Anàlisi de les Assemblees Ciutadanes pel Clima de Catalunya i Barcelona: Quins tòpics obtenen més interacció i aprovació?

Arnau Berenguer Jiménez

Eurecat - Centre Tecnològic de Catalunya

Big Data & Data Science

5 de juny de 2024

Resum

Aquest article analitza les Assemblees Ciutadanes pel Clima de Catalunya i Barcelona utilitzant tècniques de modelatge de tòpics per identificar els tòpics que generen més interacció i aprovació entre els ciutadans. S'han utilitzat diversos models, incloent *Latent Semantic Analysis*, *Latent Dirichlet Allocation* i *BERTopic*, per extreure els tòpics de les propostes recollides a través de la plataforma Decidim. Els resultats mostren que els temes relacionats amb la mobilitat i transport, l'aigua i l'energia són els que més interès i suport generen. També es discuteixen les diferències en els tòpics entre les dues assemblees i la qualitat de les propostes d'ambdues.

Paraules clau: *Topic modeling, Decidim, canvi climàtic, assemblees ciutadanes, participació ciutadana, Catalunya, Barcelona*

1 Introducció

Avui en dia, al món globalitzat en què vivim, la democràcia s'ha reinventat adoptant els avenços tecnològics de l'actualitat. Així doncs, estem contemplant noves formes de democràcia participativa i deliberativa a través dels ordinadors i d'internet. Una d'aquestes s'ha manifestat en forma de Decidim, una plataforma en línia de codi lliure i obert que permet formar part de processos participatius de manera senzilla i a través de qualsevol dispositiu amb connexió a internet. Tant la Generalitat de Catalunya com l'Ajuntament de Barcelona han desplegat instàncies de Decidim per promoure la participació

ciutadana en diversos processos participatius en forma de propostes. En el cas d'aquest article, els processos participatius que es tracten són l'Assemblea Ciutadana pel Clima de Catalunya *(Generalitat de Catalunya)* i l'Assemblea Ciutadana pel Clima de Barcelona *(Ajuntament de Barcelona)*.
Les propostes poden ser proposades per un ciutadà o per la mateixa organització a la que pertany la instància. Aquestes contenen un títol descriptiu i un cos on s'explica el contingut de la proposta. També els usuaris registrats poden comentar, esmenar, recolzar i seguir les propostes. Opcionalment, les propostes poden estar classificades a través d'una o vàries categories. L'objectiu d'aquest treball és determinar quin model determina millor els tòpics generals de cada proposta, entendre quins són els que obtenen més interacció (suports, comentaris, seguidors) i quins tenen més ràtio d'aprovació per entendre quins tòpics climàtics preocupen més a la ciutadania.
L'estructura de l'article és la següent: A la secció 2 es presenten les dades de les propostes de les assemblees amb què s'ha treballat i la metodologia que s'ha seguit per tractar-les i fer el *topic modeling*. Després, a la secció 3 es presenten els resultats obtinguts de l'estudi. Seguidament, a la secció 4 es presenten les conclusions obtingudes a partir de l'anàlisi dels resultats. Finalment, a la secció 5 s'expliquen les possibles línies futures per seguir amb aquest estudi.

2 Metodologia

2.1 Datasets

A través de les funcionalitats integrades a Decidim per exportar tot tipus de dades obertes, es va descarregar les dades d'ambdues instàncies. Per a cada instància hi havia sis datasets diferents:

- *Meetings*
- *Meeting Comments*
- *Projects*
- *Proposals*
- *Proposal Comments*
- *Results*
- *Result Comments*

De tots aquests datasets, només van ser útils per l'objectiu de l'estudi els de *Proposals* i *Proposal Comments* ja que són els que contenien informació sobre les propostes i els comentaris de cada proposta.

Un cop obtinguts els datasets necessaris, es va procedir a realitzar *Exploratory Data Analysis* a cadascun d'ells per comprendre quina era la informació que contenia cadascun i quines mides tenien.

