Topic Modeling on Food News Articles

A Samuel Pottinger

2023-12-21

Contribution: This semantic database can uncover how the topics of news articles from across the planet relate to each other and how they vary across different subpopulations. These data paint a rich picture of food conversation in news media across a geographically vast reach, offering a digital telescope spanning many (though not all) cultures and languages. This statistical vantage-point seeks to complement later qualitative explorations.

Overview: To support multi-modal exploration and to complement qualitative approaches, this short machine learning study derives quantitative information from articles about food in a mixed membership structure where each document may be found in multiple categories, tags, and keywords. Furthermore, trying both LDA as well as HDBSCAN on Word2Vec, this topic modeling effort takes steps to ensure strong global performance including ongoing monitoring of statistics by locale. This AI-empowered data pipeline results in a classification system usable for the project's next steps in interactive data visualization, creating a topical structure from which document statistics may be generated. More specifically, the resulting database of "tagged" food articles allows for comparison of how often various topics at different levels of granularity appear across regions and languages and in what context.

1 Introduction

Building on a multi-language database of news articles discussing food, this short machine learning effort attempts to tag those documents in order to describe their content semantically, generating statistics for temporal and geospatial analysis and visualization (Kanade 2022; Pottinger 2023).

1.1 Background

In data science, "topic modeling" refers to this content-based task of determining "thematic structure in large collections of texts" (Blei 2012). In "digital humanities" studies, these methods can understand natural language "corpora" and address unstructured data mathematically (Callaway et al. 2020). For example,

prior work includes use of computation to describe and contrast different bodies of discourse like comparing discussion of health during COVID-19 across different groups like industry, public, and news media (Cheung and Pottinger 2020).

1.2 Objective

Having determined the topical relationships between various articles across different geographies, the team hopes to use these quantitative insights to help inform both:

- Plans and ideation for qualitative research.
- The overall project's internationally scoped reporting on food and food justice.

Put another way, this research attempts to understand how topics of conversation vary across geographies and languages to better contextualize a global understanding of food.

2 Methods

This data science study first prepares the news article dataset for topic modeling before conducting standard text preprocessing. Afterwards, two different approaches attempt to derive topics from the resulting corpus.

2.1 Dataset

Spanning December 2022 through October 2023, the news dataset includes 26,700 entries, a database which also references discussion of "unique" and "usable" articles (Pottinger 2023). Building statistical insight from these unstructured data, this project needs to carefully consider and refine the meaning of those definitions prior to modeling.

Unique: The dataset includes some articles which appear to contain the same[1] content as others also within the corpus. That said, those "identical" records may still diverge in metadata: different URLs, dates, or countries. In this case, this study elects to merge records with identical translated title and body, addressing "duplicates" receiving over-representation in the case of "reposting" like in syndication (Beatty 2023). However, to guard for balanced representation, this method only applies to "duplicates" if they are in the same locale[2], attempting to ensure that articles do not become "invisible" within geographies and languages. In short, these "duplicates" still take part in that locale's discourse even though they appear elsewhere.

[1] Identicial in this case means same title and text content.
[2] Combination of country and language identifiers.

Usable: Due to the limited languages available in the machine translation service, the database may not offer an English version of some documents (AWS 2023). Furthermore, a country or language may have too few articles for analysis to be properly representative or some records may be missing key information. That in mind, this study filters for:

- Articles from locales with at least 50 entries.
- Only records with non-empty titles and complete metadata (date, url, source).
- Languages for which this project can access machine translation.

For locales excluded, this may indicate additional opportunities for supporting qualitative research.

Content: Many articles' publishers only offer titles to the API service beneath this news database, a phenomenon which may be more common in some locales versus others (Newsdata, n.d.). For example, Ireland sees articles with content beyond title at one tenth the rate of Canada. Therefore, this study reduces consideration to just titles in this early phase as to prevent accidental geographic bias against locales less commonly providing article bodies.

These definitions change not just the size of the "available" dataset in terms of records but also provide important context for the "shape" of those data, details further explored in discussion below.

2.2 Preprocessing

The following operations take place on titles prior to topic modeling regardless of method:

- Tokenization (NLTK 2023).
- Lower casing of all tokens.
- Removal of English "stop words" (Judah 2021).
- Conversion to lemmas (Managoli, n.d.).

Note that all methods use the "bag of words" representation (Brownlee 2019).

2.3 Latent Dirichlet Allocation

A standard approach to "mixed membership" topic modeling where a document may belong to multiple groups, Latent Dirichlet Allocation or "LDA" requires very few parameters other than the number of topics, an input variable for which this study tries 10 to 100 at increments of 10 (Blei, Ng, and Jordan 2003). Furthermore, to remove noise, this study also considers Ensemble LDA (Brigl 2018). Finally, to provide labels to these groups, this study uses manual inspection of the resulting topics.

