UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Análisis y Estudio de la Participación Ciudadana en la Plataforma DecideMadrid

Máster Universitario en Investigación e Innovación en Inteligencia Computacional y Sistemas Interactivos

Autor: BACHILLER RUBIA, Sergio

Tutor: QUIJANO SÁNCHEZ, Lara
Departamento de Ingeniería Informática

Junio 2020

Dedicatoria

"Este trabajo se lo dedico a mis padres Gloria y Damián, por ser el arquetipo de mi vida y enseñarme que todo lo que uno se propone, se logra con esfuerzo.

A los que han estado cerca, por ser un gran apoyo emocional y ser mi familia en Madrid.

Y como no, a los que paran el tiempo cuando no estamos juntos."

Sergio Bachiller Rubia

Agradecimientos

A mi familia y amigos por ser el motor para ser cada día mejor y motivarme a superarme académicamente.

A Lara por confiar en mí desde el primer momento y guiarme en este trabajo que resultará fuente de conocimiento para futuras generaciones para los fines que convenga.

A los profesores de mi alma mater, la Universidad Autónoma de Barcelona, por ser fuente de inspiración y vocación.

Índice general

Agradecimientos II

Lista de figuras V

Lista de tablas VI

1 Introducción **3**
1.1 Objetivos y tareas 4
1.2 Estructura del documento 5

2 Estado del Arte **6**
2.1 *E-Participation* 6
2.1.1 Presupuestos Participativos Online (OPB) 8
2.1.2 Compromiso Cívico 9
2.2 Caso de Estudio: DecideMadrid 10
2.3 Text Similarity 11
2.4 *Clustering* 12
2.5 Herramientas de visualizacion 13

3 Desarrollo de la propuesta **14**
3.1 Datos 14
3.2 Preprocesado de texto 15
3.3 Similitud entre documentos 16
3.3.1 Índice Jaccard 17
3.3.2 Similitud por coseno 18
3.3.3 *Word Mover's Distance* 19
3.3.4 Comparación de los métodos 20
3.4 Agrupación de documentos 22

4 Análisis **26**
4.1 Factores determinantes del éxito 26
4.2 Influencia y evolución por distrito 31

5 Aplicaciones **39**
5.1 Modelo de predicción de apoyos 39
5.2 Detección de duplicidad en nuevas propuestas 40
5.3 Herramienta de visualizacion 41

Referencias **51**

Anexos **52**

Índice de figuras

Figura 2.2.1 Diagrama del proceso de una propuesta en DecideMadrid 11

Figura 3.1.1 Diagrama E-R de los datos utilizados 14
Figura 3.3.1 *Intersection over Union for object detection* [1] 17
Figura 3.3.2 *A vector space of seven words in three contexts* [2] 18
Figura 3.3.3 *Ejemplo de uso de WMD* [3] 20
Figura 3.3.4 Histograma de las similitudes encontradas con WMD 21
Figura 3.4.1 Estructura de las comunidades más grandes 23

Figura 4.1.1 Relación del número de apoyos y la longitud de documento por comunidades 27
Figura 4.1.2 Relación del número de apoyos y la cantidad de comentarios por comunidades 28
Figura 4.1.3 Relación de métricas de controversia con el número de apoyos . . 29
Figura 4.1.4 Relación del día de publicación con el número de apoyos 29
Figura 4.1.5 Relación de apoyos con el número de categorías y distritos 30
Figura 4.2.1 Número de propuestas por distrito 31
Figura 4.2.2 Relación distrito con las categorías Medio Ambiente, Animales y Movilidad 33
Figura 4.2.3 Relación distrito con las categorías Derechos Sociales, Sostenibilidad y Asociaciones 33
Figura 4.2.4 Relación distrito con las categorías Economía, Transparencia y Seguridad y Emergencias 33
Figura 4.2.5 Relación distrito con las categorías Equidad e integración, Educación y Tercera Edad 34
Figura 4.2.6 Relación distrito con las categorías Empleo, Justicia y Delincuencia 34
Figura 4.2.7 Relación distrito con las categorías Deportes, Cultura y Ocio y Entretenimiento 34
Figura 4.2.8 Relación distrito con las categorías Política, Ayto. y Admon. Pbica. y Jóvenes 35
Figura 4.2.9 Relación distrito con las categorías Accesibilidad, Civismo, Vivienda y Participación Ciudadana. 35
Figura 4.2.10 Influencia de propuestas por categoría y comunidad 36
Figura 4.2.11 Evolución de propuestas por categoría y comunidad 37
Figura 4.2.12 Evolución de propuestas por categoría y comunidad 38

Figura 5.1.1 Comparación de las predicciones y valores reales 40
Figura 5.3.1 Participación y su evolución por distrito 42
Figura 5.3.2 Evolución de la participación, por categoría, comunidad y distrito 43
Figura 5.3.3 Mapa de calor por impacto de comunidades y categoría 44

Índice de tablas

Tabla 3.2.1 Ejemplo de preprocesado de texto 16
Tabla 3.3.1 Comparación entre los métodos de cálculo de similitud 20
Tabla 3.3.2 Desglose del histograma de similitudes WMD 21
Tabla 3.3.3 Ejemplos de propuesta por tramo de similitud 22
Tabla 3.4.1 Comunidades formadas por el método de Louvain 23
Tabla 3.4.2 Descripción de las 10 comunidades más grandes, de mayor a menor 24
Tabla 3.4.3 Similitud media en cada comunidad 25

Tabla 4.2.1 Distrito con mayor influencia de cada comunidad 32

Abstract

Citizen collaboration through digital participation can be very complex, requiring mechanisms to act in a coordinated manner in order to avoid highly demanded proposals being diluted by acting individually. In order to avoid this casuistry, throughout this work methods are proposed in order to reunify these proposals, which have a single common objective. Thanks to natural language processing techniques and graph theory, among other fields, communities of interest are extracted from a set of documents, which represent a more concrete level than the category of an urban project.

From these communities, common factors can be drawn which determine the success of a set of proposals based, for example, on the length of the document. In addition, this new level of grouping, segregates the proposals in a more exhaustive way, which opens a way to study the needs in each of the areas of the city. Such a study would not only detect inequalities in the different districts of the same city, but would also inform both the population and the municipal government of what is being demanded through visualisation tools, thus contributing to transparency and the digital transition.

Finally, it is proposed, among others, a tool that aims to avoid future duplication of proposals (based on the suggestion of other existing proposals), in an attempt to achieve consensus and sharing among the population.

Keywords: citizen participation, civic technology, e-government, text similarity, communities of interest, visualization

Resumen

La colaboración ciudadana a través de la participación digital puede llegar a ser muy compleja, necesitando mecanismos para actuar de forma coordinada con tal de evitar que propuestas muy demandadas se vean diluidas al actuar de forma individual. Para evitar esta casuística, a lo largo de este trabajo se proponen métodos con tal de reunificar estas propuestas, que tienen un único objetivo en común. Gracias a técnicas de procesado de lenguaje natural y teoría de grafos, entre otros ámbitos, se extraen comunidades de interés en un conjunto de documentos, que representan un nivel más concreto que la categoría de un proyecto urbano.

De estas comunidades se pueden extraer factores en común, que determinen el éxito de un conjunto de propuestas basado, por ejemplo, en la longitud del documento. Además, este nuevo nivel de agrupación, segrega las propuestas de una forma más exhaustiva, lo que abre una vía de estudio de las necesidades en cada una de las zonas de la ciudad. Un estudio así permitiría, además de detectar desigualdades en los diferentes distritos de una misma ciudad, poder informar tanto a la población como al gobierno municipal de lo que se demanda a través de herramientas de visualización, contribuyendo así a la transparencia y a la transición digital.

Finalmente, se propone, entre otras, una herramienta que tenga por objetivo evitar futuras duplicidades en propuestas (basado en la sugerencia de otras propuestas ya existentes), en un intento de consenso y puesta en común entre la población.

Palabras clave: participación ciudadana, tecnología civica, gobierno electrónico, similitud de texto, comunidades de interés, visualización

Capítulo 1

Introducción

En los últimos años, muchas ciudades de diferentes países han implementado plataformas online de participación ciudadana, en las que los habitantes son partícipes de las decisiones y acciones municipales a través de propuestas y debates [4]. Hasta el momento, estas plataformas han sufrido problemas como la frustración de los ciudadanos porque sus propuestas no han tenido impacto o una baja participación a causa del los excesivos trámites a la hora de identificarse como ciudadano [5].

Diferentes estudios han señalado que el nivel de participación ciudadana en plataformas electrónicas aún es bajo y señalan como posible problema el desarrollo de muchos proyectos de participación electrónica enfocados en ofrecer soluciones tecnológicas que no modelan las necesidades de los ciudadanos. Es por ello que este proyecto plantea la mejora de las plataformas de participación ciudadana a través del estudio y desarrollo de herramientas que permitan modelar mejor los intereses de la ciudadanía, visualizar y facilitar la comprensión de los problemas y las soluciones propuestas (tanto a los ciudadanos como al gobierno local) desde diferentes plataformas.

Siguiendo el modelo europeo, el Ayuntamiento de Madrid lanza DecideMadrid en 2015, convirtiéndose así en una de las primeras ciudades españolas en implementar este tipo de plataformas, la cual será de referencia en este trabajo. Durante todo el año DecideMadrid permite crear nuevas propuestas, que necesitarán del apoyo del 1 % de la población empadronada en la ciudad mayor de 16 años (27.662 en junio de 2020). Una vez conseguidos los apoyos necesarios, las propuestas son sometidas a votación ciudadana, y las propuestas que consiguen la mayoría simple de los votos son asumidas por el Ayuntamiento de Madrid como propias, con tal de llevarlas a cabo. Paralelamente, durante un período de tiempo al año, se abre la posibilidad de proponer proyectos en presupuestos participativos. El Ayuntamiento destina una parte del presupuesto municipal en esta partida, en la que los ciudadanos son los responsables de proponer y votar los proyectos con tal de hacerlos reales. A pesar de que hay ciertas acciones que el Ayuntamiento no puede llevar a cabo, porque son competencia de otro organismo como puede ser la Comunidad Autónoma de Madrid, hay muchas que son viables, como por ejemplo, la construcción de un centro de día, subvenciones que actualmente no existen o reposición de elementos de transporte.

DecideMadrid y otras plataformas como Decidim.Barcelona son de gran ayuda y facilitan la participación ciudadana, aunque se observan casos en los que no se consigue sacar el máximo potencial y la participación no es tan alta como cabría esperar, como es el caso de Madrid, en el que ninguno de los años que lleva activo ha llegado a participar más del $0,03\,\%$ de la población de la ciudad [6]. La gran cantidad de datos publicados, con contenido muchas veces similar, hace difícil navegar, comprender el contenido de estas plataformas y tomar decisiones acorde a ello. Este hecho provoca que muchos ciudadanos se rindan, decidan no participar y que muchas de las preocupaciones que ha manifestado la ciudadanía sigan sin respuesta, provocando el descontento entre aquellos que participan al no ver el triunfo de sus propuestas, tal y como indicaba [5].

Para paliar este efecto se pretende estudiar y extraer información como los intereses de cada distrito, conocer qué asuntos son preocupación constante de los ciudadanos o incluso la existencia de factores determinantes para el éxito de las propuestas, como puede ser el momento de publicación. De este modo, sería posible informar tanto a los ciudadanos que utilizan la plataforma como al gobierno local, de forma más concreta y definida, de lo que los habitantes están pidiendo, facilitando el trabajo de los gobernantes y mejorando la usabilidad de la plataforma; posiblemente derivando en una mejora en la calidad de vida de la población.

Además de los objetivos sociales y políticos que persigue este trabajo, se busca también colaborar en la transición digital hacia una ciudad inteligente, que tenga como objetivo abordar problemas urbanos y aumentar la calidad, eficiencia e interactividad de los servicios urbanos gracias a las TIC. Una transición a este modelo de ciudad satisfaría el reclamo de transparencia y buena gestión en la administración pública, contentando a los ciudadanos de forma indirecta.

1.1. Objetivos y tareas

En este trabajo se siguen cuatro ejes principales, a saber:

1. Estudio y detección de propuestas similares y/o duplicadas
2. Agrupar y estudiar dicha propuestas similares formando comunidades de interés
3. Crear una herramienta de visualización que permita tanto dibujar y visualizar las comunidades de forma concreta, sencilla e interactiva, como el total de las propuestas mediante una serie de filtros temporales, espaciales y por temática.
4. Facilitar la coproducción de propuestas similares, a través de la recuperación de propuestas similares en el momento de publicación

Los objetivos se pueden englobar en dos grupos; la solución del problema actual y la prevención a largo plazo. Para solventar el problema actual, se analiza el contenido de las propuestas o documentos, con tal de poder establecer una medida de similitud y de ese modo agrupar aquellas propuestas duplicadas y aprovechar esos grupos para conocer el impacto que tiene esa propuesta realmente. En estas agrupaciones se podrán comprobar factores que determinen el éxito de una propuesta frente a otras idénticas en contenido y de los que es posible extraer información que utilizar en beneficio de los participantes, recomendando las características que deben tener sus propuestas para triunfar.
Se analizará también la influencia de estos grupos de propuestas en los diferentes distritos de la ciudad de Madrid y la evolución de los mismos. Gracias a esto, se puede realizar un estudio social en profundidad, en el que usar variables socio demográficas para conocer y entender el motivo de las propuestas. Tras este análisis será posible detectar necesidades que se demandan asiduamente en un mismo distrito y si ha sido algo puntual o es una necesidad constante. Para acabar con la solución del problema actual, se buscará desarrollar un panel de datos, que permitirá utilizar filtros como la fecha, localización o categoría con tal de ver las propuestas que se hacen de forma clara e interactiva. Aunque esta herramienta tenga origen como sistema de soporte de decisiones para el gobierno municipal, hacerla pública contribuiría a la transparencia, haciendo partícipes a los ciudadanos de lo que ocurre en la ciudad.

De otro lado, como método de prevención a la duplicidad de documentos, se aprovecharán las técnicas utilizadas en el análisis para encontrar documentos similares a la hora de crear uno nuevo. Con esta funcionalidad, los usuarios sabrán si lo que están proponiendo ya ha sido propuesto antes o no, evitando así que se creen nuevas propuestas duplicadas.

Para el desarrollo de una solución, se utilizan los datos que ofrece el ayuntamiento a través de su portal de datos abiertos, que son cargados a una base de datos de *Amazon Web Services* con tal de facilitar el acceso desde código. Con tal de conseguir una solución apropiada al problema que se enfrenta, se utilizan técnicas de procesado de lenguaje natural, análisis y gestión de un gran volumen de datos y teoría de grafos.

Otros detalles sobre la metodología y planificación de tareas se muestran en el Anexo I.

1.2. Estructura del documento

En relación con la estructura del documento, en Capítulo 2 se presenta el estado del arte, en el que se revisan algunos conceptos y técnicas útiles en el trabajo. En el Capítulo 3 se muestra el desarrollo del trabajo, en el que se seleccionan las técnicas y se explica de forma detallada los pasos que se han tomado. Seguidamente, en el Capítulo 4 se evalúan los resultados que se han conseguido, con tal de extraer información útil que aportar a los interesados en el trabajo. A continuación, en el Capítulo 5 se comentan las herramientas y aplicaciones que han derivado de los resultados encontrados en el anterior capítulo. Finalmente, en el Capítulo 6, se extraen las conclusiones del estudio que motivan diferentes líneas de trabajo futuras.

Capítulo 2

Estado del Arte

En este capítulo primeramente se revisa la literatura referente al dominio y caso de estudio de este proyecto, las plataformas de participación ciudadana y Decide Madrid. Dicha revisión sirve para revisar necesidades existentes, motivar y señalar las aportaciones realizadas en este área. Seguidamente se incluyen revisiones de diferentes técnicas existentes de similitud de textos, clustering y visualización ya que representan la parte teórica de este proyecto.

2.1. *E-Participation*

El objetivo de la gobernanza participativa es crear una democracia funcional en la que los miembros de la comunidad influyan en las acciones de los funcionarios del gobierno local [7].