	Organitzacions	
	Generalitat de Catalunya	**Barcelona**
Proposals	(10440, 45)	(31407, 37)
Proposal Comments	(5171, 16)	(29131, 15)

Taula 1: Nombre de files i columnes a cada dataset

Un cop filtrades les dades per només quedar-nos amb les dels processos participatius de les assemblees, el nombre de files es va reduir considerablement. En aquest cas, com que l'objectiu d'aquest estudi només era entendre quins tòpics obtenien més interacció, es va decidir prescindir d'analitzar els datasets de comentaris.
Cal destacar que, a les dades de *l'Assemblea Ciutadana pel Clima de Catalunya*, la majoria de les propostes estaven en estat d'evaluació, però per contra havien obtingut interacció d'altres usuaris, i que les dades de *l'Assemblea Ciutadana pel Clima de Barcelona*, al ser penjades després dels debats presencials, la majoria tenia algun estat definit però la interacció era pràcticament nul·la.

	Organitzacions	
	Generalitat de Catalunya	**Barcelona**
Proposals	59	37
Proposal Comments	-	-

Taula 2: Nombre de files a cada dataset

Un cop escollides les dades amb les que es va treballar, el primer va ser eliminar les columnes que continguessin valors NaNs i prou. Només amb això, es va aconseguir reduir el nombre de columnes a 11 al dataset de la Generalitat i 26 al de Barcelona. Després, es van eliminar les files duplicades restants al dataset.

Seguidament, es va analitzar cada columna per separat per poder veure quins tipus de dades contenia cadascuna i extreure gràfics per representar-les. Finalment, es va analitzar la correlació entre les columnes a través d'una matriu

de correlacions i per visualitzar la correlació entre elles es van realitzar *plots* de regressions lineals.

2.2 Topic Modeling

Per a poder realitzar *Topic Modeling* per extreure els tòpics de cada proposta, primer es va determinar quins models de tòpics es farien servir. Es va determinar que serien tant models bayesians com basats en *transformers*:

***Bayesians*:**

1. *Latent Semantic Analysis*
2. *Latent Drichlet Allocation*

Basats en *transformers*:

1. *BERTopic*

Per poder extreure els tòpics, es va emprar només la columna que contenia el cos de la proposta. En el cas dels models bayesians, va caldre primer processar el text per *tokenitzar-lo* i després vectoritzar-lo en un espai de característiques *TF-IDF*. En el cas del model *BERT* només va caldre vectoritzar-lo de manera que es contés el nombre de vegades que apareixia cada *token* al conjunt de textos (*CountVectorizer*).

Llavors, es va realitzar una cerca dels hiperparàmetres més òptims per tots els models mitjançant la mitjana d'error quadràtic per *LSA i LDA* i la *silhouette* pel *BERTopic*. Després se'n va calcular la coherència per comparar com d'eficaços havien resultat els models. Finalment, es va analitzar els tòpics assignats a cada proposta per veure quins eren els que obtenien més interacció i a quin estat d'aprovació es trobaven.

3 Resultats

La mètrica de la coherència crea vectors de contingut de paraules emprant la seva co-ocurrència i després en calcula la puntuació a través de la Informació Mútua Puntual Normalitzada (*NPMI* en anglès) i la similaritat de cosinus. El valor oscil·la entre el rang [-1, 1] i un valor més proper a 1 indica que els tòpics són coherents i similars/relacionats entre sí. En canvi, valors més propers a -1 indiquen que no estan relacionats i són poc coherents. S'observa

a la taula 3 que amb gran diferència el model *BERTopic* és el que millor coherència va obtenir, amb una millora de 3.9x sobre LSA i de 3.41x sobre LDA. Va ser per això que es va decidir analitzar únicament els tòpics generats per aquest model i implementar directament només aquest per les dades de l'assemblea de Barcelona.

