AutoPersuade: A Framework for Evaluating and Explaining Persuasive Arguments[*]

Till Raphael Saenger
Princeton University
saenger@princeton.edu

Musashi Hinck
Intel Labs[†]
musashi.hinck@intel.com

Justin Grimmer
Stanford University
jgrimmer@stanford.edu

Brandon M. Stewart
Princeton University
bms4@princeton.edu

Abstract

We introduce *AutoPersuade*, a three-part framework for constructing persuasive messages. First, we curate a large dataset of arguments with human evaluations. Next, we develop a novel topic model to identify argument features that influence persuasiveness. Finally, we use this model to predict the effectiveness of new arguments and assess the causal impact of different components to provide explanations. We validate AutoPersuade through an experimental study on arguments for veganism, demonstrating its effectiveness with human studies and out-of-sample predictions.

1 Introduction

Persuasion is a common task in politics, business, government, and our daily lives. Modern tools—like A/B experiments, surveys, and focus groups—are well-equipped to identify *which* of a pre-existing set of messages is most persuasive, but provide little insight into *what about* them is compelling. Large language models (LLMs) can help to generate new plausibly persuasive messages, but they do not offer causal evidence on whether or how they have succeeded (Gomez-Uribe and Hunt, 2016; De Vaus and de Vaus, 2013; Morgan, 1996; Palmer and Spirling, 2023; Rescala et al., 2024).

In this paper, we introduce a new workflow for identifying the topical components of an argument that are persuasive, *AutoPersuade*. Our framework assists with each step of the persuasion task. Our three-step workflow is shown in Figure 1. We demonstrate this workflow in a novel study of pro-veganism persuasion.

First, we gather persuasive arguments and responses to those arguments. These arguments can be from various sources, like social media, LLMs, or manual generation. With such a collection of arguments, we discuss our exemplary case study to evaluate arguments for veganism. Using a forced-choice design with arguments randomly assigned to respondents, this case study allows us to assess which arguments survey respondents report as more persuasive. While self-reported assessments of persuasion may not correspond to actual changes in behavior (Coppock, 2023), our new framework is well suited for alternative settings that use behavioral measures of persuasion, which might be explored in future studies.

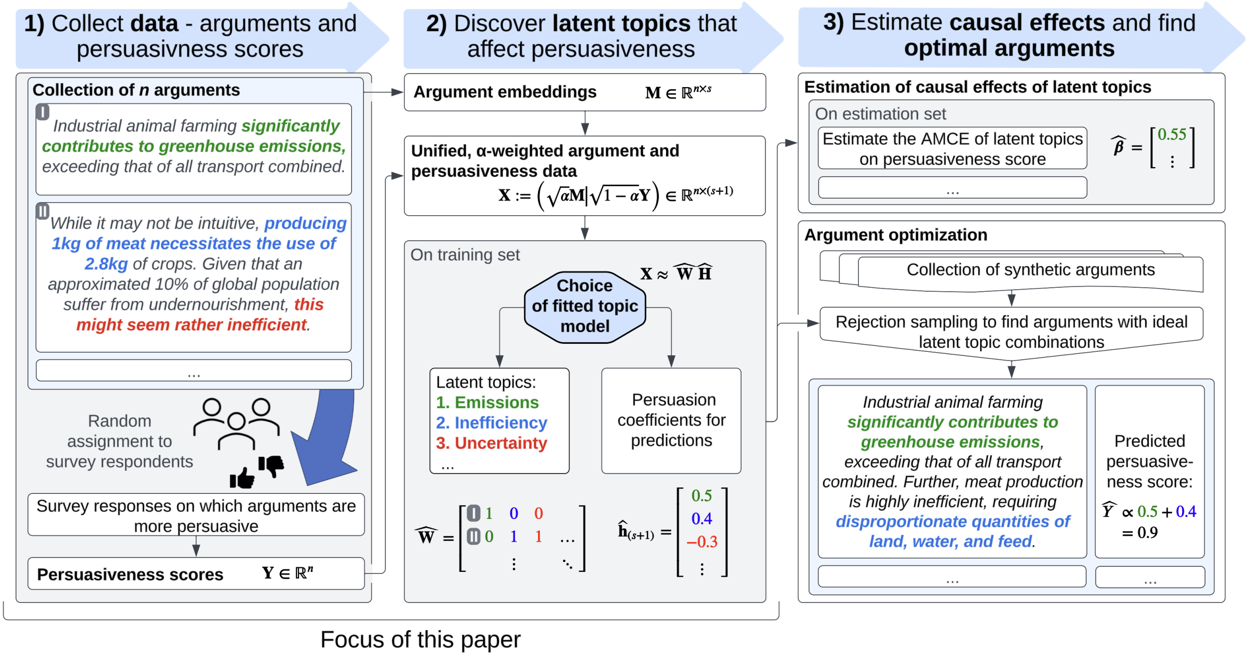
Second, with arguments and responses in hand, we introduce a new model to extract the latent features that cause arguments to be more or less persuasive. We call this model the <u>SU</u>pervised semi-<u>N</u>on-negative (SUN) topic model. The SUN topic model uses an embedding representation of the arguments and builds upon matrix factorization methods to extract latent features that characterize arguments and affect responses (Fong and Grimmer, 2016; Egami et al., 2022; Feder et al., 2022). Because our model provides interpretable output, we show how it can be used to provide easy-to-understand and actionable insights into why particular messages are (or are not) persuasive.

Third, using the output from the SUN topic model, we compute causal effects from varying the content of our arguments and assess the persuasiveness of future arguments. We target causal effects that determine how the average persuasiveness of an argument changes as the prevalence of a latent feature changes. This allows us to predict how persuasive a new argument would be if deployed to the same population. However, as we demonstrate through our experiments, our current method is better at answering questions about what is (and is not) persuasive than finding the optimal

[*] This work was supported by the New Ideas in the Social Sciences fund of Princeton University. The work described in this manuscript is subject to a pending patent application. The replication code is available here: https://github.com/TillRS/AutoPersuade.

