UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Diseño e implementación de un sistema automático de clasificación de mensajes intercambiados entre la ciudadanía y el Ayuntamiento de València

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: Marylin Leonor Mattos Barros

Tutor: Diego Álvarez

[2017-2018]

Resumen

En este documento se presenta el marco de análisis, diseño e implementación de una herramienta que permite visualizar la correlación que existe entre los temas que habla la ciudadanía en una red social y el volumen de datos publicados por el ayuntamiento en su catálogo de datos, con el fin de promover la apertura de los datos desde la demanda de la ciudadanía. El documento está dividido en seis partes, la primera ofrece una introducción al problema y su posible solución, los objetivos marcados y la metodología seleccionada. Una segunda parte contiene los avances que ha realizado el gobierno español en el marco político de los datos abiertos y el papel que desempeña la ciudadanía en el desarrollo de las políticas públicas. También, se presentan algunas técnicas de *Machine Learning* como herramientas de análisis de información en redes sociales. En la tercera y cuarta parte, se hace un análisis detallado al problema, se plantea la solución y se siguen dos metodologías de desarrollo, una para la parte que responde al problema de minería de texto y otra para la parte web. Los dos capítulos finales contienen el análisis de los resultados, y los trabajos que pueden derivar de este TFM y las conclusiones.

Palabras clave: datos abiertos, minería de texto, *Machine Learning*, análisis de contenido, Ajuntament de València, políticas públicas, Twitter, gobierno abierto.

Abstract

This work presents one tool's analysis framework, design and implementation to visualize the correlation between interest topics spoken by citizen in social networks and data published by the “ayuntamiento”. This work aims to promote the data opening, having in mind the citizens needs and demand. The project consists of 6 chapters, the first provides the introduction to the problem and possible solutions for it, the main objectives and chosen methodology as well. The second one collects advancements made by the Spanish government in its political framework regarding open data and the citizen functions in the pubic politic development. Also, Machine Learning Technics are presented as a tool to process and analyze social networks information. In the third and fourth chapters, a deeper and more detailed analysis about the problem is made, and the problem solution is presented, in addition to the development methodologies regarding data mining for text and the another for the website side. The two final chapters have the results analysis, main conclusions and possible works that might derivative from this project.

Key words: open data, text mining, Machine Learning, content analysis, Twitter.

Tabla de contenidos

Índice de Figuras

Índice de Tablas

Capítulo 1
Introducción

La historia del gobierno Abierto se remonta al siglo XVIII en Suecia, al instaurar en su legislación la "Libertad de prensa y el derecho de acceso a los archivos". Dos siglos después Estados Unidos y el Reino Unido se sumaron a la iniciativa y, a partir de ese momento muchos países permitieron el acceso de la ciudadanía a la información pública (Sánchez Trigueros, 2015)

La innovación en la gestión pública comienza en el siglo XX con la revolución digital que trajo consigo la administración electrónica incorporada por la Administración Pública (de ahora en adelante AAPP). Desde ese momento las AAPP encargadas de ejecutar las políticas públicas, han tratado de convertir la información que dispone en "materia prima digital" que pueda ser utilizada tanto por el poder político como por la ciudadanía. Por su parte, la Unión Europea a través de la Directiva 2003/98/CE, relativa a la reutilización de la información del sector ha intentado abrir el camino para regular la solicitud y uso de la información proveniente de la administración pública, pero el mundo de hoy necesita algo más, necesita que haya una comunicación bidireccional entre el gobierno y el ciudadano, y que se incorpore la sociedad en las políticas públicas para opinar e incidir en el ciclo que las conforman.

La Alianza para el Gobierno Abierto (OGP por sus siglas en inglés)[1] que cuenta hoy con 75 miembros, siendo España uno de ellos desde el año 2011, a través de la Carta Iberoamericana de Gobierno Abierto (CIGA) define al gobierno abierto como:

> *"el conjunto de mecanismos y estrategias que contribuye a la gobernanza pública y al buen gobierno, basado en los pilares de la transparencia, participación ciudadana, rendición de cuentas, colaboración e innovación, centrando e incluyendo a la ciudadanía en el proceso de toma de decisiones, así como en la formulación e implementación de políticas públicas, para fortalecer la democracia, la legitimidad de la acción pública y el bienestar colectivo"* (Centro Latinoamericano de Administración para el Desarrollo, 2016).

En este orden de ideas, tenemos ante nosotros un emergente paradigma de gobierno abierto generado a partir de una nueva forma de articular iniciativas en los principios fundamentales.

Con el objetivo de mejorar la participación ciudadana en la formulación e implementación de las políticas públicas, los gobiernos deberían hacer uso de

[1] Open Government Partnership. Se recuperó el 25 de julio del 2018 de https://www.opengovpartnership.org/

las plataformas tecnológicas y redes sociales como canales de escucha activa a las consultas, demandas o peticiones de información por parte de la ciudadanía. En el caso específico de España, en uno de los últimos estudios dispuestos "Estudio de la Demanda y uso de Gobierno Abierto en España" (Márquez Fernández, et al., 2013) se evidencia que las personas con mayor demanda de información pública son jóvenes entre 18 y 34 años con estudios superiores, con presencia en redes sociales y usuarios de Administración electrónica. En este estudio también se muestran las estadísticas de los mecanismos considerados por la ciudadanía como más acertados para ellos acceder a la información pública; donde el listado lo encabeza el "envío de la Administración a los ciudadanos que lo soliciten" con un 35,8%, seguido por las "redes sociales" con un 32,6%, en tercera posición se encuentran las "páginas web oficiales" con un 17,9%; el 13,7% restante lo conforman otros mecanismos. Estos mecanismos más que para acceder a la información deben ser considerados como mecanismos para demandar información a las AAPP; información que debería ser publicada en el Portal de datos abiertos o usadas a través de herramientas que beneficien a la mayor cantidad de ciudadanos posibles.

Ahora bien, hablando del proceso de requerir información. Aquellos que en algún momento hayan realizado solicitudes a las AAPP españolas por alguno de los canales dispuestos para ello, conocen de primera mano las demoras en que se incurren para dar solución a los distintos requerimientos. Según información proporcionada por la Dirección General de Gobernanza Pública, sobre las solicitudes realizadas al Portal de la Transparencia de la Administración del Estado de España[2], desde la puesta en marcha del portal, en diciembre del 2014 hasta el 30 de junio del 2018 tardaron en promedio 31 días en responder a las solicitudes realizadas por la ciudadanía. Esta administración indica que en el 2014 se resolvieron 591 solicitudes en un tiempo promedio de 43,4 días, en 2015 se resolvieron 2978 en 27,4 días, en 2016 se resolvieron 3167 en 26,9 días, en 2017 se resolvieron 4025 en 31,8 días y, por último, en el 2018 se han resuelto 2398 en un tiempo de 27,3 días.

La Dirección General de Gobernanza Pública hace énfasis en que en dichos plazos se encuentran incluidos las suspensiones o ampliaciones de plazos previstas en los siguientes artículos de la Ley 19/2013: 19.2 (10 días para concreción de la solicitud), 19.3 (15 días para consulta a terceros afectados) y 20.1, párrafo 2º (un mes por ampliación de plazo por volumen o complejidad de la información solicitada). Estas interrupciones y ampliaciones hacen que el plazo de resolución de algunos expedientes supere el plazo de un mes previsto en el artículo 20.1.

Si hacemos un análisis de los datos proporcionados, las AAPP ningún año han respondido dentro de los 25 días establecidos por los artículos 19.2 y 19.3 de la Ley 19/2013, respondiendo siempre dentro del mes de ampliación. Vale la

[2] Para ampliar la información proporcionada por la Dirección General de Gobernanza Pública consultar el Apéndice B de este documento.

pena resaltar que en las cifras proporcionadas no se incluye el tiempo desde que el ciudadano hace la solicitud, sino, desde que la solicitud es recibida por el órgano competente por lo que la media de tiempo desde que el ciudadano hace la solicitud hasta que es resuelta es mucho mayor de la que se nos ha proporcionado. Todos estos plazos y tardanzas pueden crear insatisfacción en la ciudadanía frente a la forma en que las AAPP atienden sus solicitudes y gestionan la información.

Por tanto, las AAPP no sólo deben ceñirse a que el ciudadano les solicite la información, sino que debería estudiar sus intereses para adelantarse a sus requerimientos y lograr convertir la información pública en un recurso cívico a disposición de la ciudadanía. Existen muchas formas de conocer los intereses de la ciudadanía, pero todas ellas tienen algo en común, saber escuchar. Entonces, por qué no ir a los lugares donde se reúnen y "espiar" sus conversaciones e impresiones. Si volvemos a lo que nos indica el estudio referente al perfil de las personas que hacen demanda de información, se puede inferir que las redes sociales pueden ser ese espacio de reunión que estamos buscando.

Según el "IAB Estudio Anual de Redes Sociales 2017" las 5 redes más utilizadas en el ámbito geográfico español son: Facebook, WhatsApp, YouTube, Twitter e Instagram. Si analizamos el tipo de contenido que se comparte a través de estas cinco redes sociales podemos descartar casi inmediatamente a WhatsApp, YouTube e Instagram por el tipo de contenido que manejan y quedarnos con Facebook y Twitter.

De Twitter podemos decir que es una plataforma de microblogs con mensajes públicos en su gran mayoría, lo que lo hace transparente en término de datos, además cuenta con una potente API (*Application programming interface*) que permite a programas informáticos obtener información. Facebook, en cambio es mucho más restrictiva; para poder obtener datos es necesario que las personas acepten que nuestra aplicación obtenga sus datos, sin contar con la poca información que se dispone de su API.

La libertad en la obtención de la información hace que nuestra investigación se decante por Twitter como la red social donde escuchar a la ciudadanía.

Lo que nos proponemos hacer es un análisis de los mensajes enviados por la ciudadanía a las AAPP a través de la red social Twitter y examinar la forma en que esta comunicación pueda influir en el ciclo de vida de apertura de los datos.

Como caso de estudio se usará el Ajuntament de València y como recurso de automatización del proceso se implementará la disciplina computacional, *Machine Learning* que permitirá automatizar el análisis de los mensajes. Con esta investigación se espera que el Ajuntament de València a la hora de elaborar sus políticas de datos abiertos tenga en cuenta que debe existir un equilibrio entre lo que necesitan las personas y la apertura de sus datos.

1.1 Motivaciones

Este trabajo fin de máster nace del interés de la cátedra Govern Obert, del Ajuntament de València y del mío propio de crear un proyecto que tuviera una utilidad real, donde la mayor beneficiada fuese la ciudadanía.

Durante la búsqueda del tema, me pareció fascinante la idea de crear un proyecto donde pudiera combinar el conocimiento técnico adquirido a lo largo de mi carrera profesional, con la creación de una herramienta que pudiera llegar a tener un impacto social. Con el desarrollo de esta herramienta no solo estaría fortaleciendo mis conocimientos, sino que a la vez estaría poniendo mi granito de arena en la creación de una sociedad más justa, donde la ciudadanía de alguna manera pueda expresar las necesidades de información que tiene e incidir en las políticas públicas.

1.2 Objetivos

1.2.1. Objetivo General

Diseñar e implementar un sistema automático de clasificación de mensajes intercambiados entre la ciudadanía y el Ajuntament de València para co-definir la política pública de transparencia y datos abiertos.

1.2.2. Objetivos Específicos

- Buscar y seleccionar fuentes de información que puedan servir de entradas al producto software a desarrollar.
- Diseñar la arquitectura de software bajo la cual se desarrollará la aplicación, incluyendo la identificación de componentes y tecnologías.
- Identificar e implementar técnicas de tratamiento de datos para que el procesamiento de la información se realice con la menor cantidad de ruido posible.
- Revisar y evaluar algoritmos de clasificación que permitan seleccionar la mejor opción a ser usada en la creación de la máquina de clasificación automática de mensajes.
- Desarrollar una máquina entrenada con la ayuda de la ciudadanía, que sea capaz de predecir automáticamente las categorías a las que pertenecen los nuevos textos ingresados en el sistema.
- Diseñar e implementar visualizaciones que permitan a los usuarios, en especial al ayuntamiento, conocer la evolución de los temas de conversación que la ciudadanía tiene hacia el Ajuntament de València y

de esta manera saber de qué información requiere la gente, antes que realicen solicitudes formales de ésta.

- Crear visualizaciones que permitan a los usuarios, en especial a los ciudadanos, comparar entre el volumen de sus conversaciones por temática hacia el Ajuntament de València y el volumen de *datasets* publicados por este.

1.3 Impacto esperado

Este proyecto pretende tener un impacto positivo tanto en la ciudadanía como en las AAPP, en aras de incentivar la apertura de los datos desde la demanda de información. Donde la ciudadanía desde sus conversaciones en las redes sociales pueda influir en la demanda de información pública y las AAPP se den cuenta que pueden utilizar como entrada del proceso de elaboración de la política pública de apertura de datos, la información producto de la escucha de los canales de comunicación de la ciudadanía.

1.4 Metodología

Antes de comenzar a explicar la metodología utilizada para desarrollar este trabajo fin de máster, es necesario indicar que el proyecto nace de un convenio de colaboración entre el Ajuntament de València y la Universitat Politècnica de València. El objetivo del convenio consistía en elaborar soluciones tecnológicas que facilitaran el acceso a la información pública, usando un diseño centrado en la ciudadanía.

Partiendo de esta premisa, nos reunimos con personal del Ajuntament de València para concretar más la idea y establecer una metodología de trabajo. Al combinar la academia y la interacción con el ayuntamiento se le da vida a cada uno de los capítulos que se narran a lo largo del documento y que se materializan a través del desarrollo de una herramienta software.

El camino a la solución del problema comienza con el estudio del estado del arte, por lo que se hace una revisión bibliográfica donde se comparan conceptos, teorías y estudios sobre el tema a tratar. Se plantea y describe la forma en la que se abordará la solución y se establecen las metodologías para desarrollar la aplicación, que será el hilo conductor a través del cual evolucionará la solución.

1.5 Estructura de la memoria

El presente documento se ha estructurado en seis capítulos. Además de bibliografía y apéndices.

- **El capítulo 2:** contiene los conceptos que enmarcan el desarrollo de este trabajo, así como el estado del arte.
- **El capítulo 3:** contiene un análisis detallado del problema, el planteamiento y diseño de la solución, así como las metodologías de desarrollo que guiarán el proceso de creación del producto software.
- **El capítulo 4:** contiene todo el proceso de construcción del producto software, dividido en dos módulos.
- **El capítulo 5:** contiene el análisis de los resultados obtenidos a través del uso de la aplicación y un análisis crítico de la herramienta, donde se muestran sus limitaciones y los posibles trabajos futuros que se pueden desarrollar a partir de ella.
- **El capítulo 6:** contiene conclusiones generales y unas conclusiones específicas, creadas a partir de los objetivos trazados al principio

1.6 Colaboraciones

El proyecto nace en el marco de la cátedra Govern Obert producto de una iniciativa surgida de la colaboración entre la Concejalía de Transparencia, Gobierno Abierto, con la cooperación del Ajuntament de València y la Universitat Politécnica de València. La cátedra Govern Obert es un espacio de trabajo colaborativo integrado por estudiantes que buscan soluciones a problemas que impiden que los ciudadanos y ciudadanas puedan influir en los asuntos que les afectan y, así, construir una sociedad más justa.

Capítulo 2
Estado del arte

Este capítulo está dividido en dos apartados principales, en el 2.1 se presenta los avances que ha realizado el gobierno español en el marco político de los datos abiertos y el papel que desempeña la ciudadanía en el desarrollo de las políticas públicas; por otra parte, en el apartado 2.2 se presentan las técnicas de *Machine Learning* como herramientas de recuperación de información en redes sociales.

2.1 Política de datos abiertos

Las políticas públicas de datos abiertos permiten que la información producida y almacenada por los organismos gubernamentales sean puestas a disposición de la ciudadanía en formatos abiertos y legibles tanto para personas como para máquinas, salvaguardando la privacidad, la confidencialidad y la seguridad. Propiciando que la información pueda ser reutilizada a través de programas informáticos que incentiven la innovación en el sector gubernamental y el sector privado.

Las políticas de datos abiertos se construyen sobre normas, leyes y decretos que regulan tanto la publicación como el acceso público a la información. A nivel legislativo los diez países pioneros en derecho a la información fueron: Suecia (1766), Finlandia (1952), Estados Unidos (1966), Dinamarca y Noruega (1970), Francia y Holanda (1978), Australia y Nueva Zelanda (1982) y, por último, Canadá (1983) (Ràfols, 2015)·Entre 1990 y 2007 más de 40 países se sumaron a la iniciativa.

A pesar del desarrollo legislativo, el inicio de la nueva era de innovación democrática se da a finales del año 2007 cuando treinta defensores del gobierno abierto (Carl Malamud, Public.Resource.Org, 2007) se reunieron en California y redactaron un conjunto de 8 principios de apertura de datos públicos (Benito Carrillo, 2016). Sin embargo, tuvieron que pasar dos años para que ocurriera un cambio significativo. Nos referimos al cambio promovido por el presidente de los Estados Unidos, Barack Obama que a través del "Memorando sobre Transparencia y Gobierno Abierto"[3] inicia una agenda de datos abiertos y declara que *"la apertura fortalecerá la democracia y promoverá la eficiencia y la efectividad en el gobierno"* (Quintanilla Mendoza & Gil García, 2013). La agenda propuesta por Obama se materializaba meses después, con el lanzamiento de la web de publicaciones de datos del gobierno estadounidense,

[3] President's Memorandum on Transparency and Open. Se recuperó el 24 de julio del 2018 de https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2009/m09-12.pdf

Data.gov[4]. Ese mismo año en España surge la iniciativa "Aporta" promovida por el Ministerio de Energía, Turismo y Agenda Digital en colaboración con el Ministerio de Hacienda y Función pública, con el fin de promover la apertura de la información en el país. Fruto de esta iniciativa, el 24 de octubre de 2011 nace datos.gob.es[5], la plataforma que organiza y gestiona el Catálogo Nacional de Datos Abiertos[6].

En septiembre del 2011 se crea la Alianza para un Gobierno Abierto con 8 países fundadores entre los que se encontraban Estados Unidos y el Reino Unido. En el acto de lanzamiento también se aprueba la Declaración para un Gobierno Abierto[7] y se anuncian los planes de acción nacionales por parte de los países fundadores.

En el año 2012, 38 países más, entre ellos España, se adhieren a la Alianza para un Gobierno Abierto, incrementando el número de encuentros entre países, cuya convicción y/o creencias se base en el poder de la transparencia y lucha contra la corrupción.

En mayo del 2013 en Estados Unidos, el presidente Obama firma la orden ejecutiva de gobierno abierto, por la que *"la información del Gobierno por defecto se publicará en abierto y lectura automática"* (Cohen, 2013) Al mes siguiente, los líderes del G8 llegan a un acuerdo sobre sus políticas de datos abiertos por lo que firman la Carta de Datos Abierto,[8] documento que establece los cinco principios fundamentales para facilitar la publicación y el acceso a la información del sector público en las naciones integrantes.

Los años siguientes han sido muy buenos para el gobierno abierto y las políticas de datos abiertos, las administraciones públicas de todo el mundo se han dado cuenta de la importancia de abrir sus datos a los ciudadanos y han ido creando o modificando leyes para darle cabida a esta nueva forma de gobernanza.

En el último informe sobre madurez de los datos abiertos en la Unión Europea[9] publicado por el Portal de Datos Europeo (EDP)[10] se muestra en cifras los indicadores de la disponibilidad de datos abiertos, donde no solo se mide la presencia de una política de datos abiertos, sino que se analiza en qué medida

---

4 Data.gov. Se recuperó el julio 24, 2018 de https://www.data.gov/

5 Datos.gob.es. Se recuperó el 24 de julio del 2018 de http://datos.gob.es/es

6 PAe - CTT - General - datos.gob.es. Se recuperó el 24 de julio del 2018 de https://administracionelectronica.gob.es/ctt/datosgob

7 Open Government Declaration | Open Government Partnership. Se recuperó el 25 de julio del 2018 de https://www.opengovpartnership.org/open-government-declaration

8 Carta Internacional de Datos Abiertos - Open Data Charter. Se recuperó el 24 de julio del 2018 de https://opendatacharter.net/principles-es/

9 Open Data Maturity in Europe n3 - 2017 - European Data Portal. 1 nov.. 2017. Se recuperó el 25 de julio del 2018 de https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

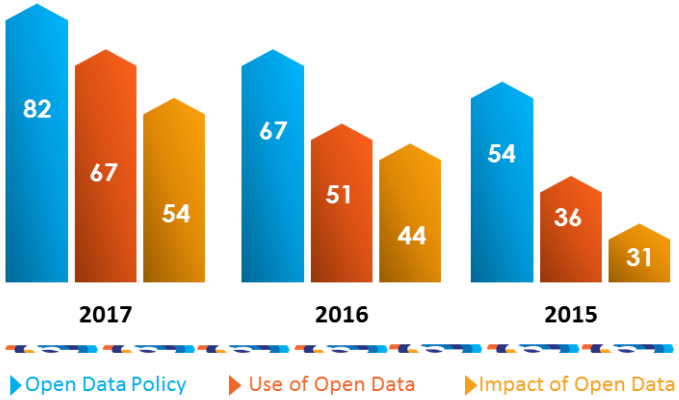
10 European Data Portal. Se recuperó el 25 de julio del 2018 de https://www.europeandataportal.eu/en/homepage.

los datos abiertos se utilizan y cómo los datos abiertos impactan a la sociedad desde una perspectiva política, social y económica.

En la Figura 1 se puede apreciar cómo el indicador de *Política de Datos Abiertos* ha pasado de un 54% inicial en el 2015 a un 82% en el 2017. Por su parte el de *Uso de Datos Abiertos* pasó de tener un grado de madurez de 36% a un 67%.

Por último, el *Impacto de los Datos Abiertos* es el que menos avance reporta en comparación con los dos anteriores, pasando de un 31% inicial en el 2015 a un 54% en el 2017.

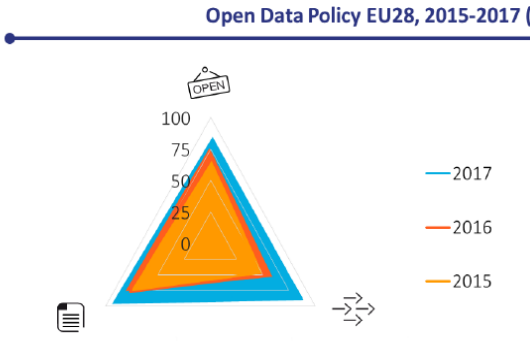
Figura 1 Disponibilidad de datos abiertos - EU28 (2015-2017, % de resultados promedio)



Fuente: (European Data Portal, 2017)

Si se pone la lupa en la evolución de las *Políticas de Datos Abiertos*, en la Figura 2 se muestra el desarrollo que ha tenido en el marco de tres subindicadores (presencia de la política, coordinación nacional y normas de licencia). Donde se evalúa en qué medida los países cuentan con una política de datos abiertos, si sus normas de licencia cumplen con los requisitos para llamarse datos abiertos y hasta qué punto existe una coordinación sobre datos abiertos.

Figura 2 Política de datos abiertos, Evolución 2015-2017, desglose por subindicador (%)

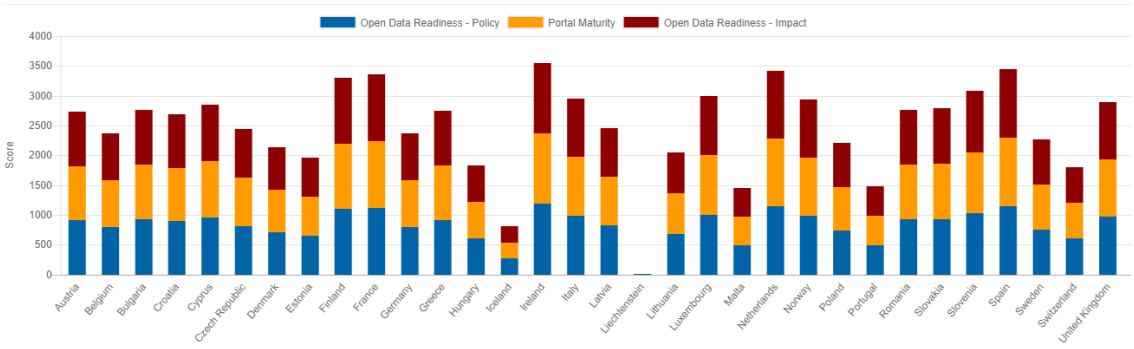


EU28 Open Data Policy	2017	2016	2015
Presence of Policy	79	71	61
National Coordination	87	52	47
Licensing norms	89	77	74

Fuente: (European Data Portal, 2017)

En relación con la situación de España (ver Figura 3), este país alcanzó la mayor puntuación posible en el indicador de *Impacto de los Datos Abiertos,* no obstante, el indicador de *Políticas de Datos Abiertos* mostró una desmejora con respecto al año 2016. Esto se debe a las normas de licenciamiento y a que el enfoque sobre cómo abrir datos no ha cambiado desde mediados de 2016. Vale la pena resaltar que, según este informe, España es uno de los referentes europeos en datos abiertos ocupando el segundo puesto del ranking con 1150 puntos, siendo superado por Irlanda por una diferencia de 15 puntos.

Figura 3 Ranking por países del nivel de madurez de los datos abiertos en Europa



Fuente: (European Data Portal, 2017)

Como el objeto de estudio de las políticas de datos abiertos, son precisamente los datos, se hace necesario dedicarle un epígrafe para definirlo y conocer sus principios.

2.1.1 Datos abiertos

Existen varias definiciones de datos abiertos, también conocidos como Open Data. Según la *Open Knowledge International* a través del portal *Open Data Handbook* se define el concepto "datos abiertos" como aquellos:

> *"datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen"* (Open Knowledge International, s.f.).

Por su parte, la Alianza para un Gobierno lo define como:

> *"la idea de que los datos deben estar libremente disponibles para que todos puedan acceder, usar y republicar como lo deseen, publicados sin restricciones de derechos de autor, patentes u otros mecanismos de control"* (The Open Government Partnership, s.f.).

La Dirección General de Gobernanza Pública de España resume la definición en que:

> *"Los datos abiertos son aquellos que se consideran datos accesibles y reutilizables, sin exigencia de permisos específicos"* (Administracion.gob.es, s.f.)

En el marco del presente proyecto, consideramos que los datos abiertos son datos que podemos utilizar y compartir sin ningún tipo de restricción, salvo las establecidas en los principios de datos abiertos, en formatos que pueden ser usados tanto por personas, como por máquinas, con el fin de poder crear programas informáticos que ayuden a sacar el mayor provecho posible a la información.

Ahora bien, si nos referimos a los datos del gobierno abierto (*Open Government Data*) debemos hacer referencia a los datos producidos por las AAPP en el desarrollo de sus funciones. Datos que son recopilados y almacenados con dinero público, por tal motivo deberían estar a disposición de la ciudadanía para su uso y promover con esto una mayor participación civil y un uso más eficiente de los recursos públicos.

Debido a que el *Open Government Data* no se estaba desarrollando de forma adecuada, durante el año 2007 la Sunlight Foundation impulsó la creación de 8 principios de apertura de datos públicos, que de ser adoptados los gobiernos del mundo pueden ser más efectivos, transparentes y relevantes para sus ciudadanos. En el año 2010 la lista de los principios fue ampliada a 10, indicando que los datos deben ser (Sunlight Foundation, 2017):

1. **Completos:** deben ser publicados sin procesar (en bruto) en la medida que esto no viole las leyes de privacidad o seguridad. En la publicación se deben incluir sus metadatos con el objetivo de que el usuario comprenda el alcance de la información.

2. **Primarios:** deben ser publicados en las fuentes de origen, con detalles en la forma de recopilación para que los usuarios puedan verificar que la información se recopiló correctamente y se registró con precisión.

3. **Oportunos:** deben ser divulgados tan pronto como sea recolectados priorizando los datos cuya utilidad es sensible al tiempo.

4. **Accesibles de forma física y electrónica:** deben ser accesibles tanto de forma física como electrónica, esta última debe permitir la descarga de información y el uso de APIs que hacen que los datos sean mucho más fáciles de acceder por programas informáticos.

5. **Procesables por Máquinas:** deben almacenarse en formatos de archivo ampliamente utilizados que se prestan fácilmente al procesamiento de la máquina.

6. **No discriminatorios:** deben permitir que cualquier persona puede acceder a los datos en cualquier momento.
7. **Uso de estándares abiertos:** deben evitar que para el uso de la información sea necesario usar una licencia de software.
8. **Sin licencia:** debe ser publicada sin restricciones de uso salvo restricciones de privacidad.
9. **Permanencia:** debe permanecer en línea, con el seguimiento de versiones adecuado y el archivo en el tiempo.
10. **Costes:** no se debe fijar un precio para poder acceder o reutilizar los datos y en el caso que se establezca este debe basarse en costes marginales y no en costes totales.