	Models		
	LSA	**LDA**	**BERTopic**
Coherència	0.19033012	0.21683089	0.74019911

Taula 3: Puntuació de coherència de cada model (dades Catalunya)

3.1 Assemblea Ciutadana pel Clima de Catalunya

A la figura 1 es poden observar els tòpics detectats i ordenats per nombre de propostes. Es destaca que el tòpic amb més propostes és *participar - agradaria - assemblea*, però això és degut a que moltes propostes eren de gent que no habia rebut la carta aleatòria per participar a l'assemblea i volia formarne part. D'altra banda, destaquen amb entre 8 i 6 propostes per cadascun els tòpics *aigua - haurien - edificis*, *co2 - caldria - canvi*, *transport - resulta - trajectes* i *aliments - pagesos - alimentària*.
A la figura 2 es pot observar el nombre de comentaris totals que ha rebut cada tòpic, destacant que el més comentat és el de *transport - resulta - trajectes* amb 6 comentaris, seguit pel de *participar - agradaria - assemblea* amb 4 comentaris. Pel què fa a la resta, només obtenen 1 sol comentari o cap.
S'observa a la figura 3 que el tòpic amb major nombre de seguidors es tracta de *aigua - haurien - edificis*, seguit de *participar - agradaria - assemblea*, que com s'ha comentat anteriorment no és representatiu de propostes reals. La majoria de propostes es troben en el rang d'entre 11 i 8 seguidors.
D'altra banda, a la figura 4, que mostra el nombre d'aprovacions (*endorsements*) que ha rebut cada tòpic, es veu amb claredat que el tòpic amb més aprovacions es tracta de *aigua - haurien - edificis*, seguit de *aliments - pagesos - alimentària*, amb 16 i 13 aprovacions respectivament. En canvi, la resta de tòpics sense comptar el de *participar - agradaria - assemblea* obtenen entre 9 i 5 endorsements, essent el valor més repetit el 5.
Per tancar l'anàlisi de les dades de *l'Assemblea Ciutadana pel Clima de Catalunya*, a la figura 5 es contempla el total d'interaccions que ha rebut cada tòpic, $I_i = E + C + F$, on I_i és el total d'interaccions del tòpic i, E és el nombre *d'endorsements*, C és el nombre de comentaris i F és el nombre de seguidors. Es veu que, el tòpic amb més interacció és *aigua - haurien - edificis*, seguit del tòpic no rellevant *participar - agradaria - assemblea*.

3.2 Assemblea Ciutadana pel Clima de Barcelona

Com s'ha mencionat anteriorment, les dades de *l'Assemblea Ciutadana pel Clima de Barcelona* careix de pràcticament interaccions ja que les propostes penjades a la web són els resultats de les propostes fetes presencialment a les trobades del procés. És per això que no se n'analitzaran les interaccions si no l'estat de les propostes únicament.
Aquest dataset, a diferència del de l'Assemblea Ciutadana pel Clima de Catalunya, contenia les diferents propostes classificades per categories, el què ha permès un anàlisi d'aquestes. A la figura 6 s'observen les diferents categories i la seva distribució en funció del nombre total de respostes:

1. *Consum i Residus*: 13 propostes
2. *Energia*: 10 propostes
3. *Mobilitat i Infraestructures*: 11 propostes

A la figura 7 s'observa el nombre de propostes de cada categoria que han estat acceptades, en avaluació o rebutjades. Es veu clarament que la categoria amb més aprovacions és la *d'Energia*, amb un 90% (9/10) aprovades. En canvi, veient les propostes de les altres dues categories, es veu que la majoria estan en avaluació encara, amb un 61% (8/13) a la categoria de *Consum i Residus* i un 73% (8/11) a la categoria de *Mobilitat i Infraestructures*.
La coherència resultant en aplicar el model *BERTopic* ha estat de 0.78983086, i els tòpics obtinguts pel model han estat:

1. *ciclistes - vianants - eixos*
2. *contaminants - malbaratament - emissions*
3. *empreses - segell - formen*
4. *energètic - consum - reducció*
5. *porta - sistema - recollida*
6. *punts - serveis - productes*
7. *públic - transport - mobilitat*
8. *recursos - reparació - cívics*
9. *tecnologies - energètiques - barcelona*
10. *tràmits - transició - comunitats*

Finalment, a la figura 8, s'observa la distribució de propostes per tòpic i estat. Destaca que el tòpic amb més propostes acceptades, però també rebutjades és el de *energètic - consum - reducció*. Empatant en nombre de propostes, també es troba el tòpic de *públic - transport - mobilitat*, amb 6 propostes, 5 en avaluació i 1 acceptada. En canvi, el tòpic amb menys propostes de tot es tracta de *punts - serveis - mobilitats*.