Cuando los ciudadanos se sienten partícipes y ven fruto en su esfuerzo participan más. La baja participación limita la cantidad de la contribución de los ciudadanos, y socava la calidad [8]. Los estudios han señalado que el nivel de participación ciudadana en las plataformas electrónicas es todavía bajo [9]. El problema puede haber surgido debido a un enfoque equivocado de muchos proyectos de participación electrónica, lo que significa que su principal preocupación es ofrecer soluciones tecnológicas en lugar de comprender las necesidades de los ciudadanos [10].

Comprender por qué los ciudadanos no están dispuestos a involucrarse en los asuntos de gobierno, e investigar la aceptación y la intención de los ciudadanos de participar es un paso esencial para analizar los niveles reales de participación de los ciudadanos. Hay una tendencia a creer que los ciudadanos se involucrarán sin la debida consideración de sus preferencias, necesidades y expectativas [11; 12]. Esto se debe probablemente a una importante noción engañosa de que los ciudadanos participarán y/o querrán participar inmediatamente cuando se les den las herramientas de participación electrónica. [12; 13]. Hay algunas pruebas de que la disponibilidad de complejas plataformas de participación electrónica, que exige a los ciudadanos grandes conocimientos técnicos, reducen considerablemente la capacidad y la voluntad de participar de los ciudadanos [10; 14].

En [15] el autor hace un llamamiento a la creación colaborativa de ciudades e insta a los diseñadores a que sitúen las perspectivas críticas en la vanguardia de su práctica y traten de desarrollar plataformas de participación ciudadana más colaborativas y accesibles, siguiendo las indicaciones de [5].

A fin de apoyar las formas plurales de participación, es necesario reconocer los diferentes tipos de compromiso que se producen en los entornos formales e informales. En parte, esto significa reconocer las numerosas motivaciones e impedimentos para participar en la vida cívica: desde el nivel de educación, pasando por los ingresos, hasta las respuestas afectivas y emocionales a cuestiones concretas [16]. La literatura sobre la participación comunitaria señala que las personas no suelen participar en procesos cívicos por obligación o altruismo, sino como resultado de la ira ante la injusticia percibida [16]. Cuando la gente está enfadada, está más dispuesta a asumir el conflicto para hacer frente a los desequilibrios en cuestiones de recursos asignación. Lo que esto sugiere es que la persuasión es, de hecho, el enemigo de la participación porque *un público persuadido no está organizado y comprometido es más pasivo que activo*. [17]. La participación de la comunidad, por lo tanto, existe dentro de un espacio de tensión entre las comunidades y los administradores públicos. Cuando las comunidades se involucran verdaderamente en los procesos cívicos, lo hacen a través de la política de conflicto [18].

Los investigadores han utilizado varias teorías y modelos para explicar y predecir la aceptación y adopción de nuevas tecnologías y sistemas por parte de los usuarios, en particular en el ámbito de la administración electrónica [19].

[20] informan que el diálogo de colaboración entre el gobierno y los ciudadanos puede aumentar la satisfacción de los ciudadanos.

[21] examina la relación entre la participación electrónica y la confianza en el gobierno local centrándose en cinco dimensiones: 1) satisfacción con las aplicaciones de la participación electrónica, 2) satisfacción con la respuesta del gobierno a los participantes electrónicos, 3) desarrollo de los participantes electrónicos mediante la participación, 4) influencia percibida en la toma de decisiones y 5) evaluación de la transparencia del gobierno. Las conclusiones revelan que la satisfacción de los participantes electrónicos con la capacidad de respuesta del gobierno está positivamente asociada a su percepción de influir en la toma de decisiones del gobierno. Existe una asociación positiva entre la percepción de los participantes electrónicos de influir en la toma de decisiones del gobierno y su evaluación de la transparencia del gobierno.

Otros investigadores abogan por un giro hacia la apertura en el diseño participativo [22] se esfuerzan cada vez más por promover el empoderamiento demostrando y entregando a las personas conjuntos de instrumentos, tecnologías y conocimientos técnicos para que los apropien, reutilicen y adopten para sus propios fines situados [23]. Las últimas revisiones de la literatura sobre participación electrónica sugieren un cambio continuo de la investigación desde un enfoque más puramente tecnológico a una visión más holística, en la que se podrían integrar otras cuestiones sociales y tecnológicas para investigar la participación de los ciudadanos [24; 25].

Estas evidencias e investigaciones sustenan la hipótesis de este trabajo en desarrollar una herramienta ágil y ligera que facilite la visualización de contenidos de plataformas de participación ciudadana, facilitando el estudio y comprensión tanto por parte de la ciudadanía como de gobiernos locales de las aportaciones realizadas a la vez que apoyando la coproducción ciudadana por medio de una herramienta que detecte similitudes y aune propuestas.

2.1.1. Presupuestos Participativos Online (OPB)

- La presupuestación participativa (PB por sus siglas en inglés) representa uno de los mecanismos más populares para involucrar a los ciudadanos en la toma de decisiones
- En PB, los ciudadanos participan en los procesos para asignar la liquidez de los presupuestos municipales o públicos en iniciativas y proyectos en diferentes ámbitos, como seguridad pública, educación, salud, transporte, etc.
- Estudios preliminares PB se han centrado principalmente en explicar los papeles de los gobiernos y los ciudadanos, la frecuencia y la tipología de la participación, y los logros y resultados obtenidos [26]
- Sin embargo, dada la relevancia y la inversión en PB, también es clave entender la forma en que funcionan estas plataformas y la naturaleza de la participación ciudadana que surge de ellas

La implementación de presupuestos participativos varía ampliamente en cuanto a la forma en que los ciudadanos pueden expresar sus ideas y propuestas y en cuanto a la forma en que estas pueden integrarse y considerarse en el presupuesto [27; 28]. Las siguientes fases pueden describirse como comunes durante un proceso de presupuestación participativa en línea [29]:

- **Información**. Los ciudadanos son informados a través de diferentes canales sobre el presupuesto, la cantidad y el procedimiento de la presupuestación
- **Participación**. La participación de los ciudadanos en la elaboración de presupuestos con el objetivo de elaborar y calificar los proyectos
- **Toma de decisiones**. El panel (normalmente un consejo municipal o de ciudad) debate sobre los proyectos y su viabilidad, tomando la decisión final

- **Rendición de cuentas**. La rendición de cuentas por las decisiones presupuestarias y su aplicación debe ser dada por el consejo municipal o de la ciudad a fin de asegurar la plausibilidad y la aceptación de los presupuestos participativos

A lo largo de estos procedimientos, las diferentes herramientas de presupuestos participativos pueden variar en:

- Canales disponibles (por ejemplo, carta, teléfono a través de un centro de llamadas, asambleas, sitio web, aplicaciones para teléfonos inteligentes).
- Herramientas utilizadas (es decir, sistema de gestión de contenidos que utiliza diferentes aplicaciones como encuestas, foros, sondeos, técnicas de visualización, etc.)
- Información personal solicitada a los participantes (solo un nombre de usuario y una contraseña, un nombre completo con datos de dirección o incluso la identificación con el padrón)
- Mecanismos de toma de decisiones (con impacto directo sobre los participantes).

Como se ha descrito anteriormente este proyecto pretende mejorar las etapas de participación y toma de decisiones en plataformas que facilitan la elaboración de presupuestos participativos. En concreto utilizando como caso de estudio los datos proporciandos por la plataforma Decide Madrid.

2.1.2. Compromiso Cívico

Los gobiernos locales son organizaciones en constante evolución. Se someten periódicamente a elecciones para cambiar de dirigentes, se renuevan los contratos de servicios para las operaciones básicas, desde los servicios informáticos hasta la recogida de basura, pasando por la recogida y el cumplimiento de las tasas de los aparcamientos públicos. Cuando ni los planes de crecimiento a largo plazo ni las operaciones diarias de los servicios críticos de la ciudad son estables, la necesidad de un compromiso comunitario robusto, accesible y significativo es crucial. El desafío aquí es crear procesos y prácticas que permanezcan más allá de este cambio organizativo para que los residentes puedan participar con tal de ayudar a dar forma a las políticas que afectan a sus barrios y vidas [30]. Es por ello que la herramienta aquí propuesta puede facilitar estas labores mediante la visualización de las propuestas de forma interactiva, por ejemplo detectando estacionalidades y tendencias en las propuestas ciudadanas que sirvan a los diferentes gobiernos a lo largo del tiempo para comprender demandas constantes no atendidas o problemas concretos estacionales.

La participación de los ciudadanos es un objetivo importante para los gobiernos, los científicos y las empresas, ya que su participación o la falta de ella puede tener un gran impacto en las cuestiones de interés común. Esto se aplica en particular a las tecnologías cívicas como las *participatory sensing* (PS). Es necesario comprender las motivaciones subyacentes de las personas para unirse, participar y abandonar la PS. Los mecanismos de incentivo actuales en la PS se centran principalmente en las recompensas; solo se ocupan de valores como el poder, el logro, la seguridad y el estímulo, dejando un vacío en los mecanismos de incentivo que afectan a otros valores [31].

Es por ello que una herramienta como la propuesta en este trabajo que permita tanto aunar objetivos agrupando propuestas similares como realizar un seguimiento visual de aquellas propuestas que han sido aceptadas o que siguen sin tener respuesta pueda servir de estímulo para la ciudadanía mejorando la calidad (propuestas más informadas) y cantidad (más propuestas debido a la facilidad de uso) de participaciones.

2.2. Caso de Estudio: DecideMadrid

DecideMadrid (`https://decide.madrid.es/`) está activo desde septiembre de 2015, y es en esta plataforma donde el Ayuntamiento dirige los presupuestos participativos anuales, que posteriormente serán asignados a diferentes proyectos. Estas propuestas y proyectos tienen como autores a más de 420,000 usuarios que se han registrado y verificado su identidad, que además han participado en debates dentro de la misma plataforma. Toda la información sobre los proyectos, comentarios y debates generados dentro de la plataforma han sido publicados como datos abiertos, con tal de que todo el mundo pueda disponer de ellos.

Procesos en DecideMadrid

Gracias a esta herramienta, la población puede crear propuestas de proyectos e iniciativas para la ciudad sobre diferentes categorías, como pueden ser urbanismo, transporte público o salud. Estas propuestas pueden ser debatidas en comentarios y votadas (esto último solo por personas que estén empadronadas en Madrid y mayores de 16 años) dentro de la misma plataforma. A pesar de que los debates que surgen no son determinantes para el Ayuntamiento, hay que tener en cuenta que sigue representando una parte de la opinión pública, por lo tanto se deben tener en cuenta. Una vez las propuestas consiguen apoyos suficientes (el 1 % de la población mayor de 16 años empadronada en la ciudad), se someten a votación y debate popular durante 45 días, donde si consiguen mayoría simple, serán asumidas por el Ayuntamiento con tal de estudiar su viabilidad y llevarla a cabo en un tiempo de cuatro meses... En cambio, si después de 30 días no se consiguen los apoyos suficientes, la propuesta se descarta. Con tal de poder implementar los proyectos propuestos por los ciudadanos, se cuenta con una partida presupuestaria que en 2019 fue de 100 millones de euros. Tras estudiar la viabilidad y la acción que han reclamado los ciudadanos, el Ayuntamiento de Madrid hará efectivo y público el gasto de la partida que ha supuesto si es que ha sido realizada, o dará los motivos por los cuales no se ha estimado viable.

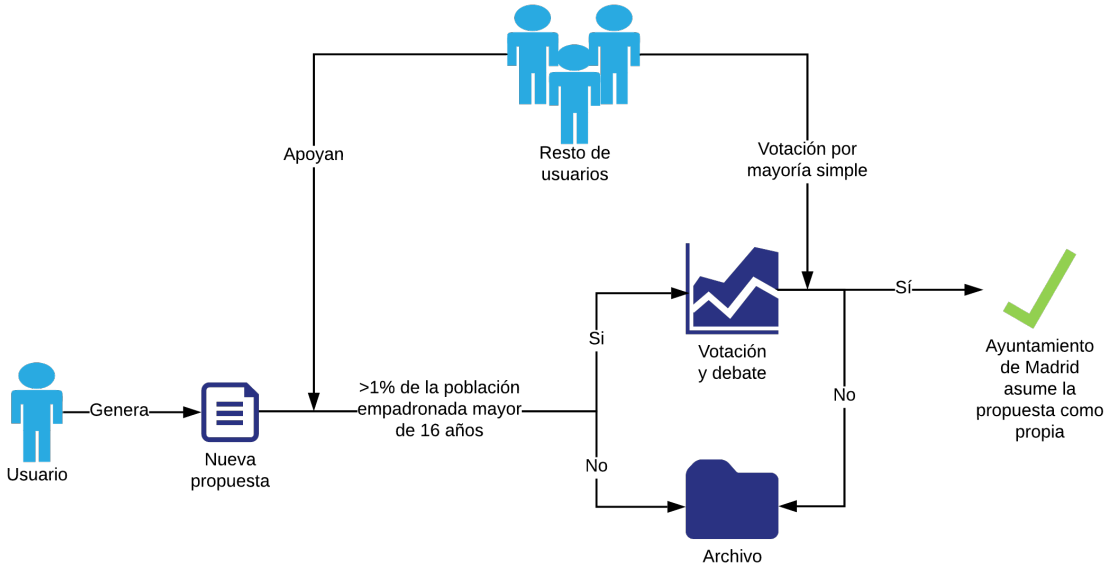


Figura 2.2.1: Diagrama del proceso de una propuesta en DecideMadrid

2.3. Text Similarity

El estudio de similitudes entre textos es un área en auge con numerosas aportaciones debido en parte a la revolución del Big Data donde numerosos textos son accesibles para su análisis en múltiples dominios [32; 33]. En [34] encontramos un trabajo bien indexado que poder usar de referencia, ya que sumariza las principales tendencias y ejemplifica limitaciones y aciertos de los métodos más populares de similitud de textos. En primer lugar, muestra como el índice Jaccard no tiene unos resultados apropiados para este problema, al no tolerar elementos similares (del mismo campo semántico). A continuación muestra diferentes técnicas en las que usa *embeddings* (o vectores que representan palabras, vector palabra), entre los que comenta que K-Means es muy sensible al número de *features*, además de tener que saber cual va a ser el número de agrupaciones. A continuación explica como el cálculo de la similitud por coseno se puede mejorar dependiendo de como se hayan conseguido los vectores palabras, consiguiendo unos resultados muy aceptables al no afectarle la longitud de los documentos, sino el ángulo.

También existen técnicas como *Latent Semantic Indexing* (LSI), en la que se busca reducir el tamaño de los *embeddings*, asumiendo que hay una espacio de menores dimensiones en el que representar todas las palabras que tiene un texto, lo que provoca una pérdida de información. Otros problemas relativos a otros campos de estudio también se han adaptado para tratar este problema, como es *Earth Mover Distance* adaptado a *Word Mover's Distance*, que pretende encontrar la mínima distancia entre dos documentos dentro de un espacio vectorial. Gracias a WMD se consigue establecer una similitud alta entre dos frases sin palabras en común, como pueden ser *'Obama speaks to the media in Illinois'* y *'The president greets the press in Chicago'*, debido a que consigue capturar la semántica y la sintaxis de los textos.

En este línea, otros enfoques han conseguido construir modelos en los que trabajar directamente con *sentence embeddings*, lo que, en términos de sintáctica, supone un avance importante [35]. Algo similar ocurre con *Bidirectional Encoder Representations from Transformers* (BERT), capaz de obtener *embeddings* que representen la posición y que rodea una palabra. Además, BERT, es el pilar fundamental de herramientas de generación de texto como GPT-2.

Métodos alternativos, como *Siamese Manhattan LSTM* (MaLSTM), calculan el nivel de similitud entre textos, gracias al entrenamiento con un conjunto de datos anteriormente etiquetados. Finalmente, en los últimos años se han empezado a construir redes semánticas, como *WordNet*, una red organizada por *sinónimos cognitivos*, que pueden utilizarse para poner en contexto un documento y estimar la similitud hacia otro concepto.
Con el objetivo de utilizar un método que mejor afine en capturar no sólo la similitud textual de textos sino su semántica en este trabajo como veremos en el siguiente capítulo se ha utilizado el WMD, ya que destaca sobre los métodos más simples y a la vez no precisa de un conjunto pre-etiquetado para ejecutarse por lo que facilita su reusabilidad en diferentes dominios y lenguajes.