[†] This work was done while MH was a postdoctoral research associate at the Data-Driven Social Science Initiative at Princeton University.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16325–16342
November 12-16, 2024 ©2024 Association for Computational Linguistics


1) Collect data - arguments and persuasivness scores
2) Discover latent topics that affect persuasiveness
3) Estimate causal effects and find optimal arguments
Collection of n arguments
I
II
...
Random assignment to survey respondents
Survey responses on which arguments are more persuasive
Persuasiveness scores $\mathbf{Y} \in \mathbb{R}^n$
Argument embeddings $\mathbf{M} \in \mathbb{R}^{n \times s}$
Unified, α-weighted argument and persuasiveness data
$\mathbf{X} := \left(\sqrt{\alpha}\mathbf{M} \middle| \sqrt{1-\alpha}\mathbf{Y}\right) \in \mathbb{R}^{n \times (s+1)}$
On training set
$\mathbf{X} \approx \widehat{\mathbf{W}}\widehat{\mathbf{H}}$
Choice of fitted topic model
Latent topics:
1. Emissions
2. Inefficiency
3. Uncertainty
...
Persuasion coefficients for predictions
$\widehat{\mathbf{W}} = \begin{bmatrix} \text{I} & 1 & 0 & 0 & \\ \text{II} & 0 & 1 & 1 & \cdots \\ & & \vdots & & \ddots \end{bmatrix}$
$\hat{\mathbf{h}}_{(s+1)} = \begin{bmatrix} 0.5 \\ 0.4 \\ -0.3 \\ \vdots \end{bmatrix}$
Estimation of causal effects of latent topics
On estimation set
Estimate the AMCE of latent topics on persuasiveness score
...
$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 0.55 \\ \vdots \end{bmatrix}$
Argument optimization
Collection of synthetic arguments
Rejection sampling to find arguments with ideal latent topic combinations
Predicted persuasive-ness score:
$\widehat{Y} \propto 0.5 + 0.4 = 0.9$
...
...
Focus of this paper

*Industrial animal farming **significantly contributes to greenhouse emissions**, exceeding that of all transport combined.*

*While it may not be intuitive, **producing 1kg of meat necessitates the use of 2.8kg** of crops. Given that an approximated 10% of global population suffer from undernourishment, **this might seem rather inefficient**.*

*Industrial animal farming **significantly contributes to greenhouse emissions**, exceeding that of all transport combined. Further, meat production is highly inefficient, requiring **disproportionate quantities of land, water, and feed**.*

Figure 1: The *AutoPersuade* workflow. After generating a collection of arguments, we collect reactions from respondents. Using these reactions and arguments, we fit the SUN topic model that discovers latent topics that simultaneously describe the documents and their persuasiveness. We then use the output from the SUN model to estimate causal effects from the existing sample and to predict the persuasiveness of new arguments. While our validation studies confirm the causal estimates of our case study, the current approach using average marginal component effects of topics appears less well-suited to finding the best argument. Improving the identification of optimal arguments in the last step might be the subject of future studies.

persuasive argument.

Our quantity of interest is in the *content* of arguments that are persuasive in particular settings rather than the *rhetorical strategies* that are effective across settings. In this sense we are pursuing a different goal than, e.g., the pioneering work of Tan et al. (2016) which considers linguistic features that are persuasive in online discourse.

2 The AutoPersuade Workflow

In this section, we describe the three-step AutoPersuade workflow (summarized in Figure 1) in detail.

2.1 Step 1: Collect Data

Arguments in this paper are a collection $\mathcal{D}$ of documents from which n distinct samples are shown to respondents from a well-defined survey population. Arguments elicit a reaction, which we collect in $\mathbf{Y}$. Unlike A/B tests, our workflow accommodates, and indeed benefits from, a large number of potential arguments (Fong and Grimmer, 2023). Ideally, the collected responses will reflect the respondent's behavior. That said, we can also use proxies for that behavior including respondent's self-reported evaluations of statements.

2.2 Step 2: Discover Topics

The SUN topic model takes the arguments and responses and extracts the latent features underlying the arguments that are driving the responses. Once estimated, this model enables us to identify why certain arguments are more or less persuasive.

To apply the SUN model we represent each argument as an element of $\mathbb{R}^s$ using a document embedding, where s depends on the dimensionality of the specific embedding used. After representing each of our arguments as an embedding, our collection of arguments is $\mathbf{M} \in \mathbb{R}^{(n \times s)}$.

Using this representation the SUN topic model builds on prior work on unsupervised topic models (Lee and Seung, 1999; Mcauliffe and Blei, 2007; Fong and Grimmer, 2016). These models typically use a bag-of-words representation that is non-negative. Because our data representation (the embeddings) includes negative values, we utilize a semi-non-negative matrix factorization that allows negative values in the data representation but still guarantees nonnegative topics. While, as with all topic models, the feature loadings produced by the SUN topic model are not intrinsically interpretable, they enable us to find lower-dimensional representations that can be interpretable when finding a good model fit and assigning appropriate topic labels. Particularly, we find that the non-negativity constraint on the loadings promotes interpretable latent factors.

To avoid issues with the definition of causal effects, we do not constrain the prevalence of the causal effects to sum to 1 (Fong and Grimmer, 2016) as would occur in classic LDA topic model variants (Blei et al., 2003). Unlike prior work (Fong and Grimmer, 2016, 2023), we do not discretize the loadings as present or absent and instead allow for topics to have a non-negative prevalence across documents. We place no constraints on the effect that the presence of a latent topic might have on the responses $\mathbf{Y}$.

SUN Model Setup: The SUN topic model discovers and estimates latent features that simultaneously explain differences in the arguments and in the responses to the arguments.[1] We build on Ding et al. (2010) to discover the latent topics for the embedding representation, $\mathbf{M}$, of the arguments, such that

$$\mathbf{M} \approx \mathbf{W}\mathbf{B} \text{ where } \mathbf{M} \in \mathbb{R}^{n \times s}, \mathbf{W} \in \mathbb{R}_{+}^{n \times J}, \quad (1)$$
$$\text{and } \mathbf{B} \in \mathbb{R}^{J \times s}$$

where $J \in \mathbb{N}^{+}$ is a user-set parameter that determines the number of topics. Here, each row of $\mathbf{W}$ captures the presence, or loadings, of each latent topic for a given document while $\mathbf{B}$ is the mapping between these latent topics and the embedding space.