4 Conclusions

A través d'aquest estudi, primerament s'ha vist que el millor model dins dels proposats per a realitzar les tasques de *Topic Modeling* és, amb diferència, el *BERTopic*. Aquest va obtenir una puntuació 3.9 vegades major que la del model LSA i 3.41 vegades major que la del model LDA.
S'han observat diferències entre els tòpics generats al dataset de *l'Assemblea Ciutadana pel Clima de Catalunya* i el de *l'Assemblea Ciutadana pel Clima de Barcelona*, ja es veu que en alguns tòpics de Catalunya apareixen preposicions o que no tenen massa sentit i en canvi, als tòpics de Barcelona es veu que estan bastant lligats i no apareix cap preposició ni paraula estranya. Aquest fenomen segurament està lligat al fet de que les propostes de Barcelona han estat redactades i revisades per la mateixa organització i en canvi, les de Catalunya qualsevol ciutadà podia proposar-les i escriure-les com creiés adient. D'altra banda, s'ha observat que en ambdós datasets existeix un tòpic força relacionat amb la mobilitat i el transport (*transport - resulta - trajectes* i *públic - transport - mobilitat*) i que és tant dels més interactuats com dels que més propostes no rebutjades té. També s'ha mostrat que la ciutadania mostra interès en altres tòpics com la reducció del consum energètic i l'energia, la contaminació i les emissions, o l'aigua. També s'ha descobert que una bona part de les propostes és gent que vol participar a l'assemblea i no ha rebut la carta aleatòria de participació, cosa que demostra que hi ha part de la ciutadania que li preocupa el canvi climàtic.
Finalment, s'ha de destacar que aquest estudi presenta la principal limitació de tenir un nombre de mostres reduït i això dificulta extreure conclusions generalitzables.

5 Línies futures

Si bé aquest estudi ha aportat informació sobre quins tòpics són els que més preocupen a la ciutadania basant-nos en el nivell d'interacció que reben i si han estat aprovats, es podria ampliar la recerca de diferents maneres. Primer

de tot, com s'ha comentat a l'apartat anterior, es podria mirar d'obtenir noves dades per ampliar el tamany del dataset i obtenir millors resultats i sobretot més generalitzables. D'altra banda, aquest estudi només ha determinat quins tòpics són els que més interessen/preocupen a la ciutadania en base a les dades recaptades, però no s'ha centrat en les característiques de cadascun o en quines característiques tenen les propostes aprovades que les diferencien de les rebutjades. Seria certament interessant entendre quines característiques defineixen una proposta aprovada i una proposta rebutjada per definir un model d'intel·ligència artificial que predigui si una proposta serà acceptada o no i entendre què fa que sigui rebutjada o no.