2.4. *Clustering*

Existiendo también númerosos algoritmos y técnicas de clustering, en esta sección destacamos algunos de ellos, en especial los que proponen un enfoque similar al finalmente adoptado en el proyecto. En [36] clasifica los diferentes métodos de agrupación para después centrarse en *K-Means*. Este algoritmo es uno de los más sencillos y rápidos, aunque es muy aleatorio en su inicialización y necesita de establecer el número de clústers.

[37] explica como agrupar elementos basados en similitud gracias al agrupamiento jerárquico aglomerativo. Este método crea un árbol jerárquico, en el que cortar en diferentes niveles para encontrar grupos.

[38; 39; 40] proponen un nuevo enfoque, en el que afrontar el problema como un grafo (aunque no especifican como lo construyen), con tal de maximizar la modularidad, encontrando comunidades de interés en otros dominios diferentes al de nuestro trabajo, como puede ser en Twitter.

2.5. Herramientas de visualizacion

En cuanto a herramientas de visualización los trabajos relacionados se centran en analizar las características desables en nuestro dominio de interés más que en ofrecer herramientas en sí. Por ejemplo, en [41] se comenta como, dado el gran número de datos de gobierno, es necesario de herramientas de visualización con tal de que los usuarios puedan extraer el máximo de información. O en[42] explica que la visualización ayuda a conocer la información, lo que mejora la capacidad organizativa y la explotación de la misma.

Según [43], algunas formas de visualización pueden ser demasiado sofisticadas, sobre todo las relacionadas con datos multivariados. Aun así, destaca que la visualización de datos mejora la interpretabilidad de los mismos, de modo que está pendiente en el estado del arte el desarrollo de herramientas más simples que faciliten esto.

Capítulo 3

Desarrollo de la propuesta

En este capítulo se describen los datos, métodos y técnicas utilizados en este proyecto así como su evolución con tal de ajustarlo al problema que abordamos.

3.1. Datos

El Ayuntamiento de Madrid, a través de su portal de datos abiertos, proporciona muchos datos relativos a la ciudad. Entre estos datos encontramos las propuestas creadas por 24.482 usuarios de DecideMadrid hasta el 11 de septiembre de 2019, que resultan en un total de 21.746 propuestas, que tienen asignadas una o más categorías como pueden ser Urbanismo, Animales, Movilidad o Seguridad y Emergencias. Otros datos, como la tabla Usuario, ha sido agregada a posteriori gracias al trabajo de [4] usando técnicas de *web scrapping*. Todos estos datos han sido cargados a una base de datos en *Amazon Web Services* con la estructura que se indica en la Figura 3.1.1.

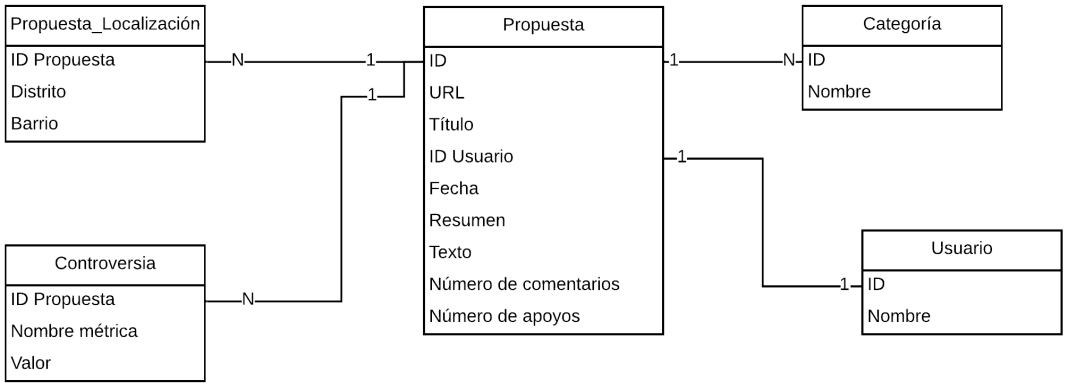


Figura 3.1.1: Diagrama E-R de los datos utilizados

Destacar que la tabla *Controversia* son una serie de medidas de controversia que hacen referencia a las de [4], proporcionado por los autores y utilizado más adelante en el estudio con tal de encontrar una correlación con el éxito de las propuestas y cómo fuente de estudio a la hora de visualizar las propuestas filtradas por este factor y por ejemplo el volumen de comentarios.

3.2. Preprocesado de texto

Con tal de garantizar que el cálculo de similitud entre dos textos sea preciso, es necesario tratar el texto adecuadamente. Realizaramos esta tarea utilizando herramientas comunes en un pipeline de procesamiento de lenguaje natural [44]. Especialmente, en nuestro caso, sabemos que los datos de los que disponemos son brutos, es decir, tal y como los usuarios lo han subido a la plataforma de DecideMadrid. Este detalle es especialmente importante, ya que las herramientas que se utilizan en este ámbito asumen que las palabras no van a tener errores ortográficos.

Así pues, el primer paso del pre procesado es la corrección de faltas en los textos. Con tal de evitar problemas aquí, se eliminan caracteres especiales, concretamente $'/\backslash., ()-$, eliminando así posibles problemas en la corrección. Para llevar esta tarea a cabo se utiliza la correspondiente versión para Python de la herramienta *Hunspell*. Esta herramienta es un corrector ortográfico que utilizan algunos programas de escritorio como pueden ser OpenOffice o Mozilla Firefox. Sin embargo, *Hunspell* necesita de unos diccionarios como fuente con tal de corregir los textos. En este caso se ha utilizado el mismo diccionario que LibreOffice, ya que es público y hemos encontrado buenos resultados. Sin embargo, nos encontramos con un problema, y es que los diccionarios no incorporan nombres propios de lugares de Madrid, como puede ser el distrito de Chamartín. Más concretamente, para este ejemplo, *Hunspell* cambia de *Chamartín* a *Cha Martín*. Este mismo problema ocurría con calles, plazas o incluso lugares de interés. Para solventar este problema, se insertan en el diccionario todos los lugares de la base de datos, un callejero de Madrid (también disponible en el portal de datos abiertos) y una lista de todas las estaciones de metro. Además, otras palabras de interés como puede ser *BiciMad*, el servicio de bicicletas públicas de la ciudad, también se incluyen. Gracias a esto, además de mantener palabras o nombres relevantes, aquellos que presenten errores serán corregidos.

Una vez los textos son ortográficamente correctos, se tratan con los principales pasos de un pre procesado de texto. Se eliminan las *stopwords* o palabras vacías, es decir, el conjunto de palabras sin significado como pueden ser preposiciones. Gracias a este paso se extraen los nombres, adjetivos y verbos que son los que aportan una información útil a la hora de encontrar similitudes. Este proceso se lleva a cabo con un diccionario de *stopwords* y con la librería *Spacy*, que aunque en español tenga pocos recursos, da buenos resultados en la eliminación de este tipo de palabras. Finalmente, con esta serie de palabras aplicamos lematización, que consiste en conseguir el lema correspondiente de una forma flexionada. A modo de ejemplificación, se utiliza el título de una propuesta para visualizar los pasos que se siguen, que se pueden comprobar en la Tabla 3.2.1.

Texto en bruto	Reparacion del pavimento de los carriles especiales para el BUS
Eliminar símbolos	Reparacion del pavimento de los carriles especiales para el BUS
Corrector ortográfico	Reparación del pavimento de los carriles especiales para el bus
Eliminar *stopwords*	Reparación pavimento carriles especiales bus
Lematización	['reparación', 'pavimentar', 'carril', 'especial', 'bus']

Tabla 3.2.1: Ejemplo de preprocesado de texto

3.3. Similitud entre documentos

Estimar la similitud entre dos textos es una tarea muy común en el ámbito de lenguaje natural. Algunas de las aplicaciones que puede tener esta medida van des de motores de búsqueda hasta herramientas de detección de plagio [34]. Como gran pilar de este trabajo, se busca encontrar una medida de similitud que nos permita identificar cuando dos documentos son extremadamente parecidos con tal de unificarlos en una sola entidad. Esta similitud se puede subdividir en dos características principales, la similitud léxica y la similitud sintáctica, es decir, que las palabras utilizadas sean de un mismo contexto y tengan un mismo significado, respectivamente. Estas dos características son importantes, y se deben tener en cuenta para el desarrollo de nuestra medida, pues debemos diferenciar dos documentos, como por ejemplo *"Fui al banco a sacar dinero"* y *"Me senté en un banco y encontré dinero"*. En este ejemplo, encontramos un gran parecido léxico, ya que las entidades que implica ambas situaciones son las mismas, una persona, banco y dinero, aunque la similitud sintáctica o de contexto será muy pequeña, porque si se pone en contexto la palabra *banco* encontraremos que en la primera hace referencia a un sitio y en la segunda a un objeto.

En nuestro problema es muy importante utilizar un método que tenga en cuenta ambas componentes, dado que la agrupación de textos que tengan poco que ver nos hará tener unos malos resultados. Tal y como explica [34], las mejor forma para llevar esta tarea es utilizando sofisticadas técnicas como pueden ser BERT o LSTM Siamesa. Sin embargo, se encuentran inconvenientes a la hora de utilizar estas técnicas, y es que el mejor modelo de BERT en español, BERTO [45], con nuestros datos no da unos resultados satisfactorios, aunque es una línea de trabajo futuro. En el caso de técnicas supervisadas como LSTM se necesita de la creación de un conjunto de datos etiquetados que resulta inviable. Por este motivo, las técnicas estudiadas y aplicadas son aquellas de las que se dispone material en el idioma que estamos trabajando y no necesitan de un etiquetado previo de muestras.

Las siguientes subsecciónes detallan los conceptos teóricos de tres medidas de similitud diferentes implementadas. A continuación se realiza una comparativa de sus resultados donde se motiva la elección de la métrica finalmente seleccionada, WMD. Finalmente se presenta un análisis del estudio de la similitud entre las propuestas de Decide Madrid siguiendo la medida seleccionada.

3.3.1. Índice Jaccard

También conocido como Intersección sobre Unión, el índice Jaccard es un estadístico usado para medir la similitud y la diversidad entre dos conjuntos. El estadístico complementario a este es la distancia Jaccard, utilizada para medir la disimilitud entre los conjuntos. El índice Jaccard viene dado por la expresion 3.3.1.

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{3.3.1}$$

Donde A y B son dos conjuntos finitos, resultando en una medida en el rango de $[0, 1]$. Por definición, si $A = B = \emptyset$, el índice Jaccard es $J(A, B) = 1$. La interpretación del estadístico, como es común en medidas de similitud, es que un índice próximo al límite superior, en este caso 1, implicará una mayor semejanza entre los conjuntos A y B. Es utilizado normalmente en tareas de visión por computador, para medir la precisión entre el resultado esperado y el obtenido por un software, tal y como se muestra en la Figura 3.3.1.

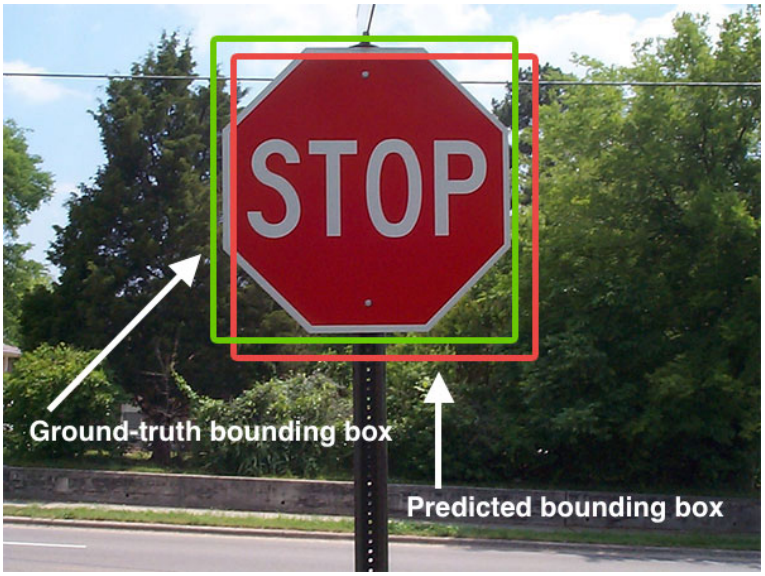


Figura 3.3.1: *Intersection over Union for object detection* [1]

A pesar de ser un estadístico que se ajusta adecuadamente a problemas del ámbito de visión por computador, en nuestro caso presenta graves inconveniente; no puede tratar la similitud léxica ni sintáctica. Esto es porque en la intersección de los dos conjuntos de palabras solo se encontrarán aquellas que sean idénticas (no es tolerante a sinónimos), y además en la unión debemos tener en cuenta que ambos conjuntos deben tener el mismo tamaño, ya que existe el caso de $A \subseteq B$, con B significativamente mayor que A y penalizar mucho, aunque sea el mismo contenido.

3.3.2. Similitud por coseno

Con el objetivo de encontrar un método para establecer una medida de similitud que sea tolerante a sinónimos y palabras de un mismo campo semántico, se prueba el método de la similitud por coseno. La premisa es la de conseguir representar un documento en un espacio vectorial de 300 dimensiones, para después calcular el coseno entre dos vectores, consiguiendo así una medida en el rango $[0, 1]$ (a pesar de que el rango de la función trigonométrica coseno sea $[-1, 1]$ se trabaja en valor absoluto) que representa como de separados están dos documentos; si los vectores son ortogonales (forman un ángulo de 90^{o}) son diferentes y si el ángulo es de 0^{o} son iguales.

La representación vectorial, llamada *embedding*, viene dada por modelos pre entrenados, en este trabajo se utiliza el modelo de Cristian Cardelino [46]. Este modelo ha sido entrenado con documentos del Europarlamento y artículos de Wikipedia entre otros. Gracias a este modelo, encontraremos unos *embeddings*, que en el espacio vectorial que se trabaja, están cerca de otras palabras sinónimas, del mismo campo semántico o de las que suelen ir acompañadas. Para poder usar este método, necesitamos representar los documentos como un único vector, y aunque hay muchas formas de transformar *word embeddings* a *document embeddings* [47], se toma la media aritmética en cada una de las dimensiones de los *word embedding* que conforman un texto.

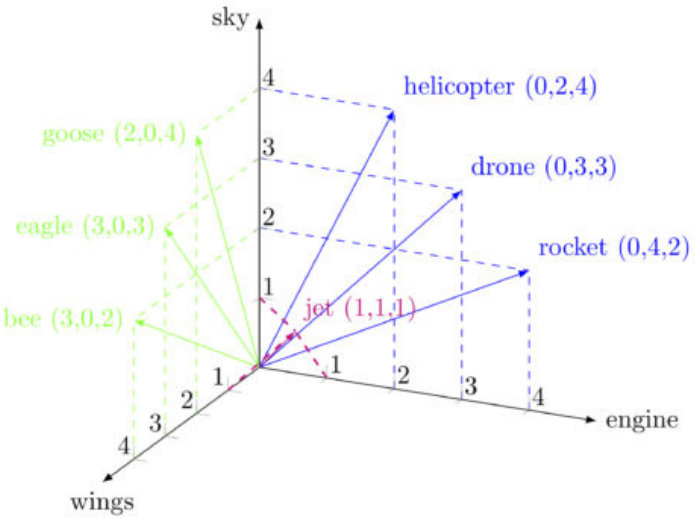


Figura 3.3.2: *A vector space of seven words in three contexts* [2]

A pesar de que este método da unos resultados aceptables al tener en cuenta la similitud léxica, podemos observar que la repetición de una palabra o más cercanas entre sí puede desviar el vector y dar un resultado erróneo. Esto se produce por tomar la media aritmética, y aunque hay algunas otras aproximaciones como tomar la *Smooth Inverse Frequency*, finalmente se decide descartar este método.

3.3.3. *Word Mover's Distance*

Inspirado en el problema de transportación *Earth Mover Distance*, WMD [48] pretende encontrar, a diferencia de los métodos anteriores, la distancia entre dos textos incluso aunque no tengan palabras en común. Se recomienda que los dos textos tengan la misma distribución, tal y como se muestra en la Figura 3.3.3 denotada con d y d'. Haciendo uso de los *word embeddings* este método va a calcular la distancia mínima necesaria para llegar de una palabra a otra.

