Exploiting Open Data to analyze discussion and controversy in online citizen participation

Iván Cantador[a,*], María E. Cortés-Cediel[b], Miriam Fernández[c]

[a] *Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain*
[b] *Facultad de Ciencias Políticas y Sociología, Universidad Complutense de Madrid, Spain*
[c] *Knowledge Media Institute, The Open University, United Kingdom*

Abstract

In this paper we propose a computational approach that applies data mining techniques to analyze the citizen participation recorded in an online digital platform. Differently to previous work, the approach exploits external knowledge extracted from Open Government Data for processing the citizens' proposals and debates of the platform, enabling to characterize targeted issues and problems, and analyze the levels of discussion, support and controversy raised by the proposals. As a result of our analysis, we derive a number of insights and conclusions of interest and value for both citizens and government stakeholders in decision and policy making tasks. Among others, we show that proposals targeting issues that affect large majorities tend to be supported by citizens and ultimately implemented by the city council, but leave aside other very important issues affecting minority groups. Our study reveals that most controversial, likely relevant, problems do not always receive sufficient attention in e-participation. Moreover, it identifies several types of controversy, related to ideological and socioeconomic factors and political attitudes.

Keywords: citizen participation, e-participation, online discussion, controversy, opinion polarization, Open Data

1. Introduction

Citizen participation is a process that allows individuals to be involved and influence on public opinion, and to be part of democratic decision and policy making. Representing one of the most effective and widespread forms of open governance, that process historically used to be triggered through physical interactions like meetings, assemblies, or working groups. Nowadays, it often occurs on the Internet, via online digital participatory platforms, where the citizens' opinions and contributions are easily shared, offering opportunities for communication, consultation and collaboration at an unprecedented scale (Held, 2006). Evidence, however, exists to suggest that, despite

*Corresponding author.
Email addresses: ivan.cantador@uam.es (Iván Cantador), mcorte04@ucm.es (María E. Cortés-Cediel), miriam.fernandez@open.ac.uk (Miriam Fernández)

current efforts to enhance citizen participation via online mediums, many governments nowadays still need to be better in touch with their societies and individual citizens (Zheng and Schachter, 2017).

In this context, two issues have been identified as factors influencing such detachment: (i) *the lack of citizen engagement with online participatory platforms* (Cortés-Cediel et al., 2019) and, (ii) *the lack of a deep understanding of the citizen-generated content in such platforms* (Fung, 2015). These issues affect the trust that both citizens and decision makers have on the effectiveness of the platforms and the usefulness of the collected information. They also impact on public decisions and actions, which are generally focused on popular citizen requests, rather than addressing more controversial and difficult topics of discussion (Ranchordás, 2017).

While various works have studied the issues that affect citizen engagement and participation via online platforms, they have mainly focused on understanding important issues around technology design (Cantador and Cortés-Cediel, 2018), and on providing solutions to address the challenges of accessing and exploring the large volumes of information accessible via the platforms (Aragón et al., 2018; Cantador et al., 2018). However, fewer works have really focused on understanding who participates in these platforms, how contributions and interactions emerge and develop, and what influence they have over government decisions and actions (Fung, 2015).

Targeting this gap, we propose a computational approach aimed to provide an in-depth analysis of online citizen participation. In particular, being our case study, we focus the analysis on Decide Madrid[1], the electronic participatory budgeting (ePB) platform of Madrid, Spain. This tool is built upon the CONSUL framework[2], which has been made open source by the city council, and, as far of November 2019, has been used by at least 130 institutions of 33 countries supporting 90 million citizens around the world. The tool allows residents to make, discuss and support (vote) proposals for the cities, thus deciding how to spend part of the city council budgets.

For large cities, the vast amount of citizen-generated content in this type of platforms (e.g., an average of around 6K proposals and 21K comments a year in Decide Madrid) challenges obtaining conclusions and insights about the underlying city problems, citizens' concerns, and citizen participation characteristics, such as the levels of discussion and controversy. For this reason, our analysis aims to provide answers to the following three research questions:

- RQ1: Are the most discussed and controversial proposals those that achieve the highest support?
- RQ2: What themes and types of proposals are more discussed, supported and controversial?
- RQ3: Which external factors may influence citizen participation, discussion and controversy?

Our approach brings two key innovations with respect to previous work on citizen participation analysis. First, it complements the citizens' proposals and debates created

[1]Decide Madrid platform, `https://decide.madrid.es/en`

[2]CONSUL e-participation framework, `http://consulproject.org`

in the platform with external knowledge extracted from *Open Government Data* collections. Integrating these sources of information allows for a better characterization of the problems and issues reported in the platform according to existing social, political, ideological, economical and environmental contexts. Second, our approach focuses on the analysis of *controversy*, in addition to discussion and support, as a way to better understand the complexities of the proposals and debates raised from citizen participation. Thus, our approach and analysis aim to support both citizens and government stakeholders to gain a clearer understanding of the processes in which public value could be created. Considering all the above, we claim the following contributions:

1. A computational approach to automatically process citizen-generated content of participatory e-platforms, where metadata (e.g., categories, topics, locations) are identified in textual contents, and Open Data is integrated to complement the information of such content.
2. A novel debate controversy metric that considers three forms of controversy in online discussions, namely the discussion content length, the opinion polarization, and the conversation structure.
3. An in-depth data-driven analysis of citizen discussion and controversy in a real e-participatory platform, which not only exploits citizen-generated content, but also external city-related statistical indicators gathered via Open Data.
4. The enrichment of a public dataset of 24.8K proposals and 86.1K comments generated by citizens in the Decide Madrid ePB platform, with thematic and geographical metadata. In doing so, we have also generated: (i) a taxonomy covering 325 city-related issues, organized into 30 thematic categories and, (ii) a comprehensive dataset of 1,500 streets and points of interest of Madrid, each of them with its district and neighborhood. All these resources have been made publicly available[3].

To the best of our knowledge, this is the first work that proposes a data-driven, large-scale analysis of the citizen proposals and debates that emerge from a participatory platform, exploiting Open Data to enrich the acquired knowledge, and considering controversy as a measure to uncover relevant topics of discussion. We believe that our approach can be applied and adapted to other e-participation tools, and that the generated resources and achieved conclusions and insights can be of great value to other researchers and practitioners in a variety of fields, such as sociology and political sciences.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 presents Decide Madrid, the case study selected for this work, as well as the developed research framework. Section 4 describes the datasets used in our study, including the data extracted from the Decide Madrid platform, and the selected Open Government Data collections. Next, Section 5 introduces the proposed controversy metric, and Section 6 presents the conducted large-scale data-driven analysis. Finally, Section 7 provides conclusions and future research lines derived from our work.

[3]Generated datasets, `http://ir.ii.uam.es/egov`

2. Related work

In this section we revise previous work related to two main aspects of our research, namely the relationships between citizen participation, ePB and Open Data, and the identification and measurement of controversy in online discussions.

2.1. Electronic citizen participation and participatory budgeting

Citizen participation is a process that allows individuals to be involved and influence on public opinion and to be part of democratic decision and policy making. According to the model proposed by the Organization for Economic Cooperation and Development, OECD (Peña-López et al., 2001) citizen participation is understood as a spectrum in which the role played by residents regarding a city-related project or initiative may range from just recipients of information (*information level*) to decision makers (i.e., at *collaboration* and *participation levels*), going through intermediate levels in which citizens are consulted, but final decisions are taken by the government (*consultation level*), or where they formally express interests and requests (*petition level*).

With the emergence of new Information and Communication Technologies (ICTs), a shift has been made from face-to-face citizen participation to an electronic citizen participation, so called e-participation (Boudjelida et al., 2016), where different technologies are used to conduct or support participation initiatives, including: ad-hoc e-platforms, mobile apps, living labs, social media and gamification, among others (Cortés-Cediel et al., 2019). In this context, one of the followed mechanisms to involve citizens in decision making (at collaboration and participation levels) via the use of ICTs has been the so called electronic Participatory Budgeting (ePB), in which online platforms allow citizens to participate in processes targeted to spend municipal or public budgets on initiatives and projects in different domains, such as housing, public safety, education, health, transport, and environment, to name a few.

Since its appearance in Porto Alegre, Brazil, in 1989, with the aim of improving redistribution, inclusion, social cohesion, and accountability (De Sousa Santos, 1998), Participatory Budgeting (PB) processes have been implemented in more than 1,500 cities around the world (Baiocchi and Ganuza, 2014). A wide range of digital tools have also been created to enable ePB[4]. In addition to such tools, several software frameworks are also used to build online PB platforms, such as: (i) CONSUL citizen participation tool[5] -an open-source framework supported by the City Council of Madrid (Spain), which is used in tens of cities in Spain, Italy, France and South America-, (ii) Stanford Participatory Budgeting tool[6] -an open-source framework used in PB digital platforms of major cities in the USA, e.g. New York, Chicago, Seattle, Oakland and Boston- and, (iii) EU Open Budgets participatory budgeting tool[7].

However, despite this growth and success, and the fact that digital platforms for PB are core elements towards the creation of public value, studies have pointed out that

[4] `https://www.demsoc.org/wp-content/uploads/2016/01/DS-Digital-Tools-paper.pdf`
[5] CONSUL citizen participation, `http://consulproject.org/en`
[6] Stanford Participatory Budgeting, `https://pbstanford.org`
[7] EU Open Budgets, `http://openbudgets.eu/tools`

the level of participation in these platforms is still low (Zheng and Schachter, 2017). Hence, existing studies have attempted to address such low participation in ePB platforms by analysing and enhancing technology design and functionalities (Cantador and Cortés-Cediel, 2018; Cantador et al., 2018; Aragón et al., 2018). As opposed to previous work, our approach deeps in the understanding of the citizens' discussions and contributions in such platforms. In this line, Boudjelida and Mellouli (2016) presented a theoretical framework for the collection, processing and analysis of contents generated in e-participation tools. However, they did not empirically assess the framework in a real-world case study. In the particular context of participatory budgeting, Boukhris et al. (2016) presented a multi-criteria decision making tool to provide decision makers with the best alternatives based on citizens' opinions, but did not evaluate the tool with Open Government Data and in a large scale scenario, as we do in this work. Hence, our empirical analysis does not only exploit the content generated by citizens, but also related external knowledge, i.e., demographic, socioeconomic and political variables.