Los datos abiertos pueden generar valor en diferentes áreas e impulsar acciones que benefician a la ciudadanía, tales como:

- **La transparencia:** la transparencia además de ser un beneficio que se obtiene con la apertura de los datos públicos es uno de los principios de un gobierno abierto, donde las personas físicas o jurídicas puedan hacer uso de los datos generados o custodiados por las AAPP ya sea a través del ejercicio de su derecho de acceso a la información o a través de la obligación de la publicidad activa.
- **La eficiencia:** se promueve la sinergia entre los organismos gubernamentales, haciendo que los datos sean más fáciles de encontrar, analizar y combinar en diferentes departamentos y agencias (Partnership, s.f.).
- **La innovación:** al dotar de información a la sociedad se generan nuevos espacios que propicien iniciativas que involucran a los diferentes actores sociales (académicos, empresarios, trabajadores, etc.) y potencian la creación de negocios y servicios innovadores que brinden valor social y comercial (Partnership, s.f.).

2.1.2 Marco legislativo en la apertura de datos en España

Las políticas de datos abiertos en España se vienen trabajando desde el año 2007 cuando se implementaron las primeras estrategias hacia la apertura de datos públicos y su reutilización.

A través de la Ley 37/2007, de 16 de noviembre, producto de la incorporación al ordenamiento jurídico Español de la Directiva 2003/98/CE, de 17 de noviembre de 2003, del Parlamento Europeo y del Consejo, se sientan las bases para poner a disposición de la sociedad por medio electrónico los documentos que se encuentran en poder del sector público y que son de libre disposición. Se autoriza a las AAPP el otorgar o no, el acceso a los documentos y se comienza a

tener en cuenta características de los datos abiertos, tales como, coste, plazos y licencia. El Real Decreto 1495/2011 de 24 de octubre, desarrolla la citada Ley, para el ámbito del sector público estatal y motiva a la creación del Catálogo Nacional de Datos Abiertos, así como, la elaboración de la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información que debido a la relevancia que tiene en este trabajo, se le dedicará un epígrafe más adelante.

Por último, según lo establecido en la Ley 19/2013 de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, se regula y garantiza tanto el derecho de la ciudadanía a solicitar información de los asuntos públicos, como la publicidad activa. Es obligación de las AAPP suministrar información de manera permanente, sin necesidad de ser solicitada. A través de esta Ley se obliga a las AAPP a divulgar la información en las correspondientes sedes electrónicas, páginas webs o portales de la transparencia de una forma clara, oportuna y accesible para que la ciudadanía pueda informarse y juzgar con mejor y más criterio, el manejo que se le están dando a los asuntos públicos. Aquí he de referirme también a la Ley 2/2015, de 2 de abril, de Transparencia, Buen Gobierno y Participación Ciudadana de la Comunitat Valenciana que en los términos previstos en el artículo 49.1.1 del Estatuto de Autonomía de la Comunitat Valenciana adopta la Ley 19/2013 y adiciona contenido, dentro de los que está la participación ciudadana en las políticas públicas de la Generalitat.

2.1.3 Norma Técnica de Interoperabilidad de Reutilización de recursos de la información[11]

Esta norma tiene como objetivo

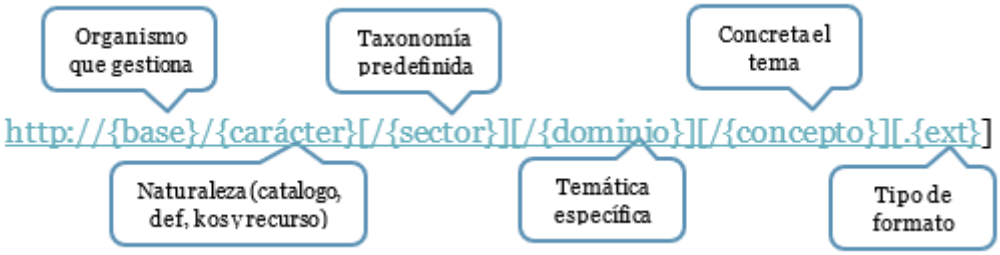
> *"facilitar y garantizar el proceso de reutilización de la información de carácter público procedente de las Administraciones públicas, asegurando la persistencia de la información, el uso de formatos, así como los términos y condiciones de uso adecuados"* (La Norma técnica de interoperabilidad de reutilización de recursos de información, 2013)

Al mismo tiempo, establece las directrices para seleccionar la información que es susceptible a ser utilizada y procede a identificarla, describirla y representarla en los formatos adecuados para ser reutilizada. Tres de las seis condiciones de uso de la Norma técnica son detallados a continuación por ser de suma importancia para el desarrollo de este trabajo:

Identificar los recursos de información*:* Debe realizarse por medio de URI (ver Figura 4) donde se establecen las referencias únicas que deben tener los conjuntos de datos y recursos de información.

[11] Para ampliar la información sobre la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información, remitirse a https://www.boe.es/boe/dias/2013/03/04/pdfs/BOE-A-2013-2380.pdf

Figura 4 Definición de un esquema URI



Fuente: Elaboración propia

Para categorizar los catálogos de recursos de información pública y sus registros se debe seleccionar un identificador del sector (primario), según lo especificado en el anexo IV de la NTI (ver Tabla 1). Esta taxonomía está definida como un esquema de conceptos identificado mediante el URI http://datos.gob.es/kos/sector-publico/sector a la que se concatenaría el identificador del sector, quedando identificado de manera unívoca cada sector como se muestra en la tabla.

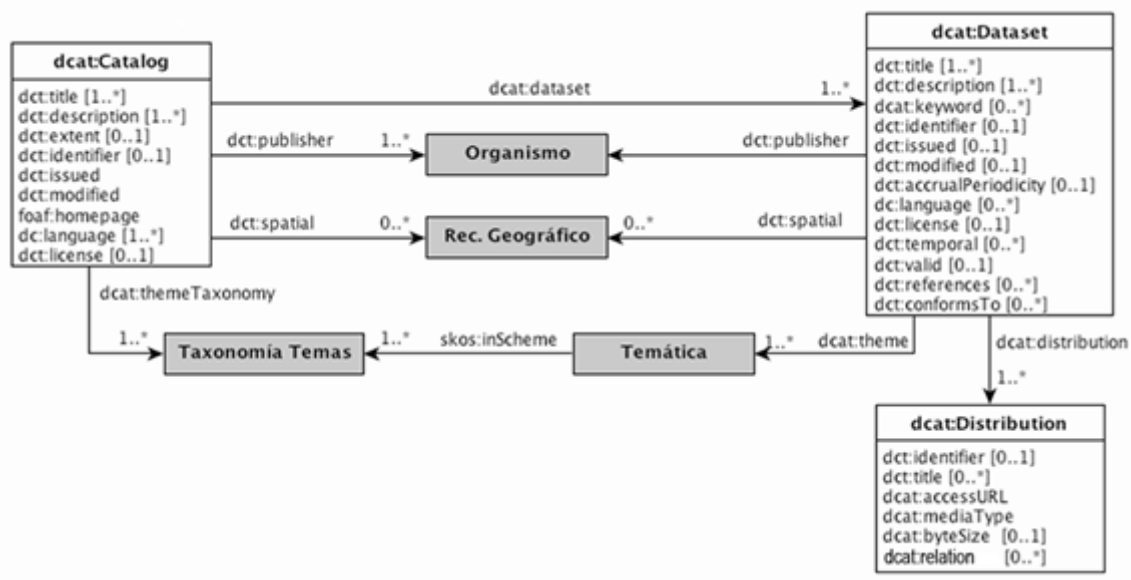
Tabla 1 Taxonomía de sectores primarios

Sector	URI
Ciencia y tecnología	http://datos.gob.es/kos/sector-publico/sector/ciencia-tecnologia
Comercio	http://datos.gob.es/kos/sector-publico/sector/comercio
Cultura y ocio	http://datos.gob.es/kos/sector-publico/sector/cultura-ocio
Demografía	http://datos.gob.es/kos/sector-publico/sector/demografia
Deporte	http://datos.gob.es/kos/sector-publico/sector/deporte
Economía	http://datos.gob.es/kos/sector-publico/sector/economia
Educación	http://datos.gob.es/kos/sector-publico/sector/educacion
Empleo	http://datos.gob.es/kos/sector-publico/sector/empleo
Energía	http://datos.gob.es/kos/sector-publico/sector/energia
Hacienda	http://datos.gob.es/kos/sector-publico/sector/hacienda
Industria	http://datos.gob.es/kos/sector-publico/sector/industria
Legislación y justicia	http://datos.gob.es/kos/sector-publico/sector/legislacion-justicia
Medio ambiente	http://datos.gob.es/kos/sector-publico/sector/medio-ambiente
Medio Rural	http://datos.gob.es/kos/sector-publico/sector/medio-rural-pesca
Salud	http://datos.gob.es/kos/sector-publico/sector/salud
Sector público	http://datos.gob.es/kos/sector-publico/sector/sector-publico
Seguridad	http://datos.gob.es/kos/sector-publico/sector/seguridad
Sociedad y bienestar	http://datos.gob.es/kos/sector-publico/sector/sociedad-bienestar
Transporte	http://datos.gob.es/kos/sector-publico/sector/transporte
Turismo	http://datos.gob.es/kos/sector-publico/sector/turismo
Urbanismo e infraestructuras.	http://datos.gob.es/kos/sector-publico/sector/urbanismo-infraestructuras
Vivienda	http://datos.gob.es/kos/sector-publico/sector/vivienda

Fuente: (Anexo IV-Norma Técnica de Interoperabilidad de Reutilización de recursos de la información. , 2013)

Describir los conjuntos de datos*:* En este punto se establece el uso de metadatos mínimos que describen el conjunto de datos. Un catálogo de documentos y recursos de información reutilizable se estandariza a través del uso del vocabulario DCAT y se representa mediante instancias de la clase *dcat:Catalog* que incluye una colección *dcat:Dataset* representado gráficamente mediante la Figura 5, donde se aprecia que los catálogos contienen conjuntos de datos y estos, a su vez, disponen de distribuciones.

Figura 5 Diagrama de clases y conceptos para la definición de metadatos



Fuente: (Anexo III-Norma Técnica de Interoperabilidad de Reutilización de recursos de la información., 2013)

La Figura 6 muestra como la instancia dcat:*Dataset* describe cada conjunto de información contenido en el catálogo y como la taxonomía del sector primario se transforma en valores de las temáticas a través de las URIs expuestas en la Tabla 1.

Figura 6 Metadatos del Catálogo de datos de datos.gob.es

```
    <dcat:keyword>Natalidad</dcat:keyword>
    <dcat:theme rdf:resource="http://datos.gob.es/kos/sector-publico/sector/demografia"/>
    <dct:references rdf:resource="http://analisis.cis.es/cisdb.jsp?ESTUDIO=2639"/>
    <dcat:theme rdf:resource="http://datos.gob.es/kos/sector-publico/sector/salud"/>
    <dct:title xml:lang="es">2639|FECUNDIDAD Y VALORES EN LA ESPAÑA DEL SIGLO XXI</dct:title>
    <dct:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2016-11-24T22:56:32+01:00</dct:issu
    <dcat:distribution rdf:resource="http://datos.gob.es/catalogo/e00136201-2638barometro-sanitario-2006-pri
    <dcat:keyword>Sexualidad</dcat:keyword>
    <dcat:theme rdf:resource="http://datos.gob.es/kos/sector-publico/sector/sociedad-bienestar"/>
    <dcat:keyword>VALORES Y ACTITUDES</dcat:keyword>
    <dct:publisher rdf:resource="http://datos.gob.es/recurso/sector-publico/org/Organismo/E00136201"/>
    <dcat:keyword>Estudio Cuantitativo</dcat:keyword>
    <dct:temporal rdf:resource="http://datos.gob.es/catalogo/e00136201-2638barometro-sanitario-2006-primera-
  </dcat:Dataset>
</dcat:dataset>
<dcat:dataset>
  <dcat:Dataset rdf:about="http://datos.gob.es/catalogo/a02002834-anejo-fotografico-sierra-de-albarracin-c31
    <dct:accrualPeriodicity rdf:resource="http://datos.gob.es/catalogo/a02002834-anejo-fotografico-sierra-de
    <dcat:theme rdf:resource="http://datos.gob.es/kos/sector-publico/sector/medio-ambiente"/>
    <dct:publisher rdf:resource="http://datos.gob.es/recurso/sector-publico/org/Organismo/A02002834"/>
    <dcat:keyword>Comarca</dcat:keyword>
    <dcat:keyword>Territorio</dcat:keyword>
```

Fuente: Elaborado a partir del link http://ondemand2.redes.ondemand.flumotion.com/redes/ondemand2/Datosabiertos/datosgobes.rdf (ultimo acceso 19 de julio de 2018)

Representar la información en el formato adecuado: En este paso se especifica la necesidad de publicar la información en formato abierto, procurando usar estándares abiertos o de uso común por la ciudadanía. No se establece el uso de un formato específico, sino que da la libertad de usar varios formatos en dependencia de los perfiles de reutilización.

2.1.4 Políticas públicas

Las políticas públicas son el conjunto de acciones y decisiones diseñadas y gestionadas por el sector público con el propósito de dar solución a problemas o necesidades presentes en la sociedad. Las acciones tienen un grado de obligatoriedad variable, generalmente su intencionalidad ha sido definida en un proceso de interlocución entre el gobierno y la ciudadanía, quien debe estar incluida en el diseño e implementación para darle legitimidad al proceso (Delgado Godoy, 2009).

El proceso o ciclo de elaboración de las políticas públicas está conformado por cinco fases principales (Delgado Godoy, 2009), tal y como se muestra en la Figura 7:

Figura 7 Ciclo de las Políticas Públicas

Fuente: Elaboración propia

La identificación y definición de problemas*:* en esta fase el gobierno detecta la existencia de un problema u oportunidad y decide actuar o no. En el proceso de identificación del problema se deben tener en cuenta tres cosas: la naturaleza, la duración y los afectados del problema; esto es importante porque permite entender las causas y la dinámica que tiene el problema, saber si un problema es pasajero o duradero, le permite decidir al Estado si actúa o no. Si un problema es transitorio el Estado probablemente no actúe frente a él, pero si es duradero, debería evaluar las posibles consecuencias que traería una no intervención gubernamental (Delgado Godoy, 2009).

Una vez identificado y seleccionado el problema se procede a establecer la agenda, que puede ser gubernamental, cuando el problema es insertado por los

decisores públicos, o sistémica, cuando los problemas son insertados por la sociedad. Cuando la sociedad identifica un problema y el Estado frente a este problema no actúa de forma coherente en su solución, la ciudadanía puede hacer uso de los llamados, modelos de incorporación en la agenda política, como son: la movilización, la oferta política, el uso de medios de comunicación, la acción anticipada del Estado y la acción corporativista o silenciosa, denominada comúnmente como *lobby*.

La formulación de políticas: los decisores públicos como organización que representan al Estado deben seleccionar la solución más adecuada, teniendo en cuenta las metas, objetivos y prioridades que se han fijado, con ayuda del plan de acción (alternativas, opciones o propuestas) y de técnicas que asisten en la toma de decisión. La última palabra o decisión la tiene el decisor público que previamente ha interactuado con los actores involucrados (públicos y/o privados).

La adopción de la decisión: la adopción de la decisión está exclusivamente en manos de los decisores públicos que son los encargados de seleccionar la solución que consideran óptima y convertirla en una política pública, por medio de un plan donde se establecen cuáles son las instituciones públicas más idóneas para desarrollar la política, así como las metas y los plazos para ejecutarla.

La implementación: en esta fase las AAPP correspondientes, movilizan recursos económicos y humanos para poner en marcha o ejecuta la política adoptada. La puesta en marcha se puede hacer desde diferentes enfoques, por un lado, se tiene, el clásico también llamado, de arriba hacia abajo, que es un enfoque jerárquico donde no hay mucha participación ciudadana y el protagonismo recae sobre lo decisores públicos; por otro lado, se tiene el enfoque por retroceso o de abajo hacia arriba, donde el protagonismo lo tienen los implementadores. En ambos enfoques se desarrollan actividades y procesos hasta que aparecen los efectos asociados con la intervención pública en cuestión.

La evaluación de los resultados: En esta fase se mide el alcance de los objetivos de la política en cuestión. La evaluación se hace en tres fases, cuando se formula el problema, durante la implementación y después de la implementación de la política pública. Con la evaluación se busca entender los resultados y los impactos de la política pública para hacer ajustes o aprender de los desaciertos y hacer mejores políticas.

Dentro del ciclo de las políticas públicas intervienen actores de la sociedad civil, los partidos políticos y sus representantes en la asamblea general y por último la administración pública.

2.1.5 El papel de la ciudadanía en las políticas públicas

Con la apertura de los datos gubernamentales se pretende que la ciudadanía tome una posición activa en los asuntos de carácter público, como lo hace a la hora de ejercer su derecho al voto para elegir la representación política, con el objetivo de influir en las políticas públicas y en las decisiones de las AAPP. Pedro Prieto Martín ve el papel de la ciudadanía como:

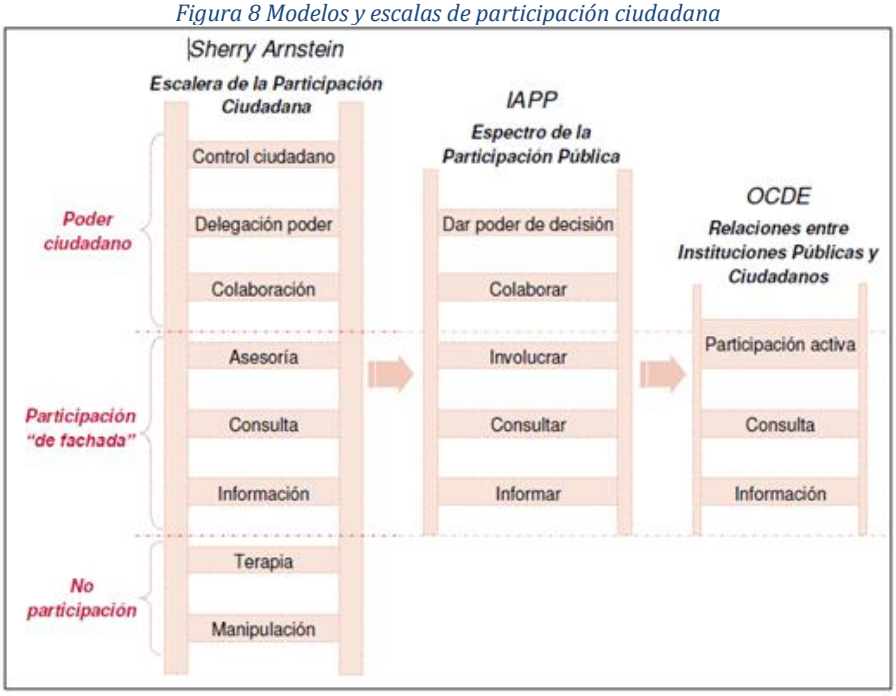
> *"Una nueva forma de institucionalización de las relaciones políticas, que se basan en una mayor implicación de los ciudadanos y sus asociaciones cívicas, tanto en la formulación como en la ejecución y control de las políticas públicas"* (Prieto Martín, 2005).

Sherry Arnstein en sus publicaciones advierte sobre la diferencia entre una participación ficticia y tener el *poder real* necesario para influir en el resultado de un proceso (Arntsein, 1969).

Según Arnstein el poder estaría dado por la distribución de poder tolerada por los organismos públicos, en el diseño de las modalidades de participación (Arntsein, 1969). A lo que Pedro Prieto se refiere como *"...implicación de los ciudadanos en el control de las políticas públicas..."*.

Para ampliar esta idea, Arnstein propuso una tipología de ocho niveles de participación representados en una escalera donde cada escalón simboliza la cantidad de poder ciudadano. A partir de este modelo varias organizaciones han creado sus propias versiones de escalera de participación ciudadana como se muestra en la Figura 8.

Figura 8 Modelos y escalas de participación ciudadana



Fuente: Las alas de Leo. La participación ciudadana del siglo XX. (Prieto Martín, 2005)

En el modelo de Arnstein se establecen ocho escalones agrupados en tres planos de participación. El primer plano denominado *"no participación"*, está formado por los dos escalones inferiores, Manipulación y Terapia. En este plano se pretende sustituir la auténtica participación por medio de la manipulación y/o aleccionamiento a los participantes. El segundo plano es el de la *Participación "de fachada"*, conformado por: Información, Consulta y Asesoría. Este plano tan solo le permite a la ciudadanía escuchar, ser escuchado y aconsejar. Por último, el plano del *Poder ciudadano* conformado por: Colaboración, Delegación de poder y Control ciudadano es el plano ideal, donde la ciudadanía se le otorga voz y voto para poder influir en las decisiones o tener el control mediante la auténtica delegación del poder.

El segundo modelo, propuesto por la Asociación Internacional por la Participación Pública (AIPP) surgió tres décadas después, a través de su "Espectro de la Participación Pública" recortó ambos extremos de la escalera dejando solo cinco niveles como se muestra en la Figura 8. Esta modificación no satisfizo a la Organización para la Cooperación y el Desarrollo Económico (OCDE), conformada por los países considerados más avanzados y desarrollados del planeta, que propuso dejar en tres los niveles de la escalera: *Información, Consulta y Participación Activa.*

Si se observa con detenimiento la Figura 8 la escalera propuesta por la OCDE ha quedado reducida a lo que Arnstein llamaba el plano de la *Participación "de fachada"*. Con este panorama poco alentador para la participación ciudadana, La Alianza para el Gobierno Abierto (OGP por sus siglas en inglés) a través de la Carta Iberoamericana de Gobierno Abierto (CIGA) manifiesta que las instituciones públicas deben procurar para sus ciudadanos una participación colaborativa por encima de las consultivas. Además, propone un modelo donde estén los niveles de Información y consulta, contemplados por el OCDE, acompañados del nivel decisorio y de cogestión o coproducción, dando vida a un nuevo modelo de participación que, es mucho mejor que el propuesto por el OCDE. Con la nueva escalera se pretende otorgar a la ciudadanía la posibilidad de ser parte del proceso de construcción de las políticas públicas a través de la colaboración, empoderamiento, consulta y toma de decisiones.

Oscar Oszlak y Ester Kaufman en el 2014 realizaron un estudio donde comparan las mejores prácticas en gobierno abierto en la vertiente de participación ciudadana. Estudio que es resumido por José Sánchez González dentro del documento *"La participación ciudadana como instrumento del gobierno abierto"* y que se muestra en la Tabla 2 (Sánchez González, 2015).

Tabla 2 Análisis comparativo de la participación ciudadana

Prácticas	Descripción
Dictar normativa sobre participación	Varios países dictan o proyectan normativa sobre participación. La misma puede ponerse en vigencia a través de una cláusula constitucional, de una ley específica o de otro tipo de normativa *(Chile, Colombia, Perú, Estonia y Argentina).*
Realizar actividades permanentes en el territorio para la resolución conjunta de problemas	El objetivo es trabajar con la gente para asegurar que sus preocupaciones y aspiraciones son consideradas y entendidas. La promesa se traduce en: "Nosotros trabajaremos con usted para asegurar que sus preocupaciones y aspiraciones estén directamente reflejadas en la toma de decisiones" *(Argentina, Guatemala y Corea del Sur).*
Invitar a los ciudadanos a expresar sus opiniones y sugerencias al gobierno	En este nivel de participación, el de consulta, con algunos componentes de involucramiento, conforme al compromiso que vayan adquiriendo los ciudadanos *(Chile, Corea del Sur, Sudáfrica, Colombia, Panamá, Tanzania, España, El Salvador, México, Finlandia, y Estonia).*
Facilitar la participación ciudadana sobre proyectos legislativos y habilitar mecanismos para petitorios y demandas	Se trata del nivel de consultas o de involucramiento. Una de las formas que puede asumir son las peticiones (e-petitions). En este último caso, son pedidos firmados online, en general a través de un formulario provisto por un sitio web *(Brasil, República Dominicana, Perú, Estonia, Brasil, Costa Rica, Gran Bretaña, Estados Unidos, Panamá).*
Abrir instancias de coparticipación ciudadana en el proceso decisorio del Estado	Mediante la coparticipación en el proceso decisorio estamos avanzando hacia reales instancias de colaboración, cuyo objetivo es: "Asociar al público en cada uno de los aspectos de la toma de decisiones, con la promesa de Nosotros buscamos su asesoramiento y lo incorporamos a las decisiones en todo lo posible" *(Canadá, Chile, Costa Rica, Honduras, Estonia, Tanzania, Filipinas, Finlandia y Perú).*
Fomentar proyectos e iniciativas de coproducción de bienes y servicios entre el Estado, el mercado y las organizaciones sociales	El gobierno abierto ha traído consigo inspiración para la generación de la coproducción en los cuales se estaría avanzando desde la colaboración hacia el empoderamiento. "Poner la toma de decisión final en manos del público" *(Corea del Sur, Argentina, Costa Rica, Brasil, Estados Unidos, Uruguay, Gran Bretaña, Panamá).*

Fuente: La participación ciudadana como instrumento del gobierno abierto (Sánchez González, 2015).

En el marco de este proyecto consideramos que las políticas públicas deben nacer y formarse dentro de la colaboración entre las AAPP y la ciudadanía, donde juntos deben apuntar a construir una sociedad más justa. Implicar a la ciudadanía en las políticas públicas es una forma de que el gobierno se abra a la toma de decisiones consensuadas.

2.2 Minería de texto en redes sociales

En este punto surgen dos conceptos que son vitales de entender para el desarrollo del proyecto. Por un lado, se encuentran las redes sociales que constituyen el medio a través del cual se desarrolla el proyecto y, por otro lado, se encuentra la minería de texto, que es el cómo se realiza. Por lo tanto, a continuación, se explicarán estos dos conceptos desde la perspectiva de uso que se le dan en el proyecto.

2.2.1 Redes sociales

Existen diferentes corrientes que se atribuyen el origen del significado de *"red social"*. Algunos autores afirman que sus inicios se dieron en la psicología y sociología, otros que en las matemáticas y un grupo restante indican que fue en la antropología.

En el área de la psicología y sociología el concepto de *"red social"* nace en la teoría de campos de Kurt Lewin, quien usando una función matemática explicaba que tanto el comportamiento de los individuos de un grupo, como la estructura misma del grupo se desarrollan en un espacio social formado por el grupo y su entorno, estableciendo así un campo de relaciones (Scott, 1991). En esta misma línea Levi Moreno introdujo el concepto de sociometría para describir la estructura conectiva de complejas redes de interacción social.

La fundamentación teórica de red social establecida por K. Lewin y L.Moreno, entre otros autores del área de la psicología y la sociología, fueron formalizados en la teoría matemática de Grafos donde cada nodo representa una persona y cada línea la relación de amistad entre los nodos.

La primera definición formal que se conoce de "*red social"* proveniente de la antropología data de 1954, cuando John Barnes realizaba un estudio en un pueblo de pescadores en Noruega a lo que dijo:

> *"La imagen que tengo es de un conjunto de puntos algunos de los cuales están unidos por líneas. Los puntos de la imagen son personas o a veces grupos, y las líneas indican qué individuos interactúan mutuamente. Podemos pensar claro está, que el conjunto de la vida social genera una red de este tipo"* (Barnes, 1954).

Por su parte, Ross Speck y Carolyn Attneave (1982) definieron la red social como:

> *"las relaciones humanas que tienen un impacto duradero en la vida de un individuo"* (Speck & Attneave, 1982).

Estas dos definiciones fueron resumidas por Garbarino como:

> *"conjunto de relaciones interconectadas entre un grupo de personas que ofrecen unos patrones y un refuerzo contingente para afrontar las soluciones de la vida cotidiana"* (Garbarino, 1983).

Aunque las definiciones provienen de diferentes áreas del saber humano, todas coinciden en que hay **individuos o grupos** (dependiendo del autor reciben el nombre de puntos o nodos), que se conectan entre sí, estableciendo **relaciones**. Estas definiciones han sentado las bases para crear los principios teóricos de lo que hoy conocemos como red social digital, también llamada red social online.

La red social online parte de la teoría de los seis grados de separación, la cual indica que una persona en la tierra está conectada a otra, por una cadena de personas que no tiene más de cinco personas intermediarias, por lo que la relación entre las dos personas no tiene más de seis enlaces (Noguera Vivo , et al., 2011).

Con la llegada de internet, las redes sociales pasaron de un entorno "real" a un entorno virtual, donde un grupo de personas que tienen algún tipo de afinidad se encuentran interconectadas. En ese mundo virtual no existe la barrera de espacio físico por lo que individuos de diferentes partes del mundo se encuentran conectados, aun cuando su idioma natal no sea el mismo. Los usos que se le pueden dar a las redes sociales se pueden resumir en las famosas "3Cs": *Comunicación*, permiten poner en común conocimientos, *Comunidad*, ayudan a encontrar e integrar comunidades, y *Cooperación*, ayudan a hacer cosas juntos.

