A Textual Analysis of Agricultural Research

Darsh Mandera, Haozhou Zhang, Poojitha Balamurugan, Boyla Mainsah, Leslie Collins, Norbert Wilson
[1]Duke University
140 Science Drive
Durham, NC 27708 USA
proceedings-questions@aaai.org

Abdalrahman Alalwan
King Saud University
Riyadh, 11362 KSA

Abstract

The adoption of machine learning techniques in agricultural economics has been limited, with traditional literature review methods still predominant. This study applies Natural Language Processing (NLP) to a dataset of articles from the American Journal of Agricultural Economics (AJAE) and Applied Economic Perspectives and Policy (AEPP) published from 2003 through June 2024, using .xml files for a more robust analysis. Expanding on Kvamsdal et al. (2021), we focused on the 'Maintext' rather than the 'Abstract' in the Latent Dirichlet Allocation (LDA) analysis. We proceeded with a novel approach of employing KeyBERT for individual article analysis, with results pooled and processed through BERTopic for topic extraction. Longitudinal analysis of LDA and BERTopic results offers new insights into thematic evolution within agricultural economics.

Introduction

The application of machine learning techniques has yet to be widely adopted in the field of agricultural economics. Previous investigations into aggregated summary of themes of studies in agricultural economics have widely followed the traditional formula of manual literature review and classification. Recent attempts have been made with a combination of Optical Character Recognition (OCR) and NLP. However, due to the limitations of OCR algorithms in practice and a restriction on the dataset, such surveys demands arduous efforts and generates noisy results. We conducted survey using NLP on a data set of published articles from two leading journals in the field of agricultural economics—AJAE and AEPP—published between 2003 and June 2024. The dataset is in the form of .xml files which offers the opportunity for a more robust analysis. We subsequently filtered and processed the dataset to extract the necessary information. We first followed the methods from Kvamsdal, et al. (2021), with a novelty of conducting the analysis on the 'Maintext' with regards to the 'Abstract' that Kvamsdal (2021) adopted. We then proceeded to use KeyBERT on individual articles then subsequently pooling the results generated and used BERTopic to extract topics. With the results of LDA and BERTopic, we performed longitudinal analysis and offered insights into the results.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

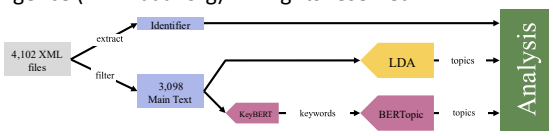


Figure 1: Flowchart of topic identification and data analysis

Data Preprocessing

Preprocessing was undertaken with .xml parsing and regular expression (RE). We extracted information that is pertinent to our analysis while eliminating potential noise.

LDA

We first preprocessed the text by tokenization, custom stopwords removal, and lemmatization. Subsequently, we generated bigrams and trigrams to capture significant multi-word expressions. This process was handled using the Gensim library (Reh˘ u˚rek & Sojka, 2010), with a list of stopwords from Jockers (2021). We then performed Term FrequencyInverse Document Frequency (TF-IDF) vectorization to adjust term importance based on document frequency. Following this, we constructed multiple LDA models based on Blei et al. (2003), varying the number of topics, and evaluated them using

coherence scores to assess topic interpretability and perplexity to measure the model's generalizability.

BERTopic

We applied KeyBERT (Grootendorst, 2020) prior to using BERTopic (Devlin et al., 2018) to enhance the accuracy of topic modeling by extracting the most relevant keywords from individual articles, ensuring the input text for BERTopic was enriched with high-importance terms to capture nuanced semantic information more effectively. Subsequently, we employed BERTopic with sentence transformers (Hugging Face, n.d.) for text embedding, UMAP for dimensionality reduction, and K-means for clustering, ensuring every article was assigned to a topic without leaving outliers. We further fine-tuned the model by adjusting the number of topics based on coherence scores and interpretability, resulting in a precise grouping of documents into semantically coherent clusters.

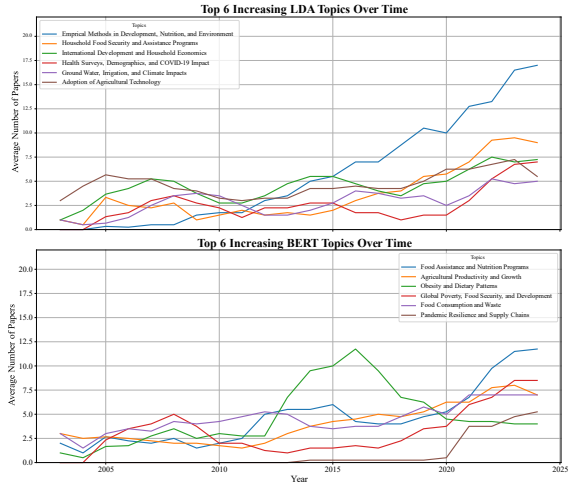


Figure 2: Trends of the top increasing topics over time, as identified by LDA and BERTopic

Results

The extracted topics from each topic model were labelled by an internal subject matter expert to ensure relevance and clarity. Our analysis found that while LDA was more effective at capturing the broader thematic content of each paper, providing a general overview of research themes, BERTopic excelled in capturing the semantic nuances of the text, resulting in more coherent and refined topics. Among the 3,098 papers in the corpus, a significant portion was attributed to specific themes such as "Advances in Microeconometrics" (332 papers), "Public Sector Impacts on Markets, Growth, and Trade" (170 papers), and "Economic Experiments: Auctions, Discrete Choice Models, and Incentives" (167 papers).

Longitude Analysis

The figures are drawn using a linear slope fitted to the panel data, we portraited the top 6 increasing and decreasing topics generated by both LDA and BERTopic. Household and Development economics and their related topics have seen increase being identified by both models. Both topic models identified a sharp growth of COVID-19 related topics after 2020. Example topics include "Health Surveys, Demographics, and COVID-19 Impact" and "Pandemic Resilience and Supply Chains". However, the decreasing topics share less similarities. The two models exhibit some similar results, such as the sharp growth of COVID-19 related topics in 2020. While LDA effectively captures the thematic content of each paper, providing an overview of general research themes, BERTopic excels in capturing the semantic nuances, offering a more refined classification of the corpus. This distinction highlights BERTopic's ability to provide deeper insights into specific issues and their interconnections within the corpus.

Conclusion

Overall, we present a novel application of NLP and topic modeling in the field of agricultural economics. Previously, a similar analysis in the field would have required a manual literature review in order to identify topics present in a corpus. Meta analyses like the one presented here remain rare in the field due to limitations in technique. While machine learning has been used in the field of agriculture economics for topic analysis, such approaches tend to focus only on the abstracts of research papers and use regression models. By contrast, we use a corpus of full-length research papers and use more sophisticated natural language processing algorithms. Thus, with this work, we create a previously unexplored space in agricultural economics by applying topic modeling algorithms to a large corpus of research papers.

Acknowledgments

We acknowledge Wiley for providing the .xml files, and Duke Data+ program for the platform that made this happen.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022. https://jmlr.csail.mit.edu/papers/v3/blei03a.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT (Version 0.3.0). Zenodo. https://doi.org/10.5281/zenodo.4461265

Hugging Face. (n.d.). all-MiniLM-L6-v2 [Software]. Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Jockers, M. (2021). Expanded stopwords list. Retrieved June 15, 2024, from https://www.matthewjockers.net/macroanalysisbook/expandedstopwords-list

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

Princeton University. (2010). About WordNet. WordNet. Princeton University. https://wordnet.princeton.edu/citingwordnet

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).