2.2. Open Government Data

The access to public information is a core element of e-government and smart governance strategies to achieve transparency and accountability (Nam and Pardo, 2011), and Open Data represent a principal instrument that enables it.

As explained by Janssen et al. (2012), Open Government Data have other benefits, which range from political and social (e.g., self-empowerment of citizens, trust in government, stimulation of knowledge developments, and improving of policy making), to economic (e.g., growth and stimulation of competitiveness and innovation, and improvement of processes, products and services) and operational (e.g., reuse of data, and improvement of public policies).

Despite all these benefits, Open Data also have a number of adoption barriers, such as the fact that data only become valuable when used, and that there is a lack of user capabilities to reuse and analyse data (Gascó-Hernández et al., 2018). Indeed, little is known about the conversion of public data into services of public value (Janssen et al., 2012), and few open government data portals provide consumption functionalities apart from simple data downloads (Attard et al., 2015). In this context, Janev et al. (2014) explored issues and challenges related to the integration and analysis of Open Data, proposing a linked data approach to modeling, merging and analyzing data, and Jetzek et al. (2014) proposed a model where various processes within an Open Data system generate sustainable value, considering contextual factors that motivate and allow stakeholders to use and create data.

Open Government Data serve citizens to have access to public information that facilitates decision making, and allow them to generate part of that information (Hivon and Titah, 2017). This enables the development of valuable e-government services, related for example with the efficient management of city resources and the provision of public services (Gagliardi et al., 2017). Differently to these applications, to the best of our knowledge, our work represents the first use of Open Data to conduct a social analysis of city problems under the perspective of the citizens concerns, opinions and suggestions.

2.3. Controversy in online discussions

Nowadays, there is a plethora of social media platforms (e.g., social networks, internet forums, and review websites) that enable users to provide opinions and participate in online discussions. These platforms facilitate the creation of different types of discussions in terms of content and structure. From a content perspective, platforms like Twitter facilitate the creation of general-purpose discussions, while others, like Reddit, cluster discussions that are more topically-focused. In terms of structure, most of the platforms capture discussions as conversational threads, where users post messages and reply to comments from others. In general, a thread is associated with a particular root post and presents a tree structure for its comments. Moreover, whereas some platforms lead to the creation of multi-threaded discussions in which information of who is replying to whom is maintained, others capture a single thread where all comments are grouped together. There are platforms that also allow users to provide metadata for the discussions, such as social tags, votes and ratings.

Due to the richness and relevance of online discussions as a resource to better understand public opinion and user needs, a variety of scientific works has emerged in the last years with the aim of investigating different aspects of the discussions, including: their topics, the events that spike them, the opinions that emerge from them, and how virally the information spreads. In this paper we focus on measuring debate controversy.

As stated by Zielinski et al. (2018), detecting controversy and controversial themes in social media through automatic methods is especially important, since presenting users with the indications and explanations of the controversy generated by the content they consume allows them to see the "wider picture" instead of leading them to obtain one-sided views. The authors summarize how controversy has been explored from multiple lenses, including social science, traditional media, social media, and Web search. They propose a formal definition and representation of controversy based on three variables: the object of discussion, the group of people that discuss it, and the distribution of opinions. In our work, the objects of discussion are the citizens' proposals within Decide Madrid, the people discussing are the residents of Madrid, and the distribution of opinions is reflected on the votes that citizens give towards the comments given on the proposals.

To measure the controversy of online discussions, various types of metrics can be found in the literature, namely (i) content-based metrics, (ii) opinion polarization-based metrics, (iii) conversation structure-based metrics, (iv) social network-based metrics, and (v) meta information-based metrics. Content-based metrics take into account the length of the messages as well as the vocabulary used in them. **Content-based metrics** are based on the use of lexicons (Mejova et al., 2014; Roitman et al., 2016), controversial topics and terms (Popescu and Pennacchiotti, 2010), language models (Jang et al., 2016), and word embeddings (Rethmeier et al., 2018). **Opinion polarization-based metrics** consider the degree of discrepancy between positive and negative opinions. Bramson et al. (2016) proposed a variety of metrics to compute this discrepancy in online discussions including spread, dispersion, coverage, regionalization and community fragmentation. Rethmeier et al. (2018), on the other hand, proposed a metric based on vote agreement ratios, where 2/3 majority of either agreeing or disagreeing votes is considered controversial. **Conversation structure-based metrics** focus on the struc-

tural characteristics of the trees that form the conversation threads, such as the number of nodes in the tree, its depth, and its level completeness. Among them, we highlight the H-index based metric presented by Gómez et al. (2008), which jointly considers a conversation tree size and depth to measure controversy. **Social network-based metrics** are based on connections between users. They are computed over the social graph of interactions, where two users are connected by an edge if they have interacted with each other. Examples of metrics of this type are those proposed by Popescu and Pennacchiotti (2010), Rad and Barbosa (2012), Lo et al. (2013) and Garimella et al. (2018). Lastly, **meta information-based metrics** make use of metadata to determine controversy. For example, Dori-Hacohen and Allan (2015) measured controversy of web pages mapping these pages to Wikipedia articles, in which topic controversy can be measured. In this context, Zielinski et al. (2018) and Rad and Barbosa (2012) computed controversy in Wikipedia by mining a variety of metadata from logs, including revisions, edits and changes on Wikipedia discussion pages.

For our study, we will consider and combine three metrics to measure the controversy of a proposal: the length of its discussion as content-based metric, a weighted ratio measuring the difference of its positive and negative comments as opinion polarization-based metric, and its H-index controversy as conversation structure-based metric. We could not consider a social network-based metric since the data used was anonymized and thus did not contain user information. We neither considered a meta information metric, since the Decide Madrid dataset does not provide user activity data, such as log records. Nonetheless, to the best of our knowledge, ours is the first attempt to measure controversy in terms of several types of features, and apply it to analyze moderated citizen debates in ad hoc e-participation tools, instead of social networks and Wikipedia, as commonly done in the literature.

3. Research overview

In this section we introduce Decide Madrid, the case study selected for our analysis, and the research framework that guided our work.

3.1. Case study

In September 2015, the City Council of Madrid launched Decide Madrid, its electronic participatory budgeting platform. Through this tool residents in Madrid can submit proposals of projects and initiatives for the city around a variety of topics, such as urbanism, public transport, healthcare, education and culture. The proposals can be both debated and voted in the platform. Debates do not trigger a specific action by the city council, but represent a useful way of gauging public opinion. Votes, on the other hand, allow citizens to show support for particular proposals. Those proposals obtaining enough supports are assessed and, if accepted, ultimately implemented by the city council. The budget allocated to these proposals was 100 million euro in 2019[8].

The selection of Decide Madrid ePB platform as a representative case study of e-participation has a twofold motivation. Firstly, participatory budgeting is among

[8] `http://www.madridforyou.es/en/decide-madrid-web-platform`

the most used citizen participation methods worldwide. From a total of around 1,400 study cases available in Participedia[9] –a collaborative wiki-based website that presents citizen participation initiatives all over the world–, more than 400 cases consisted of PB initiatives. Also, according to the Participatory Budgeting Project[10], more than 3,000 cities and municipalities worldwide have implemented PB processes. Secondly, Decide Madrid follows a standard structure and architecture of ePB tools (see e.g. the Stanford Participatory Budgeting and the EU Open Budgets tools). It consists of web pages showing proposal metadata (including user generated content, such as social tags), debates about proposals, and supports/votes for proposals.

As an illustrative example, Figure 1 shows a screenshot of the Decide Madrid web page associated to a citizen's petition for a single ticket to use any public transport (lit. "billete único para transporte púlico") in the city, i.e., bus, metro and train, and thus to ease the intermodality in a long period of time of at least 90 minutes and without increasing the ticket price. The web page provides a variety of information about the proposal, such as its title, author's nickname, description, complementary documents, and labels (expressed as free-text tags given by the author). Our data mining approach will process non only the tags, but also keywords in a proposal title in order to semantically annotate the proposal with thematic categories and topics and locations, such as city districts, neighborhoods and streets. In the example, the proposal is annotated with *public transport* and *sustainability* topics, *Mobility* and *Sustainability* categories, and *All city* location. The web page also shows the proposal status, and its number of supports and positive and negative votes. The bottom of the page contains the raised debates (discussion threads) on the proposal, which consist of trees of comments, each of them with positive and negative votes, provided by citizens registered in the platform. As we will explain, the length of the comments, the structure of the discussion threads, and the opinion polarization expressed by the positive and negative votes, are used by our approach to establish the relative levels of discussion and controversy of the proposals.

The overall participation process in Decide Madrid follows three main phases, namely submission, support and vote. In the *submission phase*, any resident can create a proposal by signing up to the platform, and filling a simple questionnaire specifying the proposal title, summary and description, as well as optional information, such as social tags. The *support phase* aims to prioritize the most interesting and relevant proposals. For such purpose, proposals that obtain support in the platform by more than 1% of residents aged 16 or over in a period of 30 days are approved; the remainder proposals are discarded and archived. Approved proposals are then commented and discussed by the citizens in the platform during a period of 45 days. Finally, in the *vote phase*, during a period of one week from its approval date, each approved proposal can be voted by residents. In case there are more people in favor than against, a proposal is accepted as a 'collective proposal' of Madrid citizens, and the city council government assumes it as its own carrying it out. To achieve this, within a maximum period of one month, the corresponding technical reports on feasibility, legality and economic cost of the proposal are published on the web. Then, citizens can access the plan to accomplish

[9] Participedia community sharing knowledge about public participation, `https://participedia.net`

[10] `https://www.participatorybudgeting.org/white-paper`

Figure 1: Screenshot of a Decide Madrid web page associated to a citizen's proposal for a single ticket to use any public transport (lit. "billete único para transporte púlico") in the city. Translations of some of the user comments on the proposal are given in Table 5.

the proposal and track its progress.

As far of November 2019, Decide Madrid has more than 420,000 registered users. The platform not only enables public participation in decision making, and constitutes a rich forum of debate where citizens discuss issues that are important to them. However, exploring the emerged discussions to gather a more in-depth knowledge of the city problems and citizens' concerns is a challenging task, which could be of vital importance to better inform decision and policy makers. Motivated by this situation, in this paper we aim to provide a computational approach able to automatically generate comprehensive analyses of the debates happening in the platform, putting a particular focus on the controversial aspects.

3.2. Research framework

To address the stated research questions, in this work, we shall analyze the citizens' proposals and debates existing in the Decide Madrid ePB platform, in terms of their themes, locations, discussion characteristics, and city-related statistical indicators.

