Comparativa Socio-Cultural entre Cataluña y el Resto de Comunidades Autónomas Españolas Basada en el Análisis de Datos de Facebook

Ignacio Martín[1], Rubén Cuevas[1], Nick Obradovich[2], and Ángel Cuevas[1]

[1]Universidad Carlos III de Madrid
[2]Massachusetts Insitute of Technology

17 de diciembre de 2017

Resumen

La situación política en Cataluña ha llevado a España a vivir una de las situaciones políticas más delicadas de su democracia. Esta crisis ha generado un gran debate político y social con multitud de opiniones defendiendo diferentes maneras de abordar y solucionar el conflicto. Este documento pretende añadir una nueva perspectiva al debate basada en el análisis de datos. Nuestro objetivo es: (i) comparar las similaridad socio-cultural entre las diecisiete Comunidades Autónomas de España, incluyendo a Cataluña, (ii) analizar posibles singularidades socio-culturales de las CCAA que se puedan derivar del análisis de datos. Para llevar a cabo nuestro estudio utilizaremos datos extraídos de Facebook, obteniendo una huella socio-cultural de cada CCAA utilizando el porcentaje de usuarios de FB en cada CCAA interesados en más de 67.000 intereses. Mediante la comparación y análisis de la huella socio-cultural de las CCAA podremos dar respuesta al objetivo planteado.

1. Introducción

En los últimos años, España ha experimentado una escalada del conflicto político existente entre Cataluña y el resto de España. Dicho conflicto ha alcanzado recientemente su momento de màxima tensión cuando el parlamento de Cataluña declaró la independencia de Cataluna de manera unilateral [1]. Este evento ha generado: (i) una reacción polìtica del Gobierno Español que, aplicando el artículo 155 de la Constitución Española, destituyó al Gobierno Catalán, pasó a tomar control de sus funciones y anunció la convocatoria de elecciones autonómicas el 21 de Diciembre de 2017 [2]; (ii) una reacción judicial que por el momento ha llevado a prisión a una gran parte del gobierno Catalán bajo la acusación de haber cometido presuntos delitos de rebelión, sedición y malversación fondos [3]; (iii) el viaje del presidente Catalán y parte de su Gobierno a Bélgica con el propósito de internacionalizar el conflicto [4].

La raíz de este conflicto radica sobre la petición de parte de la sociedad catalana y algunos partidos políticos de su derecho a decidir si quieren seguir formando parte de España u organizarse como un estado independiente. Uno de los argumentos esgrimidido por esta parte de la sociedad catalana para defender su posición ha consisitido en destacar importantes singularidades culturales e históricas que invitan a pensar en Cataluña como una nación diferente de España. Existen tanto opiniones que defienden y argumentan la postura de que Cataluña debería ser una nación distinta de España [5, 6], como opiniones que defienden que no [7, 8, 9].

Con este documento técnico buscamos abrir una nueva perspectiva al intenso debate que existe actualmente en España en relación a la situación política en Cataluña. Nuestro objetivo es analizar hasta que punto los catalanes son culturalmente similares/diferentes al resto de España. Como científicos, pretendemos dar respuesta a esta cuestión a través un análisis de datos objetivo que cuantifique la similaridad socio-cultural entre la comunidad y el resto de Comunidades Autónomas (CCAA). Por ello, este estudio carece de jucios de valor u opiniones políticas y, como tal, solamente se muestran los resultados obtenidos del análisis sin explicaciones o valoraciones cualitiativas que conlleven ningún valor político.

Nuestro estudio se soporta en datos extraídos de Facebook (FB), que es a día de hoy la red social más popular con casi 2000 millones de usuarios activos. Comparamos la similardiad cultural entre usuarios de FB viviendo en diferentes CCAA a partir de un vector de 67.000 intereses que incluye entre otros: intereses deportivos (e.g., FC Barcelona, Real Madrid, Pau Gasol), intereses musicales (e.g., Joaquin Sabina, Joan Manuel Serrat), localizaciones geográficas, etc. Para cada CA, C, y cada interés, I, obtenemos la porción de usuarios de FB residentes en C que están interesados en I. Esto genera un vector de 67.000 elementos. El valor de cada elemento del vector varía entre 0 (nigún usuario de FB en esa CA tiene interés en ese elemento) y 1 (todos los usuarios de FB en esa CCAA tienen interés en ese elemento). A este vector lo llamamos huella socio-cultural y se puede definir para cada CA. Comparando la huella socio-cultural de cada CA con el resto de CCAA podemos obtener una medida de similaridad entre cada par de CCAA.

Además de medir la similaridad entre CCAA, identificamos, cuantificamos y analizamos aquellos intereses cuya penetración es especialmente baja/alta en el vector socio-cultural de alguna CA comparado con las demás. Consideramos que estos elementos representan preferencias especialmente significativas de cada CA y nos referiremos a ellos como *singularidades* en el resto del documento.

