TECHNISCHE UNIVERSITÄT ILMENAU
Department of Economic Sciences and Media

Data Science in Digital Media: Graph Analysis and Text Mining in Practice

Emese Domahidi
Eric Tröbs

Discussion of climate change by political actors: A topic modelling analysis on Twitter

Eyson Moises Panduro Vasquez (65131): Methodology design, statistical analysis, creation of graphs, editing results section, and providing insights for discussion and conclusions.
Sobha Suseela Kesavan (65625): Managed verified Twitter account data, cleaned raw data, selected keywords, supported literature review, drafted initial version; contributed to conclusion and discussion aspects.
Martina Marcacuzco Huamani (65114): Editing introduction, literature review from relevant research, cross-checking references, refined findings, improved writing, and collaborated on ethics section and discussion of members' input.
Saim Ahmed (65300): Implementation of LDA model in Python, coding and fixing errors, supporting the analysis for the model, and contribution in the conclusion section.

Ilmenau, August 27, 2023

Contents

1 Introduction

According to Veltri and Atanasova (2015), there is a growing body of study on climate change discourse in social media, with a particular emphasis on Twitter. Given the general consensus that studying Twitter, in particular, is important for a variety of reasons—from the fact that it is now simply too important to ignore given its global reach and steadily increasing number of users and posts to the ability to provide a window on various aspects of society—this focus should not come as a surprise.

According to Irawan et al. (2020), one of the most widely used social media sites is Twitter, which enables users to publish their opinions on anything in typically condensed form. Twitter is a significant source of information for assessing people's behavior and propensity to respond to a particular political problem because of its enormous user base.

The primary objective of this term paper is to comprehensively examine and analyze the "Discussion of sustainability by political actors in different countries in Twitter". Through this study, we aim to address our specific research question (RQ): What are the most frequently discussed topics related to climate change among Democrats and Republicans legislators from California and Florida states on Twitter? The RQ was designed to explore key aspects of the chosen topic and allow us to draw meaningful conclusions about its impact and relevance.

Our research is relevant because it offers insight into the political polarization and ideological divisions that exist in the climate change debate. We offer insightful regional perspectives by concentrating on legislators from certain states with distinctive environmental settings and policy approaches. Additionally, our examination of Twitter as a platform shows how effective social media is in influencing political dialogue. The identification of commonly debated themes paves the way for educated policy-making, makes it easier to spot areas of consensus or disagreement, and directs the creation of successful climate change policies. This knowledge is beneficial to the general public as well as policymakers and scholars, enabling a thorough grasp of the political environment and the potential for cross-party cooperation on climate change concerns.

The article's clear goal is to identify the most frequently discussed issues surrounding climate change among Democratic and Republican lawmakers from the states of California and Florida. To that end, it explores and analyzes the sustainability debate on Twitter among political actors from various nations. Effectively highlighting how our study examines particular practices, the topic's importance is made clear. We have analyzed the theoretical foundations that would guide our exploratory study in order to derive the RQ.

2 Literature review

The rise of social media platforms has transformed the communication paradigm, particularly in global concerns (Georgescu & Popescul, 2015). The literature surrounding political communication's digital transformation highlights the pivotal role of social media platforms in facilitating politicians' dissemination of viewpoints, narrative shaping, and engagement with their constituencies (Jungherr, 2016). To establish a solid foundation for our research questions, an extensive review of the existing literature was conducted. This review focuses on the growing body of research concerning climate change communication on Twitter, with a particular emphasis on the discourse among political actors from divergent ideological backgrounds. Additionally, it introduces the innovative application of the Latent Dirichlet Allocation (LDA) model as a tool to effectively analyze and dissect climate change-related topics within the realm of political discourse. According to Veltri and Atanasova (2015), the study of social networks is important for understanding various aspects of society due to their global reach and the growing number of users and posts. Social networking platforms such as Twitter offer a wealth of information that allows researchers to simultaneously study the content of posts and interactions with that content. Users post their content and comment on and share (retweet) other users' posts, and social media audiences are often composed of articulated social links, making social networks crucial to understanding the content being discussed or commented on. In addition, social media data can be used to analyze the dynamics of public opinion on various issues of social importance, such as climate change. However, there is a need for a theoretically, methodologically sound, and critical - as well as ethical - use of online data.