More specifically, we will apply data processing and mining techniques on textual contents of the proposals, and will compute a number of metrics over the discussion threats of the debates. For such purpose, we will make use of external knowledge extracted from several Open Data collections provided by the City Council of Madrid.

In subsequent sections, we will detail the exploited Open Data collections, followed techniques, generated datasets, and computed discussion analysis metrics. Before that, for clarity purposes, we next depict the components and stages of our research framework, illustrated in Figure 2:

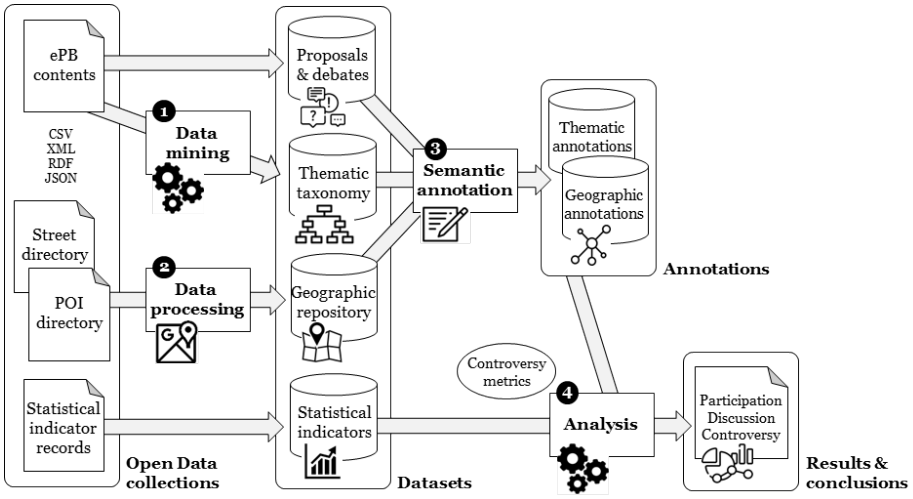

Figure 2: Proposed research framework.

- *Processing Open Data collections published by Madrid City Council*. Available in different formats (e.g., CSV, XML, RDF and JSON), we selected the following collections:
  - Citizen proposals, debates and associated metadata (e.g., proposal supports and social tags, and positive/negative comment votes) in the Decide Madrid ePB platform.
  - Geographic information of Madrid, such as its districts and neighborhoods, a street directory, and a list of touristic Points of Interest (POIs).
  - Statistical indicators of Madrid on several dimensions: demography, economy, employment, education, health, mobility, environment, and citizen participation, among others.

  As a result, we generated a number of datasets, stored into a single relational database.

- *Applying a clustering method on the ePB collection to automatically generate a thematic taxonomy* (stage 1 in the figure), later used to categorize the proposals with respect to the city issues they address.

Proposals and debates		**Thematic annotations**	
#proposals	24,867	#proposals with category	22,417 (90.1%)
#comments	86,102	#category-based annotations	4,7817
#tags	4,137	Avg. #categories/proposal	2.127
#tag assignments (TAS)	59,837	#proposals with topic	22,417 (90.1%)
#tagged proposals	18,623 (75.1%)	#topic-based annotations	55,243
Avg. #TAS/proposal	3.213	Avg. #topics/proposal	2.464
Thematic taxonomy		**Geographic annotations***	
#categories	30	#proposals with district	10,857 (43.7%)
#topics	325	#proposals with neighborhood	3,960 (15.9%)
#tags	1,826 (44.1%)	#proposals with street/POI	1,409 (5.7%)
Avg. #tags/category	80.867	#tags	235 (5.7%)
Avg. #tags/topic	5.618		
Geographic repository		#proposals with topic & district	9,785 (39.7%)
#districts	21		
#neighborhoods	129		
#streets	1,409		
#POIs	71		

* The majority of proposals without geographic annotations are applicable to the whole city, so they cannot be assigned with any particular district, neighborhood, street or POI.

Table 1: Statistics of the generated datasets. The term 'street' refers to any road type or infrastructure: street, avenue, boulevard, town square, bridge, etc.

- *Processing and integrating the street directory and POI collections into a geographic repository* (stage 2). By making use of Google Maps service, every street (i.e., street, avenue, boulevard, town square, bridge, etc.) and POI is assigned with its corresponding district and neighborhood.
- *Performing a semantic annotation process over the citizens' proposals by means of the generated thematic taxonomy and geographic repository* (stage 3). More specifically, title keywords and social tags were mapped to categories, topics, districts, neighborhoods, streets, and POIs.
- *Computing discussion and controversy metrics on the debates*.
- *Conducting a number of analyses to address the stated research questions* (stage 4). The analyses were done by mining both the generated semantic annotations and collected statistical indicators.

4. Generated datasets

In this section we briefly describe the generated datasets, which we make available online and whose statistics are shown in Table 1. In addition to the dataset with processed ePB contents, the thematic taxonomy with 30 categories and 325 topics about city issues, and the geographic repository with almost 1,500 streets and POIs of Madrid with their corresponding districts and neighborhoods, represent useful resources for a wide range of research and development purposes.

As explained in the previous section, we downloaded a number of data collections from the Open Data website[11] of Madrid City Council:

- *Decide Madrid*[12]. This collection contains 24,867 proposals and their associated 86,102 comments provided by residents during the city PB processes from 2015 to 2018. Each proposal has a title, a summary, a description, social tags (forming a set of 4,137 keywords and 59,837 annotations), and the number of received supports. The comments form discussion threads following a tree structure, and do have positive and negative votes given by the platform users.

- *Madrid street*[13] *and touristic POI*[14] *directories*. These collections contain lists of streets and points of interest and associated information. Among other data, the streets have assigned a district and neighborhood; in Madrid, there are 21 districts, each of them with several neighborhoods, being 129 the total number of neighborhoods in the city.

- *Madrid statistical indicators*[15]. This collection contains district statistics on a number of variables in different dimensions, as explained in the analysis section.

4.1. Proposals and debates dataset

When a proposal is submitted to the Decide Madrid platform, the author has to provide a corresponding title, summary and description. Optionally, as additional metadata, the author may also assign some tags to the proposal. These tags are freely chosen words that can reflect either a topic or a location, among others. Since the tags are not linked to predefined concepts or categories, their meaning is not formally represented. We thus developed data processing and mining methods to build a thematic taxonomy and a geographic repository whose elements could be mapped with title keywords and social tags. The obtained mappings would represent semantic annotations about the topics and locations of the proposals.

4.2. Thematic taxonomy

To build the thematic taxonomy, we manually set a total of 30 categories such as urban planning, sustainability, housing, health, and old age (see the full list in Table 2) which correspond to departments and service areas of Madrid city council. Then, we selected the 150 most popular tags in the Decide Madrid dataset and manually assigned each of them to the most appropriate category. As a result of this process each category has a set of seed tags associated to it.

To increase the number of tags associated to each category we automatically computed lexicographic similarities between the seed tags and the rest of the tags available in the Decide Madrid dataset. We used the Levenshtein distance to identified groups of tags corresponding to the same concept. For example, we map *asociación*, *asociacion*,

[11] Madrid Open Data, `https://datos.madrid.es/portal/site/egob`
[12] Decide Madrid collection, `https://datos.madrid.es/egob/catalogo/300312-0`
[13] Madrid street directory, `https://datos.madrid.es/egob/catalogo/213605-0`
[14] Madrid touristic POI directory, `https://datos.madrid.es/egob/catalogo/300030-10037182`
[15] Madrid statistical indicators, `https://datos.madrid.es/egob/catalogo/300087-2`

asociaciones, and *asociacionismo* for the 'association' concept. By conducting this process we automatically extended the list of tags associated to each of the categories.

Next, we built a graph based on the co-occurrences of tags in the proposals of the Decide Madrid dataset. To build this graph we considered as nodes the tags, and as edges co-occurrences between pairs of tags, where each edge was assigned a weight corresponding to the co-occurrence value between its linked tags (nodes). On the graph, we then applied the clustering method proposed by Newman and Girvan (2004), which has a criterion to automatically set an optimal number of clusters. Each cluster represents a topic, which is composed by a set of tags.

Lastly, we computed the tag overlap between each cluster (topic) and each category, and assigned each cluster to the category with which it has the highest overlap. Thus, the tags of a cluster were incorporated into the set of tags of the category.

It is important to note that a cluster could sometimes represent more than one topic associated to one or more categories. By manual inspection, we split some of the clusters (i.e., subsets of tags) to generate more accurate, detailed topics. In some cases, we also moved a cluster (topic) from one category to another. As a result of the whole process[16], a total of 1,826 tags were assigned to the 30 categories, and 325 topics were generated, each of them with an average of 5.618 tags. Tables 3 and 4 show examples of such topics.

4.3. Geographic repository

In the downloaded geographic Open Data collections, the streets were provided with their corresponding districts and neighborhoods, but POIs did not have such information. We obtained it from Google Maps platform[17]: Searching for a POI by name, Google service returned its address. Next, looking for the address in our street directory, we got the corresponding district and neighborhood.

Differently to the thematic tags, the identification of tags referring to locations was done in a quite straightforward way. We just searched for the tags in the entries (streets and POIs) of the geographic repository. To obtain search matches, all tags and entries were converted into lowercase, and their acute accents and special symbols were removed.

4.4. Proposal Annotations

Once the thematic taxonomy and geographic repository were built, we proceeded with the semantic annotation of the proposals. Table 1 shows statistics about the generated annotations. The followed method consisted of finding any entry of the above datasets as an exact matching on a title keyword or a social tag of each proposal. Similarly to the geographic matching, in this case, all titles and tags were converted into lowercase, and their acute accents and special symbols were removed.

In this context, it is important to note that we discarded the annotation of the proposals descriptions and summaries, since they also cited topics and locations distinct to

[16]The social tags mapped to entries of the geographic repository were discarded in the process.

