A Textual Analysis of Economic Speeches on Agriculture: 1919 – 2022

Team Members: Jacob Lee, Brendan Elliott, Aaron Lam
Project Manager: Neha R. Gupta
Faculty Leads: Drs. Norbert Wilson, Leslie Collins, Boyla Mainsah

Abstract

- Understanding the historical underpinnings of agricultural policies, such as the Farm Bill, and implications for food production and political developments provides insight into the relationship between agricultural policy and relevant topics such as climate change and food scarcity (Figure 1).
- The Agricultural and Applied Economics Association (AAEA) is the nation's leading economics association addressing food, agriculture, international development, rural development, and natural resources.
- This project aims to use natural language processing techniques to unearth the historical structure and direction of economic discussions that shaped agricultural policy from a dataset of AAEA research papers and speeches from 1919 to 2022*.

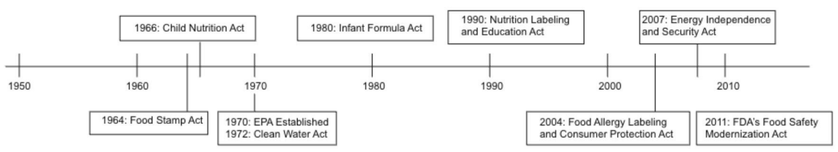


Figure 1. Timeline of significant agricultural policy changes from 1950 to present [1]

* 2010 papers intentionally omitted

Objectives and Methods

- We pre-processed a dataset of PDFs (N = 4010) of published speeches and invited papers presented at the AAEA that were downloaded from Zotero, an online bibliographical database.
- We then performed topic analysis using two topic modeling methods: Latent Dirichlet Allocation [2] and BERTopic [3].
- After training the topic models, we analyzed topic trends over time and performed author demographic analysis.

Data Pre-processing

- Pre-processing refers to downloading the PDFs and converting their contents to machine-readable text using PyTesseract [4].
- We lemmatized the text, which converted words into basic dictionary form (Figure 2) and then removed common and rare words.
- PDF formats varied, which provided a significant challenge. For example, PDFs with two columns were initially unreadable by our text-recognition code.
- The extensive runtime of the process required us to use the Duke Compute Cluster for most pre-processing.
- The data pre-processing pipeline is summarized in Figure 3.

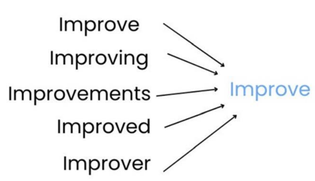


Figure 2. Lemmatization example

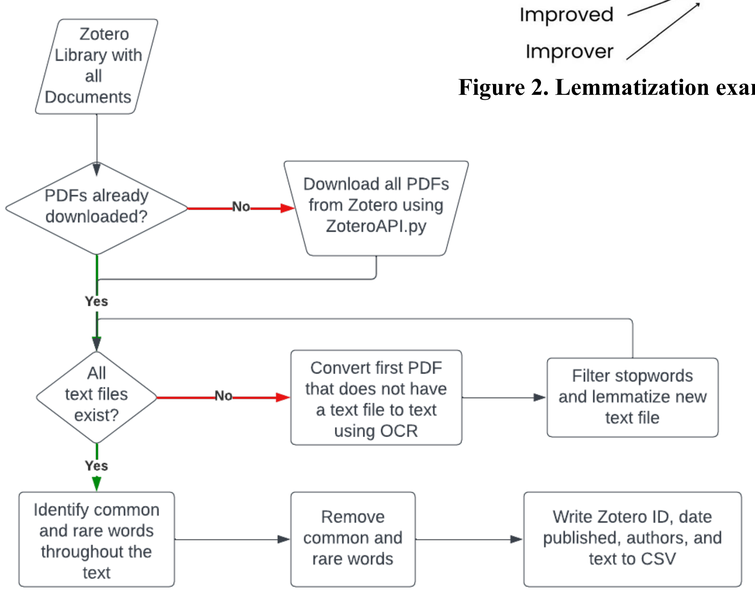


Figure 3. Flowchart of pre-processing of text data from the AAEA dataset

Latent Dirichlet Allocation (LDA)

- LDA is a common method for performing topic modelling on text, by identifying common "topics", or weighted collections of multiple words that commonly appear with each other [2].
- We use the Gensim Python library [5] to generate the model and do additional preprocessing.
  - Text is split into words, called tokens.
  - We identify bigrams and trigrams (two/three-word phrases) e.g., 'comparative advantage'.
  - We convert text into a Bag of Words, which lists how frequently each word appears. Word order is lost.
  - We use Term Frequency – Inverse Document Frequency (TF-IDF) [7] to remove words that either appear in too many documents or too infrequently overall.
- We perform coherence modeling to identify the optimal number of topics (K) (Figure 4) [6].