2. Collección de Datos

Para obtener los datos, utilizamos la herramienta *Ads Manager API* [1] que FB ofrece a sus anunciantes para llevar a cabo sus camapñas de publicidad. Este API permite a los anunciantes definir las denóminadas .ªudiencias objetivoçombinando diferentes tipos de parámetros: parámetros de localización (e.g., país, región, ciudad, código postal), parámetros demográficos (e.g., edad, género, lengua), parámetros de comportamiento (e.g., dispositivo móvil usado, frencuencia de viajes), intereses de los usuarios (cubre cientos de miles de intereses diferentes), etc. Hemos desarrollado una herramienta que genera conjuntos de audiencias automáticamente y obtiene, entre otros parámetros, el número de usuarios que FB asigna a dicha audiencia.

Las 17 CCAA españolas están disponibles como parámetros de localización en el API de FB. Para calcular la porción de usuarios de FB en una CA C interesados e un interés I obtenemos el número de usuarios que conforman dos audiencias: (i) la audiencia definida por los usuarios residentes en C mediante el parámetro de localización asociado a la CA en exclusiva. Nos referimos a esta audiencia como A_C, (ii) la audiencia definida por los usuarios residentes en C e interesados en I para la que usamos el parámetro de localización de la CA y como parámetro de interés I. Nos referios a esta audiencia como $A_{C,I}$. Divdiendo $A_{C,I}$ entre A_C, obtenemos el porcentaje de usuarios registrados en Facebook en la CA C que están interesados en el elemento I.

Facebook asigna intereses a los usarios basándose en: los likes (me gusta) en páginas de FB, los clicks que los usuarios realizan en anuncios en FB, las aplicaciones instaladas por los usuarios, cualquier cosa que los algorítmos de FB determinen como relevante para un usuario, y cualquier cosa que el usuario añada explícitamente como interés. Los usuarios de FB pueden recibir publicidad dirigida por parte de los anunciantes en base a cualquiera de los intereses asignados.

Por último, necesitamos definir un vector de intereses ($V_I = [I_1, I_2, I_3, ..., I_m]$) que permita definir la huella socio-cultural de las CCAA. Para evitar cualquier sesgo en el conjunto de intereses seleccionados podríamos utilizar dos estrategias: (i) usar un vector predefinido a priori que sea aceptado globalmente como representativo de la cultura española y catalana, (ii) usar todos los intereses disponibles en FB. Sin embargo, ninguna de estas dos alternativas es realista. Hasta donde nosotros conocemos, la primera opción simplemente no existe, mientras que la segunda opción es computacionalmente inviable. Debido a estas limitaciones, hemos adoptado alternativas basadas en fuerza bruta que minimicen lo máximo posible la presencia de sesgos sin que se dispare la complejidad de la solución. Para ello, hemos obtenido dos vectores que contienen 77.523 y 67.311 intereses a través de métodos ortogonales. Para crear el primer vector hemos descargado más de 12 millones de registros de la DBPedia [2], que a su vez han sido mapeados a 399.182 intereses de Facebook. De todos ellos, hemos seleccionado aquellos cuyo alcance en FB es superior al medio millón de usuarios en todo el mundo, lo que resulta en el vector final 77.523 intereses. Para elaborar el segundo vector, utilizamos todos los intereses que FB ha asignado a usuarios Españoles que han descargado nuestra herramienta *Facebook Data Valuation Tool (FDVT)* [10] [3]. El FDVT es una

[1] https://www.facebook.com/ads/manager/creation
[2] http://es.dbpedia.org/
[3] Facebook Data Validation Tool. See: www.fdvt.org

herramienta que proporciona a los usuarios de FB una estimación de los ingresos que generan sus perfiles a la red social en función de la publicidad dirigida que visualizan y los anuncios sobre los que hacen click. El FDVT recoge los intereses que FB ha asignado a los usuarios que se han instalado la herramienta. En total, usando esta estrategia, obtenemos un vector de 67.311 intereses provenientes de 2.101 usarios Españoles.

Los resultados de similaridad de los vectores de intereses obtenidos usando DBPedia y el FDVT difieren en menos de 0,001. Esto muestra que ambos vectores obtienen resultados muy similares en nuestra investigación, En el resto del estudio usaremos el vector de intereses generado a partir del FDVT ya que ha sido obtenido directamente de usuarios Españoles. El motivo de esta decisión es que ante unos resultados prácticamente iguales, el conjunto de intereses de FDVT proporciona mayor variedad de intereses locales y regionales, que enriquece el estudio de singularidades.