Because we want our latent topics to explain the responses $\mathbf{Y}$, we also consider

$$\mathbf{Y} \approx \mathbf{W}\boldsymbol{\gamma} \quad \text{where } \mathbf{Y} \in \mathbb{R}^{n}, \boldsymbol{\gamma} \in \mathbb{R}^{J}. \quad (2)$$

Here, $\boldsymbol{\gamma}$ captures the relationship between the latent variables and the responses. We refer to $\boldsymbol{\gamma}$ as the *persuasion coefficients*. These coefficients will be an important determinant of the causal effects we estimate later, but may not justify a causal interpretation when analyzed directly from the model.

Defining the SUN Topic Model Loss Function: The total loss function for the SUN topic model is a convex combination of a loss function for the model to explain the latent features in the arguments and a loss function for the latent features that best explain the responses.

We define the first component of our loss function corresponding to our approximation of $\mathbf{M}$ in (1) as

$$\mathcal{L}_A = \frac{1}{2} \|\mathbf{M} - \mathbf{W}\mathbf{B}\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Next, we define the loss for our approximation of the responses $\mathbf{Y}$ in (2) as

$$\mathcal{L}_R = \frac{1}{2} \|\mathbf{Y} - \mathbf{W}\boldsymbol{\gamma}\|_2^2 .$$

We combine the two loss functions to define the total loss function as

$$\mathcal{L} = \alpha\mathcal{L}_A + (1-\alpha)\mathcal{L}_R$$

where $\alpha \in (0,1)$ is a parameter that controls the share of weight placed on the argument loss function $\mathcal{L}_A$ or the response loss function $\mathcal{L}_R$. As α goes to one, the latent topics are increasingly focused on explaining the content of the arguments in the embedding space, $\mathbf{M}$. As α goes to zero, our latent topics are increasingly focused on only explaining the responses, $\mathbf{Y}$. The α parameter enables us to discover latent topic combinations that balance explaining variation in the documents and in the responses.

Further, we can perform simple algebraic manipulation, as included in the Appendix A.1, to rewrite the total loss function as

$$\mathcal{L} = \alpha\frac{1}{2} \|\mathbf{M} - \mathbf{W}\mathbf{B}\|_F^2 + (1-\alpha)\frac{1}{2} \|\mathbf{Y} - \mathbf{W}\boldsymbol{\gamma}\|_2^2$$
$$= \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \quad (3)$$

where $\mathbf{X} := \left(\sqrt{\alpha}\mathbf{M} \middle| \sqrt{1-\alpha}\mathbf{Y}\right)$ and $\mathbf{H} := \left(\sqrt{\alpha}\mathbf{B} \middle| \sqrt{1-\alpha}\boldsymbol{\gamma}\right)$. In other words, algebraic manipulation reduces the supervised problem to the well-studied problem of semi-nonnegative matrix factorization.

Estimating the SUN Topic Model: Minimizing (3) is a non-convex optimization problem, but we can use the closed form updating steps of Ding et al. (2010). We detail those straightforward updating steps in section A.2 of the Appendix. The estimation routine proceeds with iterative updates of $\mathbf{W}$ and $\mathbf{H}$ for a fixed number of steps (early stopping), or until we reach convergence. The results of this

[1] Topic model implementation is available here: https://github.com/TillRS/SUN_TopicModel.

estimation procedure are the estimated topic loadings $\widehat{\mathbf{W}}$ and the estimated relationship $\widehat{\mathbf{H}}$ between the topic loadings and the unified, scaled input data $\mathbf{X}$. Importantly, the column $\widehat{\mathbf{h}}_{(s+1)} = \sqrt{1-\alpha}\widehat{\gamma}$ captures the estimate for the scaled persuasion coefficients.

As with other topic models, each model fit only corresponds to a local minimum of the non-convex total loss function (3). This means that we cannot rely on optimization alone to choose a model for analysis. Instead, we evaluate models using both numerical and qualitative information. Quantitatively, we run 10-fold cross-validation across different choices for J and α, where we perform out-of-sample prediction using the procedure described in Section 2.3. We use numerical information about the out-of-sample predictive accuracy of the model for predicting the responses because this will correspond directly to the task of forecasting an argument's persuasiveness. We also discuss the average topic coherence metric of Mimno et al. (2011) for the chosen topic model fit in our case study. Qualitatively, we manually inspect the coherence and exclusivity of the latent topics for different model fits (Roberts et al., 2016). To perform this evaluation, we inspect elements in the training set with very high, mid, and low loadings of a given latent topic and try to identify common themes and features. If we view these topics as coherent, mutually distinct, and plausible, we deem it to be a good fit in terms of latent topics.

2.3 Step 3: Estimate Causal Effects

After fitting the SUN topic model, we use it to estimate the causal effect of changing a topic's presence in the documents and to forecast the persuasiveness of new arguments. Both of these tasks require inferring topics for a new argument not included in the original model fit. When inferring topic loadings of a new document $t \in \mathcal{D}$, we do not include its response $\mathbf{Y}_t$ as model input, as we are trying to estimate the relationship to the response or to predict it. Hence, we are trying to learn the topic loadings solely based on the embedding $\mathbf{M}_t$ after standardizing and scaling it following the training data scaling. This corresponds to using $\sqrt{\alpha}\mathbf{M}_t$ and minimizing $\mathcal{L}_A$ to infer the latent topic loadings $\widehat{\mathbf{W}}_t$ while holding $\sqrt{\alpha}\widehat{\mathbf{B}}$ fixed. We leave $\widehat{\mathbf{B}}$ unchanged to preserve the previously discovered latent topics and focus on learning the presence of these topics in the new argument t. Note that we can extract $\sqrt{\alpha}\widehat{\mathbf{B}}$ from the previously derived $\widehat{\mathbf{H}}$ and minimize $\mathcal{L}_A$ by initializing and updating $\widehat{\mathbf{W}}_t$ while holding $\widehat{\mathbf{B}}$ constant[2].

To avoid theoretical issues that arise when fitting a latent model and estimating the causal effect of those same latent topics, we use an estimation set of arguments that had not been previously used for model fitting (Egami et al., 2022).