Las redes sociales son un espacio donde se puede opinar, intercambiar comunicaciones o simplemente expresarse, que dependiendo de la configuración de privacidad que se tenga, lo que se publica puede ser visto por "amigos" (personas que están conectados con nosotros) o el público en general (personas que no pertenezcan al círculo de amistad). Por todo lo anterior, y por el elevado número de usuarios, así como, por el volumen de información que se maneja, estos medios son propicios para hacer análisis, tales como: conexiones entre personas, posturas políticas, sentimientos y opiniones sobre un tema en concreto.

El desarrollo tecnológico por su parte posibilita la extracción de la información de estos espacios, ya sea con el uso de APIs Rest proporcionadas por las mismas redes sociales, como es el caso de Facebook, Twitter o Linkedin; o con el uso de data Scraping, sea como fuese, en la actualidad resulta relativamente fácil automatizar la forma de recolectar información de las redes sociales para posteriormente analizarla.

Algo muy importante que se debe tener en cuenta a la hora de utilizar la información extraída de estos medios, es que su escritura no es formal (uso de

emoticonos, acrónimos, abreviaciones o “nuevos términos”) lo que conlleva en muchos casos a no tener la calidad deseada. También se debe tener especial cuidado en la interpretación que se hace de los textos, puesto que el contexto o el momento en el que se desarrolla una opinión puede influir en el significado que esta pueda llegar a tener. Si tomamos como ejemplo, la red social de microblogging Twitter, donde los usuarios se comunican mediante mensajes cortos, con un máximo de 140 caracteres, también entran elementos propios de esta red social, como son la posibilidad de crear un identificador para un tema de conversación, vínculos entre publicaciones o usuarios. Para ampliar la idea a continuación se definen algunos de estos elementos:

Cita (RT)*:* Si un usuario quiere citar una publicación hecha por otro y añadirle algún comentario, basta con copiar el tweet, anteponerle la sigla RT y luego poner el comentario que desea hacer.

Tema (#): Cuando un usuario quiere escribir sobre una temática o categoría global en Twitter, solo es necesario que delante de la palabra ponga la almohadilla #, convirtiendo automáticamente la palabra en un enlace temático que unirá todos los mensajes que utilicen la misma palabra con la almohadilla # delante.

Mención (@): Si un usuario quiere que otro usuario que se encuentra o no conectado con él lea lo que publica, solo necesita poner dentro de la publicación el nombre de usuario con el símbolo @ delante.

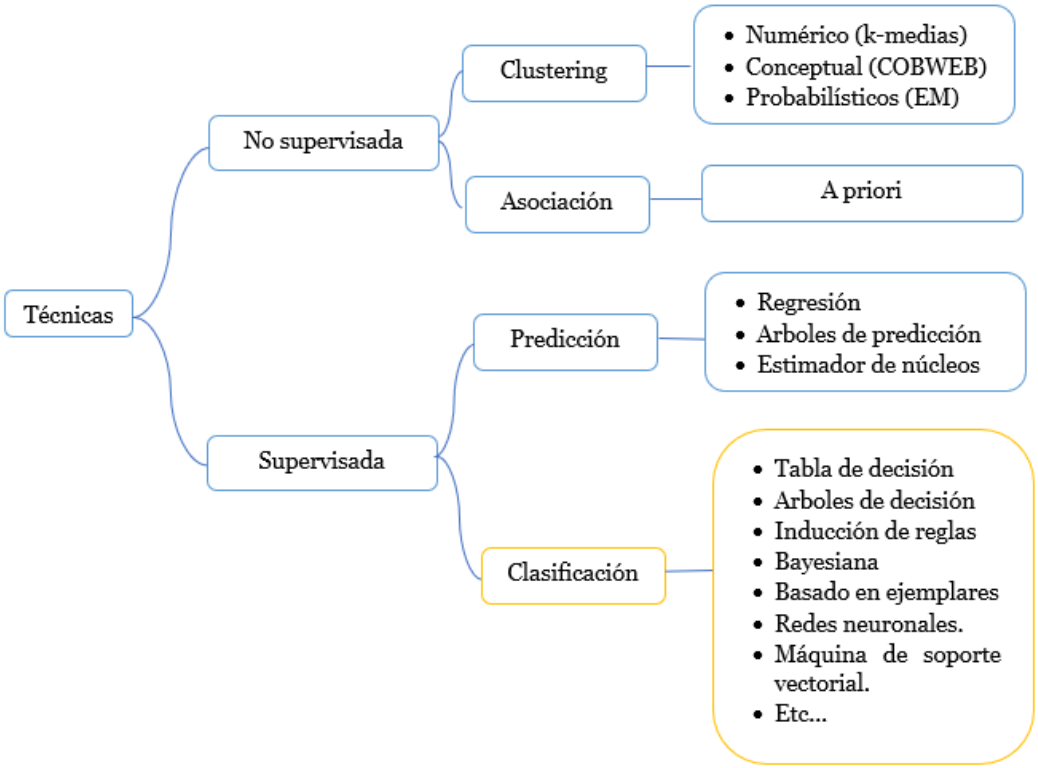
Para analizar los textos generados a través de estos medios, se deben utilizar técnicas de Machine Learning que permitan explotar al máximo la información. Este proceso se explica más al detalle en el siguiente punto.

2.2.2Machine Learning

Es un campo de investigación donde convergen la estadística, la Inteligencia Artificial y la informática para crear sistemas que “aprenden” automáticamente a través de la identificación de patrones complejos de millones de datos. El aprendizaje ocurre por inducción del conocimiento, usando técnicas que intentan obtener patrones o modelos que generalizan comportamientos a partir de información suministrada en forma de ejemplo.

Las técnicas de *Machine Learning* se dividen en dos grandes grupos, tal y como se muestran en la Figura 9.

Figura 9 Técnicas de minería de datos



Fuente: Elaborado a partir de (García Herrero, 2018) pág. 205

Una técnica materializa los conceptos de extracción de información de los datos, y en general, es implementada por varios algoritmos. Los algoritmos ejecutan paso a paso las técnicas, por lo que hay que entender los algoritmos para poder establecer cuál es la mejor técnica para utilizar.

En el **aprendizaje inductivo no supervisado** no se requiere de un corpus de entrenamiento previamente etiquetado para construir un modelo (Dipanjan, 2016). Se utilizan técnicas para agrupar documentos en conjuntos a partir de sus características, similitudes y atributos, sin necesidad de entrenar ningún modelo. Dentro de este tipo de aprendizaje tenemos dos clases de algoritmos (García Herrero, 2018):

- **Algoritmos de Clustering,** consiste en agrupar un conjunto de vectores por su similitud o distancia, ya sea a través de algoritmos matemáticos, redes neuronales o técnicas de reconocimiento de conceptos. Este tipo de algoritmos suelen utilizarse cuando hay grandes volúmenes de datos y se requiere realizar una división preliminar, para después hacer análisis sobre los clústeres.
- **Reglas de Asociación,** se usan para establecer relaciones o correlaciones entre acciones o sucesos aparentemente independientes. Son utilizadas para hacer análisis exploratorios buscando relaciones entre un conjunto de datos.

En el **aprendizaje inductivo supervisado** se requiere de un corpus de entrenamiento previamente etiquetado para construir un modelo (Dipanjan, 2016). El algoritmo aprende a partir de los datos de entrenamiento y se crea un modelo que se puede usar para predecir la clase para futuras muestras de datos, en este proceso es necesario un trabajo manual inicial para etiquetar los documentos que conforman el corpus de entrenamiento y que será la base del conocimiento del algoritmo. Dentro de este tipo de aprendizaje tenemos dos clases de algoritmos (García Herrero, 2018):

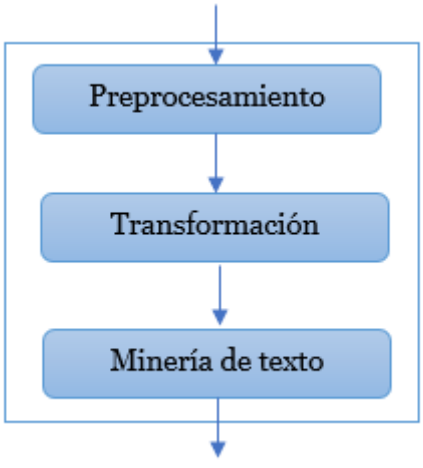
- **Algoritmos predictivos:** se usan para analizar datos, tanto actuales como históricos y predecir acontecimientos futuros. En este tipo de algoritmos se suelen utilizar métodos de regresión lineal y estimación de núcleo.
- **Algoritmos de clasificación:** son un conjunto de técnicas, algoritmos y sistemas a través de las cuales se le asigna a un documento de texto una o más clases o categorías (Dipanjan, 2016).

2.2.3Minería de texto

La minería de texto (MT), también conocido como *Knowledge discovery from text (KDT)*, busca encontrar información relevante en colección de textos donde el conocimiento no se muestra de forma explícita. Debido a que los textos carecen de atributos estructurados o que rara vez siguen un patrón específico, no se pueden implementar con ellos métodos estadísticos estándar lo que implica usar otras funciones analíticas propias del *Machine Learning* (ML) y del procesamiento de lenguaje natural (PLN), junto con técnicas de Recuperación de Información (RI) (Maimon & Rokach, 2010).

La extracción de información de texto debe seguir una serie de fases o etapas que permitan sacar el mayor provecho a la información, tal y como se muestra en la Figura 10.

Figura 10 Tres de las cinco etapas del The Knowledge Discovery of Text



Fuente: Elaboración propia

Etapa de Preprocesamiento: Durante esta fase el texto es limpiado, normalizado y transformado para que pueda ser procesado por algoritmos de Machine Learning (ML). Durante esta fase se utilizan técnicas de preprocesamiento de Lenguaje Natural y de Recuperación de información.

Los pasos principales en el preprocesamiento del texto son la **limpieza** y **normalización** que se hace a través de las siguientes técnicas:

Tokenización: proceso mediante el cual se dividen datos de texto en componentes más pequeños, llamados tokens (Manning, et al., 2009)

Ejemplo: Amigos, romanos, compatriotas, préstame tus oídos;

La práctica de esta técnica en un lenguaje como el español resulta sencillo, puesto que los espacios en blanco y los signos de puntuación marcan la separación de palabras.

***Stop words* o eliminación de palabras vacías:** proceso mediante el cual se eliminan palabras que no aportan ningún significado al texto (artículos, pronombres, preposiciones, etc) y tienen una alta frecuencia de aparición (Manning, et al., 2009).

Ejemplo: a, al, cada, de, en, etc, fui, ni, que, solo

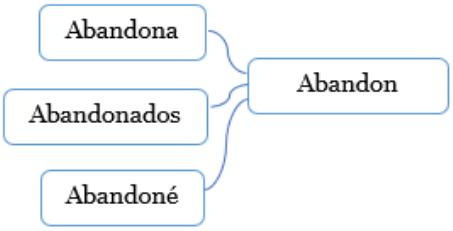
Normalización: proceso mediante el cual se transforman palabras no estándares (números, contracciones, abreviaturas, fechas y acrónimos), en inglés *non-standard "words" (NSWs)*, en palabras estándares, es decir, se escriben las palabras de la forma en la que se diría de manera oral (Sproat, et al., 2001).

Ejemplo: 24= Veinticuatro, atte=atentamente, 2%=dos por ciento patrás=para atrás, AVE= Alta Velocidad Española.

Stemming* y lematización*: estos dos procesos permiten reducir las formas flexivas y, a veces, las formas derivadas de una palabra a una forma básica común (Manning, et al., 2009).

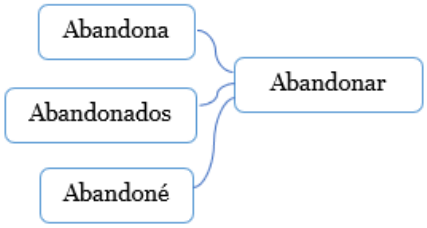
En el proceso de Stemming se reduce la palabra a su raíz a través del uso de métodos heurísticos.

Ejemplo:



La **lematización,** por su parte, utiliza un método más riguroso donde se analiza la morfología de las palabras.

Ejemplo:



Etapa de transformación: consiste en la **extracción de características**, que es el proceso mediante el cual se extraen atributos significativos de datos textuales crudos para alimentarlo en un algoritmo estadístico o de Machine Learning (ML). Este proceso también se conoce como vectorización porque usualmente la transformación final son vectores numéricos de ficheros de texto en bruto. Este proceso es realizado porque los algoritmos convencionales trabajan en vectores numéricos y no pueden trabajar directamente en datos de texto sin formato. Existen varios métodos de extracción de características entre las que se encuentran (Bautista, 2000):

- **Modelo de Bolsa de Palabras Binaria:** es de las técnicas más simples proveniente de Recuperación de Información (RI), que nos dicen si existe o no un término en el documento, si existe se asigna el valor 1 y si no existe se le asigna 0.

$$\text{tf}(t, d) = 1 \text{ si t ocurre en d, y 0 si no}$$

- **Modelo de Bolsa de Palabras basada en la frecuencia relativa:** con esta técnica se calcula la frecuencia de ocurrencia de un término en un documento, sobre el número de términos que hay dentro del documento, matemáticamente queda como:

$$D_i(t_j) = \frac{f_{ij}}{\sum_{i=1}^{m} f_{ij}}$$

  En este modelo se le asignan valores altos a los términos que aparecen con mayor frecuencia

- **Modelo de Bolsa de palabras basada en la frecuencia invertida del Documento (TF-IDF):** el valor de la frecuencia es el producto de dos métricas, la frecuencia del término *(TF)* y la frecuencia del documento *(IDF)*

  La frecuencia del término *(tf)* se puede calcular de diferentes maneras, la más sencilla es la *frecuencia bruta* que es el número de veces que aparece el término en el documento *tf(t,d) = f(t,d)*; la *binaria*, que

vimos al principio; la *frecuencia logarítmica,* dada por *tf(t,d)=1+log f(t,d) (y 0 si f(t,d)=0)*; por último tenemos la *frecuencia normalizada,* que se usa cuando los documentos son largos, la formula queda:

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(t,d) : t \in d\}}$$

La frecuencia inversa del término *(idf)* es la métrica que se usa para determinar si el término es común o no, está dada por la fórmula *idf(t)=log(N/df(t))* donde *N* es el número total de documentos de la colección y *df(t)* es el número de documentos donde aparece el término.

Hechas las consideraciones anteriores, la fórmula para calcular el tf-idf queda tfidf=*tf x idf, la* principal ventaja de este modelo frente *al Modelo de Bolsa de palabras basada en la frecuencia relativa* es que en este modelo se penaliza la frecuencia de aparición del término en el documento a partir de la frecuencia de aparición de estos términos en la colección, mientras que en el otro no.

Etapa de Minería de Texto: existe una gran cantidad de técnicas, operaciones y algoritmos provenientes del *Machine Learning* (Figura 9) usados para realizar análisis sobre texto. Los algoritmos de clasificación son algunos de ellos, por lo que en esta etapa hablaremos de dos de ellos por ser bastante efectivos para clasificar textos:

Clasificación Bayesiana, Naive Bayes (NB) y Multinomial Naive Bayes (MNB): los métodos de Naive Bayes son un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes que se utiliza específicamente para tareas de predicción y clasificación donde se tienen más de dos clases (Dipanjan, 2016).

El algoritmo *Naive Bayes*[12] se basa en el teorema de Bayes, pero bajo una suposición "ingenua" de que cada característica es independiente de las demás, es decir, asume que la probabilidad de encontrar un término en un documento es independiente de la existencia de otros términos en el documento. El *Multinomial Naive Bayes (MNB)* por su parte considera un documento como una bolsa de palabras, donde la probabilidad de observar un término dentro de una clase se estima a partir de los datos de entrenamiento, donde se calcula la frecuencia relativa de cada término en la colección de documentos de entrenamiento de esa clase (Bifet & Frank, 2010).

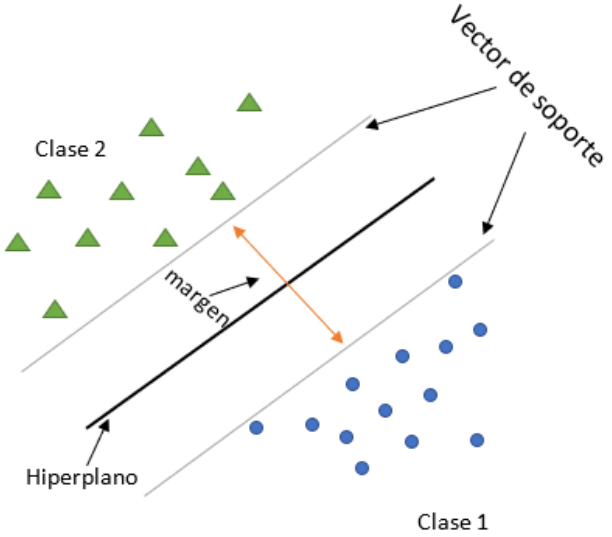
Los algoritmos de Naive Bayes son muy utilizados en la clasificación de texto, debido a su simplicidad, velocidad de aprendizaje y el no necesitar muchos datos para tener un buen aprendizaje, además de dar buenos resultados en la

[12] 1.9. Naive Bayes — scikit-learn 0.19.2 documentation. Se recuperó el 25 de julio del 2018 de http://scikit-learn.org/stable/modules/naive_bayes.html.

clasificación de documentos con clase múltiple. Tiene como desventaja la falta precisión a la hora de calcular la probabilidad de que una clase sea correcta.

La Máquina de Soporte Vectorial o *Support Vector Machine (SVM):* es un proceso típico de clasificación lineal. Este algoritmo representa los datos de entrenamiento como puntos en el espacio, de modo que, los puntos pertenecientes a cualquiera de las clases pueden separarse por un espacio entre ellos, llamado hiperplano (ver Figura 11). La asignación de la clase de los nuevos puntos de datos se hace de acuerdo con qué lado del hiperplano se ubiquen (Dipanjan, 2016). El modelo de optimización busca establecer los vectores de soporte mediante los patrones que más resaltan las distribuciones de clases.

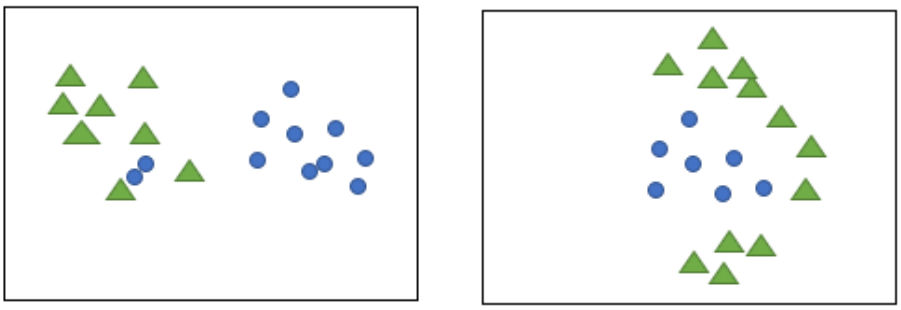
Figura 11 Descripción de los elementos de un SVM



Fuente: Elaboración propia

Cuando dos clases no son linealmente separables como se observa en la Figura 12, el método de SVM utiliza las funciones de kernel, es decir, funciones matemáticas que trasladan los datos a un espacio donde el hiperplano solución es lineal y de esta manera puede resolver el problema de clasificación. Entre las funciones de kernel más habituales se tienen, kernel con base radial (RBF), kernel polinomial y kernel sigmoide.

Figura 12 Clases no linealmente separable



Fuente: Elaboración propia

El desempeño de este algoritmo es bueno cuando se tiene una gran cantidad de características, es muy útil cuando no se cuenta con un corpus de entrenamiento muy grande y al igual que el algoritmo Bayesiano, este tipo de algoritmo da buenos resultados en la clasificación de documentos con clase múltiple. La desventaja que tiene es que puede resultar complejo hacer ajustes y optimizar el algoritmo, sobre todo a la hora de seleccionar la función de kernel más adecuada.

Los algoritmos expuestos anteriormente son utilizados para hacer clasificaciones de diferentes tipos y siguiendo los pasos básicos para aplicar algoritmos de clasificación.

Dependiendo de la cantidad de clases para predecir y de la naturaleza de la predicción el algoritmo de *NB, MNB* y *SVM* pueden ser binarios, de clase múltiple o de Multi-etiqueta:

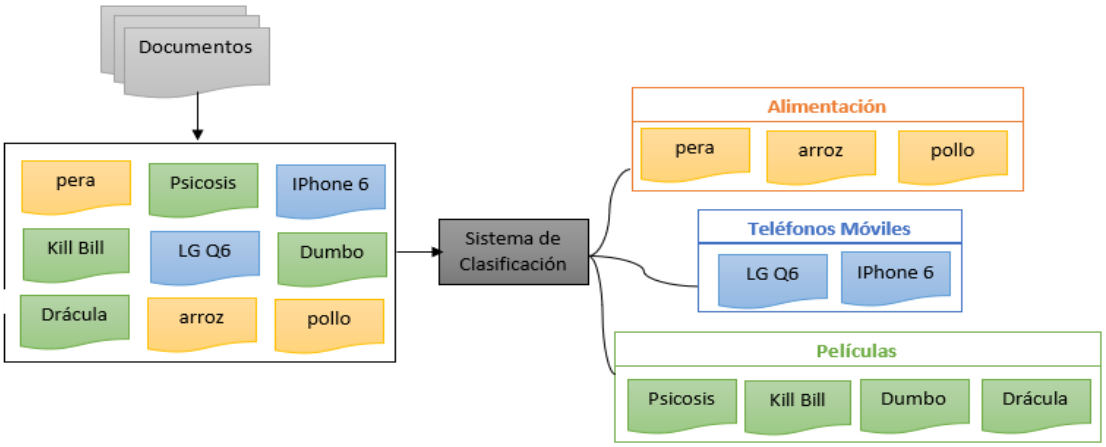
Binarios: ocurre cuando solo existen dos clases distintas entre las cuales se debe realizar la predicción.

De clase múltiple, se presenta cuando el número total de clases es más de dos, y cada predicción da una clase o categoría.

De múlti-etiqueta, se presenta cuando el número total de clases es más de dos, y cada predicción da más de una clase o categoría.

A modo de ejemplo, en la Figura 13 se muestra un sistema de clasificación de clase múltiple.

Figura 13 Descripción conceptual de la clasificación de texto



Fuente: Elaborado a partir de (Dipanjan, 2016) pág. 169

En la figura se puede ver que hay tres categorías y que existen varios documentos que representas productos, que una vez pasan por el sistema de clasificación de texto, cada elemento es asignado a una clase o categoría. En el ejemplo los documentos sólo están representados por sus nombres, pero lo normal es que haya una descripción en la que se incluyan atributos,

componentes y otras propiedades que puedan ser utilizadas como características, que permitan identificar el texto y hacer una clasificación más fácil.

Ahora bien, en el proceso de clasificación se deben realizar tres tareas fundamentales (Dipanjan, 2016):

Entrenamiento, en este punto se crea una colección de documentos etiquetados de forma manual que le permitan al algoritmo seleccionado aprender.

Evaluación, en este punto se prueba la efectividad y rendimiento del modelo, para realizar las pruebas se toma una porción del corpus de entrenamiento para entrenar el modelo y la porción restante para probar, y luego comparamos nuestra predicción con las etiquetas. Esta práctica nos permite ajustar el modelo, ya sea para incrementar el corpus de entrenamiento y darle más datos de donde aprender o simplemente cambiar de algoritmo de clasificación, que se ajuste a nuestras necesidades.

Optimización, esta tarea se realiza cuando los algoritmos no dan los resultados esperados y es necesario ajustar sus parámetros.

Con el objetivo de observar cómo se emplea la minería de texto en el mundo real, nos dimos a la tarea de buscar investigaciones que se hayan realizado en los últimos cinco años utilizando redes sociales como fuente de información y algoritmos de Machine Learning como herramienta de análisis.

De estas investigaciones nos interesaba saber las técnicas de preprocesamiento y de clasificación más utilizadas por los investigadores para trabajar con los textos de Twitter.

Un dato que nos pareció curioso fue el hecho de encontrar un gran porcentaje de investigaciones sobre análisis de sentimiento en Twitter y pocos sobre clasificaciones del tipo que hablamos en este trabajo fin de máster[13].

[13] Para ver los detalles de las investigaciones encontradas, remitirse al Apéndice G de este documento.

Capítulo 3
Desarrollo del proyecto

En este apartado se hace un análisis detallado al problema, se plantea la solución y se describe la metodología de desarrollo que servirá como hilo conductor en la creación de un producto final acorde a la solución planteada.

3.1 Análisis del problema

El proceso de requerir información a las AAPP españolas por alguno de los canales dispuestos para ello, maneja plazos que no están acorde a los tiempos en que la ciudadanía necesita la información. Según la Ley 19/2013 son 10 días para concreción de la solicitud, 15 días para consulta a terceros afectados y un mes por ampliación de plazo por volumen o complejidad de la información solicitada, sumando un total de 55 días. Estos plazos pueden crear insatisfacción en la ciudadanía frente a la forma en que las AAPP atienden y gestionan las solicitudes. También puede llevar a la pérdida del interés en requerir información e implicarse activamente en las políticas públicas de apertura de datos.

Ahora bien, si miramos el año en que más solicitudes se realizaron al Portal de la Transparencia de la Administración del Estado de España, solo se alcanza a llegar a 4025 solicitudes, que es una cifra minúscula frente a los más de 37 millones de habitantes mayores de 18 años que tiene el país y claro ejemplo de que la ciudadanía no está usando estos canales como se esperaría que lo hicieran. Ante esta situación las AAPP deben idear estrategias que impliquen a la ciudadanía una participación activa, donde no haya un requerimiento directo de información, sino que la información llegue de forma transversal.

La clave puede estar en estudiar los intereses de la ciudadanía, para adelantarse a sus requerimientos y lograr convertir la información pública en un recurso cívico a disposición de la ciudadanía. Existen muchas formas de conocer los intereses de la ciudadanía, pero todas ellas tienen algo en común, saber escuchar. Entonces por qué no ir a los lugares donde se reúnen y analizar sus conversaciones e impresiones, dándole la posibilidad de participar en la política de apertura de datos de forma indirecta y novedosa.

La información recolectada podría ser usada por las AAPP encargadas de abrir datos, para publicar en mayor o menor proporción información dependiendo de las temáticas de conversación, incrementando la probabilidad de que la información que se publica en los portales de datos abiertos sea consultada y reutilizada por muchas más personas.

3.2 Planteamiento de la solución

Para dar solución al problema planteado y después de haber realizado las consultas bibliográficas pertinentes, se ha decidido desarrollar un sistema automático de clasificación de textos proveniente de la red social de *microblogging* Twitter, usando como herramienta de análisis, algoritmos de *Machine Learning*.

La propuesta consiste en monitorizar la comunicación que la ciudadanía mantiene con los ayuntamientos a través de Twitter, usando como parámetro de selección aquellos *tweets* donde se menciona la cuenta oficial de Twitter del ayuntamiento, en este caso y para reducir el marco de estudio se usará al Ajuntament de València (@AjuntamentVLC). Con esta decisión no se pretende desarrollar una aplicación a medida para ese ayuntamiento, sino, tomarlo como piloto para crear una aplicación que permita tanto a la ciudadanía como a los ayuntamientos de cualquier comunidad autónoma sacar el mayor provecho posible a los datos e información existente.

Lo siguiente que se debe hacer, es limpiar y transformar la información recolectada para que pueda ser procesada por algoritmo de *Machine Learning*. Algoritmos que llevarán a cabo el proceso de clasificación usando como clase, aquellas categorías expuestas por la Norma técnica de Interoperabilidad de las que se hace referencia en el capítulo 2 de este documento. Estas categorías también son usadas en la catalogación de los datos abiertos publicados en los portales de datos abiertos, por lo que servirán para hacer comparaciones.

Finalizado el proceso de clasificación, los *tweets* analizados se distribuyen en 22 categorías, las mismas 22 categorías que clasifican los *datasets* publicados por el Ajuntament de València. Con estos dos conjuntos de datos, cuyo punto de unión son las categorías se pueden hacer comparaciones a partir del cálculo del número de *datasets* publicados por categoría y el número de *tweets* que pertenece a la misma categoría. De esta forma es posible evidenciar las diferencias y similitudes entre los temas manifestados por la ciudadanía y los *datasets* publicados.

El acceso a los resultados del análisis será a través de un sitio web en el que la información se representa a través de gráficas. La base de datos donde se tomará la información para generar las gráficas se actualizará automáticamente todos los días, lo que quiere decir que el proceso de clasificación se debe realizar a diario. Por tal motivo, se fijará una hora del día para descargar los *tweets* y otra para ejecutar el proceso de clasificación, teniendo en cuenta en dejar entre los dos procesos un tiempo prudencial para que no se solapen.

Finalmente se podrá conocer si los temas que ocupan a la ciudadanía corresponden con los datos publicados a través del portal de transparencia y datos abiertos de su Comunidad Autónoma.