Usando el vector de intereses derivado del FDVT generamos una huella socio-cultural para cada CA (es decir un vector con valores entre 0 y 1 para cada interés y CA). Para cada par de CCAA calculamos su similaridad usando la función cosine similarity [11] que devuelve un valor entre 0 (huellas socio-culturales totalmente diferentes) y 1 (huellas socio-culturales exactamente iguales). Además de las CCAA españolas, también obtenemos la huella socio-cultural para: (i) las dos regiones francesas más próximas a Cataluña, Languedoc−Roussillon y Midi−Pyrenees; (ii) 6 países pertenecientes a la Unión Europea: Alemania, Francia, Grecia, Italia, Portugal y Reino Unido. Nuestro objetivo es usar estas regiones y países extranjeros para validar la hipótesis incial de que las CCAA son más similares entre ellas que con países o regiones extranjera.

3. Resultados

3.1. Análisis de similaridad entre CCAA

Reg	And	Ara	Ast	IB	PV	IC	Can	C.Leon	C.Man	Cat	Ext	Gal	Mad	Mur	Nav	Rio	Val
And	1	0.975	0.968	0.966	0.972	0.963	0.972	0.984	0.987	0.968	0.976	0.976	0.98	0.978	0.97	0.971	0.98
Ara	0.975	1	0.963	0.959	0.971	0.953	0.968	0.977	0.978	0.969	0.963	0.969	0.976	0.969	0.971	0.972	0.974
Ast	0.968	0.963	1	0.951	0.962	0.951	0.97	0.976	0.971	0.956	0.959	0.971	0.966	0.96	0.958	0.959	0.964
IB	0.966	0.959	0.951	1	0.958	0.955	0.954	0.964	0.964	0.964	0.951	0.959	0.966	0.962	0.956	0.957	0.968
PV	0.972	0.971	0.962	0.958	1	0.951	0.971	0.977	0.976	0.968	0.96	0.968	0.976	0.966	0.978	0.97	0.971
IC	0.963	0.953	0.951	0.955	0.951	1	0.951	0.961	0.96	0.951	0.952	0.958	0.96	0.957	0.949	0.95	0.962
Cantabria	0.972	0.968	0.97	0.954	0.971	0.951	1	0.978	0.975	0.96	0.961	0.969	0.97	0.964	0.964	0.965	0.969
C. Leon	0.984	0.977	0.976	0.964	0.977	0.961	0.978	1	0.989	0.972	0.976	0.98	0.983	0.976	0.975	0.975	0.98
C.Mancha	0.987	0.978	0.971	0.964	0.976	0.96	0.975	0.989	1	0.972	0.976	0.978	0.986	0.979	0.975	0.975	0.982
Cataluña	0.968	0.969	0.956	0.964	0.968	0.951	0.96	0.972	0.972	1	0.957	0.963	0.974	0.964	0.965	0.963	0.971
Extr	0.976	0.963	0.959	0.951	0.96	0.952	0.961	0.976	0.976	0.957	1	0.964	0.968	0.965	0.96	0.96	0.966
Galicia	0.976	0.969	0.971	0.959	0.968	0.958	0.969	0.98	0.978	0.963	0.964	1	0.974	0.968	0.964	0.966	0.971
Madrid	0.98	0.976	0.966	0.966	0.976	0.96	0.97	0.983	0.986	0.974	0.968	0.974	1	0.974	0.974	0.971	0.979
Murcia	0.978	0.969	0.96	0.962	0.966	0.957	0.964	0.976	0.979	0.964	0.965	0.968	0.974	1	0.966	0.966	0.98
Navarra	0.97	0.971	0.958	0.956	0.978	0.949	0.964	0.975	0.975	0.965	0.96	0.964	0.974	0.966	1	0.974	0.968
La Rioja	0.971	0.972	0.959	0.957	0.97	0.95	0.965	0.975	0.975	0.963	0.96	0.966	0.971	0.966	0.974	1	0.969
Valencia	0.98	0.974	0.964	0.968	0.971	0.962	0.969	0.98	0.982	0.971	0.966	0.971	0.979	0.98	0.968	0.969	1

Tabla 1: Similaridad entre cada par de CCAA usando como métrica Cosine Similarity.

La Tabla 3.1 muestra la cosine similarity entre cada par de CCAA. La Figura 1 resume los resultados de la tabla usando una representación en forma de box plot que contiene la distribución de similaridades de cada CA. Además de las CCAA, la figura incluye las distribuciones de similaridad para las regiones y países extranjeros mencionados anteriormente. Los resultados están ordenados de izquierda a derecha de manera ascendente en función de la mediana de la cosine similarity de las localizaciones analizadas.