More formally, let $\mathbf{M}_E$ and $\mathbf{Y}_E$ denote the document embeddings and responses of the estimation set. We infer the latent topics $\mathbf{W}_E$ as described above. Then we can fit the regression

$$\mathbf{Y}_E = \widehat{\mathbf{W}}_E \boldsymbol{\beta} + \mathbf{e}$$

where $\boldsymbol{\beta}$ is the estimand of the instantaneous effect of the latent topics under the assumption that the topics are independent of $\mathbf{e}$, the error term. We focus on the effects of topics estimated using a linear model, but the AutoPersuade workflow is modular: it works with any model that infers the causal effects of underlying latent features of the arguments.

As a second application, we use the output of the SUN topic model to forecast the response to new arguments. This is particularly powerful because it enables us to quickly identify arguments that our evidence suggests will be persuasive if deployed. These can be generated effectively using LLMs. As we will show, this works well *on average* but not for optimizing the *most persuasive* arguments. This in turn makes the technique best suited to answering explanatory questions.

2.4 Designing Baselines

The challenge in evaluating and comparing our method to strong baselines is that it provides two distinct outputs: Predictions of the persuasiveness of held-out messages and, more importantly, estimates of interpretable components of the message and their causal effects. Because few methods attempt to do both tasks, we assess performance with respect to baselines that maximize each component separately.

[2] We derive the original model fit, $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{H}}$, using 100 updating steps. Matching this, we originally used 100 updating steps when inferring topic loadings for new documents to create the presented results. However, the convex topic inference problem does not consistently converge after 100 steps. Hence, we implemented an alternative approach using CVXPY (Diamond and Boyd, 2016) to infer topic loadings. While our main results remain largely unchanged, we report the results of this alternative method in Appendix B and briefly discuss some benefits of early stopping and running to convergence, respectively.

The strongest baseline for predictive performance sacrifices interpretability to maximize predictivity directly. We compare our topic model results to classic supervised baseline methods based on the same document embeddings (Lasso, Gradient Boosting, and Random Forest) which can asymptotically approximate more complex functions of the embeddings than our topic models to predict persuasiveness. Although these are older methods, they work well with relatively small datasets and they provide a direct measure of what we are giving up to obtain interpretable estimates.

It is difficult to assess how well we extract interpretable components and their causal effects (since there is neither a clear measure of interpretability nor a way to observe a causal effect). Here we design a series of human validation studies that compare our approach to the strong baseline of asking large language models to improve on existing persuasive arguments and construct new persuasive arguments.

3 Case Study: Pro-Veganism Arguments

We use a survey experiment to collect respondents' reactions to pro-veganism arguments. We ask each respondent to compare two messages and to choose the one they find more persuasive. We then summarize these pairwise contests using a Bradley-Terry model and use this as our response variable (Bradley and Terry, 1952; Newman, 2023). Finally, we use the arguments and the summarized performances to fit the SUN topic model and evaluate the persuasiveness of new potential arguments.

In particular, we start out with an original collection of arguments and responses followed by three validation studies to validate our estimates and explore the predictability of argument performances. Validation Studies 1 & 2 focused on generating and comparing synthesized and modified arguments with high persuasiveness scores to test whether we can improve on the best-performing arguments of the original argument collection. However, we find that such filtering based on the average marginal component effect of latent topics does not reliably improve arguments within the tail of best-performing arguments. Validation Study 3, on the other hand, validates our estimate of the average marginal component effect on argument persuasiveness by intervening on arguments across the entire distribution of topic loadings, not just a tail.

3.1 Curating Pro-Vegan Messages

As a first step, we created the original argument collection used in our survey evaluation. We began by curating a set of 93 root arguments, primarily sourced from longer essays on animal rights from advocacy websites. These arguments were distilled into shorter versions (approximately 280 characters) using GPT-4 (OpenAI, 2023). This distillation process involved summarizing each argument into short statements of circa 160 characters and then expanding these summaries back into 280-character versions. Each of the 93 original arguments was summarized three times, and GPT-4 was used to generate five "more persuasive" versions of the first two summaries and three "less persuasive" versions of the third summary. We prompted for less and more persuasive versions to validate if our survey-based results aligned with these instructions on how to paraphrase the arguments. Overall, this yielded a total of 1209 arguments based on the original 93 arguments.

Additionally, 100 arguments were generated solely by GPT-4 without any prior argument collection. These 100 arguments were created by prompting GPT-4 to produce ten distinct pro-veganism arguments, which were then summarized and expanded as described above. Here, we did not prompt for any of them to be "less persuasive". The complete original argument collection thus consists of 1309 pro-veganism arguments. This collection of arguments, including original sources and intermediary summaries as well as the arguments generated for later validation studies, are provided in the Supplementary Materials.

3.2 Collecting and Summarizing Responses

After curating the original collection of arguments, we divided it into a training set, comprising $2/3$ of the arguments, and an estimation set, comprising the remaining $1/3$. This division is stratified based on the 93 underlying root arguments, ensuring our training and test sets are well-balanced.

We then deployed a survey on Amazon's Mechanical Turk (MTurk) platform to evaluate the arguments. In this survey, each respondent was shown a pair of arguments about veganism and then asked to select the more persuasive argument. The argument comparisons were fully randomized, both the pairwise comparisons and the order in which the arguments appeared (displayed side by side). In total, each respondent evaluated five pairwise comparisons. We included the full survey questionnaire in the Supplementary Materials and additional details on survey results in the Appendix.

We compiled the results of 1,036 sessions with five pairwise comparisons of arguments each. This results in 5,180 total evaluations of pairs of arguments. To increase our sample size, we allowed respondents to complete multiple sessions. To summarize an argument's performance across the comparisons, we fit a Bradley-Terry (BT) model (Bradley and Terry, 1952; Newman, 2023) on the training set arguments. The BT model ranks arguments in terms of their likelihood to win a pairwise contest, and we use this summary as our response variable for the SUN topic model. We infer a test set document's response variable by applying the Bradley-Terry model to its performance in pairwise contests while keeping the training set arguments' scores fixed.

Note that we consider this argument evaluation to be a survey and not an annotation task. In that sense, disagreement is expected and appropriate, as different people are persuaded by different arguments. We measure argument persuasiveness based on average response within this population.