Al igual que el problema en el que está inspirado, tiene una complejidad $O(n^3)$ para cada una de las palabras únicas que se tratan. A diferencia de otros métodos, este consigue capturar la sintáctica y la semántica ajustándose considerablemente bien a nuestro problema. Y es que a pesar de su gran costo computacional, se han encontrado unos resultados muy buenos con este método, y es el que se ha utilizado finalmente para la continuación del trabajo.

Hay que destacar, que como el propio nombre indica, este método calcula la distancia y no la similitud, por lo que hay que aplicar una transformación. Para ello se calculan todas las distancias de la matriz de distancias M, de dimensiones NxN, (donde N denota la cantidad de documentos, 21.746 en este problema), y se encuentra la distancia máxima para aplicar la expresión en 3.3.2.

$$similarity(i,j) = \frac{1 - WMD(i,j)}{max(M)} \tag{3.3.2}$$

Con esta transformación se consigue una medida en el rango $[0, 1]$ como es común en medidas de similitud, siendo 1 el mayor grado de semejanza.

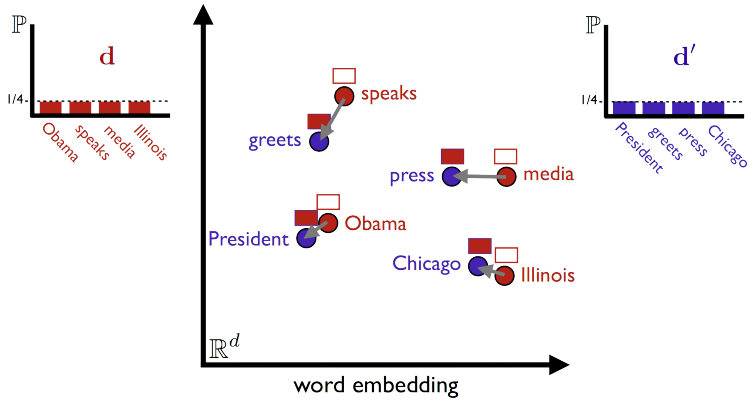


Figura 3.3.3: *Ejemplo de uso de WMD* [3]

3.3.4. Comparación de los métodos

Se realiza una comparación de los resultados obtenidos para los métodos estudiados, con tal de comprobar el comportamiento ante dos situaciones, un ejemplo en el que la similitud deba ser alta y otro en el que se determine que dos textos son diferentes.

Texto preprocesado	**Índice Jaccard**	**Similitud por Coseno**	***Word Mover's Distance***[1]
'habilitar recinto mascota' 'construir zona perro'	0,0	0,333	0,435
'habilitar recinto mascota' 'sancionar lanzar colilla suelo'	0,0	0,492	0,251

Tabla 3.3.1: Comparación entre los métodos de cálculo de similitud

Como se pueden observar en los resultados, el estadístico de Jaccard nos da un valor 0 para ambos casos, dado que ninguna de las palabras se puede encontrar también en el otro documento. En lo que la similitud por coseno respecta, podemos ver que en el ejemplo en el que debería dar un valor alto encontramos una similitud menor a la del caso en el que los textos no hablan sobre lo mismo. Para acabar, el método que hemos seleccionado da una similitud de 0,435 para el caso en el que los textos versan de lo mismo y un valor menor para el caso en el que no.

Gracias a WMD conseguimos una medida satisfactoria, que nos permite diferenciar un texto de otro, por lo que es esta medida la que se utiliza a lo largo del trabajo. Realizando un histograma de los resultados conseguidos con esta técnica obtenemos los resultados en Figura 3.3.4, que explica que hay una gran cantidad de documentos con una similitud de 0,4 o mayor.

[1]Se toma como distancia máxima la encontrada en el conjunto de datos, $max(M) = 6{,}092$.

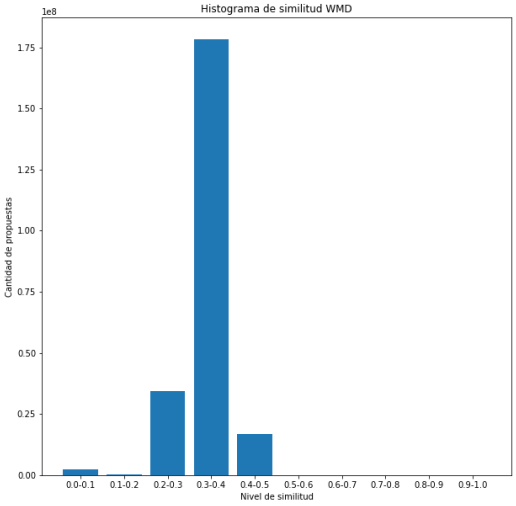


Figura 3.3.4: Histograma de las similitudes encontradas con WMD

Limite inferior	**Limite superior**[2]	**Cantidad**
0.0	0.1	2.352.964
0.1	0.2	301.085
0.2	0.3	34.304.764
0.3	0.4	178.209.626
0.4	0.5	16.912.819
0.5	0.6	106.002
0.6	0.7	3.092
0.7	0.8	0
0.8	0.9	132
0.9	1.0	0

Tabla 3.3.2: Desglose del histograma de similitudes WMD

Además, en la Tabla 3.3.3, podemos comprobar como, poco a poco en cada tramo, las propuestas en cada tramo son más similares. Se destaca el tramo $[0{,}4, 0{,}5]$, en el que se empiezan a observar similitudes considerablemente altas.

[2]No incluido

Tramo	Titulares de propuesta
[0.0, 0.1)	Poner en marcha una bolsa de empleo BiciMAD en el apeadero de Entrevías
[0.1, 0.2)	Proyecto nombre de las calles de Madrid Parking subterráneo vecinal
[0.2, 0.3)	Ovejas y huertos municipales para Casa de Campo Carril bici Corredor del Henares
[0.3, 0.4)	Eliminacion normativa municipal que permite motos en aceras Basuras de reciclaje en condiciones por un mundo y una ciudad mejor!
[0.4, 0.5)	Personas perceptoras de ayudas sociales barran las calles de Madrid Control vaciado ilegal contenedores reciclaje
[0.5, 0.6)	No a los camiones de la basura en hora punta Sistema de recogida de basuras
[0.6, 0.7)	Semaforos para peatones que originan situacion de peligro Nuevos Semaforos en Gran Via
[0.7, 0.8)	-
[0.8, 0.9)	Mas trenes circulando en el cercanias de Madrid Mas trenes circulando en el Metro de Madrid
[0.9, 1.0]	-

Tabla 3.3.3: Ejemplos de propuesta por tramo de similitud

3.4. Agrupación de documentos

Como se vió en el capítulo anterior, algunas técnicas de *clustering* de documentos normalmente utilizadas son *K-Means* y agrupación aglomerativa con dendogramas [49]. En este trabajo se propone otro enfoque, tratar la matriz de similitud M como una matriz de adyacencia [38; 39]. De este modo, y utilizando la matriz conseguida con *Word Mover's Distance*, es posible generar un grafo casi completo no dirigido al que le falta una arista, aquella que representa la máxima distancia entre dos documentos, que al aplicar la transformación 3.3.2 resulta en cero.

Con este enfoque, el problema de agrupación pasa a ser un problema de descubrimiento de comunidades en un grafo. Para resolver este problema se utiliza el método de Louvain [38; 39; 50], que pretende optimizar la modularidad del grafo localmente y asociar nodos hasta convergencia, con un buen tiempo de ejecución de $O(n \cdot log(n))$. Este método es el que ha sido escogido dado que, a diferencia de otros como K-Means, el número de comunidades no es fijo, sino que se adapta al problema. Aplicar directamente este algoritmo sobre un grafo, que podemos asumir totalmente conexo, resulta en una única comunidad que representa todo el grafo, de modo que para solucionar esto se rompen las aristas con un peso (que representa el nivel de similitud) inferior a un cierto valor. En este trabajo concretamente, se eliminan todas las aristas con un peso inferior a $min\_weight = 0{,}55$ (motivado por el tramo en el que se observan similitudes considerables), valor que ofrece unos resultados satisfactorios.

Tras eliminar las aristas con menor peso, se extraen hasta 15.113 comunidades, con una media de 1.44 elementos por comunidad. A modo de ejemplo, a continuación se muestran las comunidades más grandes que se han generado y algunos de los títulos de las propuestas que contienen.

Comunidad	**Categoría principal de las propuestas**	**Ejemplos de propuesta**
#12	Movilidad	Incluir el BiciMAD en el Abono Transportes Transbordo gratis entre metro, bus, cercanias zona 1 Abono Transporte gratuito para desempleados.
#9	Movilidad	Horario nocturno transporte publico METRO abierto toda la noche en fines de semana Reestablecer el servicio de autobuses nocturnos
#24	Animales	Sanciones significativas a duenos de perros Sancionar a la gente que deposite basuras en la calle Multar a los duenos de perros que no retiren sus excrementos
#6	Medio Ambiente	Limpiar las calles Aumentar la limpieza en Madrid Mas arboles en Madrid

Tabla 3.4.1: Comunidades formadas por el método de Louvain

Como se puede observar en la comunidad 12, la propuesta *Incluir el BiciMAD en el Abono Transportes* parece que no encaja en el contexto de las otras dos propuestas. Algo parecido ocurre también en la comunidad 6. Para explicar esto, podemos ver la estructura que tienen las comunidades en la Figura 3.4.1.

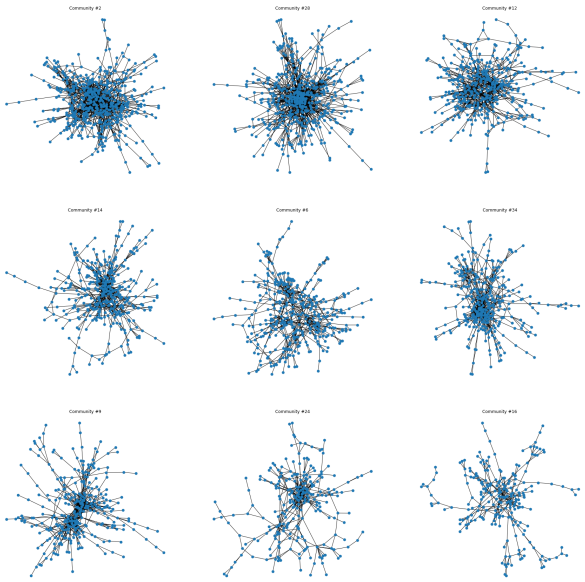


Figura 3.4.1: Estructura de las comunidades más grandes

Comunidad	Descripción
#**105**	DecideMadrid
#**2**	Alcance del transporte público
#**6**	Limpieza en espacios públicos
#**9**	Servicios públicos por la noche y fin de semana (bus, bibliotecas...)
#**12**	Transporte público gratuito para menores/jubilados/discapacitados
#**14**	Mejoras en infraestructura (institutos, fuentes, parques...)
#**16**	Pasos de peatones y semáforos
#**24**	Sanciones por ensuciar calles
#**28**	Zona azul, zona SER, aparcamiento de discapacitados...
#**34**	BiciMad

Tabla 3.4.2: Descripción de las 10 comunidades más grandes, de mayor a menor

Dada la estructura de las comunidades podemos ver que dos nodos pueden estar alejados el uno del otro, aunque por el camino que hay entre los dos podemos asegurar que tienen una base en común. En líneas generales, todas las comunidades acaban formando algo similar a una topología de estrella. En esta forma se encuentran los documentos más parecidos en el centro, mientras que aquellos que se encuentren en los extremos serán ligeramente diferentes. Aunque los nodos de los extremos no se puedan considerar duplicados entre ellos se puede garantizar que tienen una base en común, ya que si no las aristas que los conectan hubiesen sido eliminadas. Aumentando el valor de $min\_weight = 0{,}55$ se podrían conseguir comunidades más pequeñas, ya que la que las aristas más débiles serían eliminadas, consiguiendo así comunidades con una similitud media mayor.

Con tal de comprobar si este efecto también se da en otras comunidades, en la Tabla 3.4.3. Podemos ver que la similitud media es bastante consistente, lo que nos indica que este efecto aparece en todas las comunidades, y es derivado del método de agrupación que se ha realizado.

A fin de saber más sobre la consistencia de las propuestas de estas comunidades se comprueba manualmente el contenido de los documentos de las diez más grandes, en las que resulta mucho más sencillo identificar y localizar cuales son las propuestas que se hacen, con tal de actuar en consecuencia. En la Tabla 3.4.2 podemos verificar la preocupación de los ciudadanos con respecto a la propia plataforma, siendo esta la comunidad con más propuestas, como pueden ser *'Que haya opcion de NO APOYO en Madrid Decide'*, '*Publicidad Plataforma Decide Madrid*' o '*Agrupar las propuestas similares*'. Encontrar este tipo de propuestas en una comunidad tan grande **verifica nuestra hipótesis** y refuerza las indicaciones de muchos estudios preliminares, resumidos en [5]. Además, en el resto de comunidades encontramos documentos muy similares entre sí, generando subconjuntos de las existentes categorías que aportan información mucho más concreta de la que se dispone actualmente.

Comunidad	**#12**	**#9**	**#24**	**#6**	...	**Total**
Similitud media	0.3798	0.3777	0.3893	0.3755	...	0.4114

Tabla 3.4.3: Similitud media en cada comunidad

Finalmente, tras haber seguido este procedimiento, se decide que los resultados se ajustan correctamente al problema, consiguiendo una agrupación de los documentos en comunidades de propuestas estrechamente relacionadas.

Capítulo 4

Análisis

En este capítulo se realiza un análisis de las comunidades más grandes que se han encontrado, con tal de detectar la influencia en cada uno de los distritos, su evolución temporal y factores que determinen el éxito de las propuestas.

4.1. Factores determinantes del éxito

El objetivo de esta sección es analizar las comunidades con tal de encontrar aquellos atributos que puedan tener relación con el número de apoyos de la propuesta. En primer lugar, se calcula la longitud de una propuesta, que incluya la longitud del título, del resumen y del texto, con tal de encontrar la relación con el número de apoyos.

Como podemos comprobar en la Figura 4.1.1, las propuestas más extensas no son aquellas que reciben más apoyos, aunque las más cortas tampoco. Este efecto se ha encontrado en otros ámbitos, como en mensajería [51] o en herramientas de SEO, donde se recomienda cierta longitud del texto para conseguir más lectores. De igual manera, aquellas propuestas que sean suficientemente concisas atraen más lectores, aumentando así los potenciales votantes.

De la Figura 4.1.2 podemos extraer que las propuestas con más comentarios tienen más votos que aquellas que no tienen comentarios. Sin embargo, esta relación no es bidireccional, ya que podemos encontrar también muchas propuestas con suficientes apoyos y con una cantidad mínima de votos.

Con tal de comprobar que el debate o controversia que se genere en los comentarios tampoco tiene relación con el número de apoyos, se utilizan las métricas de [4]. En la Figura 4.1.3 podemos comprobar que ninguna de las métricas de controversia está relacionada con el número de apoyos. Además, los comentarios es una característica que el usuario que publica una propuesta no puede controlar, de modo que no se puede hacer ninguna recomendación acerca de estos.