2.4 Keyword clustering

Despite the power of this approach, this report also recognizes that LDA can perform poorly in corpora with short documents, similar to what one may expect with titles (Yan et al. 2013). Given this potential issue, this study also considers keyword clustering:

- Extract "keywords" using TF-IDF, trying thresholds between 0.1 to 0.5 (Karabiber 2023).
- Convert those keywords to vectors using Word2Vec pretrained on Google News (Mikolov et al. 2013; Ghosh 2011).
- Cluster those words into groups of similar or related words using cosine distance to produce what this study calls "tags" (Lakhey 2019).

Clustering takes place using either DBSCAN or HDBSCAN with this study evaluating both (Ester et al. 1996; Campello, Moulavi, and Sander 2013). Similar to LDA, the labels for each tag comes from manual review of the keywords in each of the machine's proposed groupings. Note that the literature provide examples of similar approaches in different contexts (D'Agostino 2021; Togatorop et al. 2021; Ning and Chen 2023).

2.5 Categories

To produce a small number of high level categories from either the LDA topics or the keyword cluster tags, this investigation performs "affinity diagramming" to attempt to construct a small number (less than 10) of more general themes (Dam and Siang 2022).

3 Results

Though metrics do exist, often evaluation requires "moving from exploring and interpreting a model back and forth to diagnostics and evaluation in order to decide how best to model a corpus" (Silge 2018). Due to the additional complications involved with a global dataset, this study takes this "iterative" approach which involves manual review of generated topics and tags in order to qualitatively determine the preferred approach.

3.1 LDA results

C_v plateaus at around 0.46 for 50 topics and manual review generally finds that LDA fails to yield usable results (Rijcken 2023). Indeed, consider the leading 5 words of the first two groups found in the 60 topics case:

- Topic 1: `restaurant, three, offer, new, uk`
- Topic 2: `eat, avoid, never, community, christmas`

Given poor interpretability, these results may not serve the needs of the later visualization. Ensemble LDA yielded similar unsatisfactory results.

3.2 Clustering results

Unlike for LDA, many steps in the "keyword clustering" approach do not provide clear performance metrics. Therefore, this study primarily uses qualitative evaluation to set parameters.

TF-IDF: Starting with TF-IDF for keywords, this study considers multiple thresholds but qualitative analysis selects a minimum score of 0.15. Then, this approach further reduces noise by excluding words appearing in fewer than 20 articles as well as those appearing in over half of the documents in the corpus. Note that machine translation issues or problems with unknown tokens in vectorization may cause words to fail processing. Therefore, a "manual review" file captures words with non-English characters and punctuation.

Vectorization: Next, word2vec transforms these tokens into vectors without requiring additional parameter selection. Of course, use of pretrained model may introduce bias, including in words for which vectors are not available. Therefore, analysis collects "unknown" words in the manual review file.

Clustering: Analysis finds that clusters may exhibit "varying densities" so the study quickly focuses on HDBSCAN (Pedregosa et al. 2011). For performance reasons, this investigation elects to use the hdbscan Python package (McInnes, Healy, and Astels 2016a). Regardless, with cosine similarity and a minimum samples and cluster size of 3 (sweeped 2 to 10), keyword groupings seem highly sensible as shown in these first two groups:

- Group 1: `electric, energy, fuel, gas, oil`
- Group 2: `black, blue, color, red, white`

The output dataset captures the original clusters in the `clusters_naive` table.

Manual review: For tokens requiring manual review, a team member considers each individually and attempts to associate it into one of the groups produced by HDBSCAN. In practice, this leads to minor modification of those groups to accomodate that additional population.

3.3 Tagged dataset

The resulting dataset includes 18,394 articles of which around 95% see at least one tag without further review. That said, to ensure equity for all locales, manual inspection of those untagged articles recovers further keywords, raising this to 98%. In total, research moves forward with 1,319 keywords organized into 118 tags. Finally, this study further groups those tags into 5 high level categories plus an "other" characterization.

Note that this final dataset includes 27 countries and 13 languages. Furthermore,

for equity reasons, this investigation confirms that all locales see at least 95% of their articles with at least one tag.

4 Discussion

Though not fully automated, this hybrid approach uses machine learning to provide the vast majority of topical structure extraction but leverages manual efforts in key steps to ensure strong cross-locale performance. All that said, this write up next turns to contextualizing these results including conversation of limitations and opportunities for future work.