Of the 5180 comparisons on the original argument collection, 49.9% of the arguments on the right-hand side won the pairwise comparisons, suggesting no ordering effects. Further, while we found some evidence that argument length affects performance positively, it is uncorrelated with the latent topics discovered in our argument collections and thus does not meaningfully affect our estimated effects. The estimates presented in section 3.4 are derived while controlling for argument length.

Lastly, the arguments of the original collection, prompted to be 'more convincing' achieved an average Bradley-Terry score of 1.03 while the 'less convincing' arguments averaged 0.96. As these two groups are stratified across the underlying root arguments, this indicates that the GPT-prompting and our argument performance evaluation achieve and identify the desired effects.

3.3 Fitting the SUN topic model

We represent the arguments in our collection in an embedding space using OpenAI's "text-embedding-ada-002" model (Neelakantan et al., 2022). However, we tested and found similar predictive performance using alternative embedding models, as outlined in the Appendix. Equipped with this data, we fit several instances of the SUN topic model to arrive at a final model choice. We ran 10-fold cross-validation on the training data across different topic numbers J and α values to explore how predictive performance and topic coherence and exclusivity changed under different parameter choices and local minima.

Using the output from the SUN topic model we found that 10 topics and $\alpha = 0.5$ performed well both relative to other hyperparameter choices and benchmark models as shown in Figure 2 (recalling that even matching a classic supervised baseline should be challenging because they are not constrained to working with interpretable components). We then fit multiple models with these parameter choices, but different random initializations. For this model selection step, we used 80% of our training data and checked the predictive performance on the remaining 20% to prevent us from overfitting.

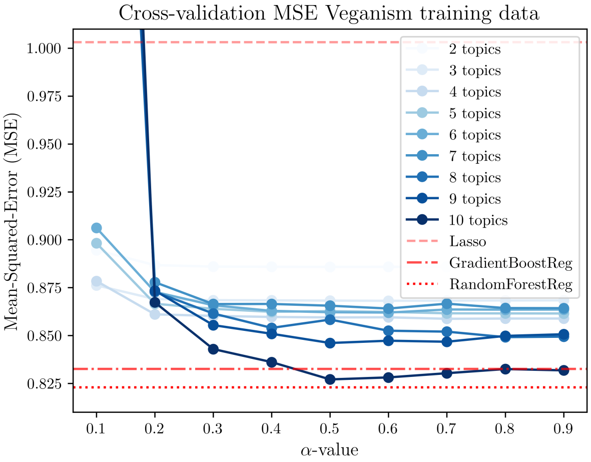


Figure 2: Out-of-sample predictive accuracy of SUN topic model for different parameter choices, as well as benchmark models on the training data. Results were calculated using 10-fold cross-validation.

Table 1 provides our human-assigned labels for our SUN topics. The model achieves an in-sample MSE of 0.82 and an out-of-sample MSE of 0.7 on our training data set. The higher in-sample MSE is likely driven by outliers.

The selection of an optimal model fit with the chosen $\alpha = 0.5$ and $J = 10$ hyperparameters primarily relies on inspecting documents with high topic loadings and manually identifying themes. However, we also explored numerical approaches for model fit selection. Specifically, we examined the topic coherence metric proposed by Mimno et al. (2011), which measures how frequently the most common words of a given topic co-occur. This metric, based on word frequencies, is typi-

	Latent Topics
(1)	Uncertainty and generalizations
(2)	Inefficient use of resources
(3)	Exploitation, suffering, and compassion
(4)	Morals, ethics, and justifications
(5)	Treatment of cows and chickens
(6)	Individual contributions and responsibility
(7)	Animal rights and speciesism
(8)	Health benefits
(9)	Addressing criticism and fallacies
(10)	Climate change and sustainability

Table 1: Labels for the discovered latent topics of the arguments for veganism.

cally used with bag-of-words representations.

For our embeddings-based topic model, we identified the most frequent words per topic by extracting the 25 documents with the highest loadings for each topic and computing TF-IDF scores (Salton and Buckley, 1988). We calculated the topic coherence metric using the 5 words with the highest TF-IDF scores for each topic.

This approach highlights the trade-off between finding the most coherent topics and identifying topics that best explain variations in the response variable. A high coherence score and a low out-of-sample prediction MSE characterize good performance. While this specific topic coherence metric is just one of many possible ways to evaluate topics selected by a topic model, and we generally emphasize the importance of the manual inspection process, Figure 5 in the Appendix demonstrates that our selected model fit presents a well-performing balance of these two metrics relative to other random model fits.

3.4 Estimating Causal Effects

Using our model fit, we inferred the latent topic loadings of the arguments in our estimation set. Combining these with the inferred Bradley-Terry scores, we estimated the average marginal component effect of the topics on a document's relative performance in the pairwise contests (Hainmueller et al., 2014; Fong and Grimmer, 2023).

Figure 3 shows the estimates and their respective confidence intervals when controlling for argument length. The latent topics associated with significant positive effects are describing the inefficient use of resources for animal products (2), highlighting the importance and impact of individual consumer choices (6), and health benefits (8). On the other hand, discussing morals and ethical justifications for meat consumption (4), animal rights and so-called speciesism (Singer, 2009) (7), and addressing criticism of veganism and its supposed fallacies (9) are associated with negative effects on our persuasiveness score. The effects of the other discovered latent topics are estimated to be close to zero.

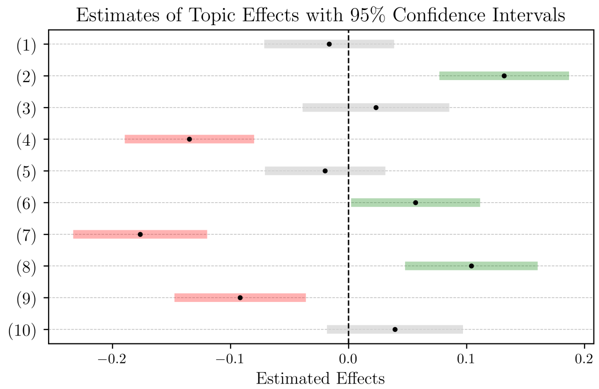


Figure 3: Estimated effects of discovered latent topics on the persuasiveness score of arguments for veganism. Refer to Table 3 in the Appendix for more details.

