**Methodology**

**Data description**

We performed a text analysis and tweets gathering of political actors of California and Florida states on the social network (Twitter). Political actors of both states have been identified through the official website of Assembly and Senate. Each state has different numbers of political representatives; for California State, the Assembly has 80 seats and Senate 40, while for Florida State, 120 and 40, respectively. The tweets collections were conducted between April 30, 2022, and April 30, 2023, using the Twitter API. Due to the extensive number of our aimed political actors -280 in total-, we determined that a sample of 20 per cent for both parties, randomly selected and equally distributed among the assembly and senate. As a sample result, we obtained 24 political representatives for California state, and for Florida, 32. The following table shows the results of the sample, as well as their respective tweet counts.

**Table 1**: Sample results

|  |  |  |
| --- | --- | --- |
| **State / party** | **# of tweets** | **# political actors** |
| **California** | 4266 | 24 |
| Democrats | 2268 | 12 |
| Republicans | 1998 | 12 |
| **Florida** | 4802 | 32 |
| Democrats | 2752 | 16 |
| Republicans | 2050 | 16 |

**Description of analysis method**

We conduct an analysis of text based on topic modelling with Latent Dirichlet allocation. LDA can be used to identify and describe latent thematic structures within collections of text documents (Blei, 2012). The aim of the LDA algorithm is to model a comprehensive representation of the corpus by inferring latent content variables, called topics (Maier et al., 2018, p. 2).

A topic is technically defined as a distribution over words: For every word in every document, the topic contains the estimated probability that this word occurs when the given topic is covered (Günter & Domahidi, 2017. p.3056).

To apply the LDA, we previously need to work in a preprocessing step according to Figure 01. This includes after data collection, removing common words, errors, or other not essential characteristics of English language, among other further steps.

**Figure 1:** *preprocessing stages*

The typical methods for preparing language involve tokenization (dividing documents into individual term components), removing punctuation and capitalization from words, eliminating stopwords as well as very common and rare terms (pruning based on frequency), and applying stemming and/or lemmatization. Stemming and lemmatization are utilized to render inflected words comparable to one another (Maier et al., 2018).

We perform these steps with python gensim package *Ldamodel* , an open-source library designed to process raw, unstructured digital text using unsupervised learning algorithms (Gensim, 2022).

*Data wrangling*

Data cleaning process were employed to clear data from unimportant characteristics and irrelevant information and pre-process a reliable data set. URLs, hashtags, punctuation, @mentions, RT, other languages rather than English, among others, were removed from the data, the tweets were tokenized to extract individual words, and lastly, stop words were removed from the dataset.

*Tokenization*

Word evaluation process is considered as categorical variable. It is therefore appropriate to use a technique of converting a text string into a sequence of tokens to make viable for analysis by an algorithm, in this case a unsupervised algorithm. Tokens are often loosely referred to as terms or words. One token is an instance of a sequence of characters in some document that are grouped together as a useful semantic unit for processing (Manning et al., 2008).

Lemmatization

In natural language processing, large paradigms imply an increased token to type ratio, greatly increasing the number of unknown words. One method to combat this issue is to lemmatize the sentence (May et al. 2019. p.2). Lemmatizing converts them to their lemma form/lexeme (e.g., “contaminating” and “contamination” become “contaminate”) (Manning & Schütze, 2003, p. 132).

Keywords Filtering

To aim our research questions is important to have a filter based on 218 key words related to climate change. These keywords have been previous defined according to the keywords of the glossary report section of the IPCC, 2018: Global Warming of 1.5°C. This stage leads to a decrease of our sample size, as follow:

**Table 2:** Filtered sample results

|  |  |  |
| --- | --- | --- |
| **State / party** | **# of tweets** | **# political actors** |
| **California** | 1479 | 24 |
| Democrats | 726 | 12 |
| Republicans | 756 | 12 |
| **Florida** | 1089 | 32 |
| Democrats | 573 | 16 |
| Republicans | 516 | 16 |

Bigrams

We have collectively referred to single words or unigrams. However, at this stage, and given that our sample and many of the climate change keywords may consist of word pairs that have a strong relationship to climate change (e.g., "sustainable development"), we are including bigrams to improve the corpus analysis.

An n-gram represents a sequence n word: a 2-gram (bigram) is at paired word sequence. It approximates the likelihood of a word based on all the previous words by using the conditional probability of the preceding word. (Jurafsky & Martin, 2023)

LDA

Latent Dirichlet Allocation is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei et al., 2003, p.2). LDA is known as an unsupervised method and considers that each ngram is associated with a specific latent topic and uses the grams to determine the topics of documents and provides probabilities associated with the topics (Inoue et al. 2022, p.3).

Considering this model concept, we decide to perform topic modeling with td-idf weights previous to the development using gensim. The TF-IDF has been widely used in the fields of information retrieval and text mining to evaluate the relationship for each word in the collection of documents (Kim & Gil, 2019. p.8).