3.3 Diseño de la solución

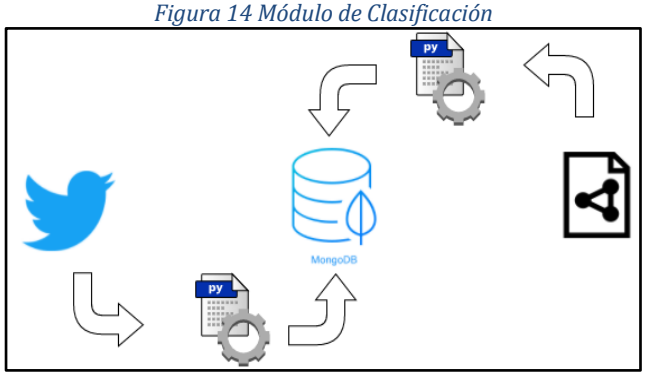
Llegado a este punto se deben seleccionar las tecnologías que van a permitir el almacenamiento y procesamiento de los datos, teniendo en cuenta tanto la naturaleza de la información a recolectar como el tipo de tratamiento que debe recibir.

Los datos que se extraigan deben ser almacenados en una base de datos teniendo en cuenta el volumen, la variedad y velocidad de estos, por lo que, la mejor opción sería utilizar una base de datos Nosql, más exactamente MongoDB[14], la cual sobresale entre las opciones existentes.

En cuanto al lenguaje de programación para procesar los datos, se selecciona Python[15] por ser un lenguaje de programación ligado al análisis de los datos, de fácil aprendizaje y con una gran comunidad de trabajo y de librerías.

Debido a que tenemos que usar técnicas de *Machine Learning* para desarrollar el sistema de clasificación y que estos deben seguir unos pasos específicos, tal y como se muestra en el apartado Minería de texto en redes sociales del capítulo 2. Decidimos abordar la solución dividiendo el proyecto/software en dos módulos. Un módulo de clasificación que contenga todos los algoritmos de descarga, tratamiento y análisis de datos (ver Figura 14).

Figura 14 Módulo de Clasificación



Fuente: Elaboración propia

Y otro módulo que permita a los usuarios finales entender los resultados obtenidos en el módulo anterior (ver Figura 15).

Figura 15 Módulo de Visualización

Fuente: Elaboración propia

[14] MongoDB. Se recuperó el 8 de agosto del 2018 de https://www.mongodb.com/.

[15] Python.org. Se recuperó el 8 de agosto del 2018 de https://www.python.org/.

El código fuente producto de los dos módulos, estará publicado en un repositorio en GitHub[16] creado para gestionar el código fuente y sujeto a la licencia *GNU General Public Licence v3.0*[17].

Una vez definido el uso de una estructura de desarrollo por módulos y sabiendo que usaremos GitHub para controlar las versiones del código fuente, es hora de establecer qué metodología de desarrollo se usarán en cada uno de los módulos.

3.4 Metodologías

Para desarrollar la solución propuesta se utilizan dos metodologías, una para el módulo de clasificación, la cual se basa en una metodología especializada en el desarrollo de sistemas de *Machine Learning* extrapolable a la minería de texto. Y otra para el módulo de visualización, donde se utiliza una metodología especializada en el desarrollo de aplicaciones web.

A partir de lo establecido en el apartado Minería de texto en redes sociales del capítulo 2, hemos seleccionado la metodología ***The Knowledge Discovery Databases (KDD)*** para desarrollar el **módulo de clasificación** en cinco etapas (Fayyad, 1996), tal y como se muestra en la Figura 16. En esta metodología el objetivo principal del proceso es extraer el conocimiento oculto de un conjunto de datos para sacar información que no está a simple vista. Su autor la define como:

> *"...el proceso no trivial de identificar patrones de datos válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles"* (Fayyad, 1996).

En esta definición hay tres conceptos claves para el desarrollo de KDD, el "proceso" que hace referencia al conjunto de tareas que se deben realizar para alcanzar las metas trazadas. Este proceso es repetible y generalizable a diferentes fuentes de datos. El "patrón" que no es más que el conocimiento que queremos extraer y los "datos" que se refieren a la fuente de datos de la que queremos extraer conocimiento.

Figura 16 Etapas que compone el proceso The Knowledge Discovery Databases (KDD)

Fuente: Elaborado a partir de (Fayyad, 1996)

[16] Repositorio del código fuente en GitHub.
https://github.com/areahackerscivics/Analisis_de_contenido_de_comunicacion_ciudadana_Tweets

[17] Para más información sobre la licencia, remitirse a https://www.gnu.org/licenses/gpl-3.0.en.html

El proceso de KDD es interactivo e iterativo y, en él se ven involucrados numerosos pasos vitales para el desarrollo de aplicaciones KDD (Fayyad, et al., 1996). En gran parte de la literatura que se refiere a realizar procesos de *Machine Learning*, se centran en los pasos que corresponden a la etapa de minería de texto, sin embargo, el éxito de dicha etapa radica en ejecutar correctamente los pasos anteriores a ella.

Etapa 1: Selección, consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables, en el que se realizará el descubrimiento.

Etapa 2: Preprocesamiento, consiste en limpiar y preprocesar los datos, usando operaciones básicas para eliminar el ruido, campos vacíos, expansión de contracciones, etc.

Etapa 3: Transformación, como su nombre lo indica consiste en transformar los datos, utilizando para ello métodos de reducción o transformación de dimensionalidad a través de la búsqueda de características útiles para representar los datos.

Etapa 4: Minería de texto, consiste en seleccionar y aplicar algoritmos de *Machine Learning*, que en nuestro caso será el algoritmo de clasificación que brinde los mejores resultados durante las pruebas de selección de algoritmo.

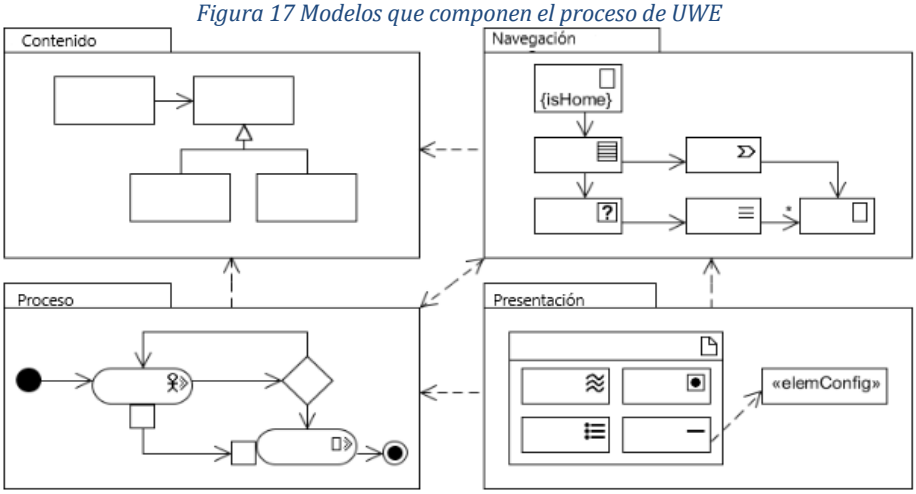
Etapa 5: Interpretación y evaluación, consiste en la interpretación y evaluación de los patrones minados.

Para terminar, el proceso KDD ha estado en continua evolución desde su aparición, incluyendo nuevos algoritmos como SVM (*Support Vector Machine*) y software de código abierto como Weka y R.

En lo que se refiere a la **metodología** utilizada para **el módulo de visualización,** se ha seleccionado la metodología de Ingeniería web basada en UML, en inglés, ***UML-based Web Engineering (UWE).*** Metodología que surge a finales de los años noventa con el objetivo de crear una forma estándar para construir análisis y diseño de modelos de sistemas web. UWE utiliza una notación de la versión "liviana" de UML, donde se incluyen *stereotypes* para modelar los diferentes aspectos de las aplicaciones web, como navegación, presentación, usuario, tarea, etc (Kraus & Koch, 2003). En la Figura 17 se puede ver algunos de estos modelos y algunos *stereotypes* que se usan con más frecuencia en esta metodología, la lista con todos los *stereotypes* que se pueden utilizar en cada modelo se encuentra en el Perfil UWE[18].

[18] Profile Overview UWE Examples. . Se recuperó el 12 de agosto del 2018 de http://uwe.pst.ifi.lmu.de/profileOverview.html

Figura 17 Modelos que componen el proceso de UWE

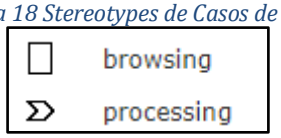


*Fuente: Elaborado a partir de UWE About UWE.*http://uwe.pst.ifi.lmu.de/aboutUwe.html

Esta metodología abarca todo el ciclo de vida de desarrollo de las aplicaciones web, proponiendo un enfoque iterativo y orientado a objetos basado en el Proceso de Desarrollo de Software Unificado y se proponen 4 etapas:

Etapa de especificación de requisitos, es el primer paso para desarrollar un sistema web, es donde se identifican y especifican los requisitos funcionales y no funcionales. Para esta etapa la metodología propone principalmente el uso de casos de uso, donde se modela una aproximación de las funcionalidades. En la Figura 18 se muestran algunos de los *stereotypes* más populares de este modelo.

Figura 18 Stereotypes de Casos de Usos

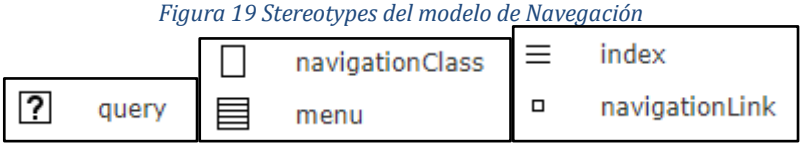


*Fuente: Profile Overview UWE Examples.*http://uwe.pst.ifi.lmu.de/profileOverview.html

Etapa de diseño, en esta etapa se realizan todos los diseños necesarios para entender el sistema web que se pretende desarrollar, se divide en tres tipos de diseños:

- **Conceptual**, lo conforman el modelo de componentes y el de objetos.
- **Navegación,** muestra cómo será la navegación en la página web, se identifican nodos y enlaces. En la Figura 19 se muestran algunos de los *stereotypes* más populares de este modelo.

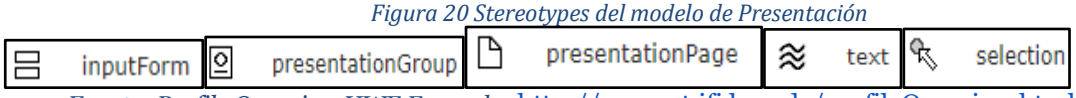
Figura 19 Stereotypes del modelo de Navegación



*Fuente: Profile Overview UWE Examples.*http://uwe.pst.ifi.lmu.de/profileOverview.html

- **Presentación,** representa de forma abstracta la interfaz de usuario. En la Figura 20 se muestran algunos de los *stereotypes* más populares de este modelo.

Figura 20 Stereotypes del modelo de Presentación



*Fuente: Profile Overview UWE Examples.*http://uwe.pst.ifi.lmu.de/profileOverview.html

Etapa de codificación del Software, en esta etapa se traduce en código fuente todo lo diseñado en la etapa anterior. No se exige el uso de un lenguaje de programación determinado ni una arquitectura específica.

Etapa de pruebas, como en cualquier otra metodología esta etapa se utiliza para verificar el correcto funcionamiento del código, en nuestro caso ejecutaremos pruebas unitarias.

Capítulo 4
Construcción de la solución

En este capítulo se materializan todas las decisiones tomadas hasta el momento, se siguen las metodologías establecidas en el capítulo anterior y se utilizan las tecnologías seleccionadas. Una vez desarrollado cada uno de los módulos, se integran para posteriormente realizar las pruebas al sistema en conjunto.

4.1 Módulo 1: Clasificación

Este módulo se encarga de realizar todo el proceso de clasificación de textos, que comienza con selección de la información y finaliza con la predicción de las clases. De acuerdo con lo establecido en la metodología seleccionada, este módulo se desarrolla en cinco etapas:

4.1.1 Etapa 1: Selección

Como en cualquier proceso de selección de fuente de información, el primer paso consistió en buscar los espacios donde la ciudadanía interactúa con el ayuntamiento y seleccionar aquellos donde la comunicación fuese de forma inmediata, donde la participación fuese masiva, generando volúmenes de datos considerable; datos preferiblemente públicos que permitieran una extracción sencilla y precisa.

Con todos estos requisitos planteados se barajaron tres posibles fuentes de información:

1. El canal de comunicación directa que tiene el Ajuntament de València con sus ciudadanos a través de la solicitud de acceso a la información.
2. La red Social Facebook.
3. La red Social Twitter.

El primero fue descartado luego que el Ajuntament de València inadmitiese nuestra solicitud, al tratarse de una petición que no se ajusta a la finalidad de la ley de transparencia[19].

La red Social Facebook fue descartada debido a sus limitaciones de acceso a la información y de no contar con una API libre y lo suficientemente documentada que permitiera extraer la información necesaria. Caso contrario ocurrió con la red Social Twitter que genera mensajes públicos en su gran mayoría, lo que la hace transparente en término de datos, así mismo cuenta con una potente API que permite a programas informáticos obtener información. Por todas estas razones se decidió seleccionar Twitter como fuente de información.

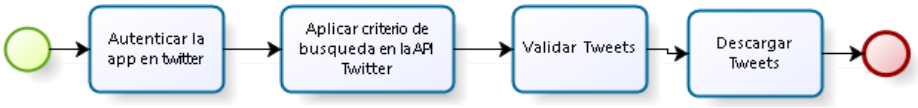
[19] Para ver la respuesta a la solicitud de información, remitirse al Apéndice A de este documento.

Para finalizar con la selección de las fuentes, sólo quedaba buscar los *datasets* publicados por el Ajuntament de València, pero esto ya es más fácil de conseguir, puesto que en su portal de datos abiertos[20] los publican. Como en cualquier sistema de clasificación se debe contar con las clases, preferiblemente estandarizadas para que puedan realizarse comparaciones, por tal motivo acudimos a la Norma Técnica de Interoperabilidad para obtenerlas.

Para extraer la información de las fuentes de información identificadas, se desarrollan tres procesos de extracción: descarga de *tweets*, lectura de catálogo y lista de categorías, este último se llevó a cabo de forma manual por la poca información a extraer. A continuación, se procede a explicar el proceso de desarrollo de los dos procesos restantes:

Proceso descarga de *tweets* (ver Figura 21), en este proceso se obtiene una muestra aleatoria de todos los tweets mediante la versión gratuita del API de Twitter.

Figura 21 Diagrama de proceso de descarga de tweet



Fuente: Elaboración propia

Para usar el API es necesario tener credenciales de acceso en esta red social (*Consumer Key "API Key", Consumer Secret "API Secret", Access Token y Access Token Secret*) y se obtienen a través del registro de la aplicación en el portal de desarrollo de Twitter. Una de las ventajas que ofrece esta API, es la de usar criterios de búsqueda, permitiendo así descargar solo los tweets que son relevantes para el proyecto, es decir, aquellos *tweets* donde se menciona la cuenta oficial de Twitter del Ajuntament de València (@AjuntamentVLC), que sería la forma en la que podemos identificar los *tweets* que son enviados por la ciudadanía al ayuntamiento o simplemente son aquellos que la ciudadanía quiere que el ayuntamiento "escuche" o preste atención. Durante este proceso se filtran los *Retweets* con el fin de reducir la mayor cantidad de ruido a los datos, al finalizar los tweets son almacenados en una base de datos (ver Figura 22) de forma automática con una periodicidad diaria.

Figura 22 Modelo de datos de la colección tweet

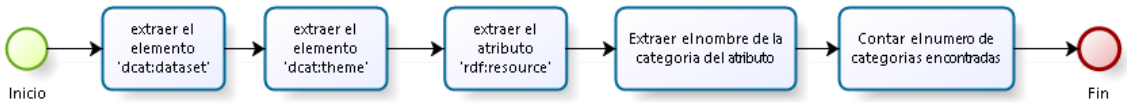
```
{
    "_id" : ObjectId("5a535455bc54f437d024d9e2"),
    "username" : "LasNaves",
    "userSeguidores" : 8767,
    "consulta" : "@AjuntamentVLC",
    "idioma" : "es",
    "dispositivo" : "Hootsuite",
    "idt" : "950323745469673473",
    "userid" : 262129793,
    "fecha_creacion" : ISODate("2018-01-08T11:10:11.000Z"),
    "fechaTweet" : ISODate("2018-01-08T12:21:51.520Z"),
    "tweet" : "Profesionales de la enseñanza se reúnen viernes y sábado en....
}
```

Fuente: Elaboración propia

[20] Portal Transparència i Dades Obertes. Se recuperó el 12 de agosto del 2018 de http://gobiernoabierto.valencia.es/va/

Proceso de lectura del catálogo, en este proceso se realiza la descarga del catálogo de datos del portal de datos abiertos del Ajuntament de València, para posteriormente realizar la lectura del archivo RDF[21]. El RDF contiene un conjunto de "etiquetas" que nos permitirán saber el número de datasets por categorías publicados en el portal. Para contar el número de dataset publicados por categorías basta con navegar a través de las etiquetas *dcat:dataset>dcat:theme* del archivo RDF del catálogo y tomar de las URI el nombre del sector y contar el número de veces que aparece cada URI dentro del archivo. El resumen del proceso se puede ver en la siguiente Figura 23.

Figura 23 Diagrama de proceso de lectura del catálogo



Fuente: Elaboración propia

Para finalizar el proceso, se almacena la información en una base de datos (ver Figura 24) de forma automática con una periodicidad mensual.

Figura 24 Modelo de datos de la colección CATÁLOGO

```
{
    "_id" : ObjectId("5a5d15265cf6265622bec13b"),
    "anyo" : "2018",
    "datos" : [
        {
            "categoria" : "ciencia tecnologia",
            "numDatasets" : 1
        },.....
        {
            "categoria" : "vivienda",
            "numDatasets" : 1
        }
    ],
    "mes" : "01",
    "FechaUpdateCatalogo" : "2018-01-05T09:06:15"
}
```

Fuente: Elaboración propia

El último proceso que se desarrolla en esta etapa de selección consiste en copiar en un archivo los nombres de cada uno de los sectores que aparecen en la Tabla 1 Taxonomía de sectores primarios, recogida en el capítulo 2, y formar una **lista de categorías**.

4.1.2 Etapa 2: Preprocesamiento

Para poder aplicar algoritmos de clasificación, primero se debe realizar un proceso de limpieza y normalización del texto con la ayuda de técnicas desarrolladas especialmente para esa tarea, como se ha mostrado en el capítulo 2. Para el desarrollo de la aplicación no utilizamos todas las técnicas explicadas, sino aquellas que hemos considerado indispensables. A continuación, mostraremos parte del código de cada una de las técnicas aplicadas:

[21] Si desea obtener el archivo RDF completo se puede descargar de http://gobiernoabierto.valencia.es/wp-content/themes/viavansi-ogov/proxyFile.php?url=http://gobiernoabierto.valencia.es/va/catalogo.rdf

Tokenización[22]**:** Para este proceso se utiliza la librería de **nltk** de Python que sirve para el procesamiento del lenguaje natural. El proceso consiste en convertir todas las letras en minúscula y dividir el texto en elementos denominados tokens usando como elemento separador los signos de puntuación y los espacios en blanco, el fragmento de código se recoge en la Figura 25.

Figura 25 Fragmento de código del proceso Tokenización

```
def tokenizar_texto(texto):
    """
    texto=texto.lower()
    palabras = nltk.word_tokenize(texto)
    return palabras
```

Fuente: Elaboración propia

Stop Words[23]**:** En esta parte se eliminan las palabras vacías, que no le aportan ningún valor al texto. Estas palabras se obtienen de una lista de palabras en español que están dentro de la librería **nltk,** el proceso consiste en tomar cada uno de los tokens y compararlos con la lista de palabras, aquellas palabras que coinciden son eliminadas, el fragmento de código se recoge en la Figura 26.

Figura 26 Fragmento de código del proceso Stop Words

```
def eliminar_stopwords(texto):
    """
    palabras=tokenizar_texto(texto)
    palabras_filtradas = [palabra for palabra in palabras if palabra not in stop_words]
    texto_filtrado = ' '.join(palabras_filtradas)
    return texto_filtrado
```

Fuente: Elaboración propia

Stemming[24]**:** En esta parte se eliminan las variantes/inflexiones de las palabras, almacenando sólo la raíz, para este proceso se usa la librería **nltk** para cargar el algoritmo de *stemming*, que toma cada uno de los tokens y lo lleva a su *stem* o raíz, el fragmento de código se recoge en la Figura 27.

Figura 27 Fragmento de código del proceso Stemming

```
def Stemmer(texto):
    """
    palabras = tokenizar_texto(texto)
    texto_stemmer = []
    for palabra in palabras:
        texto_stemmer.append(stemmer.stem(palabra))
    texto_filtrado = ' '.join(texto_stemmer)
    return texto_filtrado
```

Fuente: Elaboración propia

[22] Se ha seguido la documentación nltk.tokenize package—NLTK 3.3 documentation. Se recuperó el 8 de agosto del 2018 de https://www.nltk.org/api/nltk.tokenize.html.

[23] Se ha seguido el tutorial NLTK stop words–Python Tutorial – Pythonspot. Se recuperó el 8 de agosto del 2018 de https://pythonspot.com/nltk-stop-words/.

[24]Se ha seguido la documentación nltk.stem package — NLTK 3.3 documentation.Se recuperó el 8 de agosto del 2018 de https://www.nltk.org/api/nltk.stem.html.

Normalización: En esta parte se expanden las palabras que tienen contracciones, para este proceso creamos un mapeo de palabras que tienen contracciones versus la escritura correcta, para luego tomar cada uno de los tokens y aquellos donde haya coincidencia expandir el texto.

Figura 28 Fragmento de código del proceso de Normalización

```
def expandir_contracciones(texto, contraccion_mapping):
    """
    contracciones_patron = re.compile('({})'.format('|'.join(contraccion_mapping.keys())),
                                      flags=re.IGNORECASE|re.DOTALL)
    def expandir_match(contraccion):

        match = contraccion.group(0)
        first_char = match[0]
        contraccion_expandida = contraccion_mapping.get(match)\
                                if contraccion_mapping.get(match)\
                                else contraccion_mapping.get(match.lower())
        contraccion_expandida = first_char+contraccion_expandida[1:]
        return contraccion_expandida

    texto_expandido = contracciones_patron.sub(expandir_match, texto)
    texto_expandido = re.sub("'", "", texto_expandido)
    return texto_expandido
```

Fuente: Elaboración propia

4.1.3 Etapa 3: Transformación

Para realizar el proceso de extracción de características existen varios métodos (ver el apartado Minería de texto del capítulo 2), entre los cuales se ha seleccionado al modelo de Bolsa de Palabras con frecuencia bruta, *CountVectorizer*[25] y el modelo de Bolsa de Palabras basado en la frecuencia invertida del documento TF-IDF, que en la librería **sklearn**[26] se invoca con el método *TfidfVectorizer*. Éste último fue el que produjo los mejores resultados (ver Tabla 2 y Tabla 3).

El método *TfidfVectorizer* convierte la colección de documentos preprocesados en una matriz de funciones TF-IDF, con la función "*fit*" se calcula la media de los valores por columnas y se guardan como parámetros para que al realizar la transformación con la función *"transform"* reemplace los valores perdidos en el momento de creación de la matriz de TF-IDF, tal y como se muestra en la Figura 29.

Figura 29 Fragmento de código del proceso de creación de la Matriz TF-IDF

```
def vectorizar(textos, nombre):
    textosNormalizados = normalizar_corpus(textos)
    vectorizer = TfidfVectorizer()
    vectorizer = vectorizer.fit(textosNormalizados)
    data = vectorizer.transform(textosNormalizados)
    data = data.toarray()

    print 'Almacenando vectorizer en /VECTORIZER/vectorizer_'+ nombre +'.pickle'

    with open('../VECTORIZER/vectorizer_'+ nombre +'.pickle', 'wb') as handle:
        pickle.dump(vectorizer, handle)

    return data
```

Fuente: Elaboración propia

[25]Se ha utilizado el método sklearn.feature_extraction.text.CountVectorizer—scikit-learn 0.19.2. Se recuperó el 8 de agosto del 2018 de http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

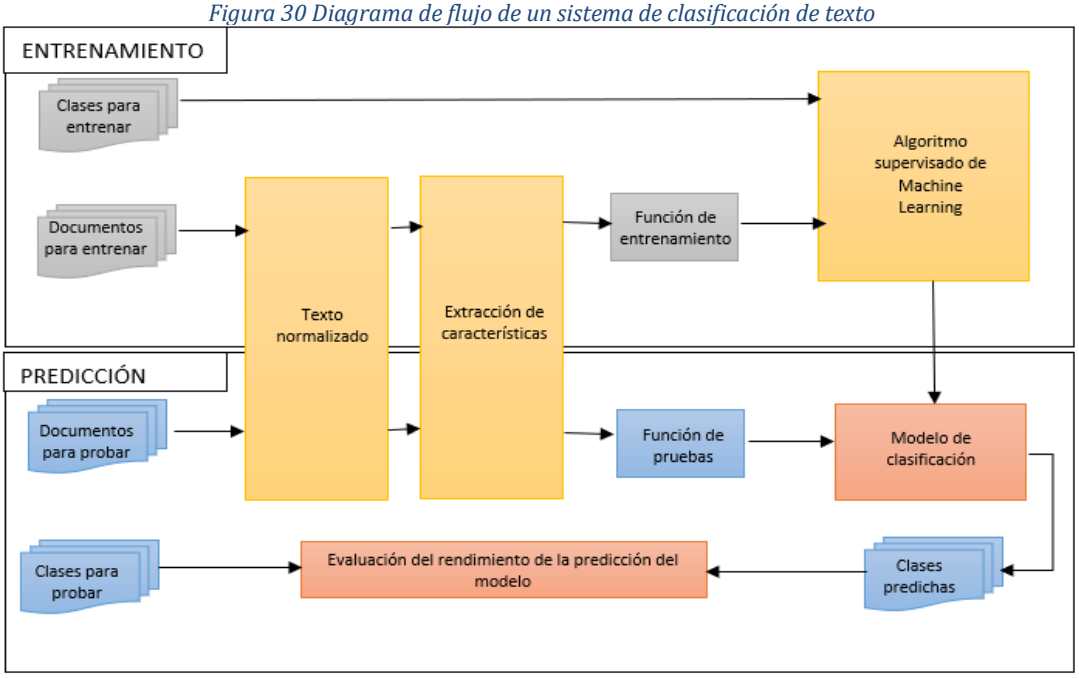
[26] Se ha utilizado la librería Scikit-learn. Se recuperó el 8 de agosto del 2018 de http://scikit-learn.org/.

Para finalizar guardamos el vector en un archivo para que pueda ser utilizado posteriormente por el algoritmo de clasificación.

4.1.4 Etapa 4: Minería de texto

Para desarrollar el sistema de clasificación se siguió el flujo de proceso para clasificar documentos, el cual nos marcó el orden en que debíamos ejecutar las tareas para programar de mejor manera el sistema. En la Figura 30 se puede apreciar los dos recorridos que se deben seguir, comenzando con el de entrenamiento y terminando con el de predicción.

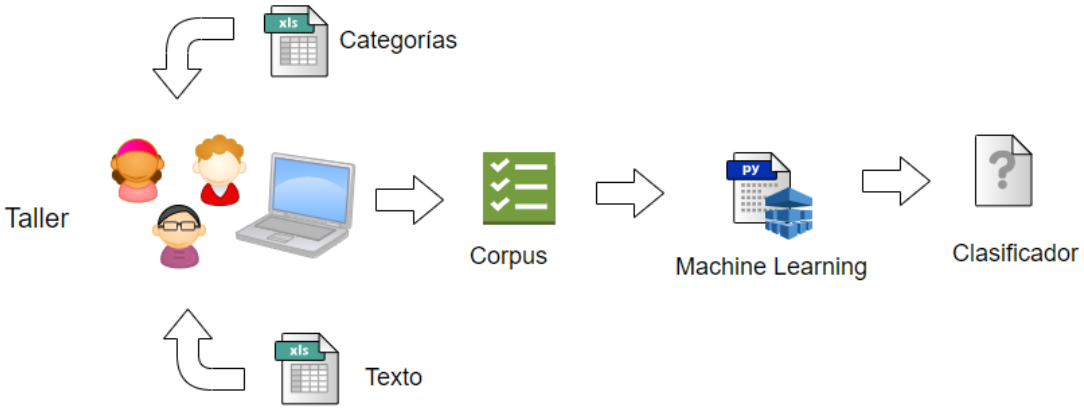
Figura 30 Diagrama de flujo de un sistema de clasificación de texto



Fuente: Elaborado a partir de (Dipanjan, 2016)

Entrenamiento: Comencemos con el flujo de trabajo para entrenar el modelo, en esta fase se le asigna de forma manual una de las veintidós (22) categorías de la lista de categorías a los textos extraídos. Para asignar las clases se usa la afinidad temática que tiene el *tweet* con la clase (ver Figura 31).