3.5 Validation Studies

The challenge of evaluating causal effects is that there is no perfect benchmark because we can never observe a causal effect (Feder et al., 2022). An observable implication of learning what we hope to learn is that we can manipulate arguments to make them more persuasive. We test that implication here across three human-validation studies.

3.5.1 Validation Study 1: Generating New Persuasive Arguments

In our first validation study, we evaluated the persuasiveness of newly synthesized and modified arguments generated from the 45 best-performing arguments in our original collection of 1309, based on Bradley-Terry scores from the initial survey experiment. We identified 20 of these top arguments as *proto-arguments* and used them to create new arguments via two approaches: synthesis and stronger emphasis.

Synthesis Arguments: We prompted GPT-4 to combine pairs of proto-arguments, generating $(20 \times 19) \times 2 = 760$ synthesis arguments.

Stronger Emphasis Arguments: We prompted GPT-4 to rewrite each proto-argument with increased emphasis on its primary latent topic, generating three new versions for each proto-argument, resulting in $3 \times 20 = 60$ stronger emphasis arguments.

For both approaches, we filtered the new arguments using rejection sampling to check whether they achieved the desired latent topic loadings and had higher predicted persuasiveness scores than the underlying proto-arguments. From these filtered arguments, we selected 30 distinct synthesis arguments and 15 distinct stronger emphasis arguments based on their predicted persuasiveness scores. We also generated 10 arguments by prompting GPT-4 to produce its "most persuasive" pro-veganism argument, resulting in a final set of 100 arguments: 45 best-performing original arguments [Original], 30 synthesis arguments [Argument Synthesis], 15 stronger emphasis arguments [Stronger Emphasis], and 10 best GPT arguments [GPT-best].

Using the same pairwise comparison strategy of the original setup, we evaluated the performance of these arguments against each other. We collected responses to 990 pairwise comparisons among the arguments from 198 unique MTurk respondents. Table 2 shows that the synthetic arguments won 54% of the time, outperforming the stronger emphasis arguments (51.8%), the GPT-best arguments (51.1%), and the original arguments (45.4%) as shown in Table 2. The confidence interval for the difference in win rates between Argument Synthesis and Original (SY − OG) is $[1.5, 15.7]$, and for the difference between Argument Synthesis and GPT-best (SY − GPT), it is $[-5.5, 12.1]$. In other words, synthesizing arguments by combining the best properties determined by our workflow beats the best of the original arguments and appears to outperform the baseline of asking GPT-4 to make its best argument.

3.5.2 Validation Study 2: Limitations of Argument Optimization

Building on Validation Study 1, Validation Study 2 aimed to assess the limitations of our argument optimization approach by directly comparing the synthesized arguments to their respective proto-arguments. No new arguments were generated for this study; instead, we focused on the 30 synthesis and 15 stronger emphasis arguments and compared them against the original arguments from which they were derived. While the order of the compared arguments (left vs. right) was still randomized, the pairings were now fixed, as each new argument was only compared against its corresponding proto-arguments.

	Win Rate (%)	95% CI
Validation Study 1		
Argument Synthesis (SY)	**54.0**	[49.9, 58.3]
Original (OG)	45.4	[41.4, 49.1]
GPT-best (GPT)	51.1	[44.4, 57.1]
Stronger Emphasis (SE)	51.8	[46.3, 56.8]
Validation Study 2		
Argument Synthesis (SY)	48.5	[44.3, 53.1]
Original (OG)	**52.0**	[47.4, 56.3]
Stronger Emphasis (SE)	45.9	[35.8, 56.4]
Validation Study 3		
Increased Topic (2) Args.	**69.7**	[65.3, 74.1]
Decreased Topic (2) Args.	31.7	[22.7, 41.4]

Table 2: Results of Validation Study 1, 2, and 3. For each study, win rates are calculated as the share of comparisons with arguments of other origins that are won. Confidence intervals are calculated based on 500 bootstraps of the individual comparison outcomes. The highest score for each study is bolded, the second-highest score is underlined.

We collected responses from 102 unique MTurk respondents, totaling 510 random pairwise comparisons. The results summarized in Table 2 indicate that the synthetic arguments did not consistently outperform their original counterparts. This suggests that while our method successfully identified persuasive topics, enhancing these topics at the higher end of the distribution did not result in significantly more persuasive arguments. The findings suggest an open challenge for maximizing argument persuasiveness.

3.5.3 Validation 3: Evaluating the Average Effects in the Population

The causal effects we estimate in our design are the effect of interventions defined over the full population of documents, the Average Marginal Component Effect (AMCE). To generate a validation based on this estimand, we focus on a single topic: Topic (2) *inefficient use of resources*. For this validation study, we intervened on randomly selected arguments from the original set of 1309 arguments. We first selected 100 random arguments with a high topic (2) loading (greater than 2.0, or the 87th percentile) and prompted GPT-4 to rewrite them to decrease the presence of topic (2). Next, we selected 200 arguments with a low topic (2) loading (less than 2.0) and prompted GPT-4 to rewrite these arguments to increase the presence of topic (2). From these alterations, we curated a collection of 80 arguments with increased topic (2) presence and 20 with decreased presence, alongside the original 100 arguments, forming the 200 arguments used for Validation Study 3.

We collected responses from 100 unique MTurk respondents, totaling 500 pairwise comparisons. Again, we only allowed for comparisons of a given new, altered argument against its underlying, original argument. Table 2 shows that manipulations of the documents behave as expected when averaged across the entire distribution.

The interventions used in the validation studies provide a face-validity check that our labels correspond well to our learned latent topics by demonstrating that adjusting arguments to emphasize a specific theme increased or decreased the associated topic loadings. A more detailed discussion of the rejection sampling process used in the validation studies is provided in the Appendix.

Together, our validation studies strongly suggest that the intervention mechanism informed by the AMCE has the expected effect, which allows us to validate topic labels and AMCE estimates on new samples. However, improving the best-performing arguments of our sample based on these AMCE insights suggests that the effect on the margin differs from the effect in the tail of the distribution. While this is great for explanation, the method is less well-suited to identifying the best argument.