Irawan et al. (2020) hold a similar perspective to Veltri & Atanasova (2015) in that they see Twitter as a useful and significant data source for examining societal trends and human behavior at a certain time and location. One of the most significant topics that Twitter users can argue is politics. Scientists or the government can track the opinion trend about a particular political topic at a specific period by extracting the data set from Twitter and assessing the statistical and analytical results.

To evaluate the content, sources, and information-sharing behavior on Twitter, Veltri and Altanasova (2015) conducted a study in which they evaluated more than 60,000 tweets connected to climate change. The analysis discovered that a tiny group of extremely active and prominent individuals dominate the dialogue about climate change on Twitter. Instead of groups or institutions, most of these users are individuals, and they frequently disseminate.

According to Veltri & Atanasova (2015), tweets regarding climate change typically have a negative tone and concentrate on the problems they cause rather than possible solutions. Semantic clustering, semantic networks, psychological process classification, sentiment analysis, and content analysis were among the analytical methods employed in the study to examine the

tweets. The study comes to the conclusion that Twitter in particular is a valuable information source for studying public opinion dynamics about climate change and other pressing societal concerns. The study also points out that this type of analysis has limits, notably when it comes to identifying higher-order meaning structures like stories and arguments/claims. Overall, the study offers insightful information about Twitter's discussion of climate change and emphasizes the significance of studying social media to comprehend diverse facets of society.

The absence of debate regarding the representativeness of the sample of tweets examined is one of the study's major flaws. It is challenging to determine whether the sample is typical of the overall population of climate change tweets on Twitter because the study doesn't explain how the sample of tweets was chosen (Veltri & Atanasova, 2015). Our study, however, provides details on how the sample of tweets was chosen, thus this is not the case. The study also notes that there are limitations to the analysis of social network data, particularly when it comes to extracting higher-order meaning structures like narratives and arguments/claims. According to the paper, future work should concentrate on creating more complex tools for examining social network data that can identify these higher-order meaning structures. Last but not least, the study doesn't go into great length about the ethical ramifications of using social network data for research. The paper acknowledges the necessity for the critical and ethical use of internet data, but it makes no recommendations or offers any specific criteria for researchers using social network data. When doing our analysis, we dealt right away with how to handle the gathering and use of data from the accounts used in an ethical manner.

Using hierarchical clustering and K-means techniques, Irawan et al. (2020) present a study on the analysis of responses to political topics in social networks. For the purpose of examining societal trends and human behavior at a certain time and location, the study employs Twitter as a useful and insightful data source. The study gives a comparison between the K-means and hierarchical clustering algorithms. Using a hierarchical histogram and word cloud clustering, the study also identifies the dominant theme or word trend of the problem. However, since our analysis doesn't use K-means we don't need to go deeper on this part.

Maier et al. (2018) present a detailed analysis of LDA thematic modeling in communication research and propose an approach to solve the issues associated with its application. The authors outline four significant concerns associated with critical methodological considerations that must be made when utilizing LDA thematic modeling. These difficulties involve unstructured text data preprocessing, algorithm parameter selection, evaluating and improving the model solution's reliability and interpretability, and validating the generated themes. To increase the importance of the validation process, the authors offer an approach that combines existing metrics and in-depth research. The methodology attempts to make LDA topic modeling more accessible to communication scholars while still ensuring disciplinary compliance.

The authors also offer a concise how-to manual for using LDA topic modeling, and they use

actual research data from an ongoing study to show how effective their strategy is. This served as the foundation for our project and served as a guide for how to use the model.

Several gaps in the body of knowledge on LDA topic modeling in communication research are addressed by Maier et al. (2018). The lack of a thorough technique to handle the difficulties of using LDA topic modeling in communication research is one of the major gaps. The authors provide an approach to increase the significance of the validation process by fusing current metrics and in-depth research. A practical user guide for utilizing LDA topic modeling in communication research is another need in the literature. The authors fill a knowledge gap about how to guarantee the reliability and validity of thematic models in communication research. The authors suggest an approach that attempts to guarantee adherence to disciplinary standards and increase the usability of LDA thematic modeling for communication researchers.