Los resultados muestran que existe una gran similaridad entre todas las CCAA, que a su vez es mucho mayor que la similaridad de cualquier CA con cualquiera de las regiones/países extranjeros considerados. Centrándonos exclusivamente en el análisis de España, Cataluña ocupa el 5º lugar entre las CCAA con menor similaridad con el resto en mediana, con un valor mediano de cosine similarity igual a 0.9645. Los dos archipiélagos españoles, Canarias y Baleares, son las CCAA que muestran una similaridad menor con el resto de CCAA con un valor de mediano de similaridad igual a 0.959 y 0.954, respectivamente. Por el contrario, dos CCAA del centro de la Meseta como Castilla y León y Castilla la Mancha son las regiones con mayor similaridad al resto de CCAA con un valor de similaridad mediano de 0.9765 y 0.976, respectivamente.

Las CCAA más similares a Cataluña en base a nuestros resultados son: Madrid (cosine similarity igual a 0,974), Castilla y León y Castilla La Mancha (ambas 0,972) y Valencia (0,971). Todas ellas

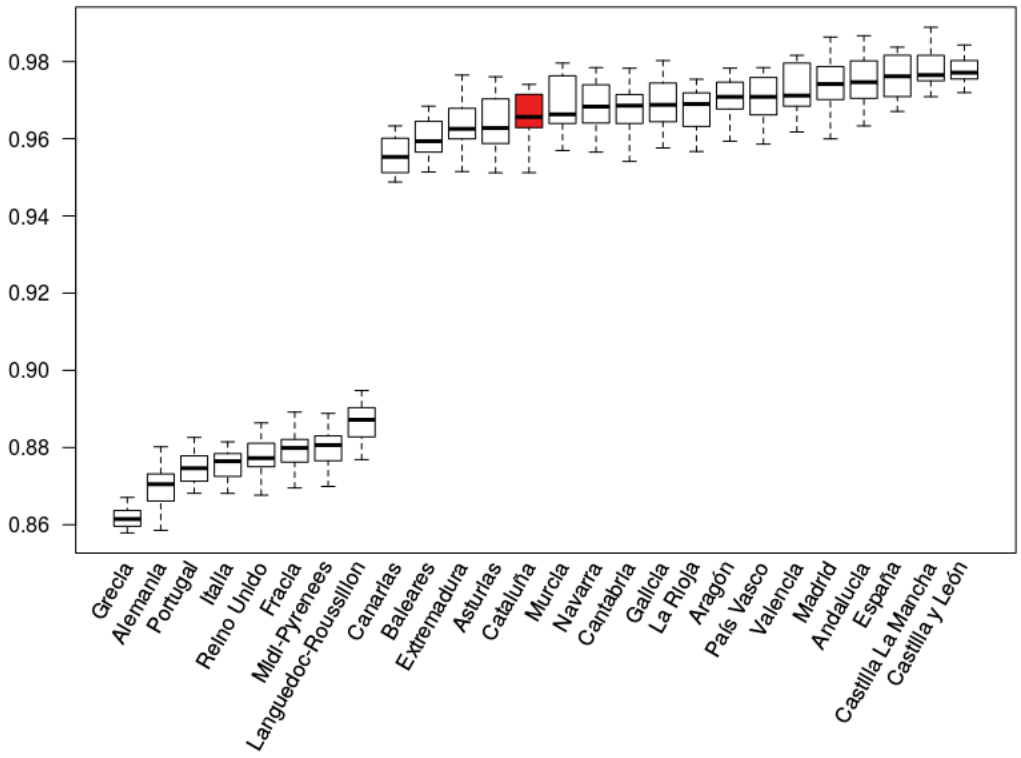

Figura 1: Boxplot que representa la distribución de similaridad de cada CA con el resto de CCAA. La figura también incluye la distribución de similaridad de países y regiones extranjeras con las CCAA españolas. Los resultados están ordenados de izquierda a derecha de manera ascendente en función de la mediana de la cosine similarity de las localizaciones analizadas. El resultado de Cataluña se resalta en color rojo.

se situan entre las 5 CCAA con una mayor similaridad en mediana al resto de CCAA. Por el contrario, las regiones que menos se parecen a Cataluña son: las Islas Canarias (0,951), Asturias (0,956) y Extremadura (0,957). Todas ellas se encuentran entre las 5 CCAA con menor similaridad al resto de CCAA.

3.2. Análisis de singularidades de las CCAA

Definimos como singularidad asociada a una CA aquellos elementos del vector socio-cultural que presentan un valor considerablemente mayor (sigularidad positiva) o considerablemente menor (singularidad negativa) en esa CA comparado con el resto de CCAA. Las singularidades positivas parecen definir intereses locales que pueden estar asociados a elementos socio-culturales diferenciadores. Por su parte las singularidades negativas definen intereses que atraen mucho menos la atención en una CCAA que en el resto de España y por tanto también pueden representar elementos socio-culturales diferenciadores.