4 Related Work

The AutoPersuade workflow builds on a burgeoning literature that examines the causal effects of texts on outcomes (Feder et al., 2022). Fong and Grimmer (2016) and Fong and Grimmer (2023) introduce a procedure for identifying the features of texts that drive responses, but their framework relies upon a more constrained topic model and was unable to infer the persuasiveness of new texts. Egami et al. (2022) provide a general guide for causal inference with texts and outline a series of identification issues. Palmer and Spirling (2023) experimentally demonstrate that LLMs can perform nearly as well as humans at producing persuasive arguments.

Other work has analyzed the linguistic characteristics of persuasive messages. Feng and Hirst (2011) categorize arguments into common schemes, while Tan et al. (2016), Habernal and Gurevych (2016), and Gleize et al. (2019) examine successful arguments in online discussions, debate forums, and Wikipedia, investigating the predictive power of structural features. Wang et al. (2019) explore personalized persuasion processes and Wachsmuth et al. (2017) propose a systematic taxonomy for argument quality. Zhang et al. (2020) explore the causal effects of conversational tendencies. However, their work does not allow for domain-specific feature discovery and causal inference with these features. Zhao et al. (2021) model the relatedness among controversial topics using embedding-based methods based on individuals' stances, integrating topic semantics from arguments and persuasion factors. Currently, controllable argument generation relies upon previously identified features or domain expertise (Saha and Srihari, 2023; Schiller et al., 2021). These procedures can augment steps 1 and 3 of the workflow.

5 Conclusion

This paper introduces AutoPersuade—a new workflow for persuasion. Our AutoPersuade approach curates arguments and collects responses to those messages, identifies the latent features that cause them to be more or less persuasive, infers the causal effects of those topics, and enables the selection of more persuasive messages from a new collection of candidate messages.

Each step of the workflow is modular and can be improved as new technologies become available. For example, better initial curation of arguments will make data collection more efficient; other interpretable models can be used to assess why some arguments are persuasive; and we can explore targeting of messages to particular people. New techniques could use the result of data and models to automatically generate more persuasive messages.

6 Limitations and Ethics

Here we briefly overview the limitations and ethical considerations of our work.

6.1 Limitations

While our workflow is quite general, there are important limitations both to the general design and to our specific version of it. The main limitation of the framework is that it must be possible to collect a credible response variable from the relevant population to be persuaded. For example, in Section 3, we collect self-reported persuasion from Mechanical Turk workers; however, self-reports might substantially differ from induced behavioral change

(Coppock, 2023) and Mechanical Turk workers may not be reflective of the population of interest. These messages must also have a sufficiently diverse set of argument features to be able to discover the ones which are most persuasive. This limitation is shared with other mechanisms of assessing messages like A/B tests.

Our specific implementation also has important limitations. We are assuming that the document embeddings preserve the relevant information that allows for persuasion and that our topic model can pick it up. A more subtle concern is driven by the scaling invariance of the matrix factorization. The numerical value of the estimated effects is relative to the range of loadings for a given topic and thus is related to the distribution of that dimension in the training data. This means that while our estimates of the directional effect of topics are robust, the magnitude may not be. This is a problem without an obvious fix because there is no natural underlying scale to latent concepts.

In the applications reported above, we restrict ourselves to estimating the average persuasiveness of features—a limitation highlighted in our validation studies. A major opportunity moving forward would be to push past this general view and consider the effects of messages personalized to individual people based on some known covariates. This would naturally induce issues of power, but these might be addressable by moving beyond our static experimental design (where documents are assigned randomly) to an adaptive design which is optimized to find the most impactful message for each subpopulation (Offer-Westort et al., 2021). While these designs, which arise out of the literature on multi-armed bandits (Slivkins et al., 2019) have been used for fixed message options, they would need to be modified to fit our setting.

6.2 Ethics

Persuasion is about convincing someone to do something they would otherwise not do. The ethical boundaries of persuasion are often viewed through the lens of what we are trying to persuade people to do. While we have chosen applications we see as ethically positive, these strategies can be used by other actors for applications we would not endorse—just like A/B testing. For example, Mathur et al. (2023) demonstrate that politicians in the US use A/B testing to optimize messages in campaign emails. One could imagine a motivated actor using email opens as a response variable and learning even more effective techniques to induce responses from voters. Whether those responses are ethically negative or positive depends on whether the email messages help voters realize their true preferences, or deceive voters into supporting a candidate they would not with better and more complete information.

It is natural to worry that more effective persuasive tools will be used to persuade the public in a way that harms general welfare. These concerns arose as the radio reached most homes, they arose again when televisions became omnipresent, when the internet reached homes, when smartphones became widely available, when social media arrived on those smartphones, and similarly, they arise now with technologies like large language models. We think this history of concern over new technology is useful because it helps contextualize the current worries as an important and common reaction as new technologies are deployed in the public. Further, while we think our method is useful as an automated way to find persuasive tactics, it is important to note that persuasion itself has its limits. It is exceedingly difficult to customize messages to audiences—say voters in an election—even with extensive marketing data (Hersh, 2015).

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Alexander Coppock. 2023. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

David De Vaus and David de Vaus. 2013. *Surveys in social research*. Routledge.

Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Chris H.Q. Ding, Tao Li, and Michael I. Jordan. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1600–1609.

Christian Fong and Justin Grimmer. 2023. Causal inference with latent treatments. *American Journal of Political Science*, 67(2):374–389.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Carlos A. Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4).

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. 2014. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1):1–30.

Eitan D Hersh. 2015. *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Arunesh Mathur, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M Stewart, and Arvind Narayanan. 2023. Manipulative tactics are the norm in political emails: Evidence from 300k emails from the 2020 us election cycle. *Big Data & Society*, 10(1):20539517221145371.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

David L. Morgan. 1996. Focus groups. *Annual Review of Sociology*, 22(1):129–152.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

M. E. J. Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

Molly Offer-Westort, Alexander Coppock, and Donald P. Green. 2021. Adaptive Experimental Design: Prospects and Applications in Political Science. *American Journal of Political Science*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alexis Palmer and Arthur Spirling. 2023. Large language models can argue in convincing and novel ways about politics: Evidence from experiments and human judgement. Technical report, Working paper), Technical report.