The research question presented in this study arose from a synthesis of the ideas, findings, and limitations observed in the literature reviewed. Finding themes, patterns, and discrepancies that called for more research was our aim. We developed research questions intended to meaningfully add to the body of knowledge in the field through an iterative process of literature analysis and synthesis, and we ultimately decided on one. RQ: What are the most frequently discussed topics related to climate change among Democrats and Republicans legislators from California and Florida states on Twitter? Given everything mentioned above on the existing literature on the subject, this research question intends to close this knowledge gap by elucidating the particular issues, top priorities, and communication tactics of lawmakers from various political parties and states. We can better understand the dynamics of climate change debates and spot potential points of agreement or disagreement by examining their discourse, which will support larger efforts to communicate and develop sensible climate change policy.

3 Methodology

3.1 Data description

We performed a text analysis of the political actors of California and Florida states on the social network (Twitter). Political actors of both states have been identified through the official websites of the 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 tweet collection was conducted between April 30, 2022, and April 30, 2023, using the Twitter API. Considering the extensive number of our targeted political actors, which amount to a total of 280, we have opted to proceed with a sample of 20 percent for each party, randomly selected and equally distributed among the assembly and the senate. The outcome of this sampling approach has yielded a sample size of 24 political representatives for California

state and 32 for Florida. The following table shows the results of the sample, as well as their respective tweet counts.

Table 1: Sample description

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

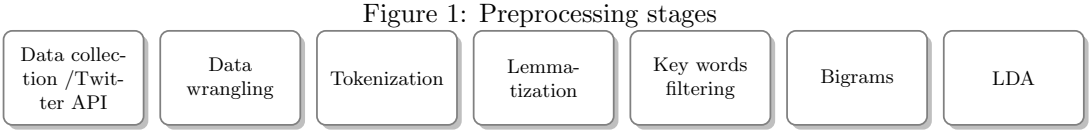
3.2 Description of analysis method

We conduct an analysis of text based on topic modeling 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 (Günther & Domahidi, 2017. p.3056).

To apply the LDA, we previously needed to work in a preprocessing step according to Figure 1. This includes after data collection, removing common words, errors, or other unessential characteristics of the 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 stop words as well as prevalent 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 the Python Gensim package Ldamodel, an open-source library designed to process raw, unstructured digital text using unsupervised learning algorithms (Gensim, 2022).

3.3 Data collection

The tweet collection was conducted using the Tweepy library and the free version of Twitter API. This API limits access to v2 endpoints, allowing for tweet posting and media upload. The access is restricted to 1,500 tweets per month with an application-level posting limit of 1 application ID. Nevertheless, we were able to obtain a significant number of tweets for our analysis.

3.4 Data wrangling

A data cleaning process was employed to clear data from unimportant characteristics and irrelevant information and pre-process a reliable data set. URLs, hashtags, punctuation, @mentions, RT, and 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.

3.5 Tokenization

The Word evaluation process is considered a categorical variable. It is therefore appropriate to use a technique of converting a text string into a sequence of tokens to make it viable for analysis by an algorithm, in this case, an 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).

3.6 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).

3.7 Keywords Filtering

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

Table 2: Filtered sample description

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

3.8 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 a 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)

3.9 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 n-gram is associated with a specific latent topic 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 decided 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 on 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 decrease until we find semantic 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" attribute from Gensim as a measure to help us determine the optimal number of topics to consider for this project.

4 Results

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

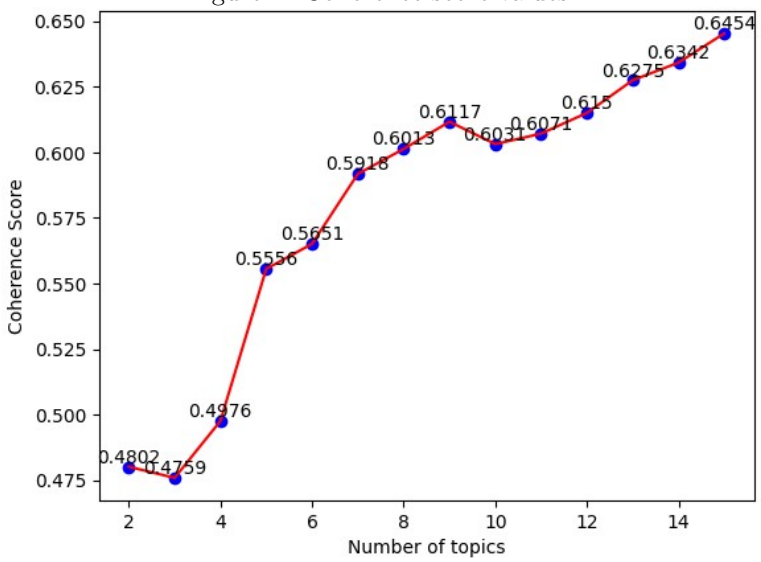
Table 3: Filtered sample description

Topics	**# of documents**	**frequent 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 2 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.