6 Apèndix

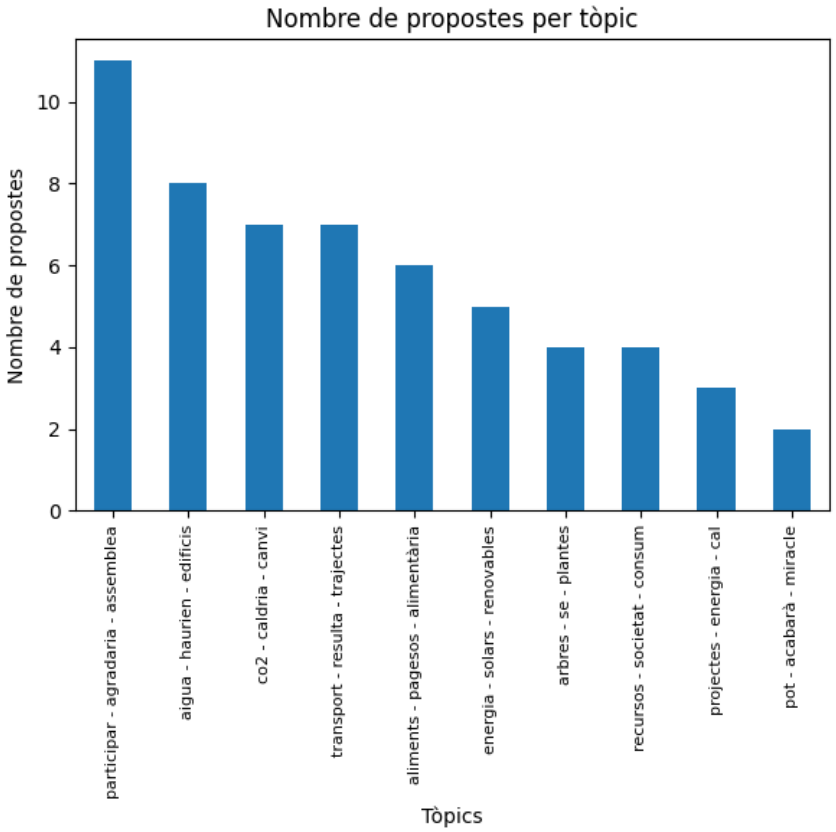

Figura 1: Nombre de propostes per tòpic (Generalitat)

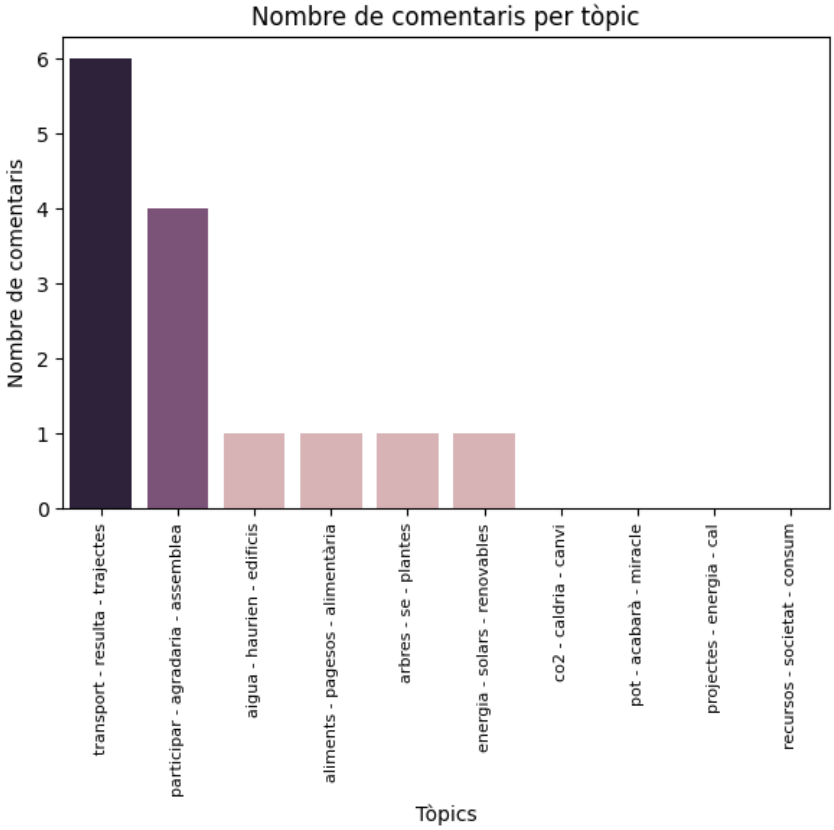

Figura 2: Nombre de comentaris per tòpic (Generalitat)