Figura 31 Modelo de alto nivel del proceso de entrenamiento



Fuente: Elaboración propia

Como el objetivo es dotar de conocimiento al sistema para que sea capaz de predecir las clases de los nuevos textos, decidimos convocar a la ciudadanía para que nos colaborara en la clasificación manual y de esta manera no "*viciar*" la máquina con una sola forma de pensar. Para desarrollar esta tarea les recomendamos a las personas que asistieron a nuestra llamada, realizar la búsqueda con la herramienta "*Search-Advance*" de Twitter y que solamente eligieron tweets significativos, lo que quiere decir que no pueden ser ambiguos, no deben estar duplicados, no deben ser *Retweets* y mucho menos tener menciones.

Durante esa jornada las personas fueron depositando los *tweets* clasificados en un archivo compartido, hasta completar entre todos un corpus de entrenamiento lo suficientemente grande que le permitiera aprender a la máquina[27]. Dentro de los tweets encontrados se reservó un porcentaje para el entrenamiento y otro para pruebas, en una proporción 80-20 por categorías, es decir, si se cuenta con 100 tweets por categoría se toman 80 para entrenamiento y 20 para pruebas. Los tweets usados para entrenar son diferentes a los tweets usados para las pruebas para evitar así el *overfitting*.

El corpus de entrenamiento lo usamos para probar varios algoritmos de clasificación y quedarnos con el que mejor resultados proporciona, al final el elegido fué, el modelo de Máquina de Soporte de Vector lineal o *linear Support Vector Machine (SVM)* que en la librería de **sklearn** se invoca con el algoritmo *SGDClassifier*[28] con el parámetro *loss=hing,* como se muestra en la Figura 32.

Figura 32 Fragmento de código del Clasificador

```
def crearClasificador(nombre,data, labels):

    nCores = multiprocessing.cpu_count()
    print 'Entrenando el modelo...'
    clasificador = SGDClassifier(loss='hinge', n_iter=100, n_jobs=nCores)
    clasificador = clasificador.fit(data, labels)

    print 'Almacenando modelo en /MODELOS/' + nombre + '.pickle'
    with open('../MODELOS/'+ nombre +'.pickle', 'wb') as handle:
        pickle.dump(clasificador, handle)
```

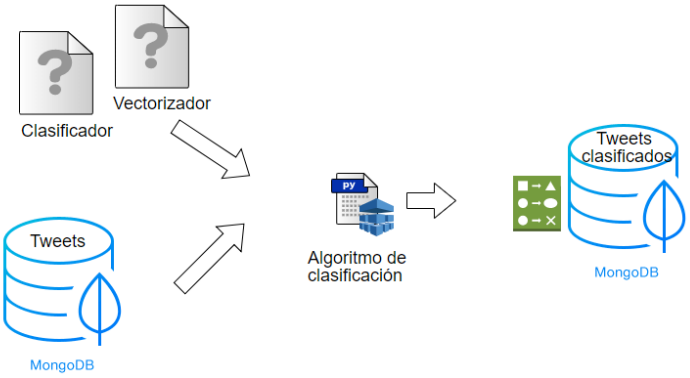
Fuente: Elaboración propia

El algoritmo *SGDClassifier* utiliza la matriz de funciones TF-IDF (ver etapa de transformación) e implementa modelos lineales regularizados con descenso de gradiente estocástica simple (SGD) que usa funciones de pérdida y penalizaciones para la clasificación. Para finalizar, guardamos el clasificador en un archivo para que pueda ser usado durante la predicción.

[27] Para ver un ejemplo del documento compartido durante el taller, remitirse al Apéndice D de este documento.
[28] sklearn.linear_model.SGDClassifier scikit-learn 0.19.2 documentation.Se recuperó el 9 de agosto del 2018 de http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html.

Predicción: En esta nueva ruta usamos tanto el modelo de clasificación que generamos durante el entrenamiento, como la matriz de funciones TF-IDF usada para crear dicho modelo y realizamos el proceso como se muestra en la Figura 33.

Figura 33 Modelo de alto nivel del proceso de clasificación



Fuente: Elaboración propia

Este recorrido comienza con la lista de clases y un conjunto de nuevos *tweets* recolectados en la etapa 1 (selección), paso seguido, se pre-procesan los textos de los *tweets*, como se indica en la etapa 2 (preprocesamiento), se transforman los textos preprocesado a través de la matriz de funciones TF-IDF generada para el entrenamiento y se crea una nueva matriz. Este recorrido se representa en forma de código fuente en la Figura 34.

Figura 34 Fragmento de código para aplicar la Matriz TF-IDF

```
def transformar(textos, nombre):
    textosNormalizados = normalizar_corpus(textos)
    vectorizador = leer_Pickle('vectorizer_'+ nombre +'.pickle')
    data = vectorizador.transform(textosNormalizados)
    data = data.toarray()
    return data
```

Fuente: Elaboración propia

El último paso en la predicción es incorporar la matriz al modelo de clasificación para hacer la predicción de las clases. Este último paso se representa en el código de la Figura 35.

Figura 35 Fragmento de código para predecir clases

```
tfidf=transformar(corpus,nombre)
SGDtfidf=leer_Pickle(nombre +'.pickle')
#print 'Se ha usado el vectorizador....'
clases=SGDtfidf.classes_ #generando las cl
puntaje=SGDtfidf.decision_function(tfidf)
```

```
for i in range(len(corpus)):
    lista= sorted(zip(puntaje[i],  clases), reverse=True)
    categoria=convNumToNom(lista[0][1])
    guardar={
        "categoria":categoria.decode('utf-8'),
        "puntaje":lista[0][0],
        "idt":idt[i], #1
        "texto":corpus[i].decode('utf-8'),
        "fecha":today,
        "fechaTweet":fechaTweet[i]
        };
```

Fuente: Elaboración propia

Finalmente están todos los *tweets* clasificados, listos para ser almacenados en una base de datos y visualizar los resultados.

4.1.5 Etapa 5: Interpretación y evaluación

El rendimiento de los modelos de clasificación por lo general se basa en qué tan bien predicen las etiquetas para los nuevos textos. Por lo general, este rendimiento se mide usando un porcentaje reservado del corpus de entrenamiento para pruebas, el cual no puede ser usado para entrenar el clasificador.

Durante el proceso de pruebas se extraen características de la misma forma que se hace durante el entrenamiento, se incorporan al modelo ya entrenado y se obtienen las predicciones, se contrastan con las etiquetas reales para ver qué tan bien o con qué precisión ha predicho el modelo. Varias métricas determinan el rendimiento de predicción de un modelo, pero nos centraremos en las métricas de *precision, recall, accuracy,* y *F1-score*. Sobre los algoritmos *Linear, Multinomial Naive Bayes (MNB), Support Vector Machine Gradient Descent (SGD), Decision Tree (DT)* y *Support Vector Machine with kernel lineal (SVM kernel=lineal)*. Los resultados se resumen en la Tabla 2, Tabla 3, Tabla 4 y Tabla 5

Tabla 2 Resultados Algoritmos de Clasificación aplicando bolsas de palabras con con TF-IDF

MÉTRICAS	**ALGORITMO**				
	Linear	MNB	SGD	DT	SVM (kernel=lineal)
Precision	76%	71%	76%	71%	75%
Recall	75%	70%	76%	70%	74%
Accuracy	75%	70%	76%	70%	74%
F1-score	75%	69%	76%	70%	74%

Fuente: Elaboración propia

Tabla 3 Resultados Algoritmos de Clasificación aplicando bolsas de palabras frecuencia bruta

MÉTRICAS	**ALGORITMO**				
	Linear	MNB	SGD	DT	SVM (kernel=lineal)
Precision	74%	71%	75%	67%	75%
Recall	74%	67%	75%	66%	75%
Accuracy	73%	67%	75%	66%	75%
F1-score	73%	65%	74%	66%	75%

Fuente: Elaboración propia

Tabla 4 Mejores resultados por tipo de vector

	Tipo de vector	
	Frecuencia Bruta	TF-IDF
Triunfos	3	15

Fuente: Elaboración propia

Tabla 5 Mejores resultados por Algoritmo

	ALGORITMO				
	Linear	MNB	SGD	DT	SVM (kernel=lineal)
Triunfos	Un empate con SGD	0	3	0	Uno y tres empates con SGD

Fuente: Elaboración propia

De los resultados obtenidos podemos decir que indiscutiblemente el modelo de bolsas de palabras TF-IDF fue el que mejores resultados proporcionó, saliendo triunfador en 15 ocasiones y empatando en 2 ocasiones, su adversario solo consiguió 3 victorias.

En cuanto a los algoritmos de clasificación, los que no ganaron en ninguna de las métricas establecidas fueron, el *Multinomial Naive Bayes (MNB)* y el *Decision Tree (DT)*. Los que mejores resultados proporcionaron fueron: el *Linear,* el *Support Vector Machine Gradient Descent (SGD)* y el *Support Vector Machine with kernel lineal (SVM kernel=lineal)*. De los tres que quedaron en el podium podemos decir que:

- *Linear* empató en el primer lugar con *SGD* para la métrica *precision* usando el vector TF-IDF.
- *SVM* con *kernel lineal* empató en el primer lugar con *SGD* para las métricas, *precision, Recall* y *Accuracy* usando el vector frecuencia bruta, con este mismo vector triunfó en la métrica *F1-score*.
- *SGD* triunfó con las métricas *Recall, Accuracy* y *F1-score* usando el vector TF-IDF.

Estos tres algoritmos implementan un clasificador de vector de soporte, pero cada uno usa un algoritmo de optimización diferente. Por tal motivo, SGD se llevó la victoria por poca diferencia porcentual y fue seleccionado como el algoritmo de clasificación de nuestra aplicación[29].

[29] El archivo de salida del código con el que se realizaron las pruebas a los algoritmos se puede ver en el Apéndice E de este documento.

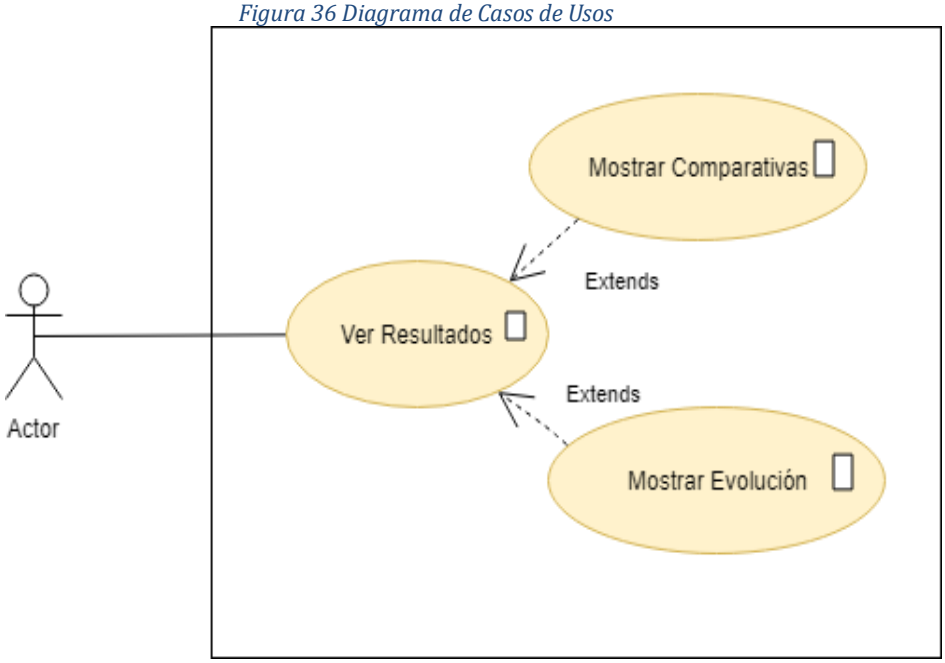
4.2 Módulo 2: Visualización

Este módulo se encarga de representar de forma visual los resultados obtenidos en el módulo anterior. De acuerdo con lo establecido en la metodología seleccionada, este módulo se desarrolla en cuatro etapas:

4.2.1 Etapa de Análisis de Requisitos

En esta sección se describe el proceso que realiza cualquier usuario al momento de entrar al sitio web, tal y como se muestra en la Figura 36. El usuario accede a la página y puede seleccionar del menú resultado la página comparativa o la página evolución. En la primera opción puede ver una comparación entre el número de tweets por categorías para un mes determinado, con respecto, al número de *datasets* por categorías publicado en el catálogo de datos abiertos del Ajuntament de València. En la segunda opción puede ver cómo ha sido la evolución de los temas de conversación en Twitter por parte de la ciudadanía hacia el Ajuntament de València.

Figura 36 Diagrama de Casos de Usos



Fuente: Elaboración propia

- El caso de uso "Ver resultados" es del *stereotype* explorar (☐«browsing»). Modela la navegación del usuario entre las visualizaciones de comparativas y evolución.
- El caso de uso "Mostrar comparativas" es del *stereotype* explorar (☐«browsing»). Ejecuta el proceso de búsqueda de los datos comparativos a través del envío de parámetros.
- El caso de uso "Mostrar evolución" es del *stereotype* explorar (☐«browsing»). Ejecuta el proceso de búsqueda de los datos de evolución a través del envío de parámetros.

4.2.2 Etapa de Diseño

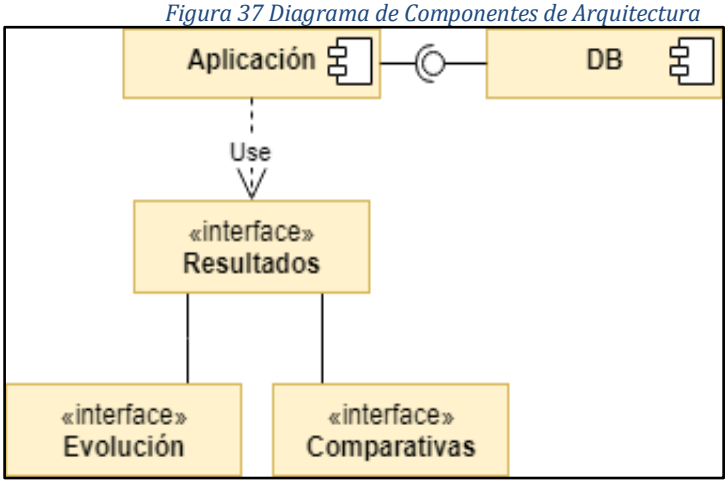
En esta sección se realizan los diseños necesarios para desarrollar el sitio web

Diseño Conceptual

Para tener una visión de lo que se va a desarrollar se diseña la arquitectura y los objetos que conforman la aplicación.

- **Diagrama de componentes de arquitectura:** se describe la arquitectura de la aplicación web. Como se observa en la Figura 37, existe una aplicación conectada a la base de datos que interactúa con la interfaz resultados.

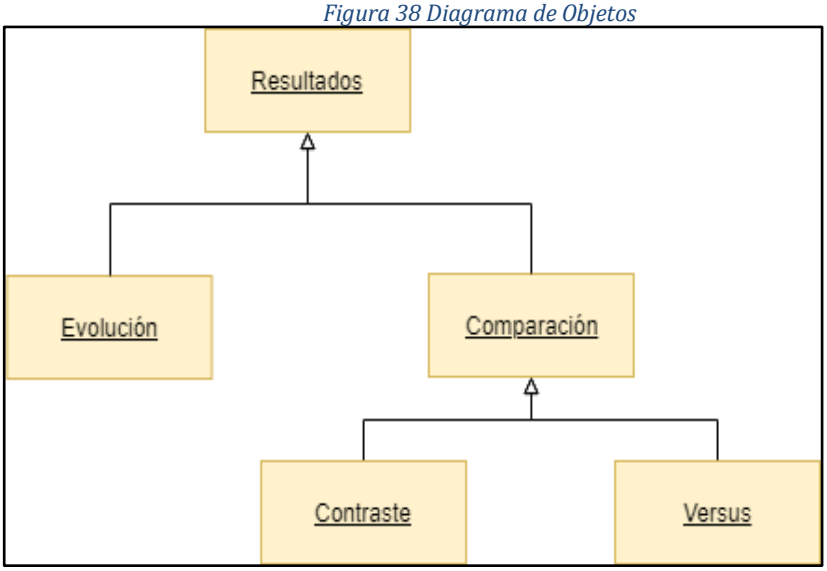
Figura 37 Diagrama de Componentes de Arquitectura



Fuente: Elaboración propia

- **Diagrama de objetos:** se describen los objetos involucrados en el módulo de visualización. Como se puede ver en la Figura 38 hay dos tipos de objetos resultado, uno de evolución y otro de comparación, este último tiene un objeto contraste y un objeto Versus.

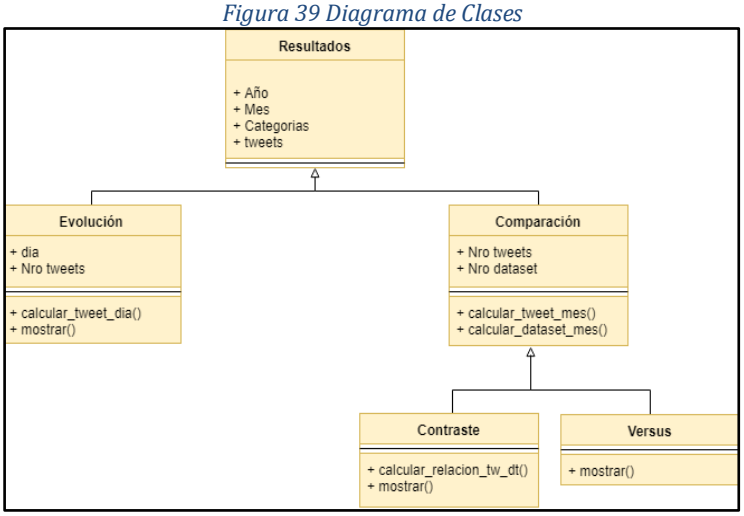
Figura 38 Diagrama de Objetos



Fuente: Elaboración propia

Al refinar el modelo conceptual llegamos al **diagrama de clases**, agregando atributos y métodos a los objetos, tal y como se muestra en Figura 39.

Figura 39 Diagrama de Clases

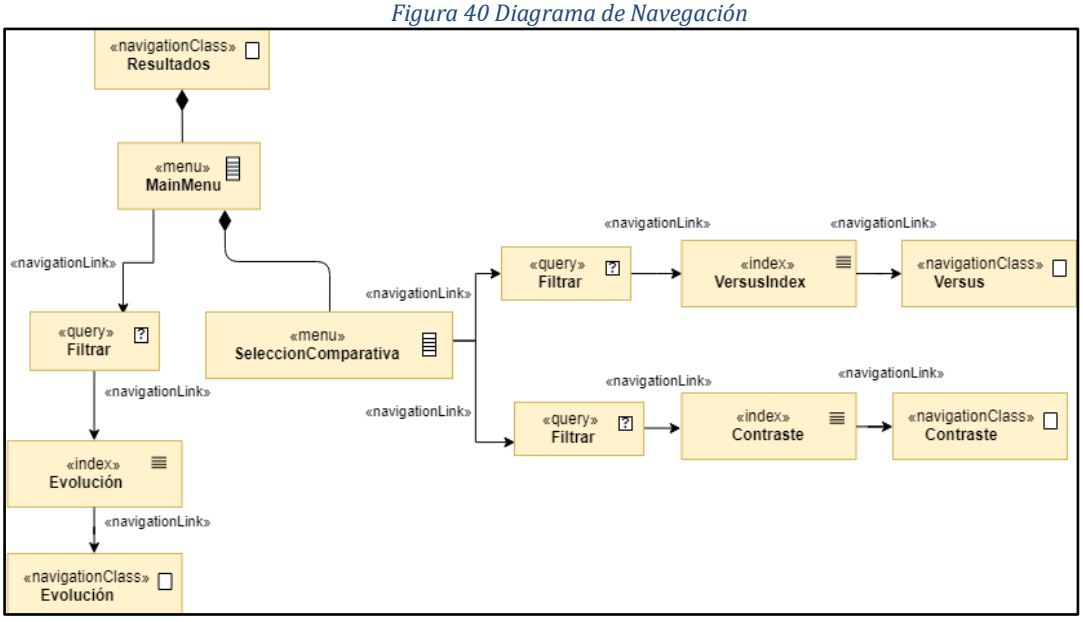


Fuente: Elaboración propia

Diseño Navegacional

En esta etapa mostramos como se puede navegar a través de nuestra aplicación web, además de establecer los nodos y los enlaces. En la Figura 40, se muestra con detalle cómo funciona la navegación en la página. Los nodos de navegación están representados por las clases "Resultados", "Versus", "Contraste" y "Evolución". Los enlaces *"navigationLink"* muestran vínculos y posibles pasos a seguir por el usuario. Los menús, "MainMenu" y "SeleccionarComparativa" indican que hay más de una alternativa de navegación. Por último, las primitivas de acceso «index» como es "VersusIndex", "ContrasteIndex" y "EvoluciónIndex" se utilizan para seleccionar los elementos con los tipos «query» cómo "Filtrar".

Figura 40 Diagrama de Navegación

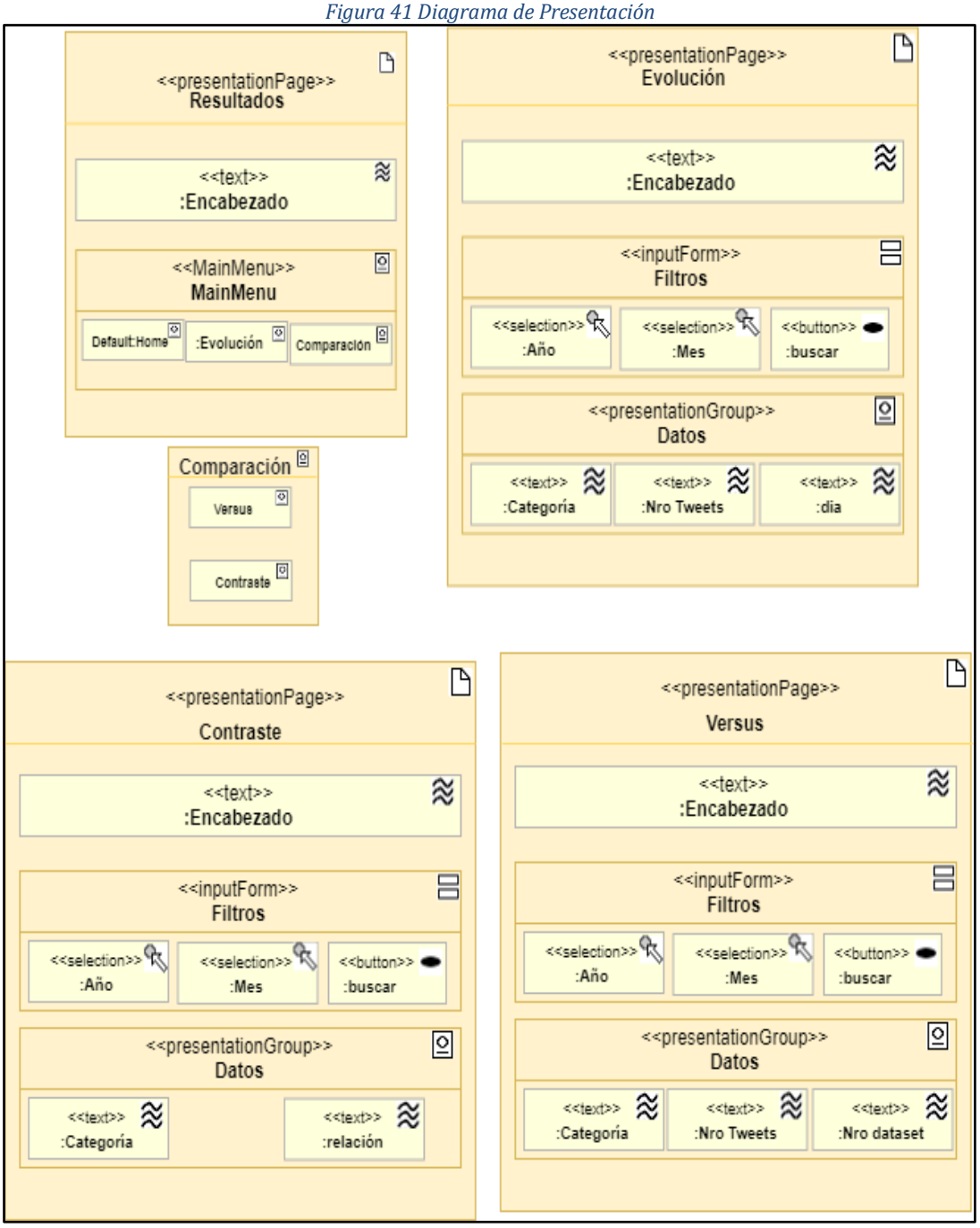


Fuente: Elaboración propia

Diseño Presentación

En la Figura 41 se modela la página de presentación "Resultados" que tiene un encabezado y un menú principal, a través del cual se puede acceder a las páginas, “Versus”, “Contraste” y “Evolución”. Estas tres últimas páginas tienen filtros que permiten mostrar la información de acuerdo con los parámetros año y mes.

Figura 41 Diagrama de Presentación



Fuente: Elaboración propia

4.2.3 Etapa de Codificación

En esta sección se hace uso de un modelo de desarrollo de software en capas, con el objetivo de separar las partes que componen el sistema y permitir que

nuestra aplicación sea escalable. En esta aplicación web usamos tres capas, una de presentación o capa Web, una de negocio o capa aplicativa y la capa de datos o de Base de Datos.

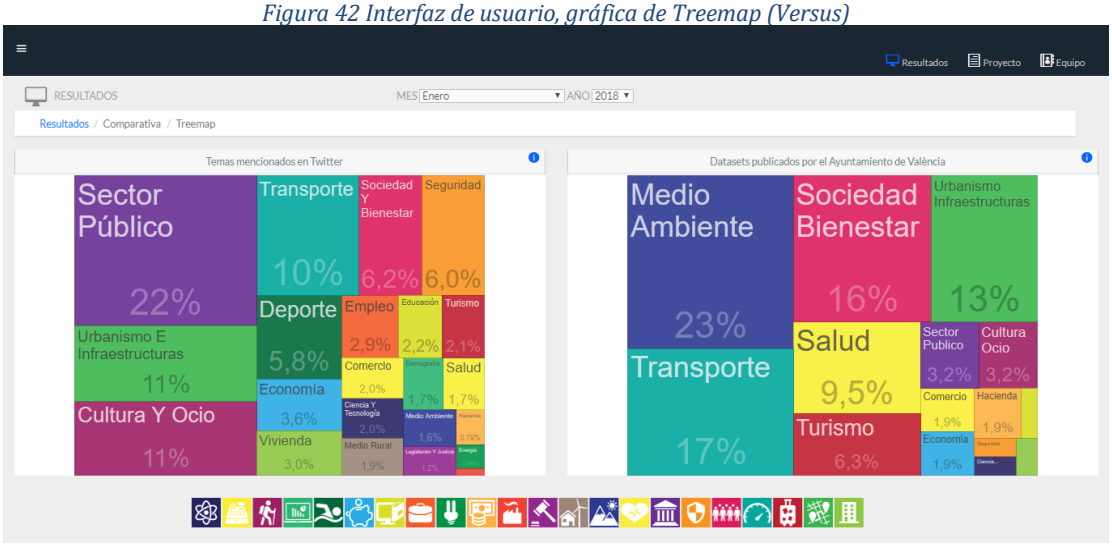
Capa de presentación

En esta capa se desarrollaron páginas HTML[30] con CSS[31] y JavaScript[32]. El HTML lo utilizamos para darle la estructura básica a nuestro sitio Web, el CSS[33] para darle los estilos, y las funciones JavaScript[34] para que se comunicara con la capa de negocio a través de peticiones a la API REST que desarrollamos.

Entre las interfaces de usuario tenemos como interfaces principales a la interfaz "Versus" en la Figura 42, "Contraste" en la Figura 43, y "Evolución" en la Figura 44.

En la pantalla que se muestra en la Figura 42, el usuario puede comparar por categoría, el porcentaje de *tweets* (gráfica izquierda) Vs el porcentaje de *datasets* publicados por el Ajuntament de València (gráfica derecha).

Figura 42 Interfaz de usuario, gráfica de Treemap (Versus)



Fuente: Elaboración propia

En la pantalla que se muestra en la *Figura 43*, el usuario puede comparar por categoría para un mes y año determinado, la relación tweets y número de *datasets* publicados por el Ajuntament de València, si el valor es positivo quiere decir que hay más oferta que demanda y si es negativo, más demanda que oferta.

[30] W3C HTML. Se recuperó el 12 de agosto del 2018 de https://www.w3.org/html/.