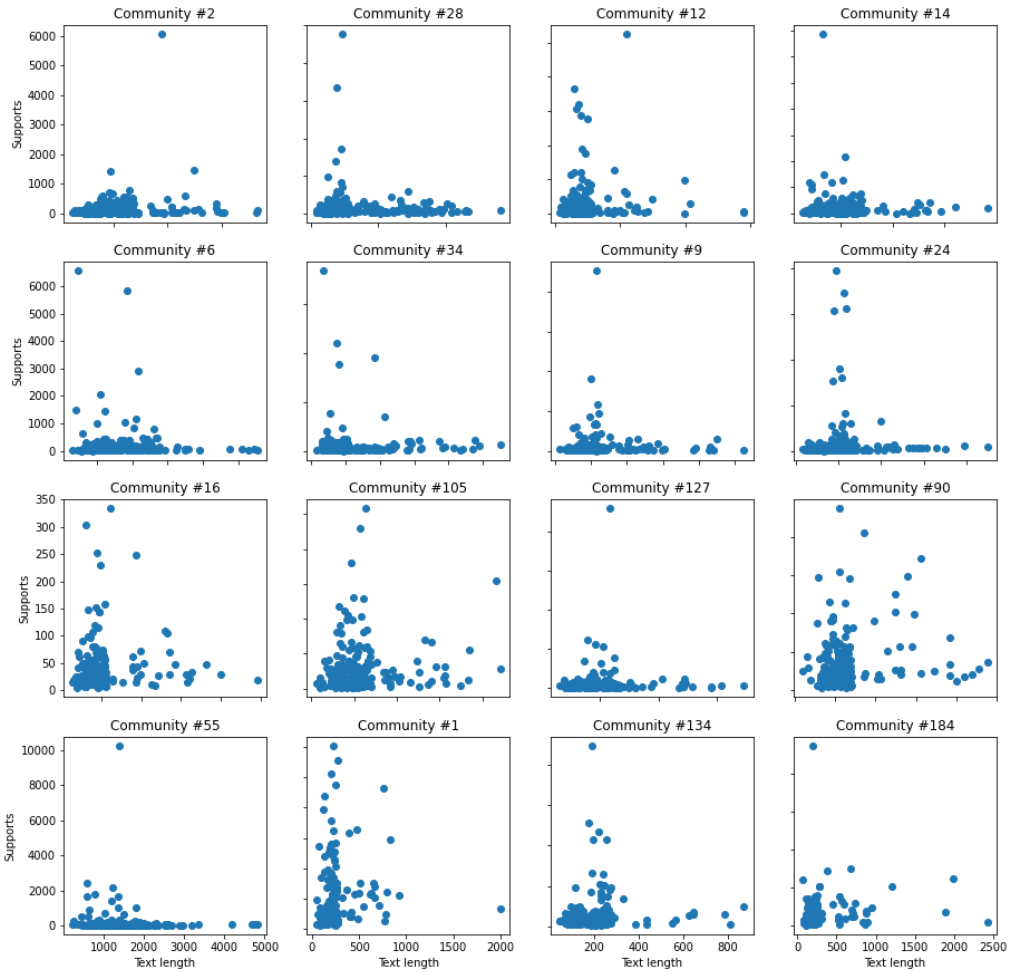


Figura 4.1.1: Relación del número de apoyos y la longitud de documento por comunidades

Es conocido por los creadores de contenido en redes sociales o redactores en blogs que el momento de publicación es crucial para llegar a más espectadores [52]. Podemos comprobar que en la plataforma DecideMadrid (Figura 4.1.4) ocurre algo parecido, las propuestas que más votos tienen son aquellas que se publican en martes, miércoles o incluso en jueves. Este fenómeno abre una línea de trabajo futuro que correlacione el éxito de publicaciones en diferentes plataformas con el horario y calendario establecidos en una sociedad, y estudiar si coincide entre plataformas de diferente ámbito, como pueden ser redes sociales y plataformas de participación ciudadana.

En la Figura 4.1.5 podemos comprobar que en lo que respecta el número de categorías que se asignan al crear una propuesta, no hay un claro patrón que lo relacione con los apoyos que reciba, aunque 4 parece tener siempre buenos resultados. Sin embargo, el número de distrito es importante, parece que aquellas propuestas que implican un número pequeño

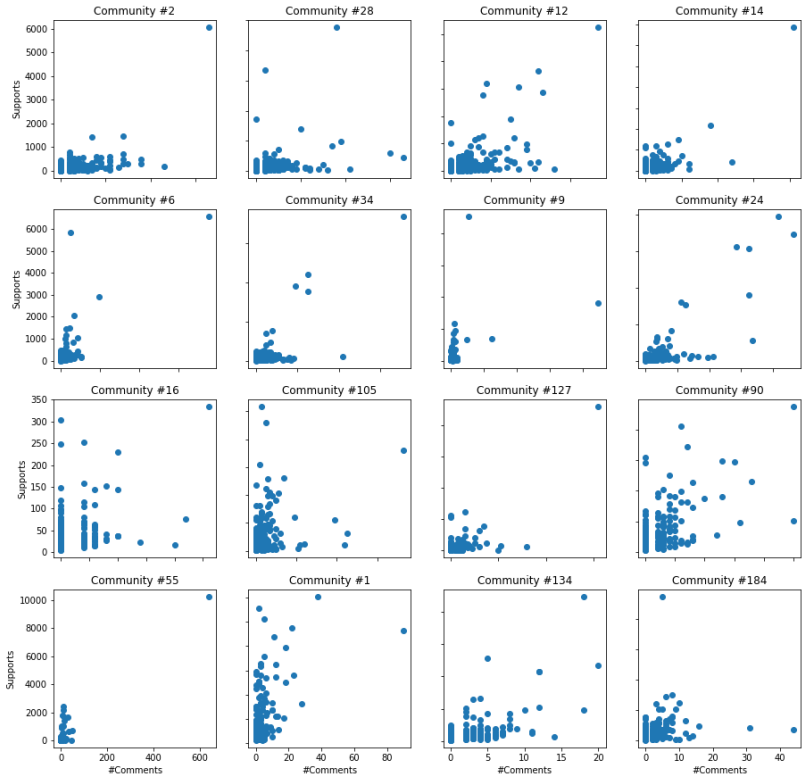


Figura 4.1.2: Relación del número de apoyos y la cantidad de comentarios por comunidades

de distritos son las que reciben más apoyos. Esto puede llegar a ser un indicador de que los votantes apoyan aquellas propuestas que les afecte de forma cercana, en su distrito o en los cercanos.

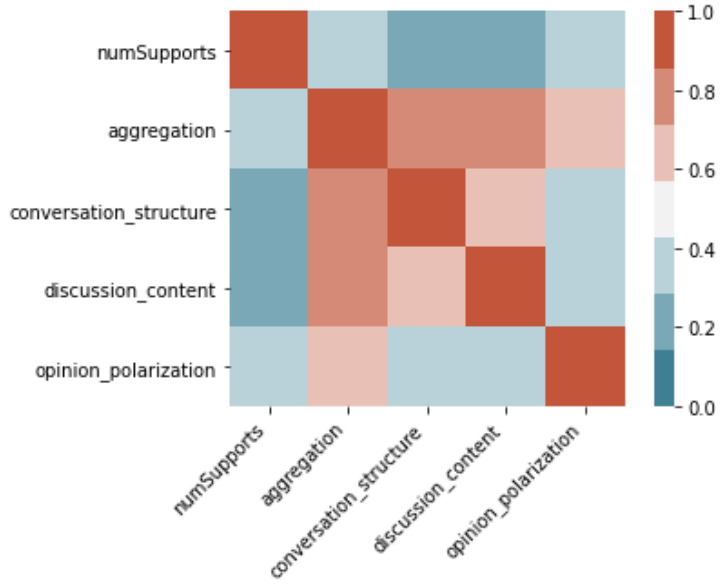


Figura 4.1.3: Relación de métricas de controversia con el número de apoyos

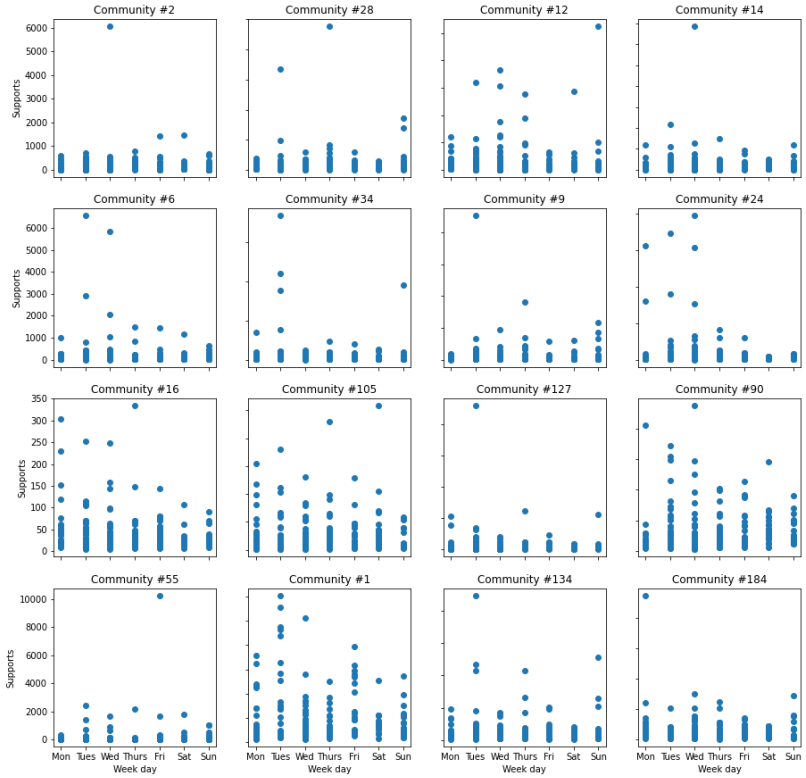
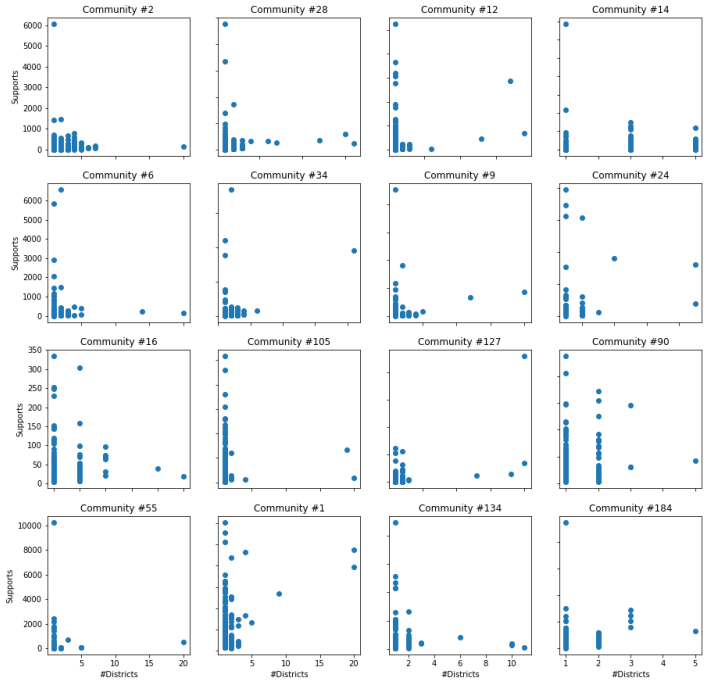
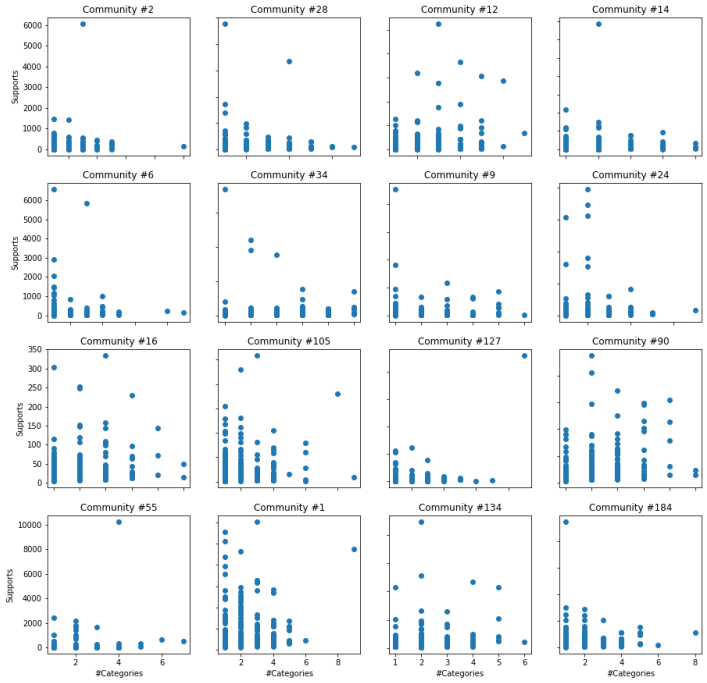


Figura 4.1.4: Relación del día de publicación con el número de apoyos



(a) Relación entre apoyos y número de distritos de una propuesta



(b) Relación entre apoyos y número de categorías de una propuesta

Figura 4.1.5: Relación de apoyos con el número de categorías y distritos

4.2. Influencia y evolución por distrito

Antes de nada, y para poner en contexto los resultados, es necesario observar el número de propuestas por distrito. Hay que destacar que en esta sección, dado que cada propuesta puede tener más de una categoría puede tener más de un distrito y más de una categoría, se trabaja con el producto cartesiano de estos atributos, es decir, para cada propuesta existen tantas muestras como combinación de categorías y distritos sea posible.

Tal y como se puede ver en la Figura 4.2.1, la distribución de propuestas es bastante

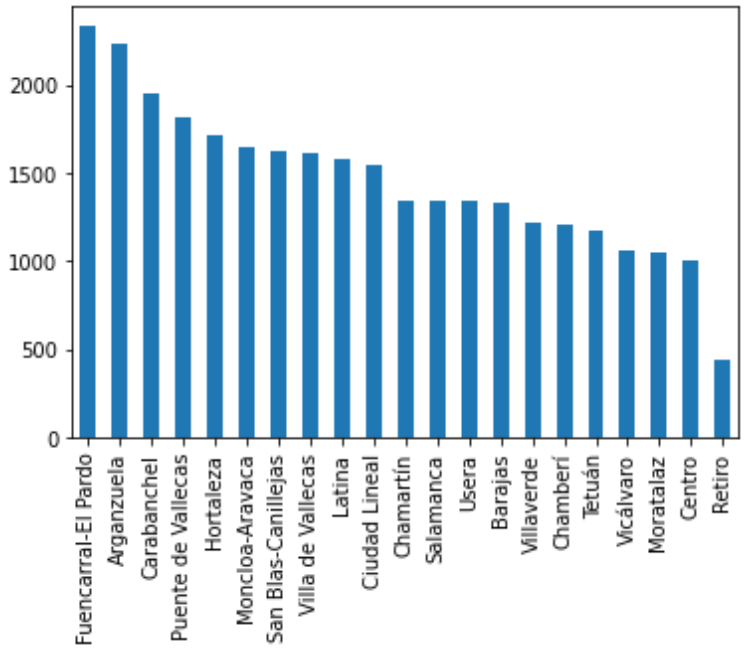


Figura 4.2.1: Número de propuestas por distrito

uniforme, aunque hay algunos distritos que despuntan. Hay algunas de estas anomalías tienen una sencilla explicación: el número de habitantes en el distrito. Por ejemplo, en Fuencarral - El Pardo hay una población de 246.021 [53] habitantes, la más poblada después de Carabanchel. Sin embargo, Retiro, a pesar de tener 119.379 habitantes, es la que menos propuestas tiene con mucha diferencia. A simple vista, podemos ver que hay cierta tendencia a que los distritos con mayores ingresos [53] son aquellos que menos propuestas hacen, ocurriendo lo contrario con los distritos con ingresos más reducidos. Este fenómeno puede servir de indicador para detectar aquellos distritos en el que no existen todas las comodidades que hay en otros, produciendo una desigualdad entre distritos.

Haciendo uso de las comunidades extraídas en el Capítulo 3 podemos, de una forma mucho más concreta que la categoría, detectar las necesidades que son más demandadas en cada distrito. Siguiendo con las comunidades mencionadas en el ejemplo de la Tabla 3.4.1, comprobamos los distritos en los que más influencia tienen cada una en la Tabla 4.2.1.

Comunidad	Descripción	Categoría principal	Distrito más afectado
#**12**	Abono de transporte	Movilidad	Barajas (13.72 %)
#**9**	Horario nocturno en transporte público	Movilidad	Moncloa-Aravaca (20.58 %)
#**24**	Sanciones por no limpiar residuos de animales	Animales	Ciudad Lineal (15.78 %)
#**6**	Limpieza en las calles	Medio Ambiente	Fuencarral-El Pardo (10.92 %)

Tabla 4.2.1: Distrito con mayor influencia de cada comunidad

Aunque a simple vista los distritos más afectados no sean los esperados, cada uno de ellos tiene una explicación de ser. Poniendo como ejemplo la comunidad 9, en la que un conjunto de propuestas que pide *Horarios nocturnos en el transporte público*, el más afectado es el distrito universitario, Moncloa-Aravaca, el cual debería tener suficiente recursos como para cubrir la gran demanda de transporte que existe en el. Y es que a pesar de la gran cantidad de estaciones de las principales líneas de metro, hay que tener en cuenta que Metro de Madrid no cuenta con servicio nocturno (igual que los servicios de Renfe Cercanías), recayendo toda la carga (teniendo en cuenta lo que esto suponen en un barrio universitario, en el que mayormente esta poblado por jóvenes que necesitan de esta clase de servicios) en los autobuses nocturnos. El distrito solo cuenta con cuatro autobuses urbanos nocturnos, todos con origen en Plaza de Cibeles, y además teniendo en cuenta los grandes lapsos de tiempo entre autobuses, resulta lógico que Moncloa-Aravaca pida una mejora en sus servicios de transporte nocturnos.