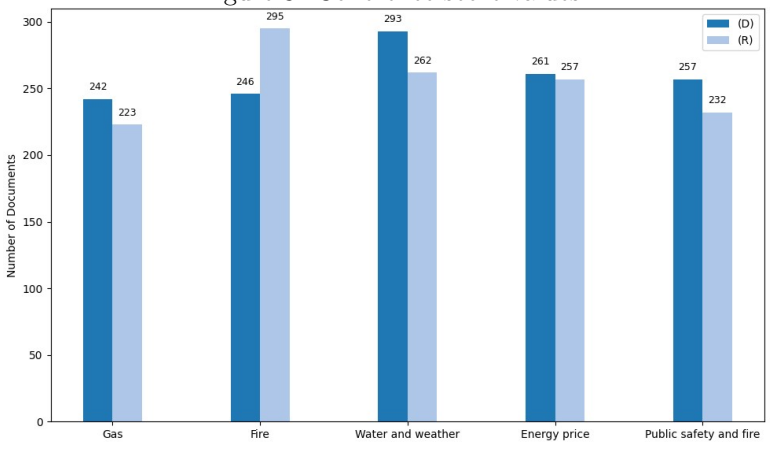
In general, as topics are generated more frequently, the degree of coherence rises and more confined topics are discovered. Grimmer (2010) noted that admitting too many topics could lead to comparable entities that are indistinguishable in any meaningful way. Evans (2014) also points out that having too few themes may lead to very wide entities that incorporate many elements that ought to be kept apart.

Figure 2: Coherence score values



Moreover, our analysis extends to identifying the pertinent subjects discussed by political actors on Twitter. Figure 3 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 3: Coherence score values



Hence, we can infer 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*.

5 Discussion

The topic model identifies 5 coherent topics linked to climate change addressed by political actors. Each topic comprises the 10 most frequent words and is based on predefined climate change keywords. We performed a manual selection of meaningful topic titles based on the coherence score, which determines the optimal number of topics, resulting in 5 topics with a coherence value of 0.5556, the first highest value between 2 and 15 iterations. Although generating more topics increases coherence and narrow results, an excessive number may have indistinguishable words. Conversely, too few topics may have diverse aspects that should remain separate.

Furthermore, our analysis extends to political discourse on Twitter. Figure 3 shows climate change issues ranked by party. *Water and climate* predominate in Democrats debates, while *fire* has importance for Republicans. *Energy prices* appear as a shared concern, with slight differences between Democrats and Republicans. This suggests that Republicans prioritize fire issues, Democrats lean toward water and climate, but both converge on energy prices. Both parties pay less attention to gas issues. This comprehensive analysis highlights each party's specific approaches and shared concerns in the climate change discourse.

Also, we have several limitations that should be taken into consideration when drawing conclusions from this research. First, the sample is relatively small, our sample was reduced to approximately 28% after filtering the keywords. Second, the nature of climate change may be variously expressed semantically with different keywords that we may not have taken into consideration. Third, the hyperparameters initially from Gensim have been taken by default, without any manipulation and the random seed may produce different values. Finally, our debate in labeling the topics with a name has been a deliberation of the members of this paper but based on the coherence index. Therefore, our results may probably differ in interpretation.

6 References

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

Casero-Ripollés, A. (2021). Influencers in the Political Conversation on Twitter: Identifying Digital Authority with Big Data. Sustainability, 13(5), 2851. DOI:10.3390/su13052851

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

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

Georgescu, M., & Popescul, D. (2015). Social Media – The New Paradigm of Collaboration and Communication for Business Environment. Procedia. Economics and Finance, 20, 277–282. DOI:10.1016/s2212-5671(15)00075-1

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Political Analysis, 18(1), 1–35.

