cuando las reglas impliquen un gran número de ítems, situaciones en las cuales los medios de visualización actuales dejan bastante que desear.

Bibliografía

[1] Abu Daher, L., Elkabani, I., and Zantout, R. Identifying influential users on twitter: A case study from paris attacks. *Applied Mathematics and Information Sciences 12* (09 2018), 1021–1032.

[2] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., and Gomes, J. B. A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications 55* (2016), 351–360.

[3] Agrawal, R., Imieliński, T., and Swami, A. Mining association rules between sets of items in large databases. In *Acm sigmod record* (1993), vol. 22, ACM, pp. 207–216.

[4] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[5] Bing, L., Chan, K. C., and Ou, C. Public sentiment analysis in twitter data for prediction of a company's stock price movements. In *2014 IEEE 11th International Conference on e-Business Engineering* (2014), IEEE, pp. 232–239.

[6] Cagliero, L., and Fiori, A. Analyzing twitter user behaviors and topic trends by exploiting dynamic rules. In *Behavior Computing*. Springer, 2012, pp. 267–287.

[7] Cagliero, L., and Fiori, A. Discovering generalized association rules from twitter. *Intelligent Data Analysis 17* (01 2013).

[8] Cambria, E., Speer, R., Havasi, C., and Hussain, A. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI Fall Symposium Series* (2010).

[9] Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., and Elgart, Z. Extracting patterns from twitter to promote biking. *IATSS Research 43*, 1 (2019), 51–59.

[10] Dehkharghani, R., Mercan, H., Javeed, A., and Saygin, Y. Sentimental causal rule discovery from twitter. *Expert Systems with Applications 41*, 10 (2014), 4950–4958.

[11] Dehkharghani, R., Mercan, H., Javeed, A., and Saygin, Y. Sentimental causal rule discovery from twitter. *Expert Systems with Applications 41*, 10 (2014), 4950–4958.

[12] Ediger, D., Jiang, K., Riedy, J., Bader, D. A., Corley, C., Farber, R., and Reynolds, W. N. Massive social network analysis: Mining twitter for social good. In *2010 39th International Conference on Parallel Processing* (2010), IEEE, pp. 583–593.

[13] Erlandsson, F., Bródka, P., Borg, A., and Johnson, H. Finding influential users in social media using association rule learning. *Entropy 18*, 5 (2016), 164.

[14] Fernandez-Basso, C., Francisco-Agra, A. J., Martin-Bautista, M. J., and Ruiz, M. D. Finding tendencies in streaming data using big data frequent itemset mining. *Knowledge-Based Systems 163* (2019), 666–674.

[15] Gundecha, P., and Liu, H. Mining social media: a brief introduction. In *New Directions in Informatics, Optimization, Logistics, and Production*. Informs, 2012, pp. 1–17.

[16] Hahsler, M., and Karpienko, R. Visualizing association rules in hierarchical groups. *Journal of Business Economics 87*, 3 (2017), 317–335.

[17] Hai, Z., Chang, K., and Kim, J.-j. Implicit feature identification via co-occurrence association rule mining. In *International Conference*

on Intelligent Text Processing and Computational Linguistics (2011), Springer, pp. 393–404.

[18] Hamed, A. A., Wu, X., and Rubin, A. A twitter recruitment intelligent system: association rule mining for smoking cessation. *Social Network Analysis and Mining 4*, 1 (2014), 212.

[19] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In *ACM sigmod record* (2000), vol. 29, ACM, pp. 1–12.

[20] Kakulapati, V., and Reddy, S. M. Mining social networks: Tollywood reviews for analyzing upc by using big data framework. In *Smart Innovations in Communication and Computational Sciences*. Springer, 2019, pp. 323–334.

[21] Kwon, K., Jeon, Y., Cho, C., Seo, J., Chung, I.-J., and Park, H. Sentiment trend analysis in social web environments. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* (2017), IEEE, pp. 261–268.

[22] Laney, D. 3D data management: Controlling data volume, velocity, and variety. Tech. rep., META Group, February 2001.

[23] Liu, B., and Zhang, L. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 2012, pp. 415–463.

[24] Mamgain, N., Pant, B., and Mittal, A. Categorical data analysis and pattern mining of top colleges in india by using twitter data. In *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)* (2016), IEEE, pp. 341–345.

[25] Mandave, P., Mane, M., and Patil, S. Data mining using association rule based on apriori algorithm and improved approach with illustration. *International Journal of Latest Trends in Engineering and Technology (IJLTET), ISSN* (2013).

[26] Margono, H., Yi, X., and Raikundalia, G. K. Mining indonesian cyber bullying patterns in social networks. In *Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147* (2014), Australian Computer Society, Inc., pp. 115–124.

[27] Martin-Bautista, M., SÃ¡nchez, D., Serrano, J., and Vila, M. Text mining using fuzzy association rules.

[28] Mosley Jr, R. C. Social media analytics: Data mining applied to insurance twitter posts. In *Casualty Actuarial Society E-Forum* (2012), vol. 2, Citeseer, p. 1.

[29] Mourtzis, D., Vlachou, E., and Milas, N. Industrial big data as a result of iot adoption in manufacturing. *Procedia cirp 55* (2016), 290–295.

[30] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems 50*, 3 (2011), 559–569.

[31] Noferesti, S., and Shamsfard, M. Resource construction and evaluation for indirect opinion mining of drug reviews. *PloS one 10*, 5 (2015), e0124993.

[32] Oktay, H., Taylor, B. J., and Jensen, D. D. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics* (2010), ACM, pp. 1–9.

[33] Pak, A., and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (2010), vol. 10, pp. 1320–1326.

[34] Petland, A. Reinventing society in the wake of big data, 2012.

[35] Phan, H. T., Nguyen, N. T., and Hwang, D. A tweet summarization method based on maximal association rules. In *International Conference on Computational Collective Intelligence* (2018), Springer, pp. 373–382.

[36] Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., and Valencia-Garcia, R. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine 2017* (2017).

[37] Serrano-Cobos, J. Big data y analítica web. estudiar las corrientes y pescar en un océano de datos. *El profesional de la información 23*, 6 (2014), 561–565.

[38] Silverstein, C., Brin, S., Motwani, R., and Ullman, J. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery 4*, 2-3 (2000), 163–192.

[39] Srikant, R., and Agrawal, R. Mining generalized association rules. *Future Generation Computer Systems 13*, 2 (1997), 161 – 180.

[40] Weiss, S. M., and Indurkhya, N. *Predictive data mining: a practical guide*. Morgan Kaufmann, 1998.

[41] Yin, Y., Kaku, I., Tang, J., and Zhu, J. Association rules mining in inventory database. In *Data Mining* (2011), Springer, pp. 9–23.

[42] Yuan, M., Ouyang, Y., Xiong, Z., and Sheng, H. Sentiment classification of web review using association rules. In *International Conference on Online Communities and Social Computing* (2013), Springer, pp. 442–450.

[43] Zainol, Z., Wani, S., Nohuddin, P. N., Noormanshah, W. M., and Marzukhi, S. Association analysis of cyberbullying on social media using apriori algorithm. *International Journal of Engineering & Technology 7*, 4.29 (2018), 72–75.

[44] Zhou, X., Tao, X., Yong, J., and Yang, Z. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (2013), IEEE, pp. 557–562.