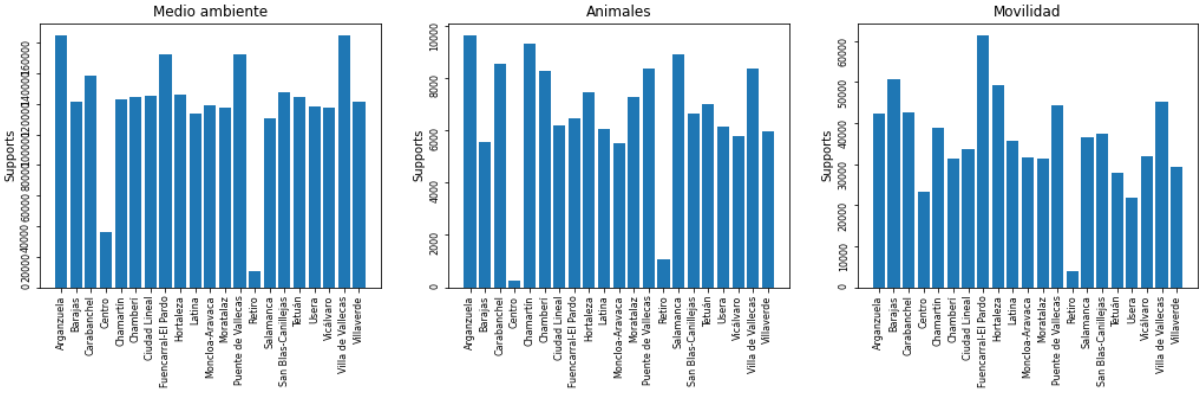


Figura 4.2.2: Relación distrito con las categorías Medio Ambiente, Animales y Movilidad

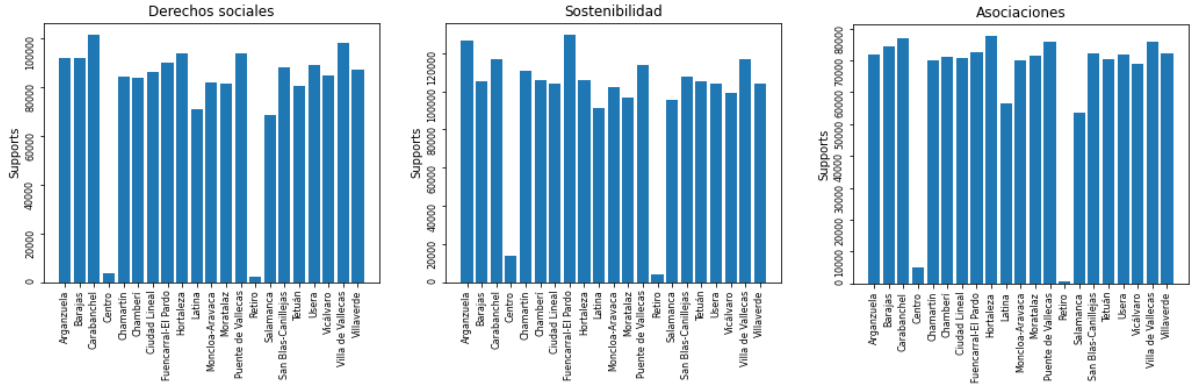


Figura 4.2.3: Relación distrito con las categorías Derechos Sociales, Sostenibilidad y Asociaciones

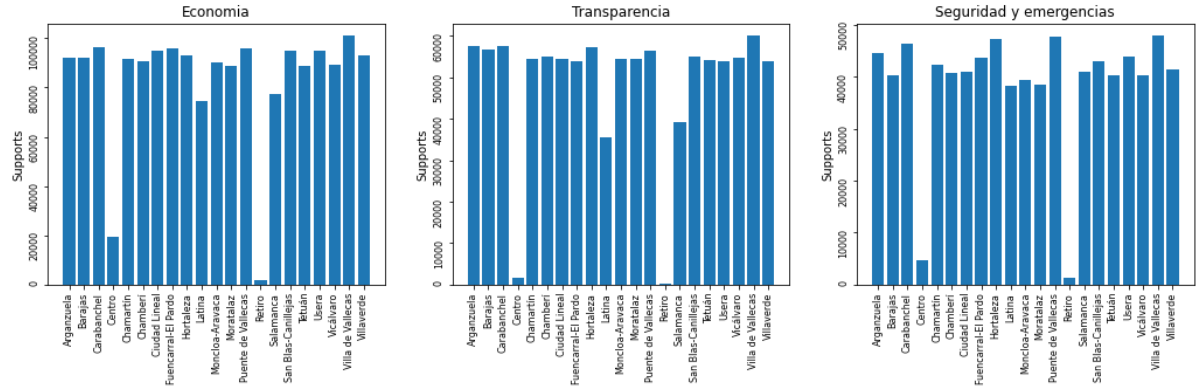


Figura 4.2.4: Relación distrito con las categorías Economía, Transparencia y Seguridad y Emergencias

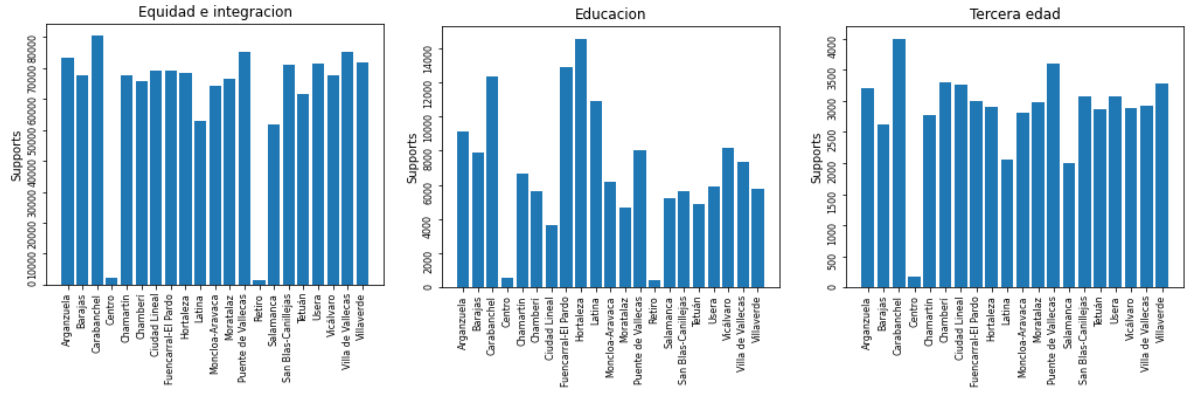


Figura 4.2.5: Relación distrito con las categorías Equidad e integración, Educación y Tercera Edad

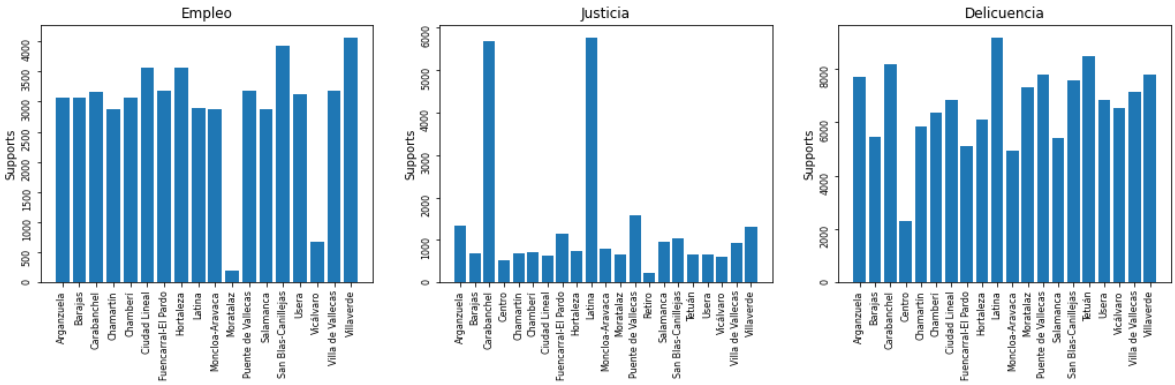


Figura 4.2.6: Relación distrito con las categorías Empleo, Justicia y Delincuencia

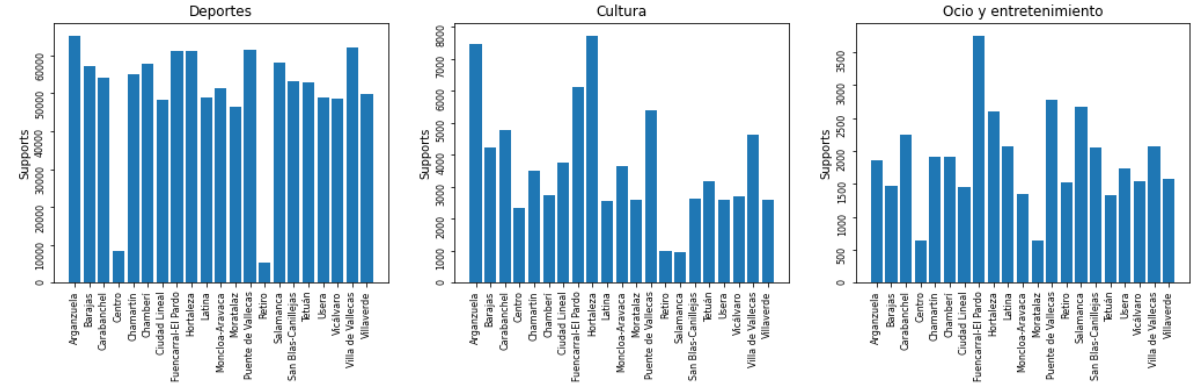


Figura 4.2.7: Relación distrito con las categorías Deportes, Cultura y Ocio y Entretenimiento

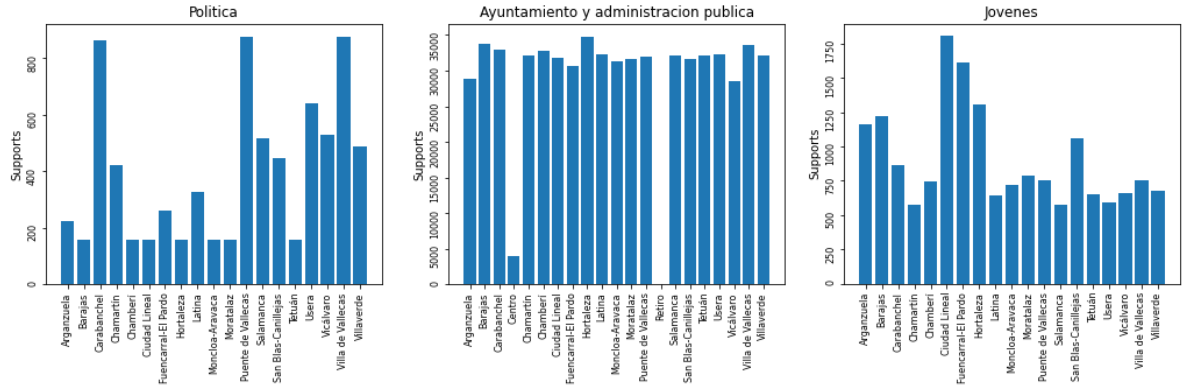


Figura 4.2.8: Relación distrito con las categorías Política, Ayto. y Admon. Pbica. y Jóvenes

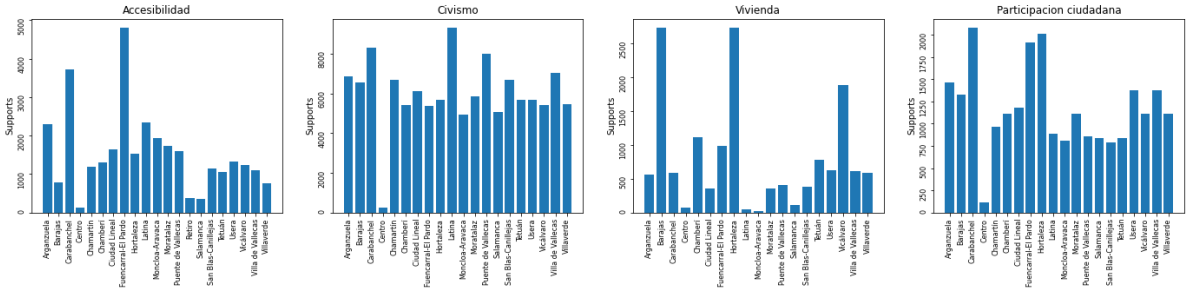


Figura 4.2.9: Relación distrito con las categorías Accesibilidad, Civismo, Vivienda y Participación Ciudadana.

Desde la Figura 4.2.2 a la Figura 4.2.9 podemos comprobar aquellos distritos que se ven más afectados por una categoría, por ejemplo, los distritos con mayores propuestas de la categoría Política son Puente de Vallecas y Villa de Vallecas, lo que hace denotar cierta preocupación en esta zona de la ciudad. De hecho, estos dos distritos, junto a otros con mayor participación en esta categoría como Usera o Carabanchel, son aquellos en los que tanto en las elecciones municipales de 2015 y 2019 han conseguido una mayoría el partido *Ahora Madrid* (posteriormente *Más Madrid)*. Contrariamente, en distritos con menores propuestas en Política, como Hortaleza o Moncloa-Aravaca, el susodicho partido no tuvo tanto impacto [54].

Con tal de mejorar la accesibilidad a los datos de influencia de propuestas en cada distrito, se desarrolla un *dasboard* en *Tableau* [55], del que se habla más en profundidad en el siguiente capítulo. Gracias a este panel, se mejora considerablemente la interpretabilidad de los datos, pudiendo ver además la evolución de categorías y comunidades por distrito.

Se pueden ver algunas imágenes de este panel en 4.2.10 y 4.2.11. Más detalladamente, en la Figura 4.2.10 podemos ver el impacto de la comunidad seleccionada a la izquierda, *Alcance del transporte público*, en la ciudad, en la que hay una clara necesidad en los distritos de la periferia. De hecho, hay expertos que indican que la M-30 al este de la ciudad supone una gran barrera, dividiendo a la población, lo que ha hecho crear iniciativas como *Parque-30* [56]. De forma contraria, a la derecha, podemos observar como en la categoría Turismo, la mayoría de propuestas provienen de los distritos céntricos, donde el número de lugares de interés es mucho mayor.

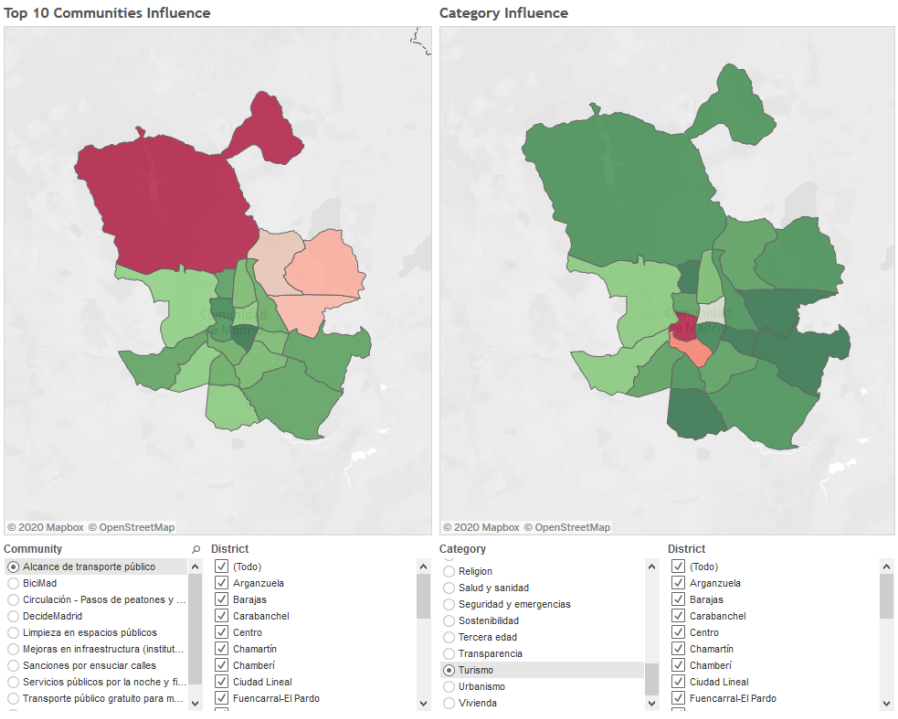


Figura 4.2.10: Influencia de propuestas por categoría y comunidad

También es posible encontrar cierta estacionalidad en la evolución de las comunidades y categorías. Siguiendo el ejemplo anterior, relativo a la comunidad que representa *Alcance del transporte público*, podemos ver que es el primer trimestre de cada año (o incluso el cuarto del año anterior) 4.2.11, cuando más impacto tiene, seguramente con tal de que se incluya en los objetivos de las organizaciones responsables para ese año.