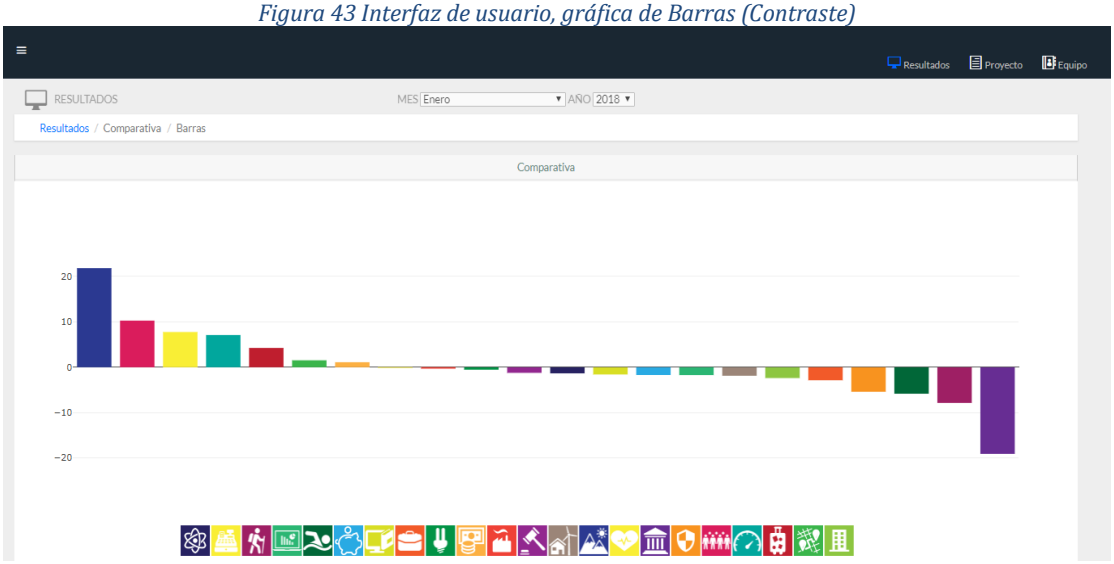
[31] CSS Snapshot 2017 World Wide Web Consortium. Se recuperó el 12 de agosto del 2018 de https://www.w3.org/TR/css-2017/.

[32] JavaScript.com. Se recuperó el 12 de agosto del 2018 de https://www.javascript.com/.

[33] CSS Snapshot 2017 - World Wide Web Consortium. Se recuperó el 12 de agosto del 2018 de https://www.w3.org/TR/css-2017/.

[34] JavaScript.com. Se recuperó el 9 de agosto del 2018 de https://www.javascript.com/.

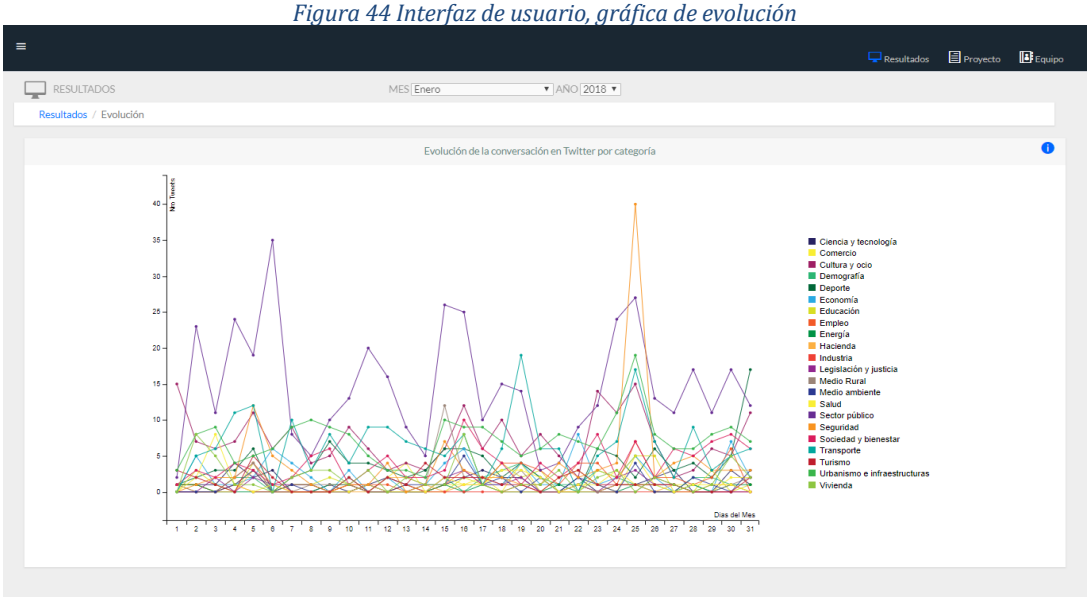
Figura 43 Interfaz de usuario, gráfica de Barras (Contraste)



Fuente: Elaboración propia

En esta pantalla, ver Figura 44, el usuario puede ver para un mes y año determinado, la evolución de las conversaciones por temáticas en Twitter.

Figura 44 Interfaz de usuario, gráfica de evolución



Fuente: Elaboración propia

Capa de negocio

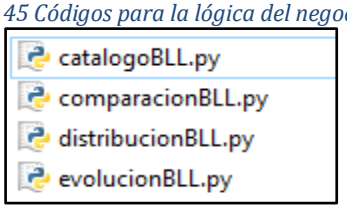
En esta capa se desarrolló la API REST usando *bottle*[35], que es un framework diseñado para apoyar el desarrollo de sitios web dinámicos, aplicaciones web y servicios web en Python.

Dentro de la API hay un programa principal llamado *inicio.py* que hace las llamadas a los códigos que contienen la lógica del negocio. En la Figura 45 se muestran los cuatro códigos desarrollados en Python que conforman la lógica de

[35] Bottle: Python Web Framework -Bottle 0.13-dev documentation Se recuperó el 12 de agosto del 2018 de https://bottlepy.org/. 2018.

nuestra aplicación. Si hacemos la analogía con el diagrama de clases diseñado en la etapa anterior, podemos decir que, la clase "Versus" la conforman funciones que están tanto en el código catalogoBLL como el código comparacionBLL. La clase "Contraste" en el código distribucionBLL y por último la clase "Evolución" en el código evolucionBLL.

Figura 45 Códigos para la lógica del negocio



Fuente: Elaboración propia

Cada vez que un usuario en la capa de presentación selecciona una opción de los parámetros mes o año, se envía una petición a esta API y el programa *inicio.py* gestiona con que código se debe trabajar dicha petición, para hacer los cálculos pertinentes y enviar la respuesta.

Capa de datos

En esta capa se especifica el modelado de la base de datos NoSql que utilizamos, por tanto, en vez de mostrar un modelo relacional, mostramos un ejemplo de cada una de las dos colecciones que conforman el proyecto. En la Figura 46 se muestra la colección TWEET, en la Figura 47 la colección CATÁLOGO y en la Figura 48 la colección TWCLASIFICADO.

Figura 46 Colección TWEET


```
{
    "_id" : ObjectId("5a535455bc54f437d024d9e2"),
    "username" : "LasNaves",
    "userSeguidores" : 8767,
    "consulta" : "@AjuntamentVLC",
    "idioma" : "es",
    "dispositivo" : "Hootsuite",
    "idt" : "950323745469673473",
    "userid" : 262129793,
    "fecha_creacion" : ISODate("2018-01-08T11:10:11.000Z"),
    "fechaTweet" : ISODate("2018-01-08T12:21:51.520Z"),
    "tweet" : "Profesionales de la enseñanza se reúnen viernes y sábado en....
}
```


Fuente: Elaboración propia

Figura 47 Colección CATALOGO


```
/* 1 */
{
    "_id" : ObjectId("5a5d15265cf6265622bec13b"),
    "anyo" : "2018",
    "datos" : [
        {
            "categoria" : "ciencia tecnologia",
            "numDatasets" : 1
        },.....
        {
            "categoria" : "vivienda",
            "numDatasets" : 1
        }
    ],
    "mes" : "01",
    "FechaUpdateCatalogo" : "2018-01-05T09:06:15"
}
```


Fuente: Elaboración propia

Figura 48 Colección TWCLASIFICADO


```
{
    "_id" : ObjectId("5a5d0c7c5cf62655696d56f0"),
    "categoria" : "Cultura y ocio",
    "idt" : "946651734776205312",
    "texto" : "@festesdeVLC @joanribo @pichiavo @perefuset
    "puntaje" : 0.374041422043131,
    "fechaTweet" : ISODate("2017-12-29T07:58:56.000Z"),
    "fecha" : ISODate("2018-01-15T20:18:04.202Z")
}
```


Fuente: Elaboración propia

Creación de Esquema: las tres colecciones TWEET, CATÁLOGO y TWCLASIFICADO se crearon en tiempo de ejecución. La colección TWEET se creó cuando se descargó el primer *tweet*, la colección CATÁLOGO cuando se leyó por primera vez de forma automática el RDF del Catálogo de datos del portal de datos abiertos del Ajuntament de València y la colección TWCLASIFICADO cuando se clasifica automáticamente el primer *tweet*. Todos estos procesos fueron desarrollados y ejecutados en el módulo de clasificación.

4.2.4 Etapa de Pruebas

En esta sección se hace referencia a las pruebas unitarias realizadas para verificar y validar la aplicación web. Para realizar pruebas a este módulo durante la codificación se fueron evaluando las sentencias desarrolladas y si alguna de estas fallaba o generaba resultados inesperados por la existencia de algún defecto, se corregía enseguida. También se realizaron pruebas sobre la interfaz de usuario para comprobar la funcionalidad, el comportamiento en la entrada y salida de datos y la integridad de la información enviada y recibida. En la Tabla 6 se puede ver un ejemplo de algunas de las pruebas realizadas.

Tabla 6 Ejemplo de caso de prueba unitaria

Identificador	VI_PAR_005
Nombre	Entradas de parámetros
Objetivo	Verificar que en las páginas “Versus”, “Contraste” y “Evolución” al seleccionar como entrada un mes que no tiene datos, el sistema no muestre error y simplemente no muestre datos.
Descripción	Se selecciona en el desplegable de mes “enero” y en el desplegable de año “2017”.
Resultado esperado	Mostrar un mensaje que indique que no hay datos para ese mes y año en específico.

Fuente: Elaboración propia

4.3 Integración y Pruebas

En esta sección se prueban todos los módulos que han sido desarrollados, con el objetivo de conocer si funcionan acorde a la lógica del negocio. En la Tabla 7 se listan las pruebas que se realizaron a los módulos de clasificación y visualización:

Tabla 7 Ejemplo de caso de prueba

Módulo	Id Prueba	Tipo	Descripción
Clasificación	CL_DT_001	Unitaria	Comprobar que durante la descarga no se están duplicando *tweets*, es decir, no se guarda dos veces el mismo tweet (idt repetido).
Clasificación	CL_DT_002	Unitaria	Comprobar que se ejecuta de forma diaria el proceso de descarga de *tweet*.
Clasificación	CL_DT_003	Integral	Comprobar que el proceso **descargatweet** se ejecuta a la hora establecida, para que no interfiera con el proceso de clasificación
Clasificación	CL_DT_004	Unitaria	Verificar que se está generando un archivo de log informando el número de tweet descargados o si ocurrió algún problema con la descarga.
Clasificación	CL_CT_001	Unitaria	Comprobar que se ejecuta de forma diaria el proceso de clasificación de tweets
Clasificación	CL_CT_002	Unitaria	Comprobar que se genera un archivo de log informando si el proceso de clasificación se ha realizado o no.
Clasificación	CL_CT_003	Unitaria	Revisar que inmediatamente después que se ejecuta el proceso de clasificación, todos los textos que cumplen las condiciones establecidas son clasificados
Clasificación	CL_CT_004	Unitaria	Realizar revisiones aleatorias de los textos clasificado para verificar que el clasificador está prediciendo de forma acertada.
Clasificación	CL_CAT_001	Unitaria	Comprobar que las lecturas al catálogo funcionan correctamente.
Clasificación	CL_CAT_002	Unitaria	Comprobar que mensualmente el programa está accediendo al catálogo de datos abiertos y almacena el número de datasets publicados para ese mes.
Visualización	VI_PAR_001	Integral	Verificar que al pasar como parámetro un mes y año determinado el sistema realmente muestra los datos para ese año y mes.
Visualización	VI_GR_001	Integral	Revisar que los *tweets* clasificados diariamente se visualicen en las gráficas que se le muestran al usuario.

Módulo	Id Prueba	Tipo	Descripción
Visualización	VI_PA_001	Unitaria	Revisar que la página web se visualiza de forma correcta en por lo menos tres navegadores diferentes.
Visualización	VI_PA_002	Unitaria	Comprobar que los links entre las diferentes pestañas funcionan de forma correcta y enlazan a las páginas que corresponden.
Visualización	VI_GR_002	Unitaria	Comprobar que por cada gráfica existe un icono (i) ubicado en la parte superior de cada gráfica que muestra información sobre dicha gráfica.
Visualización	VI_GR_003	Integral	Verificar que los cálculos presentados en las gráficas son correctos.

Fuente: Elaboración propia

Después de haber ejecutado las pruebas, tanto unitarias como de integración, se obtuvo más de un 90% de cumplimiento, los errores presentados fueron corregidos. Durante la integración de los módulos no ocurrieron muchos errores, debido en gran parte a que la salida de uno de los módulos era la entrada del otro, por lo que solo teníamos que asegurarnos que el módulo de clasificación generase los datos que debía y el módulo de visualización leyera los datos como debía.

Para concluir podemos decir que, aunque durante el desarrollo realicemos pruebas al funcionamiento del código, se pueden pasar por alto funcionalidades primordiales que pueden evidenciarse al aplicar este tipo de pruebas. Además, es vital comprobar que los módulos desarrollados por separado se compenetran y funciona en armonía una vez se ponen a trabajar juntos.

Capítulo 5
Resultados

El producto final de todo el proceso de desarrollo se resume en 4 visualizaciones repartidas en tres páginas web.

- En la primera página se muestra información porcentual de las conversaciones por temática en Twitter vs el porcentaje de *datasets* publicados por categorías en el portal de datos abiertos.
- En la segunda página se muestra la relación oferta y demanda tomando como base las temáticas de las conversaciones y los *datasets* publicados.
- En la última página se muestra la evolución de las conversaciones por temáticas en Twitter.

A continuación, se detallan dos tipos de resultados, por un lado, se muestran los resultados que nos ofrece la solución a través de las gráficas y por otro lado se muestran los puntos que quedaron pendientes en el desarrollo de la herramienta, por falta de tiempo, personal o infraestructura y que podrían mejorar sustancialmente la eficacia de la herramienta.

5.1 Resultados de la solución

En este apartado se explican cada una de las gráficas elaboradas, se realiza un análisis de cada una de ellas y se discute acerca de las conclusiones a las que se llega o se puede llegar con la información

5.1.1 *Datasets* publicados por el Ajuntament de València

Antes de comenzar a analizar la información correspondiente a la gráfica de los datasets, es importante saber que ella muestra para un mes y año determinado, los *datasets* publicados hasta ese momento y no se tiene en cuenta si han sido actualizados o no. Por lo tanto, el valor de una categoría de un mes a otro solo cambia, si se publica un nuevo *dataset* y este se vincula a cualquiera de las 22 categorías analizadas.

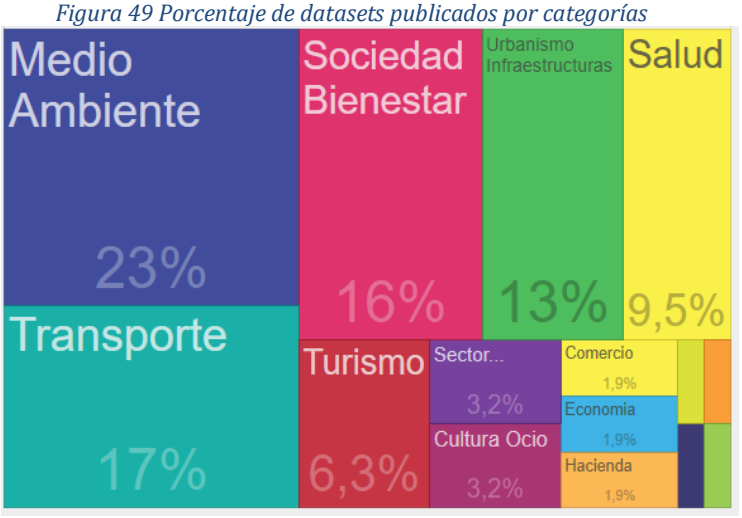
Para representar la información se utiliza la visualización en *treemap*, por ser excelentes para mostrar grandes cantidades de información en forma jerárquica. La visualización está dividida en rectángulos cuyo tamaño y orden está determinado por el valor de la variable que representan (TIBCO Software, s.f.).

Descripción de la información de la gráfica

En el portal de datos abiertos del Ajuntament de València para el mes de agosto del 2018 hay publicados 118 *datasets* repartidos entre 15 de las 22 categorías. En la Figura 49 se muestra esta información en forma porcentual. En

esta misma figura se puede apreciar como Medio ambiente es la categoría que más *datasets* tiene publicados, con un 23%, seguido de Transporte con un 17% sobre el total de *datasets*. En el caso de categorías como demografía, legislación, industria, deporte, empleo, energía, y medio rural, el valor es igual 0% por no contar con ningún dataset publicado y por tanto en la gráfica no se muestra.

Figura 49 Porcentaje de datasets publicados por categorías



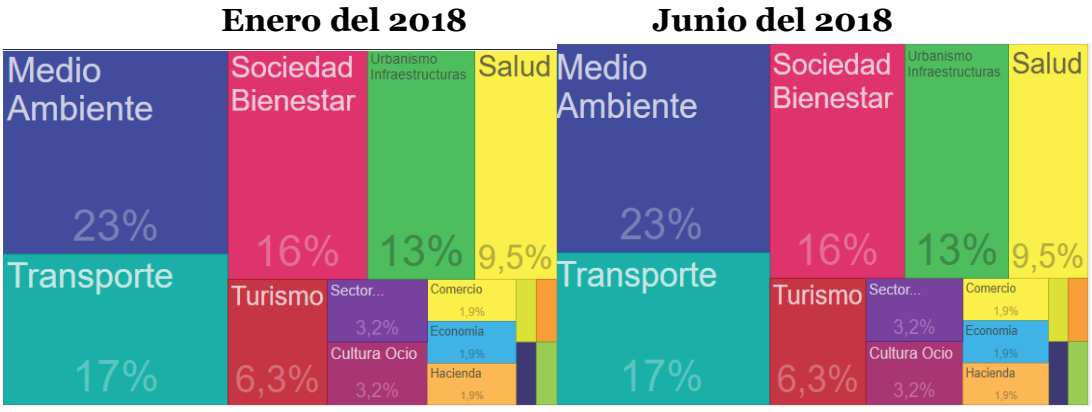
Fuente: Elaboración propia

La información proporcionada se puede contrastar accediendo al Portal de Datos Abiertos del Ajuntament de València en el apartado datos abiertos.

Comparativa temporal del volumen de datasets por categorías

Si comparamos la información de la gráfica de dataset desde enero hasta junio del presente año observamos que la información es exactamente, la misma, tal y como se muestra en la Figura 50.

Figura 50 Comparativa temporal de datasets publicados por categorías.



Fuente: Elaboración propia

Aunque en la figura solo se muestra información referente al mes inicial y al mes final de la evaluación, la comparativa la hicimos mes a mes, mirando si había algún cambio de un mes a otro, para concluir que el Ajuntament de València en lo que va del año 2018 no ha ingresado ni un solo nuevo dataset.

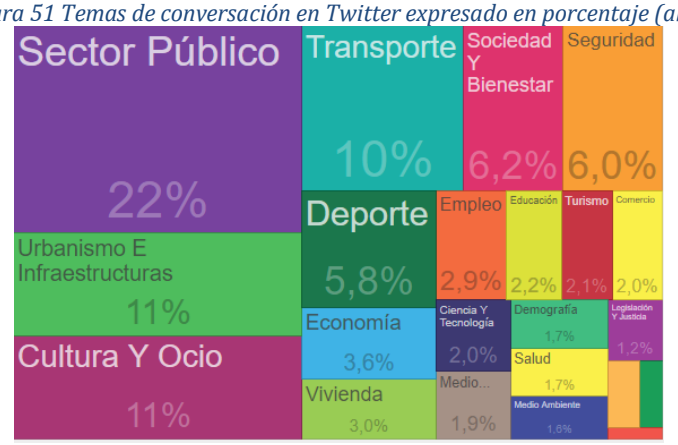
5.1.2 Conversaciones por temática en Twitter

Antes de comenzar a analizar la información correspondiente a la gráfica de las conversaciones por temática en Twitter, es importante saber que en ella se visualiza la información mensual de los tweets clasificados con la herramienta que hemos desarrollado. Para representar la información, al igual que la gráfica explicada en el punto 4.1.1 se utiliza la visualización en treemap por las mismas razones que se explican en ese punto.

Descripción de la información de la gráfica

El sistema de clasificación toma los tweets descargados y los distribuye en 22 temáticas que se visualizan en forma porcentual como se observa en la Figura 51.

Figura 51 Temas de conversación en Twitter expresado en porcentaje (abril)



Fuente: Elaboración propia

El sistema de clasificación tiene en cuenta algunos criterios para determinar si los *tweets* pertenecen a una u otra categoría. A modo de ejemplo a continuación se exponen algunos criterios sobre las tres principales temáticas que se muestran en la gráfica:

Sector Público: Esta es la categoría donde más diversidad hay. Básicamente hay información sobre Presupuestos, contrataciones, licitaciones, Organigrama institucional, Legislación interna, Función pública, etc.

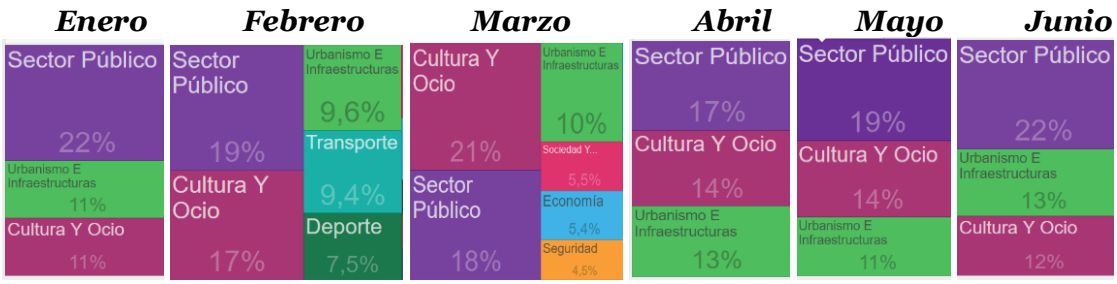
Cultura y Ocio: Es toda la información relacionada con literatura, cine, música, danza, teatro, tiempo libre, fiestas, etc.

Urbanismo e Infraestructura: Es toda la información relacionada con Saneamiento público, construcción, información cartográfica, información catastral, mapas georreferenciados, etc.

Los principales temas de conversación durante el primer semestre del 2018

Si comparamos la información mensual que se muestra en la Figura 52, podemos observar que, durante el primer semestre del año, han predominado tres temas de conversación en Twitter.

Figura 52 Principales temas de conversación durante el primer semestre del 2018.



Fuente: Elaboración propia

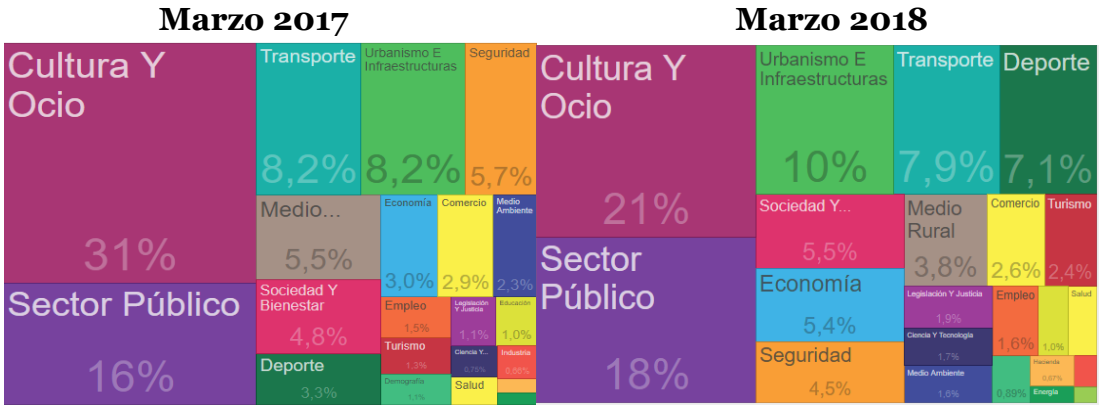
El primer lugar lo ocupó el sector público durante cinco de los seis meses analizados, exceptuando el mes de marzo, cuando el primer puesto lo ocupó cultura y ocio y el sector público pasó a un segundo lugar. El sector que más veces ocupó el segundo lugar fue Cultura y Ocio, en tres de los seis meses evaluados. En uno de los tres meses restantes ocupó el primer lugar y los otros dos meses el tercer lugar, puesto ocupado en 4 ocasiones por Urbanismo e infraestructura.

A pesar de que se demuestra un gran interés de la ciudadanía sobre estas temáticas, el Ajuntament de València solo tiene publicado en su portal de datos abiertos 5 *datasets* para sector público y 5 para cultura y ocio que representan un 3,2% sobre el total de *datasets* publicados. Para la temática que ocupa el tercer lugar, el ayuntamiento cuenta con 20 *datasets*, que representa un 13% sobre el total de *datasets* publicados. El hecho de que estas tres temáticas continuamente ocupen los tres primeros lugares, debería ser un indicio para que las AAPP revisen las políticas de apertura de datos y analicen sobre que subcategorías deberían abrir los datos.

El principal tema de conversación durante el mes de marzo

Cultura y Ocio ocupa el primer lugar en las conversaciones durante el mes de marzo del año 2017 y 2018, tal y como se muestra en la Figura 53.

Figura 53 Principal tema de conversación durante el mes de marzo del 2017 y 2018.

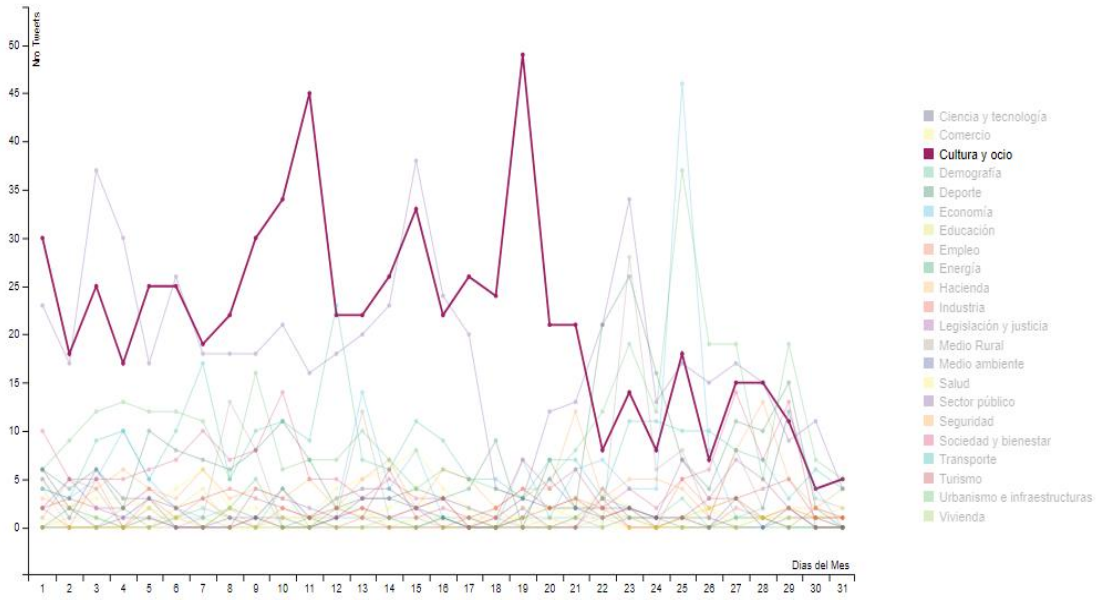


Fuente: Elaboración propia

Este comportamiento tiene su explicación en la celebración de Las Fallas, fiesta tradicional de la Comunidad Valenciana, que se celebran desde el 15 al 19

de marzo, pero oficialmente comienza el último domingo del mes de febrero, por lo que si se mira la evolución en la Figura 54 se nota como el tema de conversación desde comienzos del mes de marzo se va inclinando hacia esta temática y después del 19 va decayendo drásticamente.

Figura 54 Evolución del tema Cultura y Ocio durante el mes de marzo del 2018.



Fuente: Elaboración propia

Esta información puede resultar muy valiosa para el ayuntamiento porque a tenor del interés identificado, es previsible pensar que la apertura de *datasets* relacionados con Las Fallas podrían tener niveles elevados de uso.

Aunque tres de los cinco *datasets* publicados en la categoría cultura y ocio, son sobre Las Fallas, sería positivo publicar más información o por lo menos tenerla actualizada anualmente.

5.1.3 Demanda vs Oferta

Para representar la información se utiliza la visualización en barras, por permitir resumir el conjunto de datos por categoría. Los datos se muestran usando barras de la misma anchura, el valor de la variable determina su altura (TIBCO Software, s.f.).