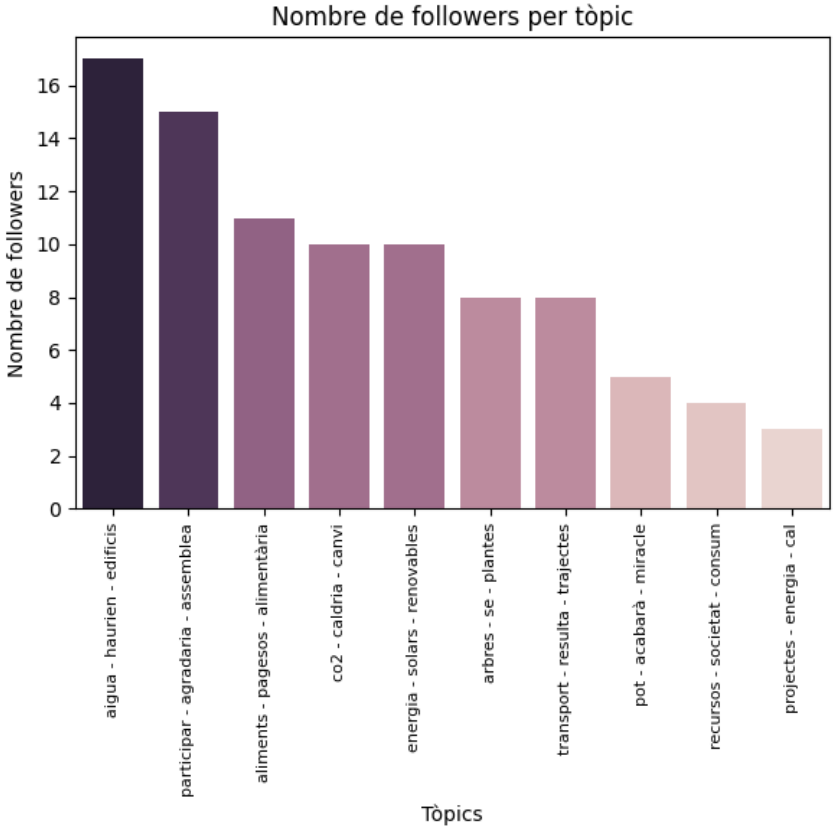

Figura 3: Nombre de seguidors per tòpic (Generalitat)

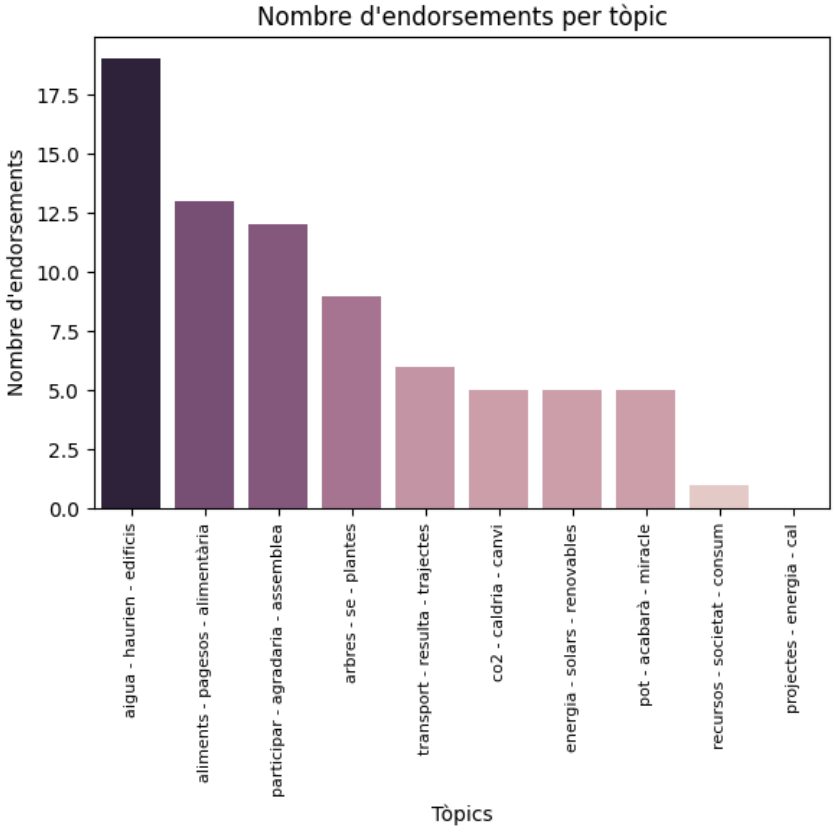

Figura 4: Nombre *d'endorsements* per tòpic (Generalitat)