Consideramos tres valores umbral multiplicativos para determinar que intereses son singularidades en cada región: x2, x5 y x10. Para las singularidades positivas, cualquier elemento del vector socio-cultural de una CA debe ser 2, 5 o 10 veces superior a la CA con el segundo valor más alto. Si tomamos como ejemplo el umbral 2, si una CA C tiene un valor 0.25 asociado al interes I en su vector de intereses socio-cultural (es decir, el 25 % de los usuarios de FB en C están interesados en I), I será considerado una singulardidad de C en el caso que el resto de CCAA tenga un valor igual o menor que 0.125 para I en sus vectores de intereses. En el caso de singularidades negativas, los umbrales se utilizan para identificar elementos del vector de intereses que sean 2, 5 o 10 veces inferiores para una CA respecto de las demás.

La Figura 2 muestra en un diagrama de barras la cantidad de singularidades positivas para cada CCAA y umbral (2x-rojo, 5x-azul y 10x-amarillo). La Figura 3 usa la misma representación para cuantificar las singularidades negativas.

En el caso de Cataluña obtenemos 1.008 (para el umbral x2), 312 (x5), 129 (x10) singularidades

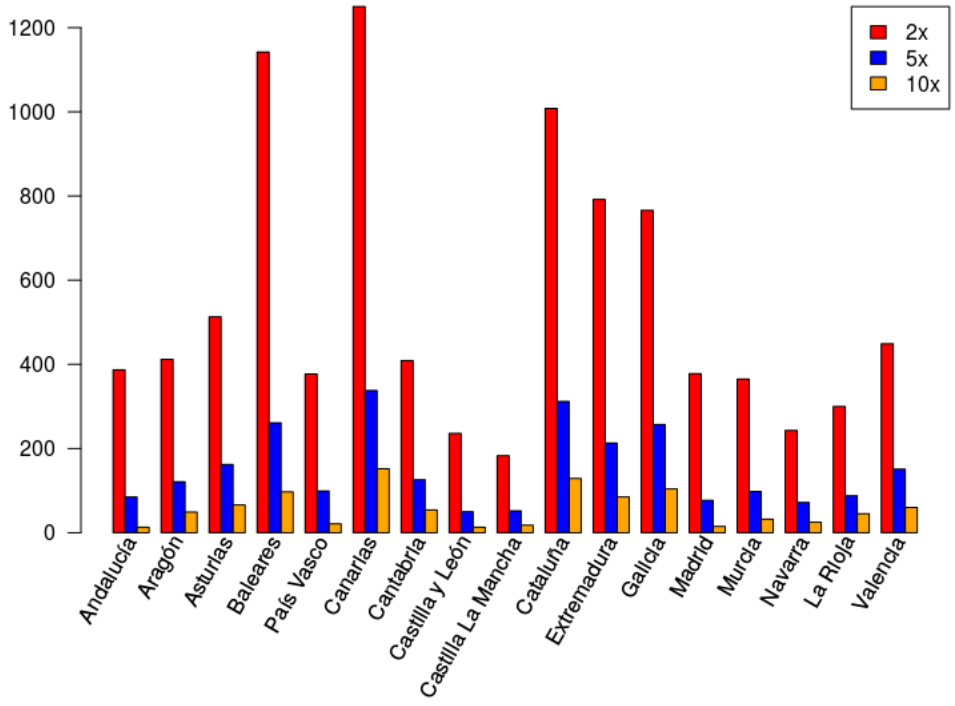

Figura 2: Número de singularidades positivas por CCAA para diferentes umbrales: 2x (rojo), 5x (azul) y 10x (amarillo).