[17]Google Maps platform, `https://cloud.google.com/maps-platform`

those of the proposals, and thus originated many wrong annotations. As a result, in addition to a very high coverage (90.1% of the proposals were assigned category/topic annotations), the annotations were very accurate: 99.05% and 100.00% precision on topic and district annotations respectively, from a manual evaluation over 3,573 annotations on 1,000 randomly selected proposals, conducted by 3 experts with a Fleiss' kappa agreement coefficient of 0.99. The wrong topic annotations identified by the assessors corresponded to (i) ambiguous words, e.g., 'banco' as bank or as bench; (ii) nouns that should be part of compound nouns, e.g., 'coche' (car) instead of 'coche de policía' (police car), 'parque' (park) instead of 'parque infantil' (playground), 'caminos' (roads) instead of 'Cuatro Caminos' (a street in Madrid); and (iii) correct words that are not the main focus of the proposal, e.g., 'city council' in 'cleaning service of the city council.' The assessors' disagreements, on the other hand, were subtle differences on the most appropriate topic annotations of the proposals, e.g., 'cycling' vs. 'BiciMAD' for a proposal mentioning municipal bicycles, and 'tourism' vs. 'immigrants' for a proposal aimed to increase tourism in the so-called Chinatown of Usera, a neighborhood in the South of Madrid, which nowadays is a commercial area.

5. Developed Controversy Metrics

In this section we present the metrics used to analyze the controversy of a citizen proposal submitted and commented in the Decide Madrid ePB platform. As already mentioned, for a given proposal, our metrics are defined in terms of the length of the comments, the opinion polarization expressed by the positive and negative votes on the comments, and the structure of the conversation trees formed by the comments threads.

The proposed metrics are motivated by the sources of information available in Decide Madrid in particular, and in ePB and other e-participation platforms in general: debate threads and votes. Since, to the best of our knowledge, there is no study that shows which type of feature is most appropriate for the addressed case study, we decided to jointly consider several forms of controversy. Hence, from a practical point of view, we make use of an aggregated metric that allows us to address the stated research questions. We could also consider the opinions expressed by citizens in the comments of the debates. In order to avoid errors derived from natural language processing and opinion mining, we leave such potential signal of controversy for future investigation. Nonetheless, based on the revision of the research literature on controversy in online discussion, we confirmed that the used sources of information had been already explored. The proposed metrics are indeed implementations and adaptations of previous metrics, as we explain next.

For a better understanding of the metrics, we refer the reader to Figure 3, which illustrates pairs of comments threads for which one of threads (the one on the left) is less controversial than the other (the one on the right).

Capturing such notions and characteristics of controversy, we decided to explore metrics of different types described in the Related Work section. Specifically, we selected the following types and metrics:

- *Discussion content-based metric*: The length of the discussion of proposal p,

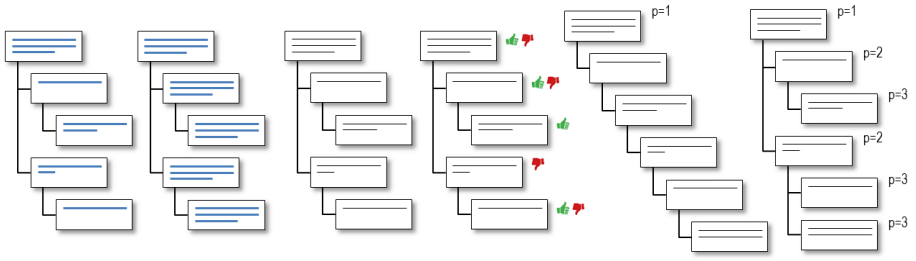

Figure 3: Illustration of the 3 developed controversy metrics on pairs of comments threads. For each metric, the thread on the left is less controversial than the thread on the right.

measured as the sum of the length of its comments c:

$$controversy_1(p) = \sum_{c \in comments(p)} length(c)$$

This metric, used as a controversy feature by several authors (Dori-Hacohen and Allan, 2015), assumes that the longer a discussion about a particular proposal, the more controversial the proposal.

- *Opinion polarization-based metric*: A weighted ratio measuring the difference of positive and negative votes for the comments of a proposal p:

$$controversy_2(p) = 1 + min(pos(p), neg(p)) \cdot \frac{min(pos(p), neg(p))}{max(pos(p), neg(p))}$$

where $pos(p) = \sum_{c \in comments(p)} posVotes(c)$ and $neg(p) = \sum_{c \in comments(p)} negVotes(c)$, being $posVotes(c)$ and $negVotes(c)$ the number of positive and negative votes given to comment c, respectively. If p has no comment vote, $controversy(p) = 0$. This metric corresponds to the "size parity" notion of controversy given by Bramson et al. (2016) which assumes that distinct opinion polarity samples are more polarized (controversial) if they have comparable sizes. To implement this notion of controversy, considering the agreement between the positive and negative votes given to comments was for example done by Rethmeier et al. (2018).

- *Conversation structure-based metric*: The adaptation of the H-index proposed by Gómez et al. (2008) for measuring discussion controversy:

$$controversy_3(p) = \sum_{n=1}^{depth(p)} H(width(p, n) \geq n) + \frac{1}{1 + |comments(p)|}$$

where H is the Heaviside step function, i.e., $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ if $x < 0$, $depth(p)$ is the depth of the discussion thread (tree) of proposal p, and $width(p, n)$ is the number of comments (nodes) at level n of the discussion thread (tree) of p. This metric measures controversy by considering both the degree of completeness and the depth of the discussion tree.

In order to assess that these metrics reflect distinct characteristics of the discussions, we computed the average linear correlation values between each pair of metrics for all the proposals of our dataset. In particular, the discussion content metric had linear correlations of 0.626 and 0.435 with the *opinion polarization* and *conversation structure* metrics, respectively, whereas the *opinion polarization* metric had a correlation of 0.342 with the *conversation structure* metric.

These positive and moderate correlation values can be considered as a signal of the existence of related, but different controversy features captured by the metrics. Hence, they also allow us to propose a new controversy metric consisting of the aggregation of the normalized scores of the three metrics:

$$controversy(p) = \frac{1}{3}\sum_{i=1}^{3}\frac{controversy_i(p)}{\arg\max_{p'} controversy_i(p')} \in [0,1]$$

This proposed metric is the one we finally used in our analysis, presented in the next section. We have to note that we tested other metrics, such as the *number of comments* and the *depth of the discussion thread*, but we discarded them due to their very high/low correlation values with the above metrics. For instance, the *number of comments* and the *discussion length* metrics showed a correlation of 0.974, and the *depth of the discussion* and the *opinion polarization* metrics showed a correlation of 0.193. We also note that, differently to ours, the metrics proposed in the literature only exploit a single type of feature.

6. Analyzing citizen participation and controversy

In this section we present the results of the analysis conducted to address the research questions that have motivated our investigation. Three key variables are considered in the analyses, namely the number of comments, the number of supports, and the degree of controversy (measured by the proposed aggregated metric) of the proposals in the Decide Madrid dataset. The presented analyses and results do not consider a yearly split of the data since we observed no meaningful differences between discussion and controversy distributions during the four analyzed years, 2015-2018. In addition, the global analysis of all proposals enables a better comprehension of the strength of the controversial themes that have arisen in the platform since its implementation.

6.1. Levels of discussion and controversy

The goal of RQ1 is to answer if the most discussed and controversial proposals are those that achieve the highest support in ePB, and are further selected by the city council for their implementation. To address this question we first analyze to what extent there is discussion in the Decide Madrid ePB platform.

Despite the arguments in favor and against (electronic) participative budgeting, and its benefits and limitations, there is a general agreement that nowadays the levels of citizen involvement in ePB processes are still low (Zheng and Schachter, 2017). Among the variety of elements that may influence this limited participation, several authors have highlighted issues related to the technical designs of the used e-platforms (Aragón et al., 2017; Cantador and Cortés-Cediel, 2018), and have proposed novel solutions in

terms of content visualization and search and recommendation mechanisms (Cantador et al., 2017, 2018), to increase the citizens' engagement.

In principle, an opposite situation may be claimed for Decide Madrid: After 4 years of operation, the platform has more than 420,000 users registered and more than 6,000 proposals recorded per year. However, we conducted a simple data analysis that does not allow us to confirm such situation. The Decide Madrid Open Data collection provided by the city council has no user information, so we cannot establish how many users really participated and how they did it, for example by measuring the activity (e.g., logins, browsed proposals, supports, comments, and votes) per user within the platform. We, in contrast, can analyze the **level of discussion** maintained by participants. In particular, Figure 4a (in logarithmic scale) shows the number of comments of each proposal in the dataset. We can observe that there exists a heavy tail distribution where a relatively low number of proposals receive a high number of comments, and a majority of proposals receive less than 10 comments. More specifically, out of a total of 24,867 proposals, 823 have 10 or more comments, 488 have 20 or more comments, and 99 have 50 or more comments. According to these numbers, we also note that the level of discussion in Decide Madrid is relatively low.

Even being more interesting, we can analyze the **level of controversy** of the proposals, as expressed in their comments threads. Figure 4b shows a scatter plot of the proposals in the dataset, with respect to their number of supports and controversy score (computed with the aggregated metric proposed in Section 5). The figure shows that, similarly to the discussion levels (measured by the number of comments per proposal), the controversy scores also follow a heavy tail distribution where a majority of the proposals have low controversy (i.e., controversy scores below 0.1), and a few proposals show high controversy (i.e., controversy scores over 0.3). In this context, we remind that most discussed proposals are not necessarily the most controversial. Our aggregated controversy metric not only considers the discussion length, but also the opinion polarization and the conversation structure associated to the comments threads.

The figure also allows us to give an answer to our first research question, claiming that **highly supported proposals are not necessarily the most controversial**. Indeed, the computed linear correlation between the level of support received by a proposal and its level of controversy is low (0.493). This situation also applies to the level of discussion: in general, highly discussed proposals do not always correspond to those that receive large number of comments. As commented in Section 5, the number of comments highly correlates (0.974) with the discussion length-based controversy metric.

In the Decide Madrid platform, proposals with low level of support are currently discarded and archived, independently on their levels of discussion and controversy. We believe, in contrast, that from a decision or policy making perspective, it would be important to take a deeper look into the controversial proposals, and understand the city issues and problems they uncover and the citizens to whom they affect. Motivated by this thought, and aiming to better understand citizen participation in ePB, we next analyze in detail the themes and types of proposals that show more/less discussion and controversy.