Since we lack information of the number of latent topics included and the optimal values of the hyperparameters to initialize the model, we decide to start randomly from 15 topics and progressively decreasing until we find linguistic coherence between and within the topics. Coherence score that corresponds well with human coherence judgments and makes it possible to identify specific semantic problems in topic models without human evaluations or external reference corpora (Mimno et al., 2011, p. 262).

We use the coherence “c\_v” from gensim as a measure to help us determine the optimal number of topics to consider for this project.

**Results**

Our resulting topic model contains 5 topics under a coherence level of 0.55 value. We set a limit of 10 most frequents words in for each topic. Table 3 presents the word probabilities of the top 10 words in each latent topic of our sample results as much the count of documents per latent topic. The topics contains at least one of the predefined keywords linked to climate change, as well there are bigrams allowing to identify better associated words. Also, the most appropriate topic titles were selected manually, thus contrasting with the linguistic understanding of human comprehension, and were assigned a name after discussion among group members, setting aside the labelling of topics from an automatically generated process.

**Table 3: The top ten associated words with each topic for climate change**

|  |  |  |
| --- | --- | --- |
| Topics | # documents | words |
| Gas | 465 | {update, acre contained, information, gas, size, mountain, available, increase, human trafficking, democrat} |
| Fire | 541 | {gas tax, fire, assembly, evacuation order, bill, public safety, California, snow, total, public} |
| Water and weather | 555 | {update, weather, service, local, state, line, water storage, prop, please, committee} |
| Energy price | 518 | {gas price, state, energy, update, Californian, fire, gallon, month, acre, yet} |
| Public safety and fire | 489 | {safety committee, bill, penalty, assembly public, governor, fire, help,  unified, mosquito fire, year} |

As previously mentioned, the degree of coherence serves as a metric for identifying and refining the ideal number of topics, as it offers a solution for limiting the quantity of topics. Figure 1 offers valuable insights into our decision-making process. We have opted for just 5 topics, driven by the fact that the coherence level at this point stands at 0.5556, the first highest value within a range spanning from 2 to 15 iterations of topics.

**Figure 1: Coherence values**

A graph with numbers and lines

Description automatically generated

Generally, when more topics we generate, the coherence level increase and more narrowed topics will be encounter. However, accepting too many topics might result in similar entities that cannot be distinguished in a meaningful way (e.g., Grimmer, 2010, pp. 12–13). At the same time, too few topics might lead to very broad entities combining different aspects that should be separated (Evans, 2014, p. 2).

Moreover, our analysis extends to identifying the pertinent subjects discussed by political actors on Twitter. Figure 2 illustrates the frequency of topics associated with climate change, categorized by party. It is evident that the topic of *"water and weather"* holds notable prominence, particularly within the Democratic party. The second most prevalent topic is *"fire"*, which exhibits greater significance for the Republican party. Interestingly, *"energy price"* emerges as a shared concern among both Democrats and Republicans, displaying only marginal differences between the two.

**Figure 2: Number of documents by topic and party**

A graph of different blue bars

Description automatically generated with medium confidence

Hence, we can deduce that Republicans place a higher emphasis on discussing fires, whereas Democrats demonstrate a greater inclination to reference "water and weather." Nevertheless, both parties converge when addressing concerns related to energy prices. The subjects that receive comparatively less attention from both parties, though still of significant relevance, include gas and public safety and fire.

**References**

Blei, D.M., et al. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022

Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. PLoS One, 9(2), 1–10.

Gensim: topic modelling for humans. (2022, December 21). <https://radimrehurek.com/gensim/intro.html#what-is-gensim>

Daniel Maier, A Waldherr, P Miltner, G Wiedemann, A Niekler, A Keinert, B Pfetsch, G Heyer, U Reber, T Häussler, H Schmid-Petri & S Adam (2018): Applying LDA topic modeling in communication research: Toward a valid and reliable methodology, Communication Methods and Measures, DOI: 10.1080/19312458.2018.1430754

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, USA, 262–272

Manning, CD., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval* (Anniversary). Cambridge University Press.

Manning, C. D., & Schütze, H. (2003). Foundations of statistical natural language processing (6. print with corr.). Cambridge, MA: MIT Press.

May, C., Cotterell, R., & Durme, B.V. (2016). Analysis of Morphology in Topic Modeling. ArXiv, abs/1608.03995.

Jurafsky, D., H. Martin, James. (2023). Speech and Language Processing. <https://web.stanford.edu/~jurafsky/slp3>

Kim, S.‑W., & Gil, J.‑M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-Centric Computing and Information Sciences, 9(1). https://doi.org/10.1186/s13673-019-0192-7

IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, 616 pp. https://doi.org/ 10.1017/9781009157940.

Günther, Elisabeth & Domahidi, Emese. (2017). What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals. International (Inoue et al., 2023)Journal of Communicati (Dan Jurafsky, 2023)on. 11. 3051-3071.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate

press releases. Political Analysis, 18(1), 1–35.

Inoue, M., Fukahori, H., Matsubara, M., Yoshinaga, N., & Tohira, H. (2023). Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing. Japan Journal of Nursing Science : JJNS, 20(2), e12520. https://doi.org/10.1111/jjns.12520 (Kim & Gil, 2019)