Figura 4.2.11: Evolución de propuestas por categoría y comunidad

Gracias a este panel de datos resulta sencillo conocer la estacionalidad de grupos de propuestas, incluso detectar picos inusuales como el que se muestra en la Figura 4.2.12 en la categoría transparencia.

El pico en la cantidad de propuestas de la categoría transparencia en abril de 2016, así como muchos otros, viene fundamentado en un acontecimiento. Fue en ese mismo mes cuando la directora general de Economía del Ayuntamiento de Madrid presentó su dimisión, después de que se filtraran documentos que la relacionaban con una sociedad en Panamá durante cinco años [57].

Así entonces, esta herramienta puede servir de ayuda a los gobernantes de la ciudad para conocer las necesidades de los ciudadanos, de una forma más concreta y discreta, lo que facilita su tarea y les permite cubrir estas necesidades y, posiblemente, aumentar la calidad de vida de los habitantes.

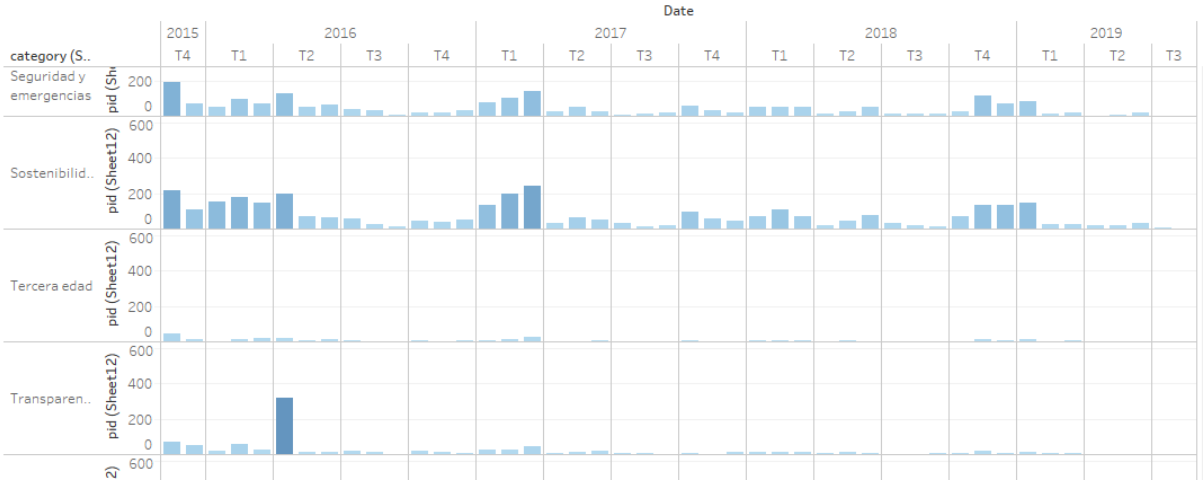


Figura 4.2.12: Evolución de propuestas por categoría y comunidad

Capítulo 5

Aplicaciones

Después de haber hecho un estudio de los datos que se disponen, se proponen herramientas que puedan servir de utilidad para la plataforma de DecideMadrid.

5.1. Modelo de predicción de apoyos

Como el plan de trabajo establece, se entrena un modelo de regresión, concretamente un *Random Forest Regressor*, con tal de hacer una predicción de los apoyos que tendrán futuras propuestas. Específicamente, los atributos de las propuestas que se utilizan son los siguientes:

- Longitud del texto
- Número de comentarios
- Día de la semana de publicación
- Distrito
- Categoría
- Comunidad

Tras separar los datos con el 30 % de ellos en test, el pre procesado de los datos es el que se utiliza normalmente en estos casos, se generan variables *dummies* y se normalizan los datos de entrenamiento y de test por separado y tras esto, se entrena el regresor con 10 estimadores.

Como se puede observar en la Figura 5.1.1, las estimaciones dibujan aproximadamente el valor real, aunque no se ajustan correctamente. Con este regresor se obtiene un $RMSE = 0{,}024$ y un valor de $R^2 = 0{,}332$. Los resultados que se encuentran no son nada buenos, y es que nos encontramos en un caso en el que hay que considerar una serie de factores, como puede ser la situación política, que no se tienen en cuenta. Dada esta situación se determina que, al igual que en otros casos en el que existen factores que no se pueden tener en cuenta (como en finanzas), no es posible hacer una predicción fiable del número de apoyos que va a tener una propuesta. Sin embargo, el regresor ha utilizado unos estimadores que encajan muy bien en los mencionados en el análisis previo. Entre los primeros podemos encontrar el número de comentarios, la longitud del texto o que el día de publicación sea martes.

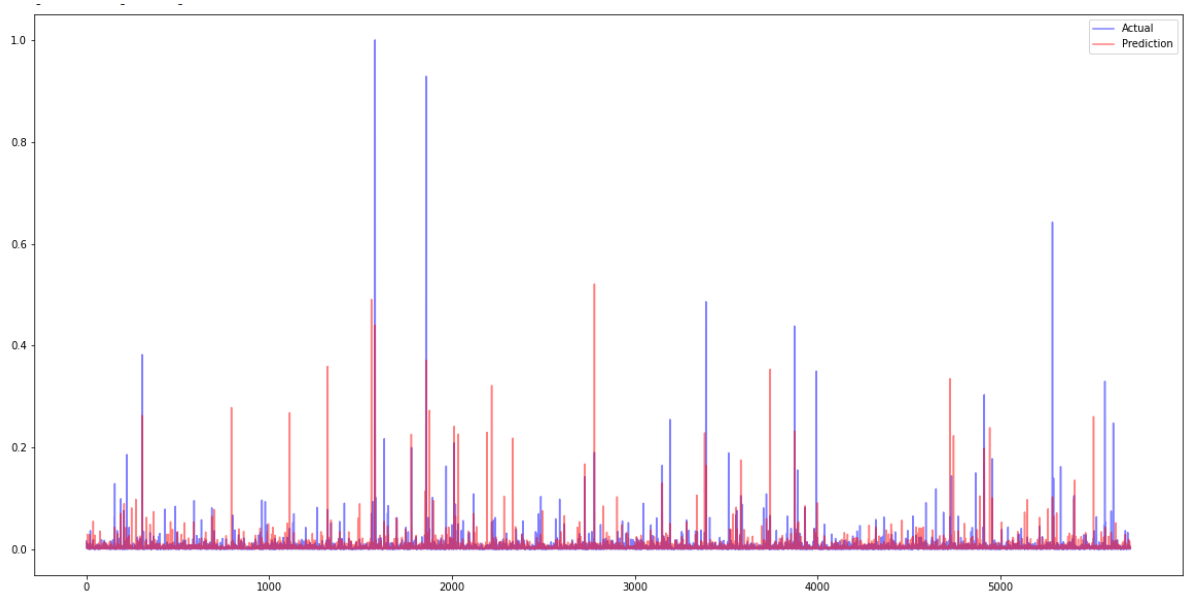


Figura 5.1.1: Comparación de las predicciones y valores reales

5.2. Detección de duplicidad en nuevas propuestas

Utilizando la medida de similitud escogida, es posible, dada una nueva propuesta, encontrar el documento más similar, con tal de proponérselo al usuario y así evitar la duplicidad de propuestas en la plataforma. Para poder llevar la implementación, sería necesario desplegar una aplicación web que contase con el modelo cargado en todo momento, con tal de reducir tiempos, ya que WMD es una técnica con una complejidad muy alta. Existe una variación de *Word Mover's Distance*, que pretende disminuir esta complejidad hasta una muy ambiciosa complejidad lineal y acelerada por GPU, llamada *Relaxed Word Mover's Distance* [58].

Por motivos técnicos no se hace la implementación de RWMD en un servidor web, sino que se utiliza WMD clásico con la implementación utilizada anteriormente en este trabajo en un entorno Python local. A modo de ejemplo, se escribe una propuesta totalmente nueva con tal de saber si es una herramienta útil:

Título: Construir un recinto para mascotas en Plaza de Castilla
Resumen: Creacion de una zona para poder pasear a los perros donde puedan estar sueltos
Texto: En una zona como Plaza de Castilla es difícil encontrar una zona donde se puedan sacar a los animales a pasear tranquilamente, donde no molesten. La propuesta sería crear una zona cerrada y/o vallada para que los perros puedan estar sueltos y jugar, con árboles, fuentes... Esto podría hacer que los animales no molesten a los peatones de la zona y que los perros puedan hacer pipi sin ensuciar la calle.

El procedimiento de tratar el texto y encontrar la propuesta más cercana lleva en total 4 minutos y 44 segundos. Este tiempo se podría mejorar considerablemente manteniendo el modelo cargado en todo momento y utilizando RWMD. La propuesta más cercana tiene un 51,37 % de similitud:

> **Título:** CREACION DE UN PARQUE CANINO-PIPICAN EN EL CANAVERAL
> **Resumen:** Construccion de un area canina donde poder soltar a los perros para que jueguen y corran tranquilamente sin molestar, ya que no existen zonas para ello y la existencia de canes en la zona es numerosa.
> **Texto:** Se soilicita la creacion de un area canina donde poder soltar a los perros y que puedan jugar y correr tranquilamente sin molestar a viandantes y sin peligros tanto para los animals como para las personas y coches que circulan por la zona. Un recinto cerrado, donde poder pasar un rato agradable tanto para duenos como para los perros de la zona, ya que son numerosos y cada vez van mas en aumento.

Aunque es cierto que los lugares para los que se pide en cada propuesta son diferentes, el nivel de similitud es más que evidente. Este sencillo proceso se podría complementar agregando otra información importante, como en este caso es el lugar, para determinar la duplicidad más allá del parecido del texto.

5.3. Herramienta de visualizacion

De acuerdo a la bibliografía [41; 42; 43], una herramienta de visualización de datos, sin demasiadas variables simultáneas, ayuda a la interpretabilidad de los datos. Además, proporcionar este tipo de herramientas a los ciudadanos fomenta la coproducción de nuevos proyectos, incentivando a otros ciudadanos a participar en este tipo de procesos. Para llevar esta tarea a cabo, se utiliza la herramienta Tableau para crear un panel de datos en tres pestañas: `https://public.tableau.com/profile/sergio8719#!/vizhome/Libro1_15923257034100/Historia`.

En la primera pestaña, mostrada en la Figura 5.3.1, podemos comprobar la participación por distrito y la evolución que ha tenido. En la parte superior podemos comprobar la participación total desde 2015 hasta 2019, en la que tal y como se ha comentado anteriormente, los distritos con más ingresos tienden a ser los que menos han participado y, aquellos con ingresos más reducidos, son los que más tienden a realizar propuestas. Además, en la parte inferior podemos ver, gracias a un selector de fechas, la evolución de la participación por distrito en un cierto rango de fechas. Un dato interesante es que la participación parece ser bastante equitativa entre distritos, colaborando todos de modo proporcional.

Figura 5.3.1: Participación y su evolución por distrito

A través de la segunda pestaña (Figura 5.3.2) es posible un análisis más profundo de la evolución de la participación añadiendo dos nuevas variables: la comunidad en la parte superior y la categoría en la parte inferior. En ambos casos, podemos ver el gran descenso del número de propuestas durante los meses de verano. Con estos gráficos de barras es posible estimar la estacionalidad de un problema a lo largo del tiempo, por ejemplo, en la categoría Animales, que aumenta de forma considerable durante los últimos meses del año, seguramente propiciado por el abandono de mascotas o por los animales que son regalados durante Navidad.

Figura 5.3.2: Evolución de la participación, por categoría, comunidad y distrito

Finalmente, en la Figura 5.3.3, vemos el impacto de las categorías y las comunidades de una forma mucho más sencilla. Concretamente, en el mapa de la izquierda podemos verificar el caso de Moncloa-Aravaca y los servicios de transporte público por la noche comentado anteriormente en este trabajo. De forma análoga, en el mapa de la izquierda, podemos comprobar el impacto de propuestas de Salud y Sanidad, que tiende a tener un efecto mayor en los distritos de la periferia.

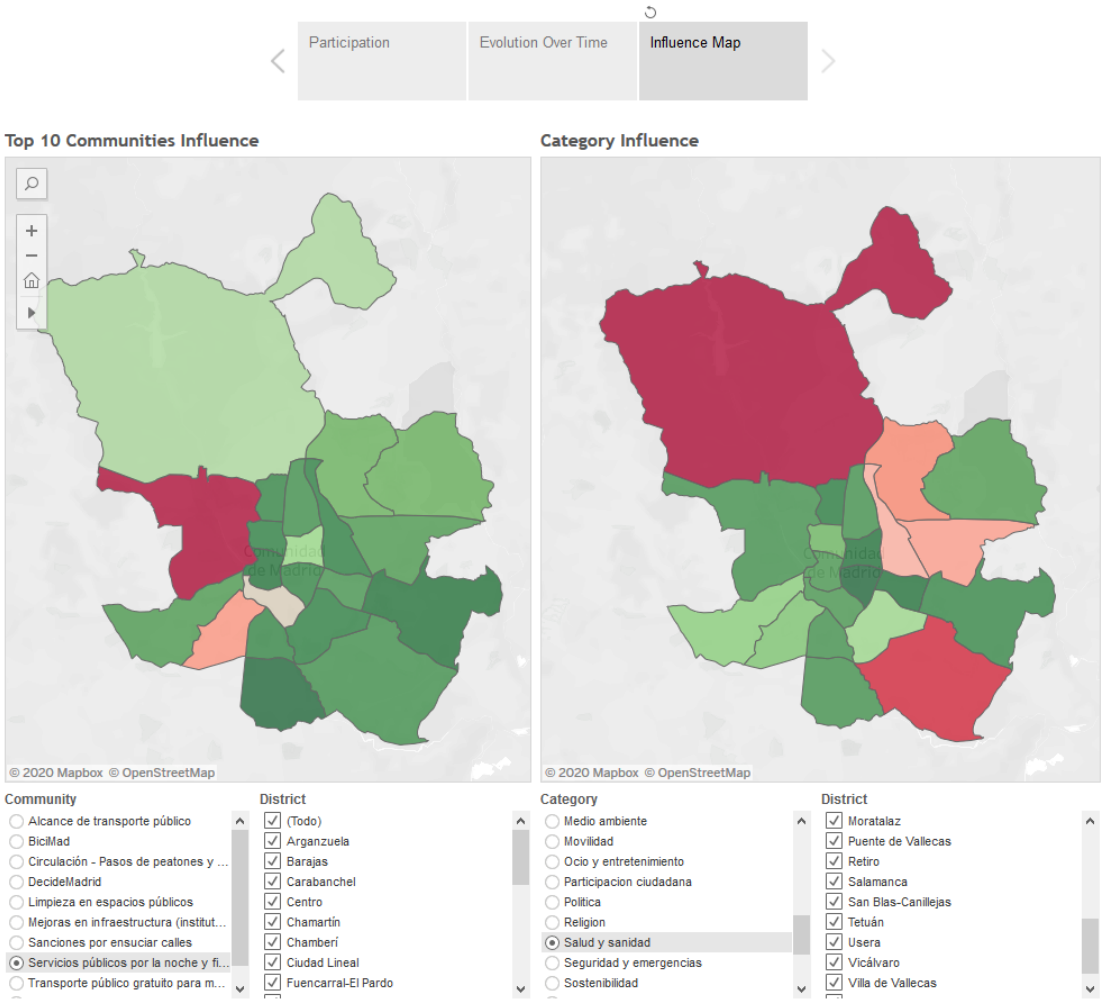


Figura 5.3.3: Mapa de calor por impacto de comunidades y categoría

Conclusiones

Durante todo este trabajo se han procesado las propuestas desde 2015 hasta 2019 de DecideMadrid, proporcionadas como datos abiertos por el Ayuntamiento de Madrid, con tal de encontrar grupos de interés en ellas. Gracias a las técnicas y métodos empleados se ha conseguido crear una serie de agrupaciones de propuestas, que permiten detectar necesidades de una forma mucho más precisa. Estas agrupaciones suponen un nivel mucho más concreto que las categorías, permitiendo informar al gobierno municipal y a los usuarios que es lo que propone el resto de ciudadanos. Además, al agrupar propuestas por similitud, posiblemente disminuirá la frustración de muchos usuarios al no ver triunfar sus propuestas debido a su duplicidad, ya que recabarán apoyos gracias a los otros miembros de su grupo (tal y como hemos visto en la Tabla 3.4.1). Aun así, se ha realizado una herramienta de uso ligero que, antes de la publicación de nuevas propuestas, sugiere la propuesta más parecida existente, de modo que se reduzca el número de propuestas duplicadas.