En esta gráfica se hace uso de los tweets clasificados por la herramienta desarrollada y los *datasets* publicados por categoría para mostrar la diferencia entre la oferta y la demanda, tal y como se observa en la Figura 55.

Figura 55 Oferta y demanda de información durante el mes de junio

Fuente: Elaboración propia

Cada una de las barras representa la diferencia entre oferta y demanda de información por categoría. En la Figura 55 se han dividido las barras en tres bloques o conjuntos:

- el conjunto azul indica las categorías que tienen mayor oferta (Medio ambiente, sociedad-bienestar, salud, transporte, turismo y hacienda),
- el rojo, las que mayor demanda presentan (sector público, cultura-ocio, deporte, seguridad, economía, empleo, etc) y
- el verde, el equilibrio, donde la oferta y la demanda se igualan, y deberían estar todas las categorías. A través de la interacción entre el ayuntamiento y la ciudadanía se puede llegar a ese equilibrio deseado.

5.1.4 Evolución de las conversaciones en Twitter

Como forma de apoyo a la visualización descrita en el punto 4.1.2, se ha creado la gráfica de evolución, que permite visualizar la información diaria de los tweets clasificados con la herramienta que hemos desarrollado. Para representar la información se utilizan gráficos de líneas por ser excelentes para mostrar tendencias a lo largo de un período de tiempo (TIBCO Software, s.f.).

Figura 56 Evolución de las tres principales temáticas durante el mes de junio

Fuente: Elaboración propia

La gráfica de línea muestra por días cuánto se ha hablado en Twitter sobre una temática. A manera de ejemplo en la Figura 56, hemos seleccionado las evoluciones de los tres principales temas de conversaciones durante el mes de junio, esta información muestra al detalle los días en que se habló más o menos sobre un tema y si se asocia con eventos que están ocurriendo se pueden sacar muchas conclusiones.

5.2 Análisis crítico de la herramienta desarrollada

A pesar de que la herramienta desarrollada genera resultados sobresalientes, recibió algunos reconocimientos[36] y nos brindó la fundamentación teórico-práctica para impartir un taller[37]. Los resultados podrían ser mejores si se le realizarán algunos ajustes. Por tal motivo, le dedicaremos el siguiente apartado a revisar algunos puntos donde vemos que se podría mejorar tanto en eficiencia como en funcionalidad el trabajo realizado.

1. Utilizar técnicas[38] más rigurosas para realizar el preprocesamiento de la información.

2. El corpus que utilizamos para realizar el entrenamiento de la máquina de clasificación es escaso, por lo que dedicarle un tiempo a clasificar manualmente *tweets* resultaría enriquecedor para el modelo. Con un corpus más robusto se podrían volver a realizar las pruebas de algoritmo y observar si con más características alguno de los algoritmos exhibe mejores métricas o si el que tenemos sigue siendo el mejor, pero con una mayor precisión.

3. Añadir a las pruebas más métricas, como por ejemplo la matriz de confusión para ver la precisión que tienen los diferentes algoritmos al realizar predicciones por clases y con base a ello poder tomar decisiones.

4. A pesar de que guardamos el puntaje que se le da a cada una de las clases, no tenemos establecido un umbral a partir del cual decidir los tweets que están bien clasificados, por lo que fijar un umbral podría ayudar a acertar en la clasificación de los *tweets*. La idea no es solo crear dicho umbral, sino monitorizarlo para ver que funciona correctamente y, llegado el caso, ajustarlo según como se vaya comportando el modelo.

5. Desarrollar un módulo de monitorización, donde la ciudadanía sea la encargada de revisar los *tweets* y de acuerdo con unas pautas se seleccionen los que se han clasificado mejor.

---

[36] Para ampliar la información con respecto al reconocimiento que recibió este proyecto, remitirse al Apéndice F de este documento.

[37] Taller de análisis de contenido de comunicación ciudadana. Impartido el 15. 12.17 por Marylin Mattos.

[38] Técnicas de imputación de valores perdidos y técnicas de identificación de ruido.

6. Actualmente no conocemos las subcategorías de las clasificaciones que realizamos. Por ejemplo, si las personas hablan sobre sector público no sabemos de forma automática si se refieren a la parte legislativa, a presupuesto, al organigrama o cualquier otra cosa que tenga que ver con esta temática, por lo que no podemos aventurarnos a puntualizar en la información que se necesitaría abrir.
7. El sistema que desarrollamos no es bidireccional, es decir, no extraemos la información de las respuestas que el ciudadano recibe del Ajuntament de València. Sería interesante agregar esa parte para saber cómo está interactuando el ayuntamiento.
8. No contamos con otro sistema con que comparar lo que clasificamos en Twitter, por lo que si además de los tweets, clasificamos las quejas y reclamaciones que llegan al Ajuntament de València serviría para hacer comparativas y ayudar a las AAPP a gestionar mejor la información que les llega.
9. Al terminar la parte de lectura de catálogo de datos nos dimos cuenta de que hay mucha información que se puede extraer y usar como, por ejemplo, las fechas de actualizaciones de los *datasets*, con lo que se puede ver la frecuencia de actualización de estos.
10. Para finalizar, el módulo de visualización tiene muy pocas gráficas, así que estaría muy bien agregarle más gráficas, como, por ejemplo, la de evolución mensual de ingreso de nuevos *datasets* al catálogo y la de frecuencia de actualización de los datasets. En este proceso resultaría fundamental seguir un diseño centrado en el usuario.

Capítulo 6
Conclusiones

El desarrollo de este trabajo fin de Máster me ha supuesto un enriquecimiento y aprendizaje tanto cívico como técnico, a la vez de reforzar lo aprendido a lo largo del estudio del Máster Oficial Universitario en Gestión de la Información.

A nivel cívico me ha permitido obtener más conocimiento sobre políticas públicas y su proceso o ciclo de elaboración, cuyas entradas son las necesidades o problemas que aquejan a la ciudadanía y las salidas son las políticas públicas formuladas. Existen varios tipos de participación ciudadana que están dados por la distribución de poder tolerada por los organismos públicos, en el diseño de las modalidades de participación. Siguiendo en esta línea de conocimiento descubrí la existencia de un "nuevo paradigma de gobierno" que realmente no es tan nuevo, puesto que, desde hace décadas, existen movimientos, leyes, normas, decretos, etc que buscan un gobierno más abierto, un gobierno transparente, donde todos podamos saber qué hacen las AAPP y cómo lo hacen. Se sabe que falta mucho camino por recorrer, pero lo más importante es que ya se comenzó a caminar en esa dirección.

A nivel técnico me documenté sobre la importancia que tiene aplicar una adecuada metodología de desarrollo en dependencia de la naturaleza del problema. Los proyectos de Machine Learning cuentan con sus propias metodologías de aplicación, por lo que no se deben aplicar metodologías tradicionales de desarrollo. Del mismo modo, aprendí la relevancia que tiene fijar técnicas de preprocesamiento de la información (limpieza y normalización) y modelos de transformación, como pasos previos a la ejecución de algoritmos de clasificación.

En cuanto a las asignaturas vistas a lo largo de mi proceso de formación en el máster, fue importante el contenido curricular de varias de ellas para el desarrollo de este proyecto. La asignatura *explotación de datos masivos* desencadenó mi curiosidad en el lenguaje de programación Python para los procesos de extracción de información y análisis de datos. En esta misma asignatura conocí de la existencia de la API Tweepy para extraer información de la red social Twitter. En la asignatura *fuentes de datos e información*, conocí sobre Open Data que era un tema totalmente extraño para mí en esos momentos. En la asignatura *almacenamiento y recuperación de información* me enseñaron sobre las formas de extraer la información de los textos o el Text Mining, a través del uso de algunas técnicas de Recuperación de Información (RI). En la asignatura *Business Intelligence*, recibí la orientación necesaria para tomar la decisión de utilizar una interfaz gráfica para representar la información. Cerrando la lista se encuentra la asignatura *tecnologías de*

información para el gobierno abierto que fue donde supe que la formación técnica, adquirida hasta ese momento podía ser usada para elaborar herramientas que ayuden a crear una sociedad más justa, donde la información sea un recurso que le permita a la ciudadanía incidir en las políticas públicas de apertura de datos.

Para finalizar, resultó muy provechoso desarrollar el proyecto en un ambiente de trabajo colaborativo dentro del marco de la Cátedra Govern Obert, donde contaba con la ayuda y orientación de personas con experiencia en proyectos cívicos. Con personas adscritas al Ajuntament de València qué nos daban el punto de vista de las AAPP y que trataron en todo momento de facilitarnos información, la cual, hubiera resultado imposible obtener por nuestros propios medios. Gracias a ellos, tuve la oportunidad de interactuar con personal técnico del Ajuntament de València que me indicaron, entre otras cosas, como funcionaba el proceso de quejas y reclamaciones una vez llega al ayuntamiento, dándome ideas para nuevos proyectos.

6.1 Conclusiones finales

En este apartado se establecen las metas alcanzadas a partir de los objetivos específicos trazados al iniciar el proyecto:

1. Después de analizar diferentes canales de comunicación que pudieran servir como fuente de información para nuestro sistema de clasificación, escogimos la red social de microblogging Twitter, por ser uno de los canales de comunicación donde hay presencia activa de ciudadanos que comparten temas que se estiman de interés colectivo. Así mismo, esta red social brinda la facilidad de extraer información publicada en abierto y que va dirigida a un órgano de la administración pública, en este caso al Ajuntament de València.

2. Decidir desarrollar el software en dos módulos fue vital para el éxito del proyecto. La modularidad del proyecto permitió fijar tareas más simples, centradas en solucionar problemas más pequeños, además poder aplicar metodologías diferentes a cada uno de los módulos, en función de la naturaleza del problema que debían resolver. Para el módulo de clasificación aplicamos una metodología de *Machine Learning*, mientras que para el módulo de visualización aplicamos una metodología propia para aplicaciones web.
   La selección de las tecnologías para desarrollar la aplicación también fue acertada, porque nos permitió reforzar conocimientos y aprender mucho más sobre cómo aplicar dichas tecnologías a problemas reales.

3. En cuanto al proceso de limpieza de los datos, considero que nos quedamos cortos en la lista que creamos para las palabras contraídas o que estaban mal escritas, debimos ser más rigurosos en la eliminación de

menciones y referencias a páginas webs y aplicar más técnicas de preprocesamiento. La librería que utilizamos para hacer esta tarea, aunque es bastante completa, aún le queda mucho por explorar en el procesamiento del lenguaje español. Pese a que, las librerías para Python desarrolladas por la comunidad para esta tarea, son muy prometedoras se nota que surgieron de personas de habla inglesa, así que es importante que las personas de habla hispana nos sumemos y tratemos de mejorar lo que hay para nuestro idioma.

Para finalizar, podemos decir que, la limpieza de los datos es más que aplicar técnicas, es analizar la forma como se escribe y aplicar procesos acordes a las necesidades, por lo que se deben invertir más horas y esfuerzos a esta parte del proceso.

4. En total analizamos cinco algoritmos de clasificación (*Linear, Multinomial Naive Bayes MNB, Support Vector Machine Gradient Descent SGD, Decision Tree DT* y *Support Vector Machine with kernel lineal SVM kernel=lineal)* y utilizamos las métricas de *precision, recall, accuracy,* y *F1-score,* para determinar el rendimiento en la predicción de cada uno de los cinco algoritmos. Para casi todas las métricas los resultados estuvieron por encima del 70% y los resultados más sobresalientes recayeron en los algoritmos que implementan un clasificador de vector de soporte, dando como ganador al algoritmo SGD por poca diferencia porcentual. Este algoritmo ha funcionado muy bien haciendo las predicciones de las clases, aunque consideramos que se debe crear un corpus de entrenamiento más robusto y realizar más pruebas, ya sea para mejorar el 76% que dio como resultado en las métricas o si alguna otra muestra mejores resultados.

5. Involucrar a la ciudadanía en la creación del corpus de entrenamiento fue crucial para dotar a la máquina de un "pensamiento" heterogéneo y no viciarla con una sola forma de pensar. Con esto, la máquina utiliza como ejemplo para extraer patrones al pensamiento de más de una persona y crear el modelo de clasificación. La creación de la base del conocimiento es uno de los pasos más importantes en la construcción de los modelos predictivos y que sea la ciudadanía la encargada de hacer esta tarea, es darle el poder de que ellos mismos sean los que determinen cómo debe "pensar" la máquina.

6. El producto final del proyecto son visualizaciones que le sirven tanto a la ciudadanía como al Ajuntament de València poder conocer la evolución de los temas de conversación que la ciudadanía tiene en Twitter, el volumen de esas conversaciones por temáticas y el volumen de datasets. Con estas gráficas pudimos darnos cuenta que:

- El Ajuntament de València no publica nuevos *datasets* por lo menos desde enero del 2018.
- Desde enero del 2018, en la red social Twitter predominan temas de conversaciones asociados al sector público, cultura-ocio y urbanismo-infraestructura.
- El mes de marzo es el más propicio para promover herramientas enfocadas en la cultura y ocio.
- La mayor oferta de información está en las categorías medio ambiente, sociedad-bienestar, salud, transporte, turismo y hacienda.
- La mayor demanda de información se encuentra en las categorías sector público, cultura-ocio, deporte, seguridad, economía y empleo.

Todo esto nos lleva a conocer los niveles de oferta y demanda de información, así como la falta de apertura de datos en algunas temáticas y la necesidad de que se tenga en cuenta estos resultados como entrada en el proceso de elaboración de la política pública de apertura de datos.

Bibliografía

Administracion.gob.es, s.f. *Datos Abiertos*. [En línea]
Available at:
https://administracion.gob.es/pag_Home/espanaAdmon/Transparencia_DatosAbiertos/datos_abiertos.html
[Último acceso: 28 Julio 2018].

Ajuntament de València, s.f. *Portal Transparència i Dades Obertes*. [En línea]
Available at: http://gobiernoabierto.valencia.es/va/
[Último acceso: 12 Agosto 2018].

Anexo III-Norma Técnica de Interoperabilidad de Reutilización de recursos de la información. (2013) BOE.

Anexo IV-Norma Técnica de Interoperabilidad de Reutilización de recursos de la información. (2013) BOE.

Anon., s.f. *PAe-Portal Administración Electrónica de España*. [En línea]
Available at: https://administracionelectronica.gob.es/ctt/datosgob
[Último acceso: 28 Julio 2018].

Arntsein, S., 1969. *A Ladder of Citizen Participation*. s.l.:Journal of the American Planning.

Athul, J., 2017. *A Data Mining System for Sentiment Classification of Indian Currency Demonetisation using Naive Bayes Classifier Algorithm*. Noida, India, s.n.

avaScript.com, s.f. *avaScript*. [En línea]
Available at: https://www.javascript.com/.
[Último acceso: 12 Agosto 2018].

Barnes, J. A., 1954. *Class and comittees in a Norwegian Island parish, Human Relations, n° 7, p. 39-58.*. [En línea]
Available at: http://journals.sagepub.com/doi/10.1177/001872675400700102
[Último acceso: 27 Julio 2018].

Bautista, M. J., 2000. *Phd Thesis-Soft Computing Models for Information Retrival,* Granada, España: Universidad de Granada.

Benito Carrillo, A., 2016. *Los 8 principios básicos de los Datos Abiertos*. [En línea]
Available at: https://www.viavansi.com/blog-xnoccio/es/8-principios-de-los-datos-abiertos/
[Último acceso: 24 Julio 2018].

Bifet, A. & Frank, E., 2010. *Sentiment knowledge discovery in twitter streaming data. In International conference on discovery science.*. s.l.:Springer.

Bográn, A. P., Alonso, J. L. . B. & García , C. F., 2013. *Análisis léxico sobre los tweets de Twitter*. Salamanca, s.n.

Carl Malamud, Public.Resource.Org, 2007. *Open Government Working Group*. [En línea]
Available at: https://public.resource.org/open_government_meeting.html
[Último acceso: 25 Julio 2018].

Centro Latinoamericano de Administración para el Desarrollo, 2016. *Carta Iberoamericana de Gobierno Abierto*. Bogotá, Colombia, s.n.

Cohen, R., 2013. *Obama Signs Open Data Executive Order: U.S. Government Data To Be Made Freely Available*. [En línea]
Available at: https://www.forbes.com/sites/reuvencohen/2013/05/09/obama-signs-open-data-executive-order-all-u-s-government-data-to-be-made-freely-available/#6aced4e45e79
[Último acceso: 24 Julio 2018].

Delgado Godoy, L., 2009. *Documentación sobre gerencia pública, del SubgrupoA2, Cuerpo Técnico, Especialidad de Gestión Administrativa, de la Junta de comunidades de Castilla-La Mancha,* Castilla-La Mancha: s.n.

Dipanjan, S., 2016. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. s.l.:Apress.

Entidad Pública Empresarial Red.es, 2011. *Iniciativa de datos abiertos del Gobierno de España*. [En línea]
Available at: http://datos.gob.es/es
[Último acceso: 28 Julio 2018].

European Data Portal, 2017. *Open Data Maturity in the Europe 2017,* s.l.: European Union.

Fayyad, U. M., 1996. *Advances in knowledge discovery and data mining*.. s.l.:AAAI Press / The MIT Press..

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. Knowledge Discovery and Data Mining:Towards a Unifying Framework. Volumen uno, p. 7.

Free Software Foundation, s.f. *GNU General Public License*. [En línea]
Available at: https://www.gnu.org/licenses/gpl-3.0.en.html
[Último acceso: 8 Agosto 2018].

Garbarino, M. S., 1983. *Sociocultural Theory in Anthropology: A Short History*. Illinois: Weveland Press, Inc.

García Herrero, j., 2018. *ciencia de datos. minería de datos tecnicas analiticas y aprendizaje estadistico en un enfoque practico*. s.l.:altaria.

Holdren, J. P., Orszag, P. & Prouty, P. F., 2009. *Memorandum for Heads of Departments and Agencies*. Washington D. C, s.n.

INRIA, s.f. *scikit-learn. Machine Learning in Python*. [En línea]
Available at: http://scikit-learn.org/.
[Último acceso: 8 Agosto 2018].

Kraus, A. & Koch, N., 2003. *A Metamodel for UWE,* s.l.: Ludwig-Maximilians-Universität München.

La Norma técnica de interoperabilidad de reutilización de recursos de información (2013) BOE.

Maimon, O. & Rokach, L., 2010. *Data mining and knowledge discovery handbook*. Segunda ed. s.l.:Springer.

Manning, C. D., Raghavan, P. & Schütze, H., 2009. *Introduction to Information Retrieval*. [En línea]
Available at: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
[Último acceso: 8 Agosto 2018].

Marcel Hellkamp, s.f. *Bottle: Python Web Framework -Bottle 0.13-dev documentation*. [En línea]
Available at: https://bottlepy.org/docs/dev/
[Último acceso: 12 Agosto 2018].

Meda, C., Bisio, F., Gastaldo, P. & Zunino, R., 2016. *Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles*. Roma, s.n.

MongoDB, Inc., s.f. *MongoDB*. [En línea]
Available at: https://www.mongodb.com/
[Último acceso: 8 Agosto 2018].

Nelson, H. M. C., Alberto, G. L. J. & Dixon, S., 2016. *Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles*. Madrid, s.n.

NLTK Project, s.f. *NLTK 3.3 documentation: Natural Language Toolkit*. [En línea]
Available at: https://www.nltk.org/index.html
[Último acceso: 8 Agosto 2018].

Noguera Vivo , J. M., Grandío Pérez , M. d. M. & Martínez Polo, . J., 2011. *Redes sociales para estudiantes de Comunicación: 50 ideas para comprender el escenario online*. s.l.:UOC.

Norma Técnica de Interoperabilidad de Reutilización de recursos de la información (2013) BOE.

Open Data Charter team, 2017. *Open Data Charter*. [En línea]
Available at: https://opendatacharter.net/principles-es/
[Último acceso: 25 Julio 2018].

Open Government Directive, 2009 . *The home of the U.S. Government's open data*. [En línea]
Available at: https://www.data.gov/
[Último acceso: 28 Julio 2018].

Open Government Partnership, 2011. *Open Government Declaration*. [En línea]
Available at: https://www.opengovpartnership.org/open-government-declaration
[Último acceso: Julio 25 2018].

Open Knowledge International, s.f. *¿Qué son los datos abiertos?-The Open Data Handbook*. [En línea]
Available at: http://opendatahandbook.org/guide/es/what-is-open-data/
[Último acceso: 25 Julio 2018].

Partnership, O. G., s.f. *Open Data*. [En línea]
Available at: https://www.opengovpartnership.org/theme/open-data
[Último acceso: 20 07 2018].

Prieto Martín, P., 2005. *Sistemas avanzados para la participación electrónica municipal: ejes conceptuales para su diseño*. [En línea]
Available at:
https://www.researchgate.net/publication/28110296_Sistemas_avanzados_para_la_participacion_electronica_municipal_ejes_conceptuales_para_su_diseno
[Último acceso: 26 Julio 2018].

Python Software Foundation, s.f. *Python*. [En línea]
Available at: https://www.python.org/
[Último acceso: 8 Agosto 2018].

Quintanilla Mendoza, G. & Gil García, J. R., 2013. *Gobierno Abierto en América Latina: Modelo Conceptual, Planes de Acción y Resultados Preliminares*. México: ©Instituto Nacional de Administración Pública, A.C..

Ràfols, F., 2015. *Sindicat de Periodistes de Catalunya*. [En línea]
Available at:
http://www.sindicatperiodistes.cat/sites/default/files/documents/comparativa_europeadicast.pdf
[Último acceso: 20 07 2018].

Sánchez González, J. J., 2015. *La participación ciudadana como instrumento del gobierno abierto*. [En línea]
Available at: http://www.redalyc.org/pdf/676/67642415003.pdf
[Último acceso: 27 Julio 2018].

Sánchez Trigueros, J., 2015. Los antecedentes del gobierno abierto: una mirada retrospectiva en la evolución de la administración pública. *Ciencia Política y Administración Pública,* XIII(23), pp. 67-84.

scikit-learn developers, s.f. *Naive Bayes*. [En línea]
Available at: http://scikit-learn.org/stable/modules/naive_bayes.html
[Último acceso: 25 Julio 2018].

Scott, J., 1991. *Social Network Analysis*. Newbury Park,London: Sage.

Speck, R. & Attneave, C., 1982. *Redes Familiares*. Amorrortu ed. s.l.:s.n.

Sproat, R. y otros, 2001. Normalization of non-standard words. *Computer Speech and Language*.

Sunlight Foundation, 2017. *Ten Principles For Opening Up Government Information*. [En línea]
Available at: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/
[Último acceso: 25 Julio 2018].

The Open Government Partnership, s.f. *Open Data*. [En línea]
Available at: https://www.opengovpartnership.org/theme/open-data
[Último acceso: 25 Julio 2018].

TIBCO Software, s.f. *Qué es gráfico de barras*. [En línea]
Available at: https://docs.tibco.com/pub/spotfire_web_player/6.0.0-november-2013/es-ES/WebHelp/GUID-6023CECC-E502-4AE1-B5C5-FFE5DAF6FAE2.html
[Último acceso: 20 Agosto 2018].

TIBCO Software, s.f. *Qué es gráfico de líneas*. [En línea]
Available at: https://docs.tibco.com/pub/spotfire_web_player/6.0.0-november-2013/es-ES/WebHelp/GUID-60AD831A-D7C3-4407-AA54-44EBD23A29D0.html
[Último acceso: 20 Agosto 2018].

TIBCO Software, s.f. *Qué es un treemap*. [En línea]
Available at: https://docs.tibco.com/pub/spotfire_web_player/6.0.0-november-2013/es-ES/WebHelp/GUID-F3F4ABDF-8418-42D3-A1C4-60B7A8121C75.html
[Último acceso: 20 Agosto 2018].

Universidad de Múnich, s.f. *Profile Overview UWE Examples*. [En línea]
Available at: http://uwe.pst.ifi.lmu.de/profileOverview.html
[Último acceso: 12 Agosto 2018].

W3C, s.f. *CSS Snapshot 2017 World Wide Web Consortium*. [En línea]
Available at: https://www.w3.org/TR/css-2017/
[Último acceso: 12 Agosto 2018].

W3C, s.f. *W3C HTML*. [En línea]
Available at: https://www.w3.org/html/
[Último acceso: 12 Agosto 2018].

Yusra, y otros, 2017. *Music interest classification of twitter users using support vector machine*. s.l., s.n.

Apéndice A
Solicitud de información al Ajuntament de València

Carta de respuesta a la solicitud realizada al portal de transparencia sobre el texto en bruto de todas las sugerencias, quejas y reclamaciones recibidas por el Ajuntament de València.

Figura 1. Carta de respuesta a la solicitud de información al Ajuntament de València

AJUNTAMENT DE VALÈNCIA

N. Eixida *N. Salida*	
Data *Fecha*	05/04/2017
Expedient *Expediente*	E-00702-2017-000024-00
Servici *Servicio*	SER. TRANSP-GOVERN OBERT
Secció *Sección*	
ASSUMPTE *ASUNTO*	RESOLUCION

Destinatari(ària) / *Destinatario(a)*

Per Resolució JM-19 de data 05/04/2017, dictada per EL REGIDOR DE TRANSPARÈNCIA, GOVERN OBERT I AUDITORIA CIUTADANA / COOPERACIÓ AL DESENVOLUPAMENT I MIGRACIÓ, en virtut de delegació conferida per RESOLUCIÓ 865 DE 6 DE SETEMBRE DE 2016, s´ha disposat:

Antecedents

Amb data 2 de març va entrar a este Servei de Transparència i Govern Obert sol·licitud d' accés a la informació 2017-5 en la que Marylin Mattos Barros demanava el text en brut de tots els suggeriments, queixes i reclamacions que haja rebut l'Ajuntament de València de tots els ciutadans durant l'any 2016.

El motiu de la sol·licitud és obtindré informació per a realitzar un projecte d'investigació en la UPV mitjançant un sistema de classificació de la informació utilitzant la taxonomia comuna.

Fonaments

1. Resulta aplicable la normativa següent:

- Llei 7/1985, de bases del règim local (L.B.R.L.).

- Text Refós de les disposicions legals vigents en matèria de règim local (R.D. Leg. 781/1986) (T.R.R.L.).

- Llei 8/2010, de la Generalitat, de Règim Local de la Comunitat Valenciana (L.R.L.C.V.).

- Llei 39/2015, de Procediment Administratiu Comú de les Administracions Públiques.

- Llei 19/2013, de 9 de desembre, de transparència, accés a la informació i bon govern.

1/6

- Llei 2/2015, de 2 d'abril, de la Generalitat, de transparència, bon govern i participació

ciutadana de la Comunitat Valenciana.

- Reglament municipal de transparència i participació ciutadana, de 28 de setembre de 2012, modificat el 24 d'abril de 2015.

2. Resulten d'especial rellevància en relació a la resolució d'aquest expedient els següents preceptes:

a) Article 15 de la Llei 19/2013, de 9 de desembre, de transparència, accés a la informació i bon govern que, en relació a la protecció de dades personals, disposa.

1. Si la informació sol·licitada continguera dades especialment protegides als quals es refereix l'apartat 2 de l'article 7 de la Llei Orgànica 15/1999, de 13 de desembre, de Protecció de Dades de Caràcter Personal, l'accés únicament es podrà autoritzar en cas que es comptara amb el consentiment exprés i per escrit de l'afectat, llevat que dit afectat haguera fet manifestament públics les dades amb anterioritat al fet que se sol·licitara l'accés.

4. No serà aplicable l'establit en els apartats anteriors si l'accés s'efectua prèvia dissociació de les dades de caràcter personal de manera que s'impedisca la identificació de les persones afectades.

b) Article 18.1,e) de la Llei 19/2013, de 9 de desembre, de Transparència, Accés a la Informació i Bon Govern, que estableix que *no s'admetran a tràmit, mitjançant resolució motivada, les sol·licituds que siguen manifestament repetitives o tinguen un caràcter abusiu no justificat amb la finalitat de transparència d'aquesta Llei.*

3. La informació que sol·liciten no sols és extensíssima, ja que és refereix a instàncies presentades per ciutadans en un període d'un any a l'Oficina de Queixes, Suggeriments i Reclamacions (més de 7.000), la qual cosa suposa un laboriós treball de cerca de dades personals que cal ocultar d'acord a la Llei de Protecció de Dades Personals, sinó que, a més a més, no s'ajusta a la finalitat de la llei de transparència (LTAIBG) com a supòsit d'accés a la informació.

La LTAIBG té per objecte, segons el seu article 1, ampliar i reforçar la transparència de l'activitat pública, a més de regular i garantir el dret d'accés a la informació relativa a aquella activitat. Ara bé, al seu Preàmbul, es posa de manifest que només quan l'acció dels responsables públics es sotmet a escrutini, quan els ciutadans poden conéixer com es prenen les decisions que els afecten, com es manegen els fons públics o baix quins criteris actuen les nostres institucions, podrem parlar de l'inici d'un procés en el qual els poders públics comencen a respondre a una societat que és crítica, exigent i que demanda participació dels poders públics.