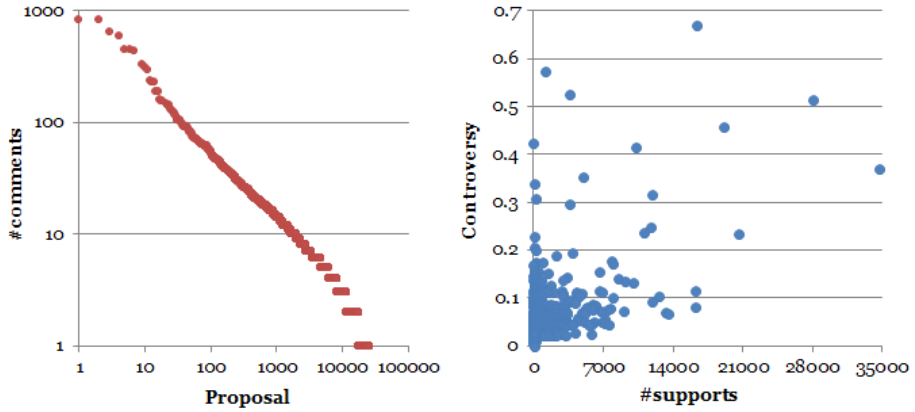

Figure 4: (a) Number of comments per proposal (the axes are in logarithmic scale); (b) Scatter plot of the proposals by their number of supports and controversy scores.

6.2. Discussed and controversial themes and proposals

By stating RQ2 we aim to find out which are the themes and types of proposals that are more discussed, supported and/or controversial in ePB, and whether they are the same.

Our computational approach was able to automatically assign categories and topics to 90.1% of the proposals of the Decide Madrid Open Data collection. Analyzing the number of proposals, average number of supports per proposal, and average controversial score per proposal of each category, we provide first answers to the above question. Table 2 shows these values. For clarity purposes, the categories are sorted by decreasing controversy score, and are clustered into 5 groups, according to their relative values of the analyzed variables.

The first group captures categories that, despite having a **low number of proposals** associated to them, receive **very high support and controversy**. These categories include religion, housing, culture and tourism. Conducting a qualitative analysis of the proposals associated to these categories, we observe a high controversy around the decision of Madrid city council to change two Christmas traditions: (i) the city placement of the nativity scenes, a special exhibition traditionally located at the Puerta de Alcalá, to a different location, and (ii) the inclusion of LGBTI (lesbian, gay, bisexual, transgender and intersex) groups in the traditional Three Wise Men parade. These actions were considered disrespectful by a fraction of the residents, especially those of Catholic faith, and derived in a series of proposals within Decide Madrid to maintain the traditions. Other controversial proposals related to religion focused on debating the public funding and tax benefits provided to Catholic institutions. Controversies around housing included the prices of house rental, the creation of social housing, and the annual property taxes. Regarding culture, the most controversial proposals were about bullfighting and whether it should be forbidden. Lastly, with respect to tourism, the majority of proposals targeted the creation of a tourist tax for visitors staying in hotels, and critiqued the lack of the government control of houses rented to tourists. With all

Group	Category	#proposals	Avg. #supports	Avg. controversy
1	Religion	84	828.345	0.052
	Housing	195	346.267	0.038
	Culture	333	213.652	0.037
	Tourism	105	196.743	0.036
2	Laws & legislation	309	172.421	0.036
	Social rights	669	167.543	0.035
	Public administration	492	197.004	0.035
	Citizen participation	402	82.948	0.035
	Politics	157	167.898	0.035
	Civics	89	112.798	0.035
	Equity & integration	598	173.709	0.034
	Delinquency	460	229.165	0.034
	Transparency	280	186.089	0.034
3	Animals	787	205.618	0.036
4	Mobility	4,338	157.877	0.033
	Urban planning	2,764	128.669	0.033
	Sports	1,533	150.763	0.033
	Economy	1,076	183.618	0.033
	Sustainability	1,000	242.613	0.033
	Environment	3,757	154.707	0.032
5	Entertainment	435	108.552	0.034
	Education	784	148.852	0.033
	Family & childhood	517	142.915	0.033
	Associations	239	131.573	0.033
	Old age	245	156.020	0.033
	Security & emergencies	614	140.368	0.032
	Health	395	188.934	0.032
	Employment	198	144.530	0.032
	Accessibility	359	121.830	0.032
	Youth	74	113.203	0.031

Table 2: Thematic categories grouped by average level of support and controversy of their proposals: group 1 refers to categories having a relatively low number of proposals with very high support and controversy; group 2 refers to categories having a moderate number or proposals with high controversy; group 3 refers to a category having a high number of proposals with high support and very high controversy; group 4 refers to categories having many proposals with moderate support and controversy; and group 5 refers to categories with moderate/low number of proposals, support and controversy.

the above, we could claim that citizens' ideological differences take an important role in this group of controversial categories.

The second group is associated to a number of categories with a **moderate number of proposals and support, but high controversy**. Within this group of categories, we can identify two sets of categories: (i) transparency, politics, citizen participation, public administration, laws and legislation, which are related to governance, and (ii) social rights, civics, equity, integration and delinquency, which are related to social and civic rights, obligations and movements. This group shows the relevance that participants gave to political and social issues. The number of proposals and supports in these categories are not as many as in other categories, but reflect highly controversial issues and topics. Again, by means of a preliminary qualitative analysis, we found out that the majority of the proposals belonging to this group represented citizens' complaints about vandalism, political corruption, gender violence, and remunicipalization plans, as well as the (bad) situation and needs of homeless people, immigrants, refugees, and social services. In any case, it is interesting to note how these related categories have come up with similar support and controversy patterns through the automatic data processing and mining processes of our approach. Citizens' political and socioeconomic factors, as well as NIMBY –Not in Not In My Back Yard (Dear, 1992)– reasons, seem to be predominant in the controversy of the proposals of this group.

The third group captures a category that receives a **high number of proposals with high support and controversy**. These proposals focus on animals, mainly dogs. Many of these proposals are complaints about dog fouling, its negative impact on the city, and potential solutions, including increasing the budget for cleaning, creating dog parks, introducing penalties for the owners, etc. Various proposals also targeted the need of having dogs on leash. There were also proposals addressing animal protection issues, e.g., more restrictions to adopt a pet, penalties to abandon them, and dog shelters. It is interesting that, while these topics do not attract much attention in the local and national media, they nonetheless constitute a major concern for the citizens.

The fourth group is composed of categories with a **very high number of moderately supported and controversial proposals**. They range from mobility and urban planning to economy, sustainability and environment, addressing very diverse topics: improving the public transport (a unique ticket, more frequency at night, its expansion to other areas), decreasing the price of parking, reducing pollution, optimizing energy consumption, increasing the number of green spaces, improving the cleaning of the city, avoiding food waste by supermarkets, obliging banks to rent owning empty houses, and including vegan and vegetarian options within the food menus of public schools, to name a few. This group contains the largest percentage of the proposals in the platform, entailing topics that most concern and agreement received from citizens.

The fifth and last group includes categories with a **moderate/low number of proposals, support and controversy**. These categories refer to education, health, family, childhood, youth, old age, and employment. Proposals in these categories include proposals such as opening schools during summer time, better conditions for hospitals and doctors, increasing the number of available places in public nursing homes as well as reducing their cost, the creation of more youth centers, and training opportunities during unemployment. As we can observe, these proposals either discuss low controversial topics, such as those related with the general social good, which are "easy" to support,

Proposal	Topics	#supports	#comments	Controversy
Single ticket for all public transport	public transport	34,722	816	0.371
Massive tree planting in Madrid	sustainability	20,596	310	0.235
Valdemingómez incinerator - NO!	natural environment	16,300	140	0.114
Free calls to 010 city council info. and 092 local police phones	city council, police	16,188	92	0.081
Eliminating abusive salaries of ex-public officials	transparency	13,574	32	0.068
Preventing supermarkets from throwing food	commercial centers	13,237	40	0.069
Right to play: for a more child-friendly Madrid	children	12,560	145	0.104
Let's reject "Ley Mordaza" (an expression freedom limiting law)	laws, social rights	11,943	62	0.093
Fines for those banks having empty houses	banks	10,018	74	0.132
Better timetables for night service	public transport	9,032	53	0.072

Table 3: Examples of proposals with high number of supports.

or the well-being of minorities (e.g., people with disabilities, senior citizens). We note that categories targeting minorities tend to gather less citizen engagement, particularly in terms of suggested proposals, than other categories.

These five groups do not only allow for a better identification of the problems and issues that most concern the citizens, but also reflect interesting insights about which are the underlying reasons and motivations that raise more/less support and controversy. In particular, ideological, political, socioeconomic and NIMBY factors have naturally emerged in that respect.

Going into a more specific analysis, Table 3 shows 10 **top proposals in terms of received support**, alongside their topics (automatically identified and belonging to the considered 30 categories), the number of comments they received, and their level of controversy.

As we can see, although these proposals got a high level of support, they tend to present low levels of controversy (below 0.1). Among the controversial proposals, we can highlight 'Massive tree planting in Madrid' and 'Valdemingómez incinerator - NO!.' These two proposals may be considered as NIMBY, since they refer to urban plans of the city council that received strong opposition of residents in their local areas. In the first case, plantation of trees was suggested to avoid the usage of land as waste dump . In the second case, residents had demanded the city council to close a waste incinerator. On the other hand, the controversy of the first proposal (a single ticket for all public transport) captures discussions around how to implement the measure, examples to follow, and the potential increment of its cost for residents living in the city outskirts.

Uncontroversial proposals refer to initiatives that may have a positive impact on a large fraction of the population, and therefore are easy to support, such as free calls to local police phone numbers, preventing supermarkets from throwing food, eliminating abusive salaries of ex-public officials, or the elimination of the so called 'Ley Mordaza' or Gag Law, introduced in 2015 to put constraints on the freedom of assembly and expression[18].