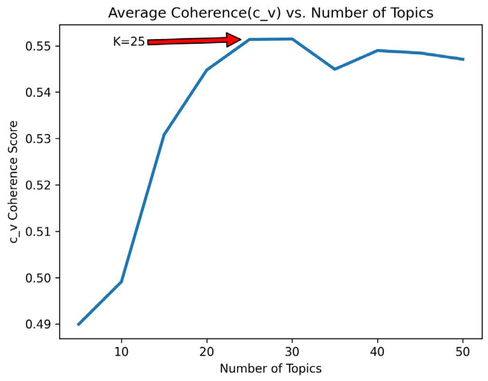
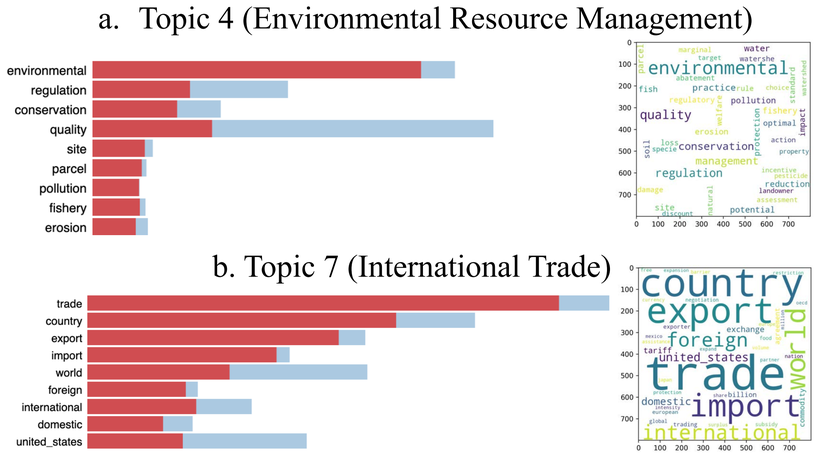


Figure 4. Coherence vs. Number of Topics in the LDA Model

- We consult a subject expert (Dr. Wilson) to manually name topics in order to aid in analysis. The results were generally viewed positively in terms of interpretability and coherence.

a. Topic 4 (Environmental Resource Management)



b. Topic 7 (International Trade)

Figure 5. Most relevant words and word clouds for two example topics from the LDA topic model. The overall bar size indicates the total frequency of the term in the corpus. The red subsection indicates the term's frequency in the topic

Longitudinal Prevalence of LDA Topics

- We analyze topic trends over time. For example, we find that Topic 1 (Agribusiness Strategy) increased considerably throughout.

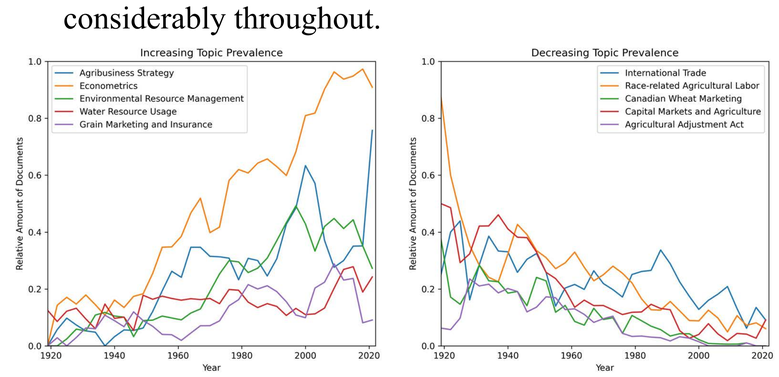


Figure 6. Selected topics obtained from the LDA topic model with (a) increasing and (b) decreasing prevalence

Author Demographic Analysis

- We used the GenderGuesser algorithm [7] to predict the authors' gender and found no significant trend from 1980 to the present (Figure 7).
- We used the Bayesian Improved Surname Geocoding (BISG) model [8] to predict the race/ethnicity of authors. In general, there was an increase in predicted Asian/Pacific Islander authors in the 2000s (Figure 8).