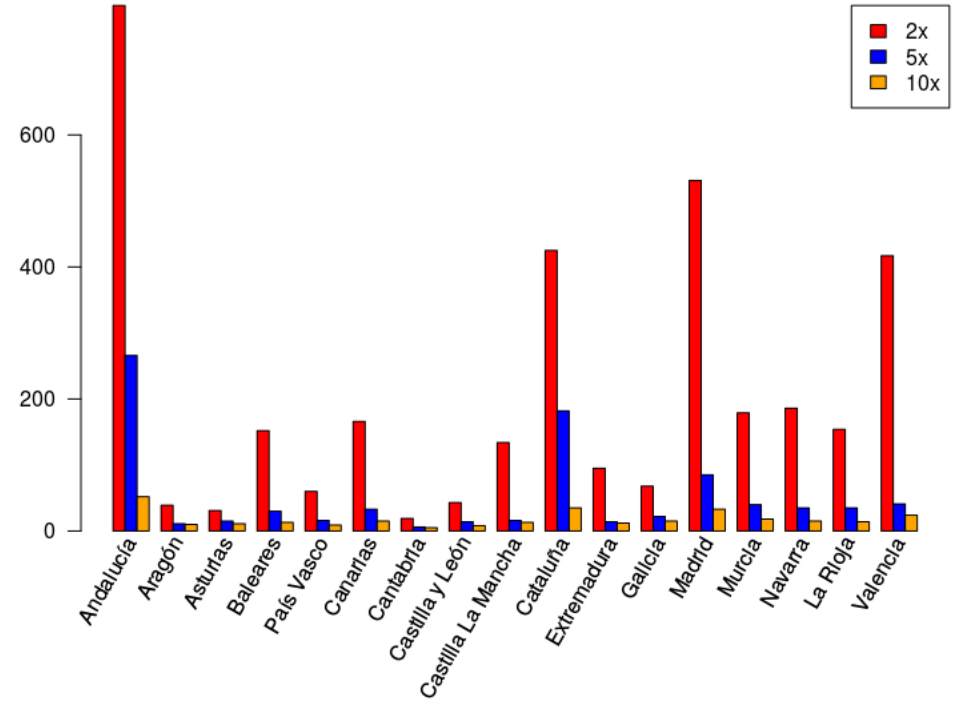

Figura 3: Número de singularidades negativas por CCAA para diferentes umbrales: 2x (rojo), 5x (azul) y 10x (amarillo)

positivas. Estos resultados posicionan a Cataluña como la 2ª o 3ª CA (dependiendo del umbral elegido) con mayor número de singularidades positivas. El en caso de singularidades negativas Cataluña cuenta con 425 (x2), 182 (x5) y 35 (x10), y de nuevo se posiciona como la 2ª o 3ª CA con mayor número de singularidades negativas.

Threshold	2x		5x		10x	
	Mean	std	Mean	std	Mean	std
Andalucía	1.11	2.92	0.99	1.34	0.84	0.89
Aragón	1.03	4.29	2.22	7.48	2.12	3.75
Asturias	1.43	4.92	2.51	6.59	2.81	4.79
Baleares	0.66	3.15	1.65	6.04	2.33	6.74
País Vasco	1.13	3.76	0.86	1.19	0.83	0.81
Canarias	1.03	3.33	2.28	5.60	3.82	7.80
Cantabria	1.19	4.52	2.14	7.86	1.65	3.83
Castilla y León	1.03	2.45	1.76	3.49	2.54	3.92
Castilla La Mancha	0.92	2.52	1.59	3.97	1.78	2.81
Cataluña	0.90	2.77	1.16	2.44	1.39	2.52
Extremadura	0.89	3.50	1.90	6.34	2.95	8.19
Galicia	0.92	3.06	1.60	4.27	2.24	5.11
Madrid	0.64	2.11	0.65	0.75	0.48	0.36
Murcia	1.10	4.25	2.73	7.86	4.67	8.49
Navarra	1.37	5.22	1.25	2.37	1.41	1.80
La Rioja	1.49	4.85	2.87	8.37	3.03	8.30
Valencia	0.95	3.43	1.28	2.34	1.32	1.73
España	1.05	0.24	1.73	0.66	2.3	1.1

Tabla 2: Penetración media de elementos diferenciales

Hemos extendido el análisis para además de cuantificar el volumen de singularidades obtener su penetración en la CA donde aparecen. Definimos penetración como el porcentaje de usuarios de FB en la CA a la que pertenece la singularidad analizada con interés en dicha singularidad. Hemos realizado ese análisis para las singularidades positivas. La Tabla 2 muestra la penetración media y la desviación típica asociadas a las singularidades en cada CCAA para los distintos umbrales. Además mostramos la penetración media en toda España como valor de referencia.

La penetración media de las singularidades en Cataluña para los distintos umbrales es 0.9 % (x2), 1.16 % (x5) and 1.39 % (x10). Obersvando los resultados de la tabla y usando como referencia la media para toda España podemos concluir que la penetración en Cataluña es menor que la media Española. En concreto, si analizamos los resultados para el umbral 2 hay 13 CCAA cuyas singularidades presentan una mayor penetración que las de Cataluña. Solamente Madrid y las Islas Baleares presentan una penetración menor para sus singularidades. Los resultados para el umbral x5 son los mismos, pero en este caso las tres CCAA cuyas singularides alcanzan menor penetración media comparadas con Cataluña son: Madrid, País Vasco y Andalucía. En el caso del umbral x10, obtenemos la misma lista pero añadiendo la Comunidad Valenciana.

Para entender mejor en que consisten las singularidades asociadas a cada CCAA hemos listado en la Tabla 3 el nombre de las singularidades más relevantes en cada CA. Para ello, hemos seleccionado aquellas singularidades positivas de cada CCAA que alcanzan al menos una penetración mayor de un 5 %. Los resultados más relevantes de este análisis son:

- La lista de singularidades es consistente ya que la mayoría de elementos en la lista están relacionados con la CA a la que han sido asignados. Este resultado confirma la validez de los algoritmos de FB en la asignación de intereses a los usuarios, al menos de forma agregada por región. Sin embargo, dichos algoritmos no son perfectos, ya que podemos encontrar singularidades que a priori no tienen ninguna relación con la CA en las que aparecen. Algunos ejemplos son: Última Hora (Paraguay) en Baleares, El Nacional (Caracas) en Cataluña, San Mateo (California) en La Rioja, Oregón en Aragón.