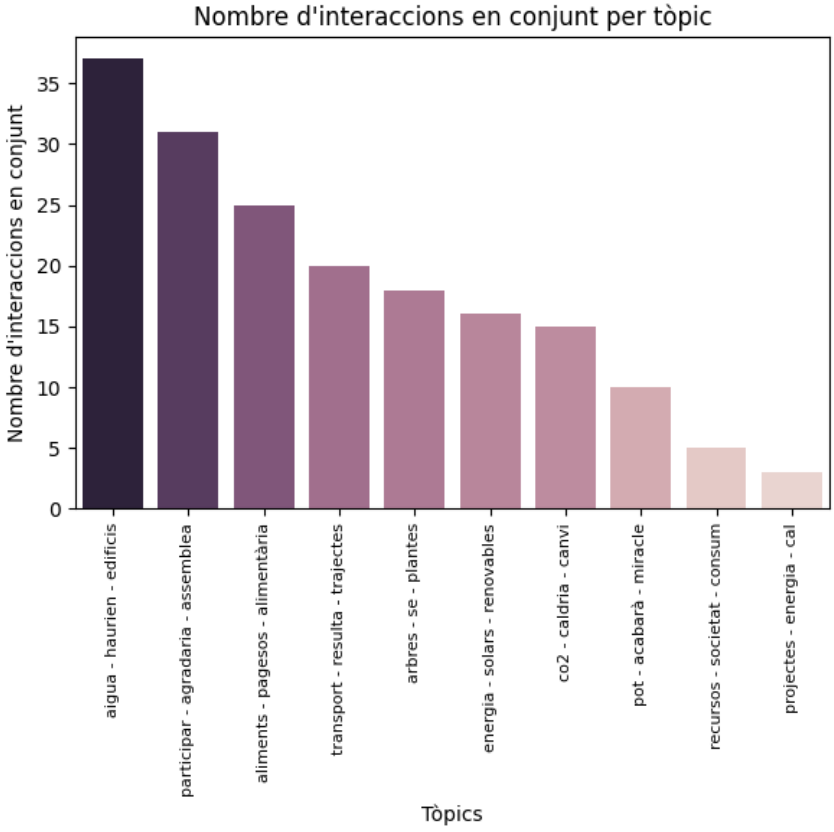

Figura 5: Interaccions totals per tòpic (Generalitat)

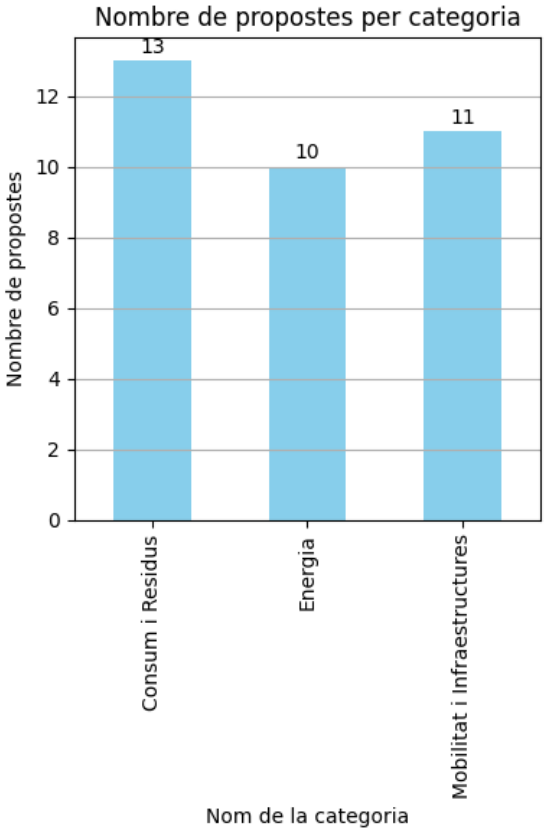

Figura 6: Nombre de propostes per categoria (Barcelona)