Table 4 shows 10 **proposals that show top levels of controversy**. As we can observe, while some of these proposals are not attracting a high level of support, such as the prohibition of outdoor smoking, or the establishment of a bank holiday in honour of Santiago, patron saint of Spain, the topics they cover are of sensitive nature. The

[18] https://www.theguardian.com/world/2015/mar/12/spain-security-law-protesters-freedom-expression

Proposal	Topics	#supports	#comments	Controversy
Eliminating bullfights and its subsidies	bullfighting, subsidies	16,327	591	0.670
Beneficial remunicipalization for all	remunicipalization	1,246	436	0.573
Stop ARTEfacto Valdebebas (an urbanistic plan)	urbanism	3,628	448	0.524
Madrid - 100% sustainable	sustainability	28,097	836	0.515
Application of IBI (a land value tax) to properties of the Church	church, ibi	19,136	449	0.457
Bank holiday to celebrate Santiago patron saint	festivals	21	333	0.423
Creating a real safe bike lane network in Madrid	bike lanes	10,255	639	0.415
Respect the tradition of nativity scenes in Christmas	nativity scenes	4,977	227	0.353
Prohibiting outdoor smoking, respecting non-smokers' rights	banned smoking law	190	132	0.308
Allowing dog access to public transport	dogs, public transport	3,615	227	0.295

Table 4: Examples of proposals with high controversy.

establishment of a new bank holiday to celebrate a Catholic saint, modifying the location of the nativity scenes, or applying land value taxes to properties of the Catholic church, are controversial proposals that confront two poles of the Spanish society: the more traditional groups, rooted on the Catholic faith, and the liberal groups that advocate for a secular state. Other controversial topics include: animal protection (and the ban on bullfight), animal tolerance (allowing dogs to access public transport), citizens' health and safety (prohibiting smoking in public outdoor areas, or the creation of a safe bike network), and environmental measures ('Madrid 100% sustainable') - which required, among other measures, removing cars from the city center to reduce the alarming levels of pollution. A topic of controversy is also the use of land by the city council, including remunicipalization (changes from private to public ownership) and the management of NIMBY urbanistic plans, like 'ARTEfacto Valdebebas,' aimed to provide social housing with common spaces and partially oriented to to vulnerable groups.

As a complement of controversy metrics, an analysis of the content of the comments may lead an in-depth characterization of the controversies. This type of analysis is out of the scope of this paper, and is envisioned as future work, where natural language processing and opinion mining techniques will be needed. Nonetheless, in the following, we present three representative examples of proposals in Decide Madrid that reflect some of the above mentioned **types and reasons of controversy** in citizen participation. Some real citizen comments on the proposals are given in Table 5.

The first example is a non controversial proposal, where a citizen suggests a single ticket to take any public transport, i.e., bus, metro and train. Even being accepted by the majority, the citizens' comments about the proposal raise problems and issues to be taken into account, such the presumably high price of the ticket and the fact that it should be managed by both the city council and the Autonomous Community of Madrid. A tool to automatically extract, relate, aggregate and summarize the arguments given in the debates would be a very valuable resource for both citizens and government in decision and policy making tasks.

The second example is a proposal with an ideological (religious) controversy. A citizen suggests that the Catholic Church should pay the IBI municipal tax applied to properties. In this case, the comments on the proposal are not focused on how to implement the proposal, but mainly express personal opinions about whether or not the proposal should be implemented. In the conversations, nonetheless, some objective facts are stated by citizens, such as the fact that Caritas, a Spanish NGO, does not belong to the Catholic church, and provides a very small percentage of its donations to

the religious institution. The automatic finding of evidences for these facts in reliable (online) resources would enable their verification and would help on personal decisions on this type of proposals.

The third and last example is a proposal aimed to stop the ARTEfacto urbanistic plan designed by the city council for Valdebebas neighborhood. This case represents a NIMBY proposal that raises high controversy in a minority group: the neighborhood inhabitants. The vast majority of the comments has complaints against the proposal and the related implementation process followed by the city council. Dense discussion and no opinion polarization are the main characteristics of the comments threads. Linking this type of proposals with external sources, such as online social networks and news media, may help on a better understanding of not only the proposals, but also their (potential) impact and controversy.

Regardless the future research lines suggested above, what we have done in this work is linking the computed metrics with a variety of city statistics provided by the city council as Open Data. Our goal is to provide first insights about external factors that may influence citizen participation, discussion and controversy in ePB. We present the corresponding analysis in the next section.

6.3. External factors influencing participation, discussion and controversy

With RQ3 we intended to conduct a preliminary analysis on which external factors may influence the existence of more (less) citizen participation, discussion and controversy in ePB.

To address the question, we make use of Open Data to study whether different **city-related statistical indicators** could be associated to our analysis variables, namely the quantity, support and controversy of proposals in Decide Madrid. Table 6 lists some of the collected and analyzed indicators, which capture multiple important aspects of Madrid, such as population, level of education, unemployment, per-capita income, and home topology. All these indicators are broken down per district, a key aspect for the analysis presented next.

The analysis is based on measuring the linear correlation existing between the values of certain indicator and the values of our analysis variables by district, i.e., the number of proposals, average support, and average controversy, per district. We note that correlation does not imply causation, but can be considered as a signal of which indicators (external factors) may have certain relation, influence or impact on the levels of participation, discussion and controversy. We also note that we consider indicators in isolation, and a deeper analysis observing the potential correlations among indicators and the contrast between them is left for future work.

Table 7 shows an **example of correlation computation for the number of 'Participants in Decide Madrid'**, a specific citizen participation indicator. In the table on the left, we show the values of the analysis variables (i.e., number of proposals, and average number of supports and controversy) for 10 districts. In the middle table, for the same districts, we show the number of participants in Decide Madrid (i.e., the value of the indicator per district, collected from Open Data). Lastly, in the table on the right, we show the 3 correlations computed between the values of each column of the first table (analysis variables) and the values of the second table (external indicator) for all the 21 districts in Madrid. As expected, the above indicator highly correlates with the

Single ticket for all public transport
- In Sydney, there is an electronic system that allows you to use a single ticket for taking any type of transport. The ticket... - Great system! The Oyster card in London works similarly, but it is very expensive. It is operated by a private company... - Certainly, it would be a good system, but the key point is its cost. Which would be the price of a one-way ticket? - This is a competency of the Autonomous Community, not the City city council. I would also ask for...
Application of IBI to properties of the Church
- I disagree. The Church's properties are available to citizens, for example, schools with sports facilities that... - That is true. Also, Caritas centers help people with needs... - Caritas is not the Catholic church. Only 1% of donations collected by Caritas are given to the church; the rest are private donations... - Non only the Catholic, but all religious confessions should pay IBI. By the way, in 2002, the Patronage Law was extended to NGOs. Church and NGOs give a very important service, but... - Obviously the IBI must be paid by all religious confessions, and it is the Catholic church the one that has 99.9% of the real estate. NGOs fulfill a social function, and they should be exempt of the tax. Caritas, which does a good job, is precisely listed as an NGO,...
Stop ARTEfacto Valdebebas
- The neighborhood has no public nursery school, no health center, a few bus stops... and now this experiment is the priority of the city council. I do not give credit. - Unfortunately, health centers and schools depend on the Community of Madrid... - Yes, I know, but I hoped the priority of the city council would be claiming this for the neighborhood, as well as taking care of its waste, municipal library and sport center. From whom are these competencies? - Incredible! 40 million euros for 31 social rental homes... How much does a house cost? Is this really so necessary to do without transparency? - Who has established that 27,064 supports are necessary for this? Proposals from other groups with less than 10% of that amount have been approved!

Table 5: Examples of proposals and comments with high controversy.

Type of indicator	Examples of indicators
Demography	Population (total, by gender, and by age group: 0-14, 15-29, 30-44, 45-64, 65-79, 80+, 65+), percentages of children and young/middle age/old people, numbers of immigrants and emigrants, birth and death rates, life expectancy, percentages of family structures
Economy	Per capita incomes, average pensions
Employment	Labor force participation rates, unemployment rates, long-term unemployment rates
Education	Population at each educational stage (preschool, elementary/middle/high school, graduate studies, postgraduate studies), public and private school enrollment rates, education levels of ≤25 years old people
Health	Sedentary lifestyle levels, overweight rates, tobacco consumption rates, medicine consumption rates, illness rates, population with disabilities
Housing	Building status and home topology (principal, secondary, inhabited) rates
Social vulnerability	Poverty or social exclusion rates, low and very low salary rates
Security	Number of police interventions (by type), number of arrests (by type of crime)
Public services	Number of social/cultural/educational/health care/sports/local commerce centers, number of people attended by each service
Environment	Air pollution measurements, average amount of collected waste
Quality of life	Average satisfaction scores about life quality, neighborhood living, and public services (by type), level of security perception
Citizen participation	Number of election votes by political party, number of associations (total, per association type: neighborhood, cultural, religious), number of participants in Decide Madrid

Table 6: Types and some examples of evaluated annual indicators, all of them per district and many available by gender and age group.

number of proposals: the districts with a higher (lower) number of participants in Decide Madrid are the ones that have a higher (lower) number of proposals in our dataset. This computation and interpretation can be applied to all the collected indicators.

District	#proposals	Avg. #supports	Avg. controversy	Participants in Decide Madrid
Fuencarral-El Pardo	850	341.020	0.034	5,308,000
Hortaleza	614	440.964	0.037	5,237,667
Arganzuela	828	327.920	0.035	5,048.667
Latina	577	354.246	0.036	4,748,333
Centro	779	142.092	0.033	4,597,333
Salamanca	461	425.315	0.038	2,340,333
Villaverde	422	507.519	0.038	2,190,000
Moratalaz	355	593.746	0.039	1,681,333
Viclvaro	363	586.030	0.038	1,653.667
Barajas	464	530.591	0.038	1,289,333

Variable	Correlation
Proposals	0.782
Avg. supports	-0.532
Avg. Controversy	-0.649

Table 7: Computing correlation values for the 'Participants in Decide Madrid' indicator. (i) Variables: Number of proposals, average number of supports and average controversy for 10 example districts; (ii) 'Participants in Decide Madrid' indicator values for those districts; (iii) correlation between the above variables and indicator computed for all 21 districts. Districts are sorted by decreasing indicator value.

Indicator	Proposal corr.	Support corr.	Controversy corr.
PSOE votes in 2015 municipal elections	0.508	-0.212	-0.314
LGBT rights organizations	0.274	-0.438	-0.378
Podemos votes in 2015 municipal elections	0.555	-0.255	-0.388
IU votes in 2015 municipal elections	0.553	-0.283	-0.396
Neighborhood associations	0.581	-0.324	-0.396
Consumer organizations	0.229	-0.665	-0.492
Ahora Madrid votes in 2015 municipal elections	0.641	-0.399	-0.503
Environment and ecology associations	0.363	-0.584	-0.512
Associations	0.608	-0.548	-0.592
Participants in Decide Madrid	0.782	-0.532	-0.649
Labor force replacement rate	-0.185	0.587	0.351
Animal rights organizations	-0.190	0.160	0.242
Child-related associations	-0.180	0.123	0.239
Population: 85+ years old	-0.341	-0.093	0.172
Population: 0-15 years old	0.025	0.426	0.153
Human rights organizations	0.081	0.055	0.140
Youth index	0.080	0.379	0.103
People with disabilities	0.236	0.089	0.086
Population: 65+ years old	-0.309	-0.169	0.053
Birth rate	0.198	0.169	0.039

Table 8: Correlation values between some indicators and i) number of proposals, ii) average number of supports per proposal, and iii) average controversy per proposal, computed for all districts.