- La mayoría de singularidades se refieren a localizaciones geográficas (ciudades, regiones, aeropuertos, etc) situadas en la CA a las que pertenecen. El porcentaje medio de singularidades ligadas a localizaciones por CA es del 55 % con una desviación típica del 22 %.

Region	Intereses
Andalucía	Andalucía, Málaga, Granada, Cádiz, Estaciones de esquí, Marbella, Almería Provincia de Cádiz, Huelva, Provincia de Málaga, Córdoba Jerez de la Frontera, Costa del Sol, Fuengirola
Aragón	Zaragoza, Aragón, Huesca, mbar, Francisco de Goya, Oregon
Asturias	Asturias, Gijón, Oviedo, Nueva España, Avilés, Sidra, Sporting de Gijón, Independent Online (Sudáfrica), Real Oviedo, David Villa, Academy (Colegio Inglés), Estilista personal
Baleares	Majorca, Islas Baleares, Palma de Mallorca, RCD Mallorca, Ibiza, Última Hora (Paraguay), Minorca, Formentera, Yate, Pachá (discoteca)
País Vasco	Bilbao, País Vasco, Vizcaya, Vitoria, Athletic Bilbao, Euskadi Ta Askatasuna, Taberna, Baracaldo, athletic club, Álava, Real Sociedad de Fútbol
Canarias	Tenerife, Santa Cruz de Tenerife, Gran Canaria, Océano Atlántico Las Palmas de Gran Canaria, Provincia de Las Palmas, Fuerteventura Lanzarote, San Cristóbal de La Laguna, Africa, Unión Deportiva Las Palmas, La Palma, Rudy, Arona (Italia), Santa Cruz, Murga, Volcano, Western European Summer Time, Guerra Civil Española, Hard Rock Cafe, Cantera, Informe, Candelaria, CD Tenerife B, Psittacoidea, Palmas, Nicky Jam, Película biográfica, Department
Cantabria	Cantabria, Santander, Laredo, Racing de Santander, Posada (establecimiento)
Castilla La Mancha	Ciudad Real, Castilla-La Mancha, Toledo, Toledo (Ohio), Guadalajara (México), Cuenca
Castilla León	Valladolid, Salamanca, León, Segovia, Zamora, Ávila, Club León
Cataluña	Cataluña, Cataluña (Albéniz), Gatos, Provincia de Barcelona, Circuito de Barcelona-Catalua, Anexo:Comarcas de Catalua, Catalán, Barcelona-El Prat Airport, Referendum, FC Barcelona Balonmano, El Nacional (Caracas), Tarragona, TV3 (Catalonia), Carles Puyol, Estrella Damm, Esquerra Republicana de Catalunya, Girona, Sabadell, Badalona, Democracy, FC Barcelona B, Mataró, El Peridico de Catalunya, Costa Brava, Tarrasa, Catalunya Experience, TV3 (Malaysia)
Extremadura	Extremadura, Badajoz, Cáceres, Province of Cáceres, Hoy (programa de televisión), República Dominicana, Mérida, Plasencia, Piacenza, Chupa-chups, Santo Domingo, Alcántara, Secretario, Uber (empresa)
Galicia	La Voz de Galicia, Vigo, A Corñua, Pontevedra, Ourense, Santiago de Compostela, Instituto Nacional de Estadística (España), Lugo, Celta de Vigo, Gallego, Tempo, Porto, Galitzia, Haití
Madrid	Comunidad de Madrid, Waze, Aeropuerto Adolfo Suárez Madrid-Barajas
Murcia	Región de Murcia, Cartagena, San Javier, Traducción, Águilas, Regiones de Francia, Arte urbano
Navarra	Navarra, Pamplona, Euskera, Agencia de información, Club Atlético Osasuna, Diario de Notcias
La Rioja	La Rioja (España), Rioja (vino), Tienda de conveniencia, Provincias de Argentina, Laurel, San Mateo (California), Via
Valencia	Valencia, Comunidad Valenciana, Valencia Club de Fútbol, Valenciano, Castellón de la Plana, Provincia de Castellón, Costa Blanca, Gandía, Fallas de Valencia, Valencia CF Mestalla, Manuel de Falla

Tabla 3: Lista de hechos diferenciales por CCAA con una penetración (porcentaje de usuarios de Facebook) superior al 5 %

- Diez de las diecisiete CCAA tienen al menos una singularidad relacionada con un equipo de fútbol local.
- Dos CCAA, Galicia y Navarra, incluyen una singularidad relacionada con medios de comunicación locales.
- En algunos casos encontramos singularidades con un componente cultural importante. Por ejemplo: "Sidra.^{en} Asturias, "Yate.^{en} Baleares, "Taberna.^{en} el País Vasco, "Murga.^{en} las Islas

Canarias, Rioja (Vino).[en] La Rioja y "Fallas.[en] la Comunidad Valenciana.