4.1 Structure

Affinity diagramming finds 5 high level categories with the following example tags:

- **Economy and Industry**: Agriculture, Commerce, Economy, Hospitality, Labor
- **Environment and Resources**: Energy, Environment (includes climate), Water, Weather and Disaster, Nature
- **Food and Materials**: Alcohol, Coffee, Food Security, Food Service, Grocery
- **Health and Body**: Allergy, Body, Diet, Mental Health, Healthcare
- **People and Society**: Aid, Crime, Governance, Game and Sport, Media

In addition to these general groups, some tags also sit in an "other" category.

4.2 Addressing bias

Some of these technologies and processes introduce bias.

- Machine translation only supports some but not all languages (AWS 2023).
- Algorithms like Word2Vec may reflect bias of the underlying corpus (Petreski and Hashim 2023).
- This study manually reviews all 1,319 keywords post-translation to ensure reasonableness of categorization and to combat bias. However, despite best efforts made to conduct follow up translation and web search of terms, this team also carries its own backgrounds and limitations.

Later work may consider further refinement of methods for both vectorization and translation as well as an expanded team with more diverse language skills and cultural contexts.

4.3 Future modeling efforts

Articles may appear in multiple keywords, tags, and categories. This achieves a form of mixed-membership modeling for documents (Blei 2015). However, note

that this study uses the HDBSCAN approach due to poor LDA performance, a modeling decision likely necessitated by short texts (Yan et al. 2013). While offering a usable topical structure, this alternative approach means keywords can only appear in a single tag and tags can only appear in a single category. This poses issues for tokens like "mcdonalds" which currently appears in "business (individual)" alongside other named businesses even though it could also appear in food service. An artifact of HDBSCAN itself, additional study could work to provide alternative methods to circumvent this structure (McInnes, Healy, and Astels 2016b). Specifically, later work may consider other methods like BTM or "soft clustering" approaches (Yan et al. 2013; McInnes, Healy, and Astels 2016b). Of course, note that switching to these alternatives may complicate other areas of the project including the visualization.

4.4 N-grams

Addressing a dataset of titles which generally consist of short pieces of text, this study uses single word tokens in a "bag of words" approach (Brownlee 2019). Well supported by underlying libraries, this simplifies a number of pipeline steps. That said, future work could consider n-grams which are sets of adjacent words which may offer more context to the meanings of individual tokens (Kumar 2017). For example, a manual check of "security" indicates that the word overwhelmingly means "food security" in the context of this dataset but 2-grams could see "food security" as a discrete "keyword" to categorize. While this "larger window" of tokens for modeling carries its own challenges and potential pitfalls, later study may expand context.

4.5 Entities

This approach reveals a large number of named entities as keywords now tagged in the resulting semantic database. These identified topics include both individuals as well as government agencies and private sector companies. That in mind, as potential work for future study, this investigation notes that these tended to mostly track "high profile" entities like celebrities or national-level organizations. This may reflect an actual observation of the global conversations manifested into this database. However, it may also emerge from limitations of short text samples or infrequent usage (names only appearing in one or two articles). Though it carries its own limitations, later work could consider using named entity recognition algorithms for the specific purpose (Pakhale 2023).

5 Conclusion

Though this study takes great care in critically addressing potential areas of bias and using manual efforts when necessary to attempt to try to mitigate those important concerns, any topical structure likely proves inadequate, especially

when addressing a global dataset. Even still, this resulting data product enables structured quantitative comparison of food articles across different regions and languages. Despite its limitations, this invites a much broader and larger population to participate in research than feasible in qualitative perspectives alone. Indeed, paired with this computational way of seeing, traditional design research and this "digital humanities" approach provide complementary information (Blei 2012; Callaway et al. 2020).

Works cited

AWS. 2023. "Amazon Translate Features." Amazon Web Services. https://aws.amazon.com/translate/details/.

Beatty, Hannah. 2023. "Content Syndication Explained + How to Properly Repost Articles." Impulse Creative. https://impulsecreative.com/blog/blog/how-to-properly-repost-articles.

Blei, David. 2012. "Topic Modeling and Digital Humanities." Journal of Digital Humanities. https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

———. 2015. "Mixed-Membership Models (and an Introduction to Variational Inference)." Columbia University. https://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixed-membership.pdf.

Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation." *J. Mach. Learn. Res.* 3 (March): 993–1022.

Brigl, Tobias. 2018. "Ensemble LDA." Technische Hochschule Ingolstadt. https://www.sezanzeb.de/machine_learning/ensemble_LDA/.