En conseqüència la finalitat de la llei és que puga facilitar-se informació pública de tal forma que permeta:

- sotmetre a escrutini l'acció dels responsables públics,

- conéixer com es prenen les decisions públiques,

- conéixer com es manegen els fons públics,

- conéixer baix quins criteris actuen les institucions públiques.

Com s'ha dit abans, en el present supòsit el motiu de la sol·licitud és obtenir informació per a un projecte d'investigació el resultat del qual pot ser molt interessant, però que no té encaix en allò que disposa la Llei de Transparència com a sol·licitud d'accés a la informació, no sols pels motius abans assenyalats, sinó perquè resulta materialment impossible obtindré la informació que es demana en els terminis que marca la llei de transparència per a resoldre les sol·licituds. Cal considerar que els documents que es demanen són en gran part manuscrits que, a més a més, contenen informació personal protegida, la qual cosa implica que s'ha de dissociar la mateixa per a poder facilitar-la. El problema radica en què es tracta de més de 7.000 documents i això impedeix que puga fer-se eixe treball en un termini tan breu com l'assenyalat a la llei de transparència.

No obstant això, la voluntat de la Regidoria de Transparència, Govern Obert i Auditoria Ciutadana és aprofundir en el coneixement dels interessos de la ciutadania per aconseguir un Govern Obert. Tenint en compte que, a més a més, aquesta sol·licitud vol utilitzar-se per realitzar un treball d'investigació en el marc de la Càtedra de Govern Obert que té l'Ajuntament de València amb la Universitat Politècnica de València, des del Servei de Transparència i GO van a fer-se gestions per trobar formules que possibiliten un coneixement més exacte dels suggeriments, queixes i reclamacions, així com altres informacions provinents d'altres canals municipals com ara: Telèfon 010, Oficines d'Atenció Ciutadana, Registre, Parla amb Joan Ribó, etc., contemplant tant la gestió del coneixement com la protecció de dades personals.

Antecedentes

Con fecha 2 de marzo ha tenido entrada en este Servicio de Transparencia y Gobierno Abierto solicitud de acceso a la información 2017-5 a través de la cual Marylin Mattos Barros pide el texto en bruto de todas las sugerencias, quejas y reclamaciones que haya recibido el Ayuntamiento de Valencia de todos los ciudadanos durante el año 2016.

El motivo de la solicitud es obtener información para realizar un proyecto de investigación en la UPV mediante un sistema de clasificación de la información utilizando la taxonomía común.

Fundamentos

1. Resulta aplicable la normativa siguiente:

- Ley 7/1985, de bases del régimen local (L.B.R.L.).

- Texto Refundido de las disposiciones legales vigentes en materia de régimen local (R.D. Leg. 781/1986) (T.R.R.L.).

- Ley 8/2010, de la Generalidad, de Régimen Local de la Comunidad Valenciana (L.R.L.C.V.).

- Ley 39/2015, de Procedimiento Administrativo Común de las Administraciones Públicas.

- Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información y buen gobierno.

- Ley 2/2015, de 2 de abril, de la Generalidad, de transparencia, buen gobierno y participación ciudadana de la Comunidad Valenciana.

- Reglamento municipal de transparencia y participación ciudadana, de 28 de septiembre de 2012, modificado el 24 de abril de 2015.

2. Resultan de especial relevancia en relación a la resolución de este expediente los siguientes preceptos:

a) Artículo 15 de la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información y buen gobierno que, en relación a la protección de datos personales, dispone.

1. Si la información solicitada contuviera datos especialmente protegidos a los que se refiere el apartado 2 del artículo 7 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, el acceso únicamente se podrá autorizar en caso de que se contase con el consentimiento expreso y por escrito del afectado, a menos que dicho afectado hubiese hecho manifiestamente públicos los datos con anterioridad a que se solicitase el acceso.

4. No será aplicable lo establecido en los apartados anteriores si el acceso se efectúa previa disociación de los datos de carácter personal de modo que se impida la identificación de las personas afectadas.

b) Artículo 18.1,e) de la Ley 19/2013, de 9 de diciembre, de Transparencia, Acceso a la Información y Buen Gobierno, que establece que *no se admitirán a trámite, mediannte resolución motivada, las solicitudes que sean manifiestamente repetitivas o tengan un carácter abusivo no justificado con la finalidad de transparencia de esta Ley.*

3. La información que solicitan no solo es extensíssima, ya que se refiere a instancias presentadas por ciudadanos en un período de un año a la Oficina de Quejas, Sugerencias y Reclamaciones (más de 7.000), lo cual supone un laborioso trabajo de búsqueda de datos personales que hay que ocultar de acuerdo a la Ley de Protección de Datos Personales, sino que, además , no se ajusta a la finalidad de la ley de transparencia (LTAIBG) como supuesto de acceso a la información.

La LTAIBG tiene por objeto, según su artículo 1, ampliar y reforzar la transparencia de la actividad pública, además de regular y garantizar el derecho de acceso a la información relativa a

aquella actividad. Ahora bien, en su Preámbulo, se pone de manifiesto que solo cuando la acción de los responsables públicos se somete a escrutinio, cuando los ciudadanos pueden conocer cómo se toman las decisiones que les afectan, cómo se maneja el fondo público o bajo qué criterios actúan nuestras instituciones, podremos hablar del inicio de un proceso en el que los poderes públicos comienzan a responder a una sociedad que es crítica, exigente y que demanda participación de los poderes públicos.

En consecuencia la finalidad de la ley es que pueda facilitarse información pública de tal forma que permita:

- someter a escrutinio la acción de los responsables públicos,

- conocer cómo se toman las decisiones públicas,

- conocer cómo se maneja el fondo público,

- conocer bajo qué criterios actúan las instituciones públicas.

Como se ha dicho antes, en el presente supuesto el motivo de la solicitud es obtener información para un proyecto de investigación el resultado del cual puede ser muy interesante, pero que no tiene cabida en aquello que dispone la Ley de Transparencia como solicitud de acceso a la información, no solo por los motivos antes señalados, sino porque resulta materialmente imposible obtener la información que se pide en los plazos que marca la ley de transparencia para resolver las solicitudes. Hay que considerar que los documentos que se piden son en gran parte manuscritos que, además contienen información personal protegida, lo cual implica que se debe disociar la misma para poder facilitarla. El problema radica en que se trata de más de 7.000 documentos y eso impide que pueda hacerse ese trabajo en un plazo tan breve como el señalado a la ley de transparencia.

No obstante, la voluntad de la Delegación de Transparencia, Gobierno Abierto y Auditoría Ciudadana es ahondar en el conocimiento de los intereses de la ciudadanía para lograr un Gobierno Abierto. Teniendo en cuenta que, además, esta solicitud quiere utilizarse para realizar un trabajo de investigación en el marco de la Cátedra de Gobierno Abierto que tiene el Ajuntament de València con la Universidad Politécnica de Valencia, desde el Servicio de Transparencia y Gobierno Abierto se van a efectuar gestiones para encontrar formas que posibiliten un conocimiento más exacto de las sugerencias, quejas y reclamaciones, así como otras informaciones provenientes de otros canales municipales tales como: Teléfono 010, Oficinas de Atención Ciudadana, Registro, Habla con Joan Ribó, etc., contemplando tanto la gestión del conocimiento como la protección de datos personales.

Per tot allò que s'ha exposat, ES RESOL: /

Por todo lo expuesto, SE RESUELVE:

Primer. Inadmetre la sol·licitud efectuada per Marylin Mattos Barros, en tractar-se d'una petició que no s'ajusta a la finalitat de transparència de la llei.

Segon: Realitzar gestions des del Servei de Transparència i Govern Obert amb l'Oficina de Queixes, Suggeriments i Reclamacions amb l'objecte d'estudiar la viabilitat de crear algun sistema per a accedir a la informació d'aquelles sol·licituds que s'hagen fet via web. Comunicar a la sol·licitant la possibilitat de tindre accés a la informació que poguérem obtenir per aquesta via.

Tercer. Notificar la present resolució a la sol·licitant als efectes legals oportuns.

Primero. Inadmitir la solicitud efectuada por Marylin Mattos Barros, al tratarse de una petición que no se ajusta a la finalidad de transparencia de la ley.

Segundo: Realizar gestiones desde el Servicio de Transparencia y Gobierno Abierto con la Oficina de Quejas, Sugerencias y Reclamaciones con el objeto de estudiar la viabilidad de crear algún sistema para acceder a la información de aquellas solicitudes que se hayan hecho vía web. Comunicar a la solicitante la posibilidad de tener acceso a la información que pudiéramos obtener por esta vía.

Tercero. Notificar la presente resolución a la solicitante a los efectos legales oportunos.

Contra l'acte administratiu més amunt transcrit, el qual és definitiu pel que fa a la via administrativa, i d'acord amb les disposicions establertes en la Llei del Procediment Administratiu Comú i la Llei de Transparència, Accés a la informació i Bon Govern, vostè podrà interposar un dels recursos següents:

a) Amb caràcter potestatiu, Reclamació davant el Consell de Transparència, Accés a la informació i Bon Govern, dins el termini d'un mes comptador des de l'endemà de la recepció de la present notificació.

b) Recurs contenciós administratiu davant els Jutjats de la Jurisdicció Contenciós-administrativa de València dins el termini de dos mesos comptadors des de l'endemà de la recepció de la present notificació. Tot això sense perjui de poder exercitar qualsevol altre recurs o acció que estime procedent.

Contra el acto administrativo anteriormente transcrito, que es definitivo en cuanto a la vía administrativa, y de acuerdo con las disposiciones establecidas en la Ley del Procedimiento Administrativo Común y la Ley de Transparencia, acceso a la información y Buen Gobierno, usted podrá interponer los siguientes recursos:

a) Con carácter potestativo, Reclamación ante el Consejo de Transparencia, acceso a la información y Buen Gobierno, en el plazo de un mes a contar desde el día siguiente a la recepción de la presente notificación.

b) Recurso contencioso-administrativo ante los Juzgados de lo Contencioso-administrativo de Valencia dentro del plazo de dos meses contados desde el día siguiente al de la recepción de esta notificación. Todo ello sin perjuicio de que pueda ejercitar cualquier otro recurso o acción que estime procedente.

PROTECCIÓ DE DADES / *PROTECCIÓN DE DATOS*

Les dades de caràcter personal que apareixen en esta comunicació formen part d'un fitxer propietat de l'Ajuntament de València. De conformitat amb la Llei Orgànica 15/1999, de Protecció de Dades de Caràcter Personal, vostè pot exercitar els drets d'accés, rectificació, cancel·lació i oposició mitjançant instància presentada davant el Registre Gral. d'Entrada de l'Ajuntament de València.

Los datos de carácter personal que aparecen en esta comunicación forman parte de un fichero propiedad del Ayuntamiento de Valencia. De conformidad con la Ley Orgánica 15/1999, de Protección de Datos de Carácter Personal, Ud. puede ejercitar los derechos de acceso, rectificación, cancelación y oposición, mediante instancia presentada ante el Registro Gral. de Entrada del Ayuntamiento de València.

6/6

Apéndice B
Solicitud de información al Ministerio de Política Territorial y Función Pública

Carta de respuesta a la solicitud de información, para conocer la evolución mensual del tiempo medio que tardan en dar respuesta a las solicitudes, desde el 2014 hasta lo que va del 2018 que se hacen a través del portal de transparencia de la administración general del estado español.

Figura 2. Carta de respuesta a la solicitud de información al Ministerio de política territorial y función pública

MINISTERIO
DE POLÍTICA TERRITORIAL
Y FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO
DE FUNCIÓN PÚBLICA

DIRECCIÓN GENERAL
DE GOBERNANZA PÚBLICA

Expediente: 001-026568

Nombre:

NIF:

Correo electrónico:

Con fecha 19 de julio de 2018 tuvo entrada en la UIT del extinto Ministerio de Presidencia y Administraciones Territoriales solicitud de acceso a la información pública al amparo de la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, presentada por Dña. Marylin Mattos Barros, solicitud que quedó registrada con el número 001-026568:

> *"Quisiera una evolución mensual del tiempo medio que tardan en dar respuesta a las solicitudes, si es posible desde el 2014 hasta lo que va del 2018.".*

Con fecha 23 de julio de 2018 esta solicitud se recibió en la Dirección General de Gobernanza Pública, del Ministerio de Política Territorial y Función Pública, fecha a partir de la cual empieza a contar el plazo de un mes previsto para su resolución en el artículo 20.1 de la Ley 19/2013, de 9 de diciembre,.

Una vez analizada la solicitud, esta Dirección General de Gobernanza Pública **concede el acceso a la información** solicitada disponible, que se adjunta en ficheros Excel y CSV reutilizables, en los que se indica la evolución anual del plazo medio en días naturales que va desde que la solicitud presentada por un ciudadano entra en el órgano competente para su resolución hasta que se produce la notificación de dicha resolución.

La información facilitada abarca desde la puesta en marcha del Portal de la Transparencia, en diciembre de 2014, hasta el 30 de junio de 2018. No se dispone actualmente de la evolución mensual de los plazos de resolución sino únicamente de su evolución anual.

Por otra parte, es preciso señalar que en dichos plazos medios están incluidas las suspensiones o ampliaciones de plazos previstas en los siguientes artículos de la Ley

secretaria.dgpp@correo.gob.es

MARÍA DE MOLINA 50
28071 MADRID
TEL 91 273 32 46/47

19/2013: 19.2 (10 días para concreción de la solicitud), 19.3 (15 días para consulta a terceros afectados) y 20.1, párrafo 2º (un mes por ampliación de plazo por volumen o complejidad de la información solicitada). Estas interrupciones y ampliaciones hacen que el plazo de resolución de algunos expedientes supere el plazo de un mes previsto asimismo en el artículo 20.1.

Contra la presente resolución, que pone fin a la vía administrativa, podrá interponerse recurso contencioso-administrativo ante la Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia de Madrid (Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas, y Ley 29/1998, de 13 de julio, reguladora de la jurisdicción contencioso-administrativa), en el plazo de dos meses o, previa y potestativamente, reclamación ante el Consejo de Transparencia y Buen Gobierno en el plazo de un mes; en ambos casos, el plazo se contará desde el día siguiente al de la notificación de la presente resolución.

La Directora General de Gobernanza Pública
María Pía Junquera Temprano

Figura 3. CSV Adjuntado en la solicitud

Año	Nº Total de solicitudes resueltas	Plazo de resolución en días naturales
2018 (hasta 30 junio)	2398	27,3
2017	4025	31,8
2016	3167	26,9
2015	2978	27,4
2014 (diciembre)	591	43,4

Figura 4. XLSX Adjuntado en la solicitud

Año	Nº Total de solicitudes resueltas	Plazo de resolución en días naturales
2018 (hasta 30 junio)	2398	27,3
2017	4025	31,8
2016	3167	26,9
2015	2978	27,4
2014 (diciembre)	591	43,4
Nota:		
El plazo máximo de un mes para la resolución previsto en el Art. 20.1 de la Ley 19/2013		
puede verse suspendido o ampliado de acuerdo con los siguientes artículos de dicha Ley:		
Art. 19.2 Concreción de la solicitud		
Art. 19.3 Terceros afectados		
Art. 20.1 Volumen o complejidad de la información solicitada		

Apéndice C
Segunda solicitud de información al Ajuntament de València

Carta de respuesta a la solicitud de información, para conocer la evolución mensual del tiempo medio que tardan en dar respuesta a las solicitudes, desde el 2014 hasta lo que va del 2018 que se hacen a través del portal de transparencia del Ajuntament de València.

Figura 5. Segunda carta de respuesta a la solicitud de información al Ajuntament de València

Servicio de la Sociedad de la Información

Oficina de Información

Estimada Sra:

Todas las preguntas que la ciudadanía transmite a este Servicio Sociedad de la Información se gestionan durante la jornada laboral en la que se reciben o en la siguiente.

Los días laborables, durante la mañana, se contestan todas las preguntas recibidas durante la tarde y la noche anterior y las que nos llegan durante el horario laboral.

En el concepto gestionar, se incluyen las respuestas directas a las preguntas recibidas y las consultas que tenemos que hacer a algunos departamentos municipales, pero siempre informando al ciudadano que hemos recibido su pregunta y que la hemos trasladado al departamento pertinente para su contestación.

Aunque en muchas ocasiones los departamentos nos informan de las respuestas ofrecidas a los ciudadanos, no siempre tenemos información sobre los tiempos que se tardan en responder.

No disponemos de una estadística científica que determine número de preguntas y tiempos de resolución. Con la nueva web que estamos a punto de estrenar, se pretende poder ofrecer este tipo de información

Atentamente:

EL JEFE DE LA OFICINA DE INFORMACIÓN

Si usted necesita contestar a este correo o desea hacer alguna matización a nuestra respuesta, por favor pulse en el siguiente enlace
http://www.valencia.es/sugerencias

Apéndice D
Taller de clasificación manual de *tweets*

Captura de una de las hojas del archivo excel compartido durante el taller de clasificación manual de *tweets,* en el que colaboraron varios ciudadanos de la Comunitat Valenciana.

Figura 6. Captura del documento compartido durante el taller de clasificación

	A	B	C	D	E
1				**Salud**	
2				**Palabras claves:**Sanidad, Medicina, enfermedad, hospital, centro de salud, especialidades, cáncer, pediatría, traumatología, cardiología,...	
3	**Nro**	**usuario**	**Fecha**	**Texto**	
4	1	@aecc_es	27/03/2017	El jueves 30, en la Rueda de Prensa, denunciaremos la situación de desigualdad en España respecto al cribado de cáncer de colon.	
5	2	@EmyHedz	21/03/2017	Ley De Eutanasia,los que de abstuvieron o votaron en contra,no saben lo duro que es ver sufrir a alguien por culpa de una enfermedad terminal	
6	3	@EspanaRDN24	27/03/2017	Sanidad confirma un nuevo caso de Hepatitis A en Albacete, elevándose a 14 los afectados	
7	4	@redaccionmedica	27/03/2017	La sanidad española pierde un puesto en accesibilidad y es quinta de Europa	
8	5	@democraciareal	26/03/2017	Este es el verdadero objetivo de los recortes: desmantelar la sanidad pública y fomentar la contracción de seguros privados.	
9	6	@HectorTrejo	28/03/2017	Tu salud debe ser tan prioridad como lo es tu carrera y tus relaciones.	
10	7	@abc_salud	28/03/2017	Hasta dos terceras partes de los casos de cáncer son causados por mutaciones aleatorias e impredecibles	
11	8	@doctormacias	27/03/2017	Si algún médico cree que algunas vacunas son parte de una conspiración del gobierno, no le confíes tu salud.	
12	9	@cotipi	27/03/2017	Alimentos con fructosa industrial (jarabe de maíz) son aún peores que los con sacarosa (azúcar) para la salud de los niños	
13	10	@dalesmm	25/03/2017	Medio millón de mujeres abortan por año en la clandestinidad ¿cuánto más hay que discutir la necesidad del aborto legal? Es salud pública.	
14	11	@DyTEspana	24/03/2017	Hoy todos los pacientes pueden tener su donante ideal. Dr. José Luis Diez Martín, jefe del Servicio de Hematología Hospital Gregorio Marañón	
15	12	@Darias35	19/12/2016	Perder una hora de mi vida en un jodido centro de salud para conseguir una jodida receta, en pleno año 2016, es muy español.	
16	13	@biperezu	26/11/2016	Pásese por urgencias pediátricas de cualquier centro de salud español y verá lo que le diagnostican. Un virus.	
17	14	@silviamsol	19/04/2016	cualquier español debería ser atendido en cualquier centro de salud u hospital de cualquier comunidad sin problemas.	
18	15	@DanielSam74	11/04/2013	Esperando en un Centro de Salud de Gran Canaria x una receta. El sistema no esta unificado en todo el territorio español. Que vergüenza	
19	16	@CardiologiaSVC	28/03/2017	Caminar a paso ligero alarga la vida. ¿Te apuntas a una ruta saludable?	
20	17	@aprendepedia	28/03/2017	El neumococo es la causa más frecuente de infecciones invasivas en los niños, algunas son localizadas (otitis media y sinusitis)	
21	18	@Pediatria	18/03/2017	Récord de casos de paperas en Estados Unidos	
22	19	@ElUniversal	28/03/2017	La ceguera causada por glaucoma es irreversible	
23	20	@bitacoramedica	24/03/2017	Día Mundial de la lucha contra la Tuberculosis! Aprendamos cuáles son los principales síntomas de esta enfermedad	

Apéndice E
Pruebas de los algoritmos de clasificación

Salida del código fuente encargado de realizar las pruebas a los diferentes algoritmos de clasificación.

```
LinearSVC(C=0.5, class_weight=None, dual=False, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l1', random_state=None, tol=1e-05,
     verbose=0)
accuracy: 75.0
Precision: 76.0
Recall: 75.0
F1 Score: 75.0
LinearSVC(C=0.5, class_weight=None, dual=False, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l1', random_state=None, tol=1e-05,
     verbose=0)
accuracy: 73.0
Precision: 74.0
Recall: 73.0
F1 Score: 74.0
LinearSVC(C=0.5, class_weight=None, dual=False, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l1', random_state=None, tol=1e-05,
     verbose=0)
accuracy: 73.0
Precision: 74.0
Recall: 73.0
F1 Score: 73.0
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
accuracy: 70.0
Precision: 71.0
Recall: 70.0
F1 Score: 69.0
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
accuracy: 67.0
Precision: 71.0
Recall: 67.0
F1 Score: 65.0
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
accuracy: 67.0
Precision: 71.0
Recall: 67.0
F1 Score: 65.0
SGDClassifier(alpha=0.0001, average=False, class_weight=None,
epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', n_iter=100, n_jobs=1,
       penalty='l2', power_t=0.5, random_state=None, shuffle=True,
       verbose=0, warm_start=False)
accuracy: 76.0
Precision: 76.0
Recall: 76.0
F1 Score: 76.0
SGDClassifier(alpha=0.0001, average=False, class_weight=None,
```

```
epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', n_iter=100, n_jobs=1,
       penalty='l2', power_t=0.5, random_state=None, shuffle=True,
       verbose=0, warm_start=False)
accuracy: 75.0
Precision: 75.0
Recall: 75.0
F1 Score: 74.0
SGDClassifier(alpha=0.0001, average=False, class_weight=None,
epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', n_iter=100, n_jobs=1,
       penalty='l2', power_t=0.5, random_state=None, shuffle=True,
       verbose=0, warm_start=False)
accuracy: 75.0
Precision: 75.0
Recall: 75.0
F1 Score: 74.0
DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
accuracy: 70.0
Precision: 71.0
Recall: 70.0
F1 Score: 70.0
DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
accuracy: 66.0
Precision: 67.0
Recall: 66.0
F1 Score: 66.0
DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
accuracy: 66.0
Precision: 67.0
Recall: 66.0
F1 Score: 66.0
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma='auto',
kernel='linear',
  max_iter=-1, probability=False, random_state=None,
shrinking=True,
  tol=0.001, verbose=False)
accuracy: 74.0
Precision: 75.0
Recall: 74.0
```

```
F1 Score: 74.0
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma='auto',
kernel='linear',
  max_iter=-1, probability=False, random_state=None,
shrinking=True,
  tol=0.001, verbose=False)
accuracy: 75.0
Precision: 75.0
Recall: 75.0
F1 Score: 75.0
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma='auto',
kernel='linear',
  max_iter=-1, probability=False, random_state=None,
shrinking=True,
  tol=0.001, verbose=False)
accuracy: 75.0
Precision: 75.0
Recall: 75.0
F1 Score: 75.0
RandomForestClassifier(bootstrap=True, class_weight=None,
criterion='gini',
            max_depth=None, max_features='auto',
max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=25, n_jobs=1, oob_score=False,
random_state=None,
            verbose=0, warm_start=False)
accuracy: 73.0
Precision: 74.0
Recall: 73.0
F1 Score: 74.0
```


Apéndice F
Reconocimientos

El presente proyecto fue seleccionado entre las diez (10) mejores ideas a nivel nacional en la primera edición del "Desafío Aporta 2017 : El valor del dato para la administración", promovida por el Ministerio de Energía, Turismo y Agenda Digital, cuyo objetivo es impulsar nuevas ideas asociadas a prototipos que supongan mejoras en la eficiencia de la Administración, fomentar la reutilización directa de conjuntos de datos abiertos públicos en pro de la mejora del sector público, incentivando el talento, la capacidad técnica y la creatividad de los participantes.

Para la segunda fase del programa se presentó la aplicación en las instalaciones de Red.es en la ciudad de Madrid, donde expliqué el potencial de nuestro proyecto en el sentido de analizar los canales de comunicación entre la ciudadanía y el ayuntamiento con el fin de identificar sus intereses, lo que permitiría impulsar la apertura de datos desde la demanda de la ciudadanía.

El jurado estaba conformado por representantes de la Iniciativa Aporta, así como por organismos y profesionales del sector de los datos abiertos, la economía digital, la academia y las entidades colaboradoras en el evento, en las siguientes figuras se muestran algunos de los jurados y algunos momentos en los que estaba exponiendo la herramienta.

Figura 7. Jurado del Desafío Aporta 2017

Figura 8. Momento en el que se explica el problema (Desafío Aporta 2017)

Figura 9. Momento en el que se explica la solución (Desafío Aporta 2017)

Apéndice G
Investigaciones donde se aplica minería de texto a Twitter

En esta sección se muestra una tabla resumen con las publicaciones que se han realizado en los últimos 5 años, utilizando las redes sociales como fuente de información y algoritmos de Machine Learning como herramienta de análisis.

La búsqueda se realizó en la red social de investigadores, *ResearchGate,* la cual cuenta con un buscador semántico de artículos de revistas científicas. Se utilizaron como criterios de búsqueda combinaciones de las palabras, Twitter, Machine Learning, Text Mining y Minería de Texto. Como era de esperarse la mayoría de los resultados obtenidos eran publicaciones hechas en el idioma inglés o hacían análisis a textos en este idioma.

En los artículos encontrados se hizo un proceso de selección por relevancia para nuestra investigación, porque, aunque se quería conocer sobre qué temática se hacen investigaciones usando los textos de Twitter, se pretendía saber también las técnicas y los algoritmos de clasificación más utilizados.

Tabla 1. Investigaciones sobre Minería de Texto

#	Nombre	Objetivo	Conclusiones
1	Análisis Léxico sobre los Tweets de Twitter (2013) **Autores**: Astrid Paola Bográn, José Luis Alonso Berrocal, Luis Carlos García	Desarrollar una aplicación que sea capaz de realizar un análisis léxico sobre los tweets de noticias relacionadas con ciencia y tecnología para realizar un análisis tanto cuantitativos como cualitativo. **Algoritmo**: Naïve Bayes	El uso de Stemming para el preprocesamiento de los datos mejoró el desempeño del clasificador. El clasificador Bayesiano presenta un excelente rendimiento, debido principalmente a su robustez frente al ruido
2	Machine Learning Techniques applied to Twitter Spammers Detection (2014) **Autores:** Claudia Meda, Federica Bisio, Paolo Gastaldo, Rodolfo Zunino	Evaluar el rendimiento de algoritmos de Machine Learning en la detección de spam a través del análisis del comportamiento del usuario y los contenidos de los tweets que comparte. **Algoritmo:** SVM, ELM, RFs	El algoritmo Random Forest es más efectivo en comparación de la Support Vector Machine y Extreme Learning Machines
3	Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles (2016) **Autores:** Jaime Guzmán, Carlos Henriquez Miranda, Dixon Salcedo	Construir un sistema de minería de opiniones en español sobre comentarios dados por clientes de diferentes hoteles. El sistema trabaja bajo el enfoque léxico utilizando la adaptación al español de las normas afectivas para las palabras, en inglés (ANEW). **Algoritmos** de Machine Learning (no especifica cual)	Se ha logrado un sistema más que aceptable alcanzando una precisión del 94% para el dominio de turismo, específicamente hoteles. La calidad del recurso ANEW permitirá realizar nuevos experimentos de análisis de sentimientos enfocados a diferentes dominios empleando la metodología propuesta.

4	A Data Mining System for Sentiment Classification of Indian Currency Demonetization using Naive Bayes Classifier Algorithm (2017) **Autor:** Athul Jayaram	Comprender el sentimiento real de la población de la India durante el período de demonización **Algoritmo**: Naive Bayes	Los ciudadanos de la India están teniendo un sentimiento negativo hacia la desmonetización de la moneda india.
5	Music interest classification of twitter users using support vector machine (2017) **Autores:** Yusra, Muhammad Fikry, Bambang Riyanto Trilaksono, Rado Yendra, Ahmad Fudholi	Clasificar el interés musical de los usuarios de Twitter en idioma indonesio en tres géneros musical (jazz, pop o rock) y Analizar sus sentimientos (positivo, negativo o neutral) **Algoritmo**: Support Vector Machine (SVM)	La precisión para la categoría de música es más alta que la del sentimiento. El número de tweets neutrales es más alto que los tweets positivos y negativos, los tweets neutros para Rock son más altos que los de Jazz y Pop Esta investigación puede servir para hacer publicidad en redes sociales de acuerdo al género musical.

Fuente: Elaboración propia