Table 8 shows **correlation values computed for some indicators**. The left part of the table contains indicators having high positive correlation with the number of proposals, but high negative correlation with controversy. The right part of the table contains indicators having a low or negative correlation with the number of proposals, but a high positive correlation with controversy.

From the left part, interesting patterns that these numbers uncover include: (i) districts with a high number of residents participating in Decide Madrid are also districts with a high number of proposals (0.782 correlation), (ii) districts with a high number of associations, neighborhood associations, and consumer organizations are also more proactive generating proposals, (iii) districts with a more liberal/socialist stand, i.e., PSOE (*Partido Socialista Obrero Español*, Podemos and IU (*Izquierda Unida*), also generate more proposals and, (iv) districts with higher environmental engagement, i.e., more environment and ecology associations, also generate more proposals. All these indicators, however, present a negative correlation with the average support and controversy. In particular, the higher the number of participants in the Decide Madrid platform, associations or voters for the current government in a particular district, the lower the controversy of the proposals coming from that district. A similar trend can be found with respect to the number of supports. Particularly strong is the negative correlation between the number of consumer organizations, associations, and environment associations, with the average support received by the proposals, indicating that the more of these organizations a district has, the lower the number of supports that the proposals from that district will receive.

In the right part of the table, we can observe that districts generating more controversial proposals include: (i) districts with higher labor force replacement rate (i.e., with a higher percentage of young people), (ii) districts with higher number of vulnerable residents, i.e., senior, junior and disable residents, (iii) districts with higher birth

rate and more child-related associations, and (iv) districts with higher engagement on human and animal rights. While these are the districts generating more controversial proposals, they are not those that have more proposals. In particular, the more senior citizens a district has, the less the number of proposals it generates and the less support proposals obtain. This may be due to the technological illiteracy of more senior citizens, which could be affecting their voices and opinions being reflected in the Decide Madrid platform. On the other hand, districts with a high number of young residents obtain more support and controversy for their proposals.

7. Conclusions and future work

Motivated by the need for better understanding the nature of the citizen participation that emerges in smart governance applications, in this paper we have proposed a data-driven, large-scale approach to analyze the rich information embedded in the citizens' proposals and debates created online in the Decide Madrid electronic participatory budgeting platform.

Our final goal is to build a tool to deeply visualize, analyze and understand the problems, concerns and proposals related to a city that are expressed by its residents on Web forums, such as e-participatory platforms and online social networks. Representing a first stage to achieve this goal, our approach focuses on levels of discussion and controversy, and exploits Open Data, additionally to topic and location information, commonly considered in the state of the art.

Three key research questions have driven this work: exploring whether the most discussed and controversial proposals are those that achieve the highest support, the themes and types of proposals that are more discussed and controversial, and the demographic and socioeconomic indicators that may influence citizen participation, discussion and controversy. On addressing these questions our work provides multiple socio-technical contributions including: (i) a large dataset with thematic and geographical annotations of citizen proposals, (ii) a novel computational approach that exploits Open Data to automatically process and enrich citizen-generated contents, and (iii) an in-depth analysis of debate discussion and controversy in a digital platform for citizen participation.

Several observations have emerged from this analysis, which are relevant to highlight here. First, a high number of controversial proposals are rooted on the historical political and ideological division of the Spanish society (left vs. right political ideology, religious vs. secular, traditionalist vs. progressive). This type of controversy may not be so prominent within the discussions of ePB platforms in other countries. Many other proposals are mostly focused on addressing issues that residents find annoying, or that affect them directly, e.g., animal fouling and noises, cost of public transport, and NIMBY government plans. Less activity is observed on proposing initiatives to target social good or to benefit minorities, e.g., disabled people, senior citizens, immigrants, or even the youth. In this sense, proposals targeting issues that affect a large majority of the population of Madrid tend to be supported by citizens and ultimately implemented by the city council, but leave aside other very important issues affecting minority groups. Our study has revealed that most controversial, likely relevant, problems do not always receive the sufficient attention in e-participation.

We also note that there may be a bias on the voices expressed within ePB platforms. Thanks to the exploitation of Open Government Data, our analysis has shown how districts with liberal/socialist ideologies are more active in their participation, and how districts with more senior citizens present a relatively low number of proposals. In this context, it is important to understand who is behind the proposals, discussions, and supports, and whether decisions are being taken by considering a diverse and representative fraction of the city population, or just particular subgroups. And if that is the case, what are the barriers behind low participation (technology literacy, lack of access to technology, etc.) and which mechanisms could be put in place to ensure no subgroups are excluded?

It is also worth to answer whether most citizens maintain an over-time engagement, or whether they engage once and then become dormant. Demographic and log data to address these questions are not publicly accessible for privacy regulations, but could be taken into account by the teams behind ePB platforms to enhance their reach and the overtime engagement of citizens. Among the variety of elements that may influence the levels of participation, several authors have discussed issues related to technical aspects of the used e-platforms (Aragón et al., 2017; Cantador and Cortés-Cediel, 2018), proposing novel solutions for content visualization, search and recommendation (Cantador et al., 2017, 2018), to increase the citizens' engagement.

Along with these aspects, ethical implications of how the information is presented and analyzed should be taken into consideration. The information needs, understanding capabilities, and exploitation purposes of the target stakeholders –e.g., citizens, businesses, governments and politicians– are manifold, and thus should be considered differently. This, together with the necessity of avoiding biases in the presentation of analysis results, are ethical issues of high relevance that we leave as future work.

Regarding our technical contributions, we remark that while this work constitutes a significant step towards the automatic processing and understanding of citizen-generated contents in e-participation, multiple elements can still be improved. First, Natural Language Processing and Opinion Mining techniques may be applied to analyze the citizens comments on proposals. Among others, opinion lexicons (Hubert et al., 2018), controversy vocabularies (Mejova et al., 2014; Roitman et al., 2016), language models (Jang et al., 2016) and word embeddings (Rethmeier et al., 2018), and argument extraction methods and tools (Shum et al., 2008; Lytos et al., 2019; Dutta et al., 2019) could be used to extract statements and claims in favor and against each proposal, and hence, achieving a better understanding of the most important and urgent citizen needs, as well as the underlying discussions and controversies. Moreover, applying machine learning to automatically classify and predict the relevance and levels of discussion and controversy (Jang et al., 2016; Rad and Barbosa, 2012) of citizen proposals represents a research line whose results would be of special interest for government decision and policy making. Another interesting research direction is, in our humble opinion, the automatic categorization of controversy. As shown in our analysis, different types of controversy coexist in ePB platforms, mainly related to ideological, political, socioeconomic and NIMBY factors. The automatic classification of controversy could help determining more effective ways of creating consensus. For such task, the consideration of theories, models and resources existing in sociology and political sciences, e.g., (Toulmin, 2003), are envisioned as essential.

In this work we have focused on measuring debate controversy without considering its evolution over time. The study of temporal dynamics of controversial themes has been addressed in the literature, especially in the context of online social network analysis (Yardi and Boyd, 2010; Smith et al., 2013). In addition to analyzing the changes of discussion and controversy within and between (1-year) participatory budgeting processes, we envision as a promising research line linking and comparing citizen participation in ePB platforms and in online social networks, such as Twitter (Ma et al., 2016; Alizadeh et al., 2019; Driss et al., 2019; Vargas-Calderón and Camargo, 2019). This may raise enriched insights about the characteristics of participants, proposals and discussed issues, as well as the types and levels of controversy. We also are interested in analyzing the so-called *bandwagon effect*, which has been shown to be a suitable mechanism to explain the construction of major opinions in online environments over time (Lee et al., 2018). In this respect, we hypothesize that the support of certain citizen proposals, and their corresponding discussion and controversy levels, could be affected by related external events and trending topics in the media and on the Web.

Due to the lack of effective information retrieval and filtering mechanisms in current ePB platforms, there is duplication among proposals, and some of them target same issues. The automatic identification of similar proposals and their grouping could provide more accurate analysis and results. In our study, even if two proposals addressed the same issue, they were treated in isolation. Providing mechanisms to automatically identify similar proposals is part of future work, and content-based similarities we already used in (Cantador et al., 2017) may be very valuable.

We also note that our analysis was focused on one ePB platform. Conducting analogous studies across various ePB platforms, and identifying similar patterns, as well as divergences, could help us to pinpoint the socio-technical aspects that may enhance ePB in general, and the aspects that are rooted in the different societies and cultures, and for which ePB platforms may need specific adaptations. For such purpose, previous datasets we generated in (Cantador et al., 2018), from ePB processes of New York, Miami and Cambridge, could be considered.

Despite considering a single case study, we believe that the proposed analysis approach can be adapted and used for other platforms and cities, and the reported results may be of interest for a variety of stakeholders and researchers in disciplines distinct to computer science, such as sociology and political sciences. As shown in our analysis, a number of generic indicators (e.g., demography, economy, employment, education, health, housing, social vulnerability, security, public services, and environment) provided as Open Data allows identifying relationships between characteristics of the citizens, neighborhoods and districts, themes of the proposals, and discussion and controversy levels.

To conclude, we want to highlight that, to the best of our knowledge, our approach represents a first attempt in the research literature to exploit Open Data in order to deep into the analysis and understanding of the proposals and debates existing in ePB, as a representative case of e-participation. While there is ample room for further investigation, this work opens a novel and exciting interdisciplinary line of research, in which computer, social and political sciences can cooperate towards the realization of Smart Governance.

8. Acknowledgements

This work was supported by the Spanish Ministries of Economy, Industry and Competitiveness (TIN2016-80630-P) and Science, Innovation and Universities (CAS18/00035).

9. References

Alizadeh, T., Sarkar, S., Burgoyne, S., 2019. Capturing citizen voice online: Enabling smart participatory local government. Cities 95, 102400. doi:10.1016/j.cities.2019.102400.

Aragón, P., Bermejo, Y., Gómez, V., Kaltenbrunner, A., 2018. Interactive discovery system for direct democracy, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE. pp. 601–604. doi:10.1109/ASONAM.2018.8508554.