Brownlee, Jason. 2019. "A Gentle Introduction to the Bag-of-Words Model." https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

Callaway, Elizabeth, Jeffrey Turner, Heather Stone, and Adam Halstrom. 2020. "The Push and Pull of Digital Humanities: Topic Modeling the "What Is Digital Humanities?" Genre." *Digit. Humanit. Q.* 14. https://api.semanticscholar.org/CorpusID:218646118.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates." In *Advances in Knowledge Discovery and Data Mining*, edited by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, 7819:160–72. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14.

Cheung, Joanne, and A Samuel Pottinger. 2020. "Natural Language Processing: A New Way to Listen to People about Health." IDEO CoLab Ventures. https://medium.com/ideo-colab/natural-language-processing-a-new-way-to-listen-to-people-about-health-8cfe9bece1ce.

D'Agostino, Andrea. 2021. "Text Clustering with TF-IDF in Python." MLearning-ai. https://medium.com/mlearning-ai/text-clustering-with-tf-

idf-in-python-c94cd26a31e7.

Dam, Rikke, and Teo Siang. 2022. "Affinity Diagrams: How to Cluster Your Ideas and Reveal Insights." Interaction Design Foundation. https://www.interaction-design.org/literature/article/affinity-diagrams-learn-how-to-cluster-and-bundle-ideas-and-facts.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–31. KDD'96. AAAI Press.

Ghosh, Sugata. 2011. "Google Word2Vec." Kaggle. https://www.kaggle.com/datasets/sugataghosh/google-word2vec.

Judah, Banjoko. 2021. "Removing Stop Words with NLTK Library in Python." https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1.

Kanade, Vijay. 2022. "What Is Semantic Analysis? Definition, Examples, and Applications in 2022." Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-semantic-analysis/.

Karabiber, Faith. 2023. "TF-IDF — Term Frequency-Inverse Document Frequency." LearnDataSci. https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/.

Kumar, Prachi. 2017. "An Introduction to n-Grams: What Are They and Why Do We Need Them?" XRDS. https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/.

Lakhey, Munesh. 2019. "Word2vec Made Easy." Towards Data Science. https://towardsdatascience.com/word2vec-made-easy-139a31a4b8ae.

Managoli, Girish. n.d. "An Advanced Guide to NLP Analysis with Python and NLTK." Red Hat, Inc. https://opensource.com/article/20/8/nlp-python-nltk.

McInnes, Leland, John Healy, and Steve Astels. 2016a. "Hdbscan." hdbscan. https://hdbscan.readthedocs.io/en/latest/index.html.

———. 2016b. "How Soft Cluistering for HDBSCAN Works." hdbscan. https://hdbscan.readthedocs.io/en/latest/index.html.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." https://doi.org/10.48550/ARXIV.1301.3781.

Newsdata. n.d. "Newsdata.io." Bytesview Analytics.

Ning, Hui, and Zhenyu Chen. 2023. "Fusion of the Word2vec Word Embedding Model and Cluster Analysis for the Communication of Music Intangible Cultural Heritage." *Scientific Reports* 13 (1): 22717. https://doi.org/10.1038/s41598-023-49619-8.

NLTK. 2023. "Nltk.tokenize.word_tokenize." NLTK Project. https://www.nltk.org/api/nltk.tokenize.word_tokenize.html.

Pakhale, Kalyani. 2023. "Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges." https://doi.org/10.48550/ARXIV.2309.14084.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M.

Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Petreski, Davor, and Ibrahim C. Hashim. 2023. "Word Embeddings Are Biased. But Whose Bias Are They Reflecting?" *AI & SOCIETY* 38 (2): 975–82. https://doi.org/10.1007/s00146-022-01443-w.

Pottinger, A Samuel. 2023. "Reflections on an International News Dataset Centered on Food." Eric; Wendy Schmidt Center Data Science; Environment. https://docs.google.com/document/d/1hhUdw7RzZdtIZojg96ROWk1YtMrf9D_8-kcZ7pjQuTE/edit.

Rijcken, Emil. 2023. "CV Topic Coherence Explained." Towards Data Science. https://towardsdatascience.com/c%E1%B5%A5-topic-coherence-explained-fc70e2a85227.

Silge, Julia. 2018. "Training, Evaluating, and Interpreting Topic Models." https://juliasilge.com/blog/evaluating-stm/.

Togatorop, Parmonangan R., Rosa Siagian, Yolanda Nainggolan, and Kaleb Simanungkalit. 2021. "Implementation of Ontology-Based on Word2Vec and DBSCAN for Part-of-Speech." In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, 51–56. SIET '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3427423.3427431.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. "A Biterm Topic Model for Short Texts." In *Proceedings of the 22nd International Conference on World Wide Web*, 1445–56. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/2488388.2488514.