Agrupar propuestas ha permitido también encontrar factores que determinan el éxito de una propuesta frente a otras. Estos factores se han encontrado de forma muy similar en todas las comunidades, lo que permite hacer recomendaciones a los usuarios con tal de que mejoren el resultado de sus propuestas. Sin embargo, se ha determinado que un modelo de predicción de apoyos no es viable, debido a la existencia de diferentes factores sociopolíticos que no se han tenido en cuenta.

El análisis de estas comunidades ha mostrado que, tal y como se ha visto en varias ocasiones durante este trabajo, el tipo de propuestas que se realizan por distrito tienen una estrecha relación con factores como la situación geográfica, ingresos o partido político ganador, entre otras variables demográficas. De igual forma, se observa como todos los distritos actúan al unísono cuando, por ejemplo, se da un caso de corrupción, tal y como hemos visto en el caso de la Figura 4.2.12. Para ultimar el análisis, se ha desarrollado un panel de visualización de datos público que facilitará, tanto a gobernantes como ciudadanos, la interpretabilidad de toda esta información, contribuyendo así al derecho de la información pública y empujando a Madrid hacia la transición digital.

Aunque el resultado obtenido de las agrupaciones es satisfactorio, otro método para el cálculo de la similitud es una de las futuras líneas de trabajo. Entre las diferentes opciones, se propone el uso de BERT, que gracias a lo que ofrece podría diferenciar de forma más significativas dos documentos. Además, el método de agrupamiento supone comunidades con topología aproximadamente de estrella (con la distancia entre los documentos extremos que ello implica), de modo que otra línea de trabajo sería encontrar el valor óptimo de *min_ weight*, con tal de eliminar las aristas más débiles. En

relación con el análisis, hay resultados, como la máxima popularidad de propuestas presentadas en martes o miércoles, que abren nuevas vías de estudio. Este estudio se podría llegar a correlacionar con otras aplicaciones, comprobando si se produce el mismo efecto en otras plataformas de la misma índole.

Respecto las aplicaciones que se proponen, si se consiguiesen otras características sobre las propuestas se podría mejorar el modelo de predicción de apoyos, que adicionalmente podrían ser utilizados como filtros en el panel de visualización de datos.

Se considera que estas líneas de trabajo futuro, conjuntamente con las herramientas y resultados explicados en este trabajo, representan la puerta a la transición digital, con tal de convertir las ciudades en un modelo en el que se utilicen las TIC para garantizar la transparencia de la información publica y facilitar la coproducción ciudadana.

Referencias

[1] Adrian Rosebrock, "Intersection over Union (IoU) for object detection - PyImageSearch." `https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/`, Jun 2016.

[2] Guillaume Desagulier, "Word embeddings: the (very) basics – Around the word." `https://corpling.hypotheses.org/495`, Abr 2018.

[3] Vlad Niculae, "Word Mover's Distance in Python." `http://vene.ro/blog/word-movers-distance-in-python.html`, 2015.

[4] I. Cantador, M. E. Cortés-cediel, and M. Fernández, "Mining Open Data to Analyze Citizen Participation and Controversy in Electronic Participatory Budgeting," pp. 1–17, 2019.

[5] D.-G. F. I. Policies, "Potential And Challenges of E-Participation In The European Union," 2012.

[6] DecideMadrid, "Resultados de los presupuestos participativos 2019." `https://decide.madrid.es/presupuestos`, 2018.

[7] F. Fischer, "Participatory governance as deliberative empowerment: The cultural politics of discursive space," *The American review of public administration*, vol. 36, no. 1, pp. 19–40, 2006.

[8] L. Nyiri, D. Osimo, R. Özcivelek, C. Centeno, and M. Cabrera, "Public procurement for the promotion of r&d and innovation in ict," *Institute for Prospective Technological Studies. Luxembourg: Office for Official Publications of the European Communities*, 2007.

[9] Y. Zheng and H. L. Schachter, "Explaining citizens' e-participation use: The role of perceived advantages," *Public Organization Review*, vol. 17, no. 3, pp. 409–428, 2017.

[10] I. Susha and Å. Grönlund, "Context clues for the stall of the citizens' initiative: lessons for opening up e-participation development practice," *Government Information Quarterly*, vol. 31, no. 3, pp. 454–465, 2014.

[11] O. Publishing, *Focus on citizens: Public engagement for better policy and services*. Organisation for Economic Co-operation and Development, 2009.

[12] E. Sanchez-Nielsen and D. Lee, "eparticipation in practice in europe: The case of"puzzled by policy: Helping you be part of eu"," in *2013 46th Hawaii International Conference on System Sciences*, pp. 1870–1879, IEEE, 2013.

[13] N. Edelmann and P. Cruickshank, "Introducing psychological factors into e-participation research," in *E-governance and Civic Engagement: Factors and Determinants of E-Democracy*, pp. 338–361, IGI Global, 2012.

[14] P. Prieto-Martín, L. de Marcos, and J. J. Martínez, "A critical analysis of eu-funded eparticipation," in *Empowering Open and Collaborative Governance*, pp. 241–262, Springer, 2012.

[15] L. Forlano, "Decentering the human in the design of collaborative cities," *Design Issues*, vol. 32, no. 3, pp. 42–54, 2016.

[16] N. P. Emler, "Explaining political participation: Integrating levels of analysis," in *Political and Civic Engagement*, pp. 168–183, Routledge, 2014.

[17] D. Mathews, "Community change through true public action," *National Civic Review*, vol. 83, no. 4, pp. 400–404, 1994.

[18] G. M. Neumark, "Public administration and politics, a cultural clash: The case of tenth and monroe," 2011.

[19] N. P. Rana, Y. K. Dwivedi, and M. D. Williams, "Evaluating alternative theoretical models for examining citizen centric adoption of e-government," *Transforming Government: People, Process and Policy*, vol. 7, no. 1, pp. 27–49, 2013.

[20] N. Edelmann, J. Höchtl, and M. Sachs, "Collaboration for open innovation processes in public administrations," in *Empowering open and collaborative governance*, pp. 21–37, Springer, 2012.

[21] S. Kim and J. Lee, "E-participation, transparency, and trust in local government," *Public Administration Review*, vol. 72, no. 6, pp. 819–828, 2012.

[22] S. Marttila and A. Botero, "The'openness turn'in co-design. from usability, sociability and designability towards openness," *Smeds & Irrmann (eds) CO-CREATE*, pp. 99–111, 2013.

[23] G. A. Caldwell and M. Foth, "Diy media architecture: open and participatory approaches to community engagement," in *Proceedings of the 2nd Media Architecture Biennale Conference: World Cities*, pp. 1–10, ACM, 2014.

[24] M. Asad, C. A. Le Dantec, B. Nielsen, and K. Diedrick, "Creating a sociotechnical api: Designing city-scale community engagement," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2295–2306, ACM, 2017.

[25] I. Susha and Å. Grönlund, "eparticipation research: Systematizing the field," *Government Information Quarterly*, vol. 29, no. 3, pp. 373–382, 2012.

[26] A. Shah, *Participatory budgeting*. The World Bank, 2007.

[27] A. Novy and B. Leubolt, "Participatory budgeting in porto alegre: social innovation and the dialectical relationship of state and civil society," *Urban studies*, vol. 42, no. 11, pp. 2023–2036, 2005.

[28] S. Scherer and M. A. Wimmer, "Reference process model for participatory budgeting in germany," in *International Conference on Electronic Participation*, pp. 97–111, Springer, 2012.

[29] S. Scherer and M. A. Wimmer, "Trust in e-participation: Literature review and emerging research needs," in *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, pp. 61–70, ACM, 2014.

[30] D. R. Insua, G. E. Kersten, J. Rios, and C. Grima, "Towards decision support for participatory democracy," in *Handbook on Decision Support Systems 2*, pp. 651–685, Springer, 2008.

[31] F. Restuccia, S. K. Das, and J. Payton, "Incentive mechanisms for participatory sensing: Survey and research challenges," *ACM Transactions on Sensor Networks (TOSN)*, vol. 12, no. 2, p. 13, 2016.

[32] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text similarity in vector space models: a comparative study," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 659–666, IEEE, 2019.

[33] R. Ibrahim, S. Zeebaree, and K. Jacksi, "Survey on semantic similarity based on document clustering," *Adv. Sci. Technol. Eng. Syst. J*, vol. 4, no. 5, pp. 115–122, 2019.

[34] Adrien Sieg, "Text Similarities : Estimate the degree of similarity between two texts." `https://medium.com/@adriensieg/text-similarities-da019229c894`, jul 2014.

[35] UKPLab, "Sentence Embeddings with BERT & XLNet - Github." `https://github.com/UKPLab/sentence-transformers`.

[36] Samir Kunwar, "Text Documents Clustering using K-Means Algorithm - CodeProject." `https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm`.

[37] Alboukadel Kassambara, "Agglomerative Hierarchical Clustering - Datanovia." `https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/`.

[38] M. Nik-Bakht* and T. E. El-diraby, "Communities of interest–interest of communities: Social and semantic analysis of communities in infrastructure discussion networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 31, no. 1, pp. 34–49, 2016.

[39] M. N. Aldelaimi, M. A. Hossain, and M. F. Alhamid, "Building dynamic communities of interest for internet of things in smart cities," *Sensors*, vol. 20, no. 10, p. 2986, 2020.

[40] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowledge-Based Systems*, vol. 163, pp. 404–415, 2019.

[41] A. Eberhardt and M. S. Silveira, "Show me the data! A systematic mapping on open government data visualization," *ACM International Conference Proceeding Series*, 2018.

[42] S. Park and J. R. Gil-Garcia, "Understanding transparency and accountability in open government ecosystems: The case of health data visualizations in a state government," *ACM International Conference Proceeding Series*, vol. Part F128275, pp. 39–47, 2017.

[43] R. Barcellos, J. Viterbo, L. Miranda, F. Bernardini, C. Maciel, and D. Trevisan, "Transparency in practice: Using visualization to enhance the interpretability of open data," *ACM International Conference Proceeding Series*, vol. Part F128275, pp. 139–148, 2017.

[44] H. Amini, F. Farahnak, and L. Kosseim, "Natural language processing: an overview," *Frontiers*, 2020.

[45] "dccuchile/beto: BETO - Spanish version of the BERT model." `https://github.com/dccuchile/beto?fbclid=IwAR0WhmvOoAAMiXiT{_}ZmVUSq9sqdjAgEIcNm{_}viL-asx8PE{_}MfM2CzkkruUM`.

[46] C. Cardellino, "Spanish Billion Words Corpus and Embeddings." `https://crscardellino.github.io/SBWCE/`, Agosto 2019.

[47] S. Palachy, "Document Embedding Techniques." `https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d`, Sep 2019.

[48] D. Nag, B. Das, P. S. Dash, S. Sen, S. Paul, S. Verma, and S. K. Haldar, "Use of phenolic resin in coke making at Tata Steel," *Ironmaking and Steelmaking*, vol. 44, no. 7, pp. 526–531, 2017.

[49] I. Mokriš and L. Skovajsová, "Comparison of two document clustering techniques which use neural networks," *ICCC 2008 - IEEE 6th International Conference on Computational Cybernetics, Proceedings*, pp. 75–78, 2008.

[50] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics Theory and Experiment*, vol. 2008, 04 2008.

[51] M. Hester, S. Werner, C. Greenwald, and J. Gunning, "Exploring the effects of text length and difficulty on rsvp reading," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, pp. 1294–1298, 09 2016.

[52] Elizabeth Arens, "The Best Times to Post on Social Media in 2020 | Sprout Social." `https://sproutsocial.com/insights/best-times-to-post-on-social-media/`, mar 2020.

[53] Ayuntamiento de Madrid, "Distritos en cifras (Información de Distritos) - Ayuntamiento de Madrid." `https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Distritos-en-cifras/Distritos-en-cifras-Informacion-de-Distritos-/?vgnextfmt=default{&}vgnextoid=74b33ece5284c310VgnVCM1000000b205a0aRCRD{&}\vgnextchannel=27002d05cb71b310VgnVCM100`, 2020.

[54] RTVE, "Resultados Elecciones Municipales de Comunidad de Madrid 2015 en Centro - RTVE.es." http://resultados-elecciones.rtve.es/autonomicas-municipales/municipales/comunidad-de-madrid/madrid/madrid/centro/.

[55] S. Bachiller, "DecideMadrid - Sergio | Tableau Public." https://public.tableau.com/profile/sergio8719{#}!/vizhome/Libro1{_}15923257034100/Historia, 2020.

[56] J. Rosell, "Parque-30, la radical idea que quiere eliminar 13 kilómetros de la M-30 - EL ESPAÑOL." https://www.elespanol.com/espana/madrid/20200527/parque-30-radical-ideal-quiere-eliminar-kilometros-m-30/492951702{_}0.html.

[57] La Razón, "Dimite la número 2 de Economía de Madrid por los «papeles de Panamá»." https://www.larazon.es/local/madrid/dimite-la-numero-2\-de-economia-del-ayuntamiento-madrid-por-los-papeles-de-panama\-CH12516260/.

[58] K. Atasu, T. Parnell, C. Dünner, M. Sifalakis, H. Pozidis, V. Vassiliadis, M. Vlachos, C. Berrospi, and A. Labbi, "Linear-complexity relaxed word mover's distance with gpu acceleration," 11 2017.

Anexos

Anexo I: Metodología y tareas

Para el desarrollo del trabajo se sigue una metodología ágil, que permite añadir, eliminar o modificar tareas durante el transcurso del trabajo si es necesario. Estas características son necesarias por las limitaciones técnicas o de tiempo, o incluso si alguna de las tareas resulta inviable una vez llegue el momento. De hecho, las tareas han sufrido cambios desde la propuesta del trabajo, al encontrarnos situaciones que han supuesto dificultades en el desarrollo.

Hay que destacar que la planificación no contempla la redacción de la memoria, la presentacion oral u otras actividades inherentes a la realización del Trabajo de Fin de Máster.

Tareas	Horas
T1. Inicio del proyecto	
T1.1. Búsqueda de bibliografía en trabajos similares	10
T1.2. Extracción y preparación de los datos	40
T2. Análisis de los datos	
T2.1. Análisis de la participación	10
T2.2. Cálculo y estudio de la similitud entre propuestas	60
T2.3. Búsqueda de propuestas similares en tiempo real	30
T2.4. Clustering y estudio de las agrupaciones	40
T2.5. Búsqueda de correlacinoes y factores determinantes	20
T3. Análisis predictivo	
T3.1. Modelo predictivo de apoyos	30
T3.2. Análisis de próximas necesidades	10
T4. Presentación de los datos	
T4.1. Documentación de Tableau	5
T4.2. Desarrollo de herramienta de visualización	20
T5. Extracción de conclusiones	5
TOTAL	280