Aragón, P., Kaltenbrunner, A., Calleja-López, A., Pereira, A., Monterde, A., Barandiaran, X.E., Gómez, V., 2017. Deliberative Platform Design: The case study of the online discussions in Decidim Barcelona, in: Proceedings of the 9th International Conference on Social Informatics, Springer. pp. 277–287. doi:10.1007/978-3-319-67256-4_22.

Attard, J., Orlandi, F., Scerri, S., Auer, S., 2015. A systematic review of Open Government Data initiatives. Government Information Quarterly 32, 399–418. doi:10.1016/j.giq.2015.07.006.

Baiocchi, G., Ganuza, E., 2014. Participatory budgeting as if emancipation mattered. Politics & Society 42, 29–50. doi:10.1177/0032329213512978.

Boudjelida, A., Mellouli, S., 2016. A multidimensional analysis approach for electronic citizens participation, in: Proceedings of the 17th International Conference on Digital Government Research, ACM. pp. 49–57. doi:10.1145/2912160.2912195.

Boudjelida, A., Mellouli, S., Lee, J., 2016. Electronic citizens participation: Systematic review, in: Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ACM. pp. 31–39. doi:10.1145/2910019.2910097.

Boukhris, I., Ayachi, R., Elouedi, Z., Mellouli, S., Amor, N.B., 2016. Decision model for policy makers in the context of citizens engagement: Application on participatory budgeting. Social Science Computer Review 34, 740–756. doi:10.1177/0894439315618882.

Bramson, A., Grim, P., Singer, D.J., Fisher, S., Berger, W., Sack, G., Flocken, C., 2016. Disambiguation of social polarization concepts and measures. The Journal of Mathematical Sociology 40, 80–111. doi:10.1080/0022250X.2016.1147443.

Cantador, I., Bellogín, A., Cortés-Cediel, M.E., Gil, O., 2017. Personalized recommendations in e-participation: Offline experiments for the 'Decide madrid' platform, in: Proceedings of the 1st International Workshop on Recommender Systems for Citizens, ACM. pp. 5:1–5:6. doi:10.1145/3127325.3127330.

Cantador, I., Cortés-Cediel, M.E., 2018. Towards increasing citizen engagement in participatory budgeting digital tools, in: Proceedings of the 19th International Conference on Digital Government Research, ACM. pp. 91:1–91:2. doi:10.1145/3209281.3209389.

Cantador, I., Cortés-Cediel, M.E., Fernández, M., Alani, H., 2018. What's going on in my city? Recommender systems and electronic participatory budgeting, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM. pp. 219–223. doi:10.1145/3240323.3240391.

Cortés-Cediel, M.E., Cantador, I., Bolívar, M.P.R., 2019. Analyzing citizen participation and engagement in european smart cities. Social Science Computer Review , 0894439319877478doi:10.1177/0894439319877478.

De Sousa Santos, B., 1998. Participatory budgeting in Porto Alegre: Toward a redistributive democracy. Politics & society 26, 461–510. doi:10.1177/0032329298026004003.

Dear, M., 1992. Understanding and overcoming the NIMBY syndrome. Journal of the American Planning Association 58, 288–300. doi:10.1080/01944369208975808.

Dori-Hacohen, S., Allan, J., 2015. Automated controversy detection on the web, in: Proceedings of the 37th European Conference on Information Retrieval, Springer. pp. 423–434. doi:10.1007/978-3-319-16354-3_46.

Driss, O.B., Mellouli, S., Trabelsi, Z., 2019. From citizens to government policymakers: Social media data analysis. Government Information Quarterly 36, 560–570. doi:10.1016/j.giq.2019.05.002.

Dutta, S., Das, D., Chakraborty, T., 2019. Changing views: Persuasion modeling and argument extraction from online discussions. Information Processing & Management , 102085doi:10.1016/j.ipm.2019.102085.

Fung, A., 2015. Putting the public back into governance: The challenges of citizen participation and its future. Public Administration Review 75, 513–522. doi:doi.org/10.1111/puar.12361.

Gagliardi, D., Schina, L., Sarcinella, M.L., Mangialardi, G., Niglia, F., Corallo, A., 2017. Information and communication technologies and public participation: Interactive maps and value added for citizens. Government Information Quarterly 34, 153–166. doi:10.1016/j.giq.2016.09.002.

Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M., 2018. Quantifying controversy on social media. ACM Transactions on Social Computing 1, 3:1–3:27. doi:10.1145/2835776.2835792.

Gascó-Hernández, M., Martin, E.G., Reggi, L., Pyo, S., Luna-Reyes, L.F., 2018. Promoting the use of open government data: Cases of training and engagement. Government Information Quarterly 35, 233–242. doi:10.1016/j.giq.2018.01.003.

Gómez, V., Kaltenbrunner, A., López, V., 2008. Statistical analysis of the social network and discussion threads in Slashdot, in: Proceedings of the 17th International Conference on World Wide Web, ACM. pp. 645–654. doi:10.1145/1367497.1367585.

Held, D., 2006. Models of democracy. Stanford University Press.

Hivon, J., Titah, R., 2017. Conceptualizing citizen participation in Open Data use at the city level. Transforming Government: People, Process and Policy 11, 99–118. doi:10.1108/TG-12-2015-0053.

Hubert, R.B., Estevez, E., Maguitman, A., Janowski, T., 2018. Examining government-citizen interactions on Twitter using visual and sentiment analysis, in: Proceedings of the 19th International Conference on Digital Government Research, ACM. pp. 55:1–55:10. doi:10.1145/3209281.3209356.

Janev, V., Mijović, V., Paunović, D., Milošević, U., 2014. Modeling, fusion and exploration of regional statistics and indicators with Linked Data tools, in: Proceedings of the 3rd International Conference on Electronic Government and the Information Systems Perspective, Springer. pp. 208–221. doi:10.1007/978-3-319-10178-1_17.

Jang, M., Foley, J., Dori-Hacohen, S., Allan, J., 2016. Probabilistic approaches to controversy detection, in: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, ACM. pp. 2069–2072. doi:10.1145/2983323.2983911.

Janssen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of Open Data and Open Government. Information Systems Management 29, 258–268. doi:10.1080/10580530.2012.716740.

Jetzek, T., Avital, M., Bjørn-Andersen, N., 2014. Generating sustainable value from Open Data in a sharing society, in: Proceedings of the 10th International Working Conference on Transfer and Diffusion of IT, Springer. pp. 62–82. doi:10.1007/978-3-662-43459-8_5.

Lee, S., Ha, T., Lee, D., Kim, J.H., 2018. Understanding the majority opinion formation process in online environments: An exploratory approach to Facebook. Information Processing & Management 54, 1115–1128. doi:10.1016/j.ipm.2018.08.002.

Lo, D., Surian, D., Prasetyo, P.K., Zhang, K., Lim, E.P., 2013. Mining direct antagonistic communities in signed social networks. Information Processing & Management 49, 773–791. doi:10.1016/j.ipm.2012.12.009.

Lytos, A., Lagkas, T., Sarigiannidis, P., Bontcheva, K., 2019. The evolution of argumentation mining: From models to social media and emerging tools. Information Processing & Management 56, 102055. doi:10.1016/j.ipm.2019.102055.

Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H., 2016. Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. Information Processing & Management 52, 430–445. doi:10.1016/j.ipm.2015.10.004.

Mejova, Y., Zhang, A.X., Diakopoulos, N., Castillo, C., 2014. Controversy and sentiment in online news. ArXiv preprint arXiv:1409.8152.

Nam, T., Pardo, T.A., 2011. Conceptualizing smart city with dimensions of technology, people, and institutions, in: Proceedings of the 12th International Conference on Digital Government Research, ACM. pp. 282–291. doi:10.1145/2037556.2037602.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69, 026113. doi:10.1103/PhysRevE.69.026113.

Peña-López, I., et al., 2001. Citizens as Partners. OECD Handbook on Information, Consultation and Public Participation in Policy-Making. doi:10.1787/9789264195578-en.

Popescu, A.M., Pennacchiotti, M., 2010. Detecting controversial events from Twitter, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM. pp. 1873–1876. doi:10.1145/1871437.1871751.

Rad, H.S., Barbosa, D., 2012. Identifying controversial articles in Wikipedia: A comparative study, in: Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration, ACM. pp. 7:1–7:10. doi:10.1145/2462932.2462942.

Ranchordás, S., 2017. Digital agoras: Democratic legitimacy, online participation and the case of uber-petitions. The Theory and Practice of Legislation 5, 31–54. doi:10.1080/20508840.2017.1279431.

Rethmeier, N., Hübner, M., Hennig, L., 2018. Learning comment controversy prediction in web discussions using incidentally supervised multi-task CNNs, in: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 316–321. doi:10.18653/v1/W18-6246.

Roitman, H., Hummel, S., Rabinovich, E., Sznajder, B., Slonim, N., Aharoni, E., 2016. On the retrieval of Wikipedia articles containing claims on controversial topics, in: Proceedings of the 25th International Conference Companion on World Wide Web, ACM. pp. 991–996. doi:10.1145/2872518.2891115.

Shum, S.B., et al., 2008. Cohere: Towards web 2.0 argumentation, in: Proceedings of the 2nd International Conference on Computational Models of Argument, pp. 97–108.

Smith, L.M., Zhu, L., Lerman, K., Kozareva, Z., 2013. The role of social media in the discussion of controversial topics, in: Proceedings of the 2013 International Conference on Social Computing, IEEE. pp. 236–243. doi:10.1109/SocialCom.2013.41.

Toulmin, S.E., 2003. The uses of argument. Cambridge University Press. doi:10.1017/CBO9780511840005.

Vargas-Calderón, V., Camargo, J.E., 2019. Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. Cities 92, 187–196. doi:10.1016/j.cities.2019.03.019.

Yardi, S., Boyd, D., 2010. Dynamic debates: An analysis of group polarization over time on Twitter. Bulletin of Science, Technology & Society 30, 316–327. doi:10.1177/0270467610380011.

Zheng, Y., Schachter, H.L., 2017. Explaining citizens' e-participation use: The role of perceived advantages. Public Organization Review 17, 409–428. doi:10.1007/s11115-016-0346-2.

Zielinski, K., Nielek, R., Wierzbicki, A., Jatowt, A., 2018. Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries. Information Processing & Management 54, 14–36. doi:10.1016/j.ipm.2017.08.005.