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

Huang, C., Liang, W., Lin, S., Tseng, T. B., Wang, Y., & Wu, K. (2020). Detection of potential controversial issues for social sustainability: Case of Green energy. Sustainability, 12(19), 8057. DOI:10.3390/su12198057

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. DOI:10.1111/jjns.12520

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. DOI: 10.1017/9781009157940

Irawan, E., Mantoro, T., Ayu, M. A., Bhakti, M. a. C., & Permana, I. (2020). Analyzing Reactions on Political Issues in Social Media Using Hierarchical and K-Means Clustering Methods. Media-Tech Lab, Department of Computer Science, Sampoerna University, Jakarta, Indonesia. DOI:10.1109/icced51276.2020.9415839

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). DOI:10.1186/s13673-019-0192-7

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

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

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

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. Cyberpsychology, Behavior and Social Networking, 16(9), 708–713.

Panagopoulos, C. (2010). Affect, social pressure and prosocial motivation: Field experimental evidence of the mobilizing effects of pride, shame and publicizing voting behavior. Political Behavior, 32(3), 369–386.

Richardson, J., Grose, J., Nelmes, P., Parra, G., & Linares, M. (2016). Tweet if you want to be sustainable: a thematic analysis of a Twitter chat to discuss sustainability in nurse education. Journal of Advanced Nursing, 72(5), 1086–1096. DOI:10.1111/jan.12900.

Veltri, G., & Atanasova, D. (2015). Climate change on Twitter: Content, media ecology and information sharing behaviour. Public Understanding of Science, 26(6), 721–737.

Appendix

A Ethical review

In the context of our project, we have identified several ethical considerations that are crucial to address to ensure the integrity and responsible conduct of our research.

A.1 Informed consent

A key component of ethical human subjects research is informed consent. Explicit informed consent is not required because the data are already available to the public and our project involves the analysis of public Twitter data. We do understand the significance of respecting users' context and intent, though (Moreno et al., 2013). Since we used the publicly available, freely accessible social network "Twitter," there shouldn't be any privacy objections in our instance. Considering that the data is public and not private. Informed consent is not necessary for the collection of information, according to Moreno et al. (2013).

A.2 Privacy

A basic human right is the right to privacy. It is significant to remember that ongoing surveillance infringes on individual liberties. The idea of data privacy suggests disclosure under controlled circumstances rather than non-disclosure, which should not be mistaken with it. The right to privacy is not the same as a right to secrecy or control, as Nissenbaum (2010) correctly notes. Instead, it ensures that personal information flows in a legitimate manner.

Our study using data from Twitter that is available to the public shows that privacy is not as important. Nevertheless, we have gone out of our way to anonymize the data and avoid mentioning specific tweets or persons in our study. Rather than focusing on particular statements, our goal is to detect broad trends among political actors.

This is consistent with the viewpoints of researchers like Panagopoulos (2010) and Salganik (2017), who cast doubt on the polifacétic nature of privacy. With the help of our analysis, we are able to address the issues raised by these authors and demonstrate that we used public databases whose information was created with the public, in this case, the voting public, in mind.

A.3 Statistics

Although statistics are essential to data science, we are aware that they can be misused in visualization. In order to avoid using statistics improperly, we only ever used them to represent results rather than to manipulate them. Additionally, we have used acceptable approaches and

explicitly explained the statistical techniques used to guarantee the correctness and reliability of our statistical studies. Because of this, we have correctly cited the sources of the statistical software, libraries, and tools that we utilized to analyze the data. We want others to be able to independently verify our analyses, thus we clearly document our processes.

A.4 Bias

We aim to mitigate this issue because we are aware that sample selection and language processing methods may introduce bias into our analyses. We freely accept the limits resulting from potential bias in our conclusions and clearly specify our sampling criteria, assuring a rigorous data collection approach.

A.5 Other

A.5.1 Data source

Although we opted to use the Twitter API as opposed to scraping data, it is crucial to understand the potential limitations and biases of the information obtained through the API.

A.5.2 Transparency

We shared our trial's code, techniques, and data sources to ensure transparency. This will make it possible for others to confirm our findings and evaluate the accuracy of our analysis.

A.5.3 Ethics

Throughout the research process, we adhered to the ethical standards established by the university and the courses we took. This entailed consulting with our faculty advisor for assistance and making sure that our research is carried out in an ethical and responsible manner.