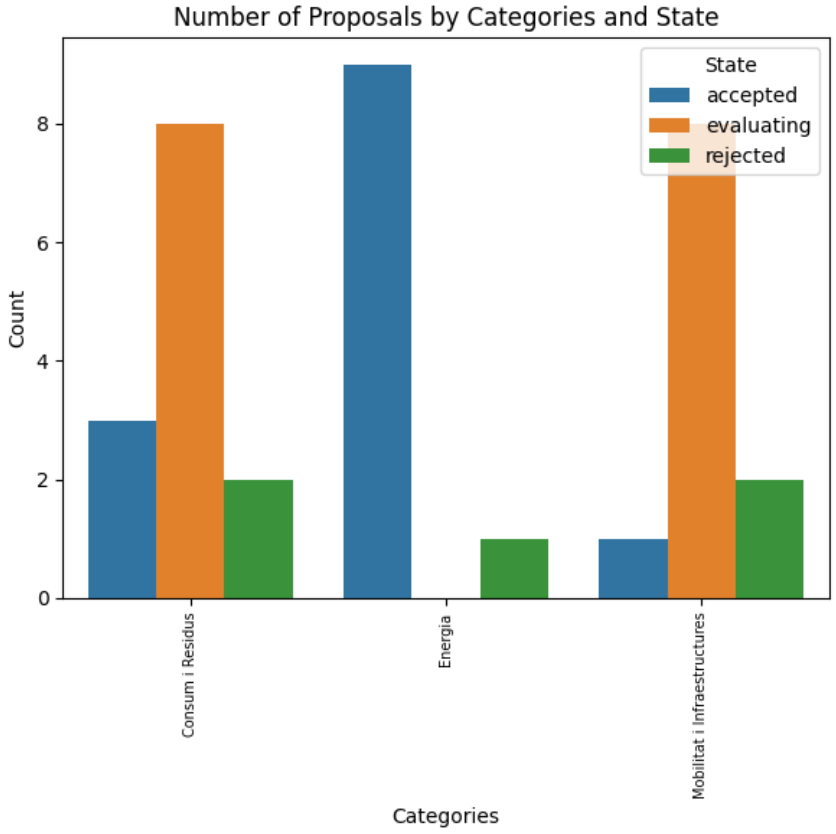

Figura 7: Propostes per categoria i estat d'aprovació (Barcelona)

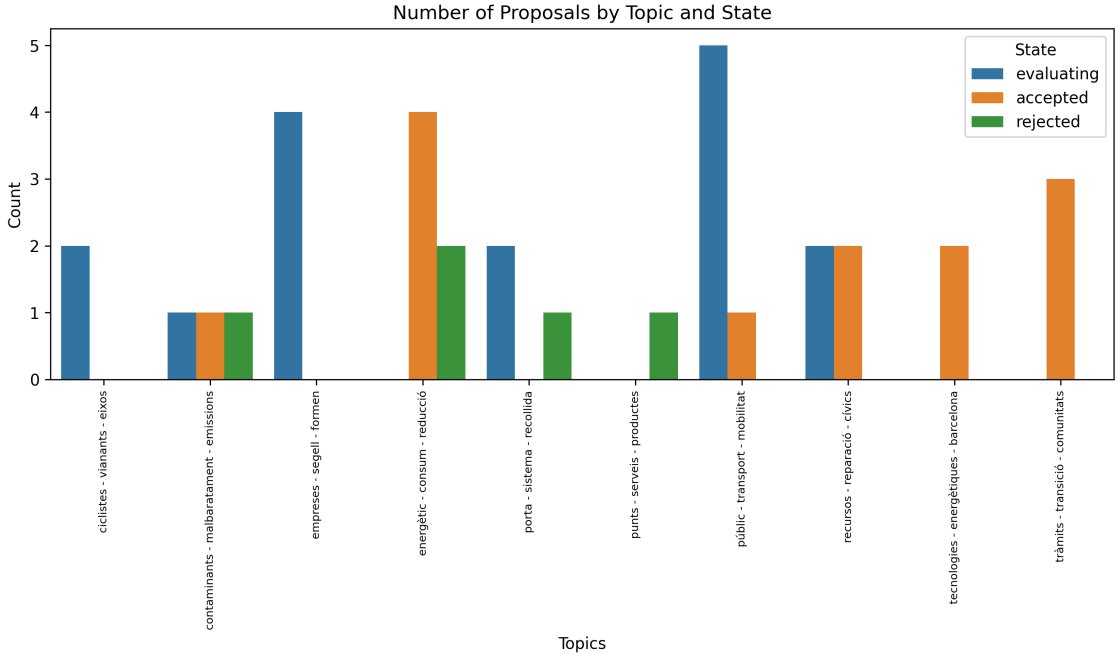

Figura 8: Propostes per tòpic i estat d'aprovació (Barcelona)