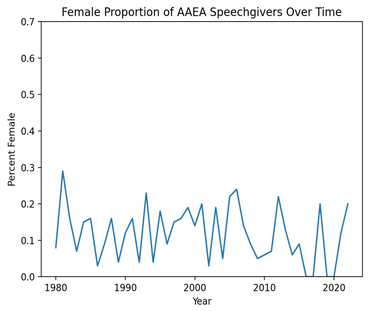


Figure 7. Predicted authors' gender distribution (1980-2022)

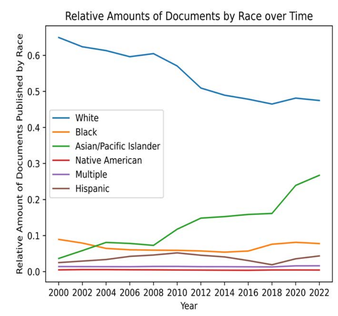


Figure 8. Predicted authors' race distribution (2000-2022)

BERTopic Model

- Given the large dataset, we also used BERTopic [3], allowing hierarchical topic modeling. BERTopic indicates which topics are subtopics of broader subjects and topics.
- For example, we see a hierarchical grouping of climate change and sustainability topics isolated from other agricultural topics (Figure 9).

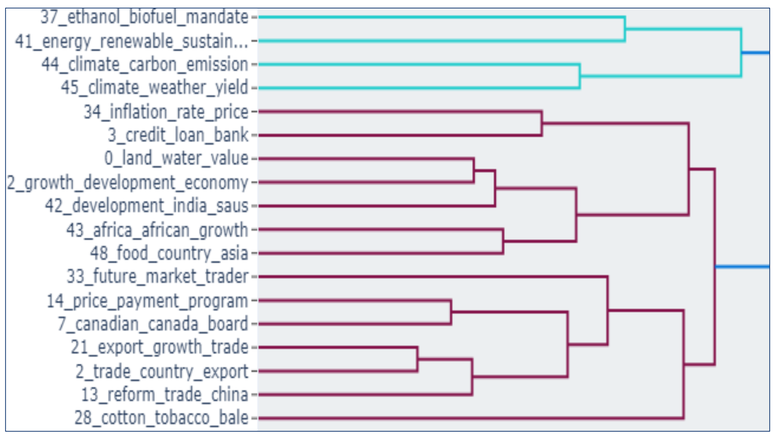


Figure 9. Subsection of BERTopic Hierarchy Chart

Conclusion and Limitations

- We see a significant rise in econometric methods over time. We see a moderate increase in discussion of environmental resource economics. We see a steep decrease in discussion of race-related agricultural labor.
- BISG and GenderGuesser predicted poorly with older data. Therefore, we looked at recent data when using these programs
- We observe an increase in predicted Asian/Pacific Islander authors from 2000-2022 and a stagnation of women authors.
- For future study, we would like to perform more in-depth analysis of topics and the demographic relationships with topics. Another point of study is a more in-depth comparison between LDA and BERTopic.

References

1. National Research Council. 2015. A framework for assessing effects of the food system. The National Academies Press, 47–47. https://doi.org/10.17226/18846
2. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
3. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
4. Ooms J (2023). tesseract: Open Source OCR Engine. https://docs.ropensci.org/tesseract/ https://github.com/ropensci/tesseract (devel).
5. Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).
6. Kvamsdal, S. F., Belik, I., Hopland, A. O., & Li, Y. (2021). A machine learning analysis of the recent environmental and Resource Economics Literature. *Environmental and Resource Economics*, *79*(1), 93–115. https://doi.org/10.1007/s10640-021-00554-0
7. Gender API [Internet]. Germany [cited 12 Dec 2020]. Available from: https://gender-api.com/en/
8. Elliott M, Morrison P, Fremont A, McCaffrey D, Pantoja P, Lurie N. Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. Health Services and Outcomes Research Methodology. 2009;9(2):69–83