Para el caso concreto de Cataluña, encontramos muchos de los puntos anteriores entre sus singularidades. Incluye 11 singularidades relacionadas con algún tipo de localización; 4 singularidades relacionadas con deportes (Circuito de Barcelona-Cataluña, FC Barcelona Handbol, FC Barcelona B and Carles Puyol); una singularidad ligada a un elemento cultural muy importante como es la lengua Catalana; dos elementos relacionados con medios de comunicación (TV3 (Catalonia) y El Periódico de Catalunya); y un interes que, en principio, no está relacionado con Cataluña (El Nacional (Caracas)). Además, cabe destacar que Cataluña presenta cuatro categorías concretas entre sus singularidades que no aparecen en ninguna otra CCAA: (i) una cerveza local (Estrella Damn), un partido político (Esquerra Republicana de Catalunya) y dos elementos con una importante componente político: Referendum y Democracia.

4. Conclusión

En este documento analizamos datos de Facebook para contribuir desde un punto de vista diferente al intenso debate que está teniendo lugar en España en relación al conflicto político existente en Cataluña. Para ello analizamos: (i) la similaridad socio-cultural existente entre las diecisiete CCAA Españolas usando un vector formado por más de 67.000 intereses extraídos de 2.101 usuarios Españoles de Facebook; (ii) las singularidades presentes en cada CA. Los resultados obtenidos muestran que: (i) Todas las CCAA muestran una gran similaridad entre ellas; (ii) Cataluña se sitúa como la 5ª CA con menor similaridad socio-cultural (mediana) con el resto de CCAA; (iii) Catauña se posiciona como la 2ª CA con mayor número de sigularidades, pero ocupa la 14ª posición en el porcentaje de población al que interesan esas singularidades; (iv) La mayor parte de singularides en las CCAA están relacionadas con localizaciones geográficas o deporte. Cataluña es la única CA que presenta singularidades con un importante componente político (Esquerra Republicana de Catalunya, Referendum and Democracia).

Referencias

[1] Reuters, "Catalan parliament declares independence from spain," https://www.reuters.com/article/us-spain-politics-catalonia-independence/catalan-parliament-declares-independence-from-spain-idUSKBN1CW1WO, accessed: 2017-11-20.

[2] CNN, "Catalonia declares independence from spain as political crisis deepens," http://edition.cnn.com/2017/10/27/europe/gallery/catalonia-independence/index.html, accessed: 2017-11-20.

[3] The-Washington-Post, "Catalonia's leaders are jailed a week after the region declared independence," https://www.washingtonpost.com/world/europe/catalonias-leaders-are-jailed-a-week-after-the-region-declared-independence/2017/11/02/9148bce2-bff1-11e7-8444-a0d4f04b89eb_story.html?utm_term=.d7db290e7919, accessed: 2017-11-20.

[4] El-País, "Former catalan premier, ministers due in belgian court on november 17," https://elpais.com/elpais/2017/11/06/inenglish/1509950966_541140.html, accessed: 2017-11-20.

[5] Periodista-Digital, "Cataluña es un pueblo aparte. es una nación independiente," http://www.periodistadigital.com/religion/opinion/2017/09/17/religion-iglesia-espana-opinion-josep-miquel-bausset-cataluna-es-un-pueblo-aparte-es-una-nacion-independiente-azorin-cataluna-y-el-sentido-comun.shtml, accessed: 2017-11-20.

[6] S. Harris, "Catalonia is not spain: A historical perspective," 4 Cats Books. ISBN13: 9781502512307.

[7] El-País, "Cuando la historia se cuenta para convencer," https://politica.elpais.com/politica/2017/10/13/actualidad/1507911804_253957.html, accessed: 2017-11-20.

[8] "Lo que todo independentista catalán debería saber," https://paraindependentistacatalan.blogspot.com.es/p/es-viable-una-cataluna-independiente.html, accessed: 2017-11-20.

[9] "¿es cataluña una nación?" www.expansion.com/opinion/2015/09/11/55f34554e2704e86618b4598.html, accessed: 2017-11-20.

[10] J. González Cabañas, A. Cuevas, and R. Cuevas, "Fdvt: Data valuation tool for facebook users," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 3799–3809. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025903

[11] P. R. Christopher D. Manning and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.