PyMF. Python matrix factorization module. `https://github.com/pzoccante/pymf`. Accessed on February 15, 2024.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? *arXiv preprint arXiv:2404.00750*.

Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Sougata Saha and Rohini Srihari. 2023. ArgU: A controllable factual argument generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8373–8388, Toronto, Canada. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Peter Singer. 2009. Speciesism and moral status. *Metaphilosophy*, 40(3-4):567–581.

Aleksandrs Slivkins et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the causal effects of conversational tendencies. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Xinran Zhao, Esin Durmus, Hongming Zhang, and Claire Cardie. 2021. Leveraging topic relatedness for argument persuasion. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4401–4407, Online. Association for Computational Linguistics.

A Appendix

A.1 Derivation of the Total Loss Function

$$
\begin{aligned}
\mathcal{L} &= \alpha\mathcal{L}_A + (1-\alpha)\mathcal{L}_R \\
&= \alpha\frac{1}{2}\left\|\mathbf{M}-\mathbf{W}\mathbf{B}\right\|_F^2 + (1-\alpha)\frac{1}{2}\left\|\mathbf{Y}-\mathbf{W}\boldsymbol{\gamma}\right\|_2^2 \\
&= \frac{\alpha}{2}\operatorname{tr}\left((\mathbf{M}-\mathbf{W}\mathbf{B})^T(\mathbf{M}-\mathbf{W}\mathbf{B})\right) \\
&\quad + \frac{1-\alpha}{2}\left((\mathbf{Y}-\mathbf{W}\boldsymbol{\gamma})^T(\mathbf{Y}-\mathbf{W}\boldsymbol{\gamma})\right) \\
&= \frac{\alpha}{2}\left[\|\mathbf{M}\|_F^2 - 2\operatorname{tr}\left(\mathbf{M}^T\mathbf{W}\mathbf{B}\right) + \|\mathbf{W}\mathbf{B}\|_F^2\right] \\
&\quad + \frac{1-\alpha}{2}\left[\|\mathbf{Y}\|_2^2 - 2\mathbf{Y}^T\mathbf{W}\boldsymbol{\gamma} + \|\mathbf{W}\boldsymbol{\gamma}\|_2^2\right] \\
&= \frac{1}{2}\|\left(\sqrt{\alpha}\mathbf{M}|\sqrt{1-\alpha}\mathbf{Y}\right) \\
&\quad - \mathbf{W}\left(\sqrt{\alpha}\mathbf{B}|\sqrt{1-\alpha}\boldsymbol{\gamma}\right)\|_F^2 \\
&= \frac{1}{2}\left\|\mathbf{X}-\mathbf{W}\mathbf{H}\right\|_F^2
\end{aligned}
$$

where $\mathbf{X} := \left(\sqrt{\alpha}\mathbf{M}|\sqrt{1-\alpha}\mathbf{Y}\right)$ and $\mathbf{H} := \left(\sqrt{\alpha}\mathbf{B}|\sqrt{1-\alpha}\boldsymbol{\gamma}\right)$.

A.2 Semi-nonnegative Matrix Factorization

Following Ding et al. (2010), the closed form updating steps for the semi-nonnegative matrix factorization to minimize the total loss function of (3) are:

(S0) Initialize $\mathbf{W}$. Do a K-means clustering. This gives cluster indicators $\mathbf{W}$: $\mathbf{W}_{ik} = 1$ if $\mathbf{x}_i$ belongs to cluster k. Otherwise, $\mathbf{W}_{ik} = 0$. Add a small constant to all elements of $\mathbf{W}$. Following Ding et al. (2010), we use 0.2.

(S1) Update $\mathbf{H}$ (while fixing $\mathbf{W}$) using the rule

$$
\begin{aligned}
\mathbf{H} &= [\mathbf{X}^T\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}]^T \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}
\end{aligned}
$$

Note $\mathbf{W}^T\mathbf{W}$ is a $k \times k$ positive semidefinite matrix. The inversion of this small matrix is trivial. In most cases, $\mathbf{W}^T\mathbf{W}$ is nonsingular. When $\mathbf{W}^T\mathbf{W}$ is singular, we take the pseudoinverse.

(S2) Update $\mathbf{W}$ (while fixing $\mathbf{H}$) using

$$
\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik}\sqrt{\frac{(\mathbf{X}\mathbf{H}^T)_{ik}^{+} + \left[\mathbf{W}\left(\mathbf{H}\mathbf{H}^T\right)^{-}\right]_{ik}}{(\mathbf{X}\mathbf{H}^T)_{ik}^{-} + \left[\mathbf{W}\left(\mathbf{H}\mathbf{H}^T\right)^{+}\right]_{ik}}}
$$

where we separate the positive and negative parts of a matrix M as

$$
\begin{aligned}
M_{ik}^{+} &= \left(|M_{ik}| + M_{ik}\right)/2, \\
M_{ik}^{-} &= \left(|M_{ik}| - M_{ik}\right)/2.
\end{aligned}
$$

Note that our implementation of this matrix factorization builds on PyMF and GitHub Copilot was used for the coding parts of this research.

A.3 Data Collection and Preparation

A.3.1 Response Quality Control

To ensure the quality of survey responses, we conducted a series of small pilot studies on Amazon Mechanical Turk (MTurk). Participants were initially selected based on their past acceptance rates and the number of completed tasks, but the results revealed mixed levels of response quality. However, we observed improved attention and response quality when we restricted the sample to English-speaking adults residing in the U.S. who held 'MTurk Master' status, a designation granted to users with a track record of consistently high-quality work.

For the main experiment, we collected 1,038 responses from MTurk Masters. Only two responses were rejected for incorrectly identifying the questionnaire's focus as arguments for political participation. All other respondents correctly recognized that the questionnaire concerned arguments for adopting a vegetarian/vegan diet or arguments against animal cruelty, resulting in 1,036 valid responses. This provided a total of 5,180 pairwise comparisons for evaluating the arguments.

We employed a pairwise forced-choice design, where participants compared two arguments at a time. This setup was chosen over ranking multiple arguments to reduce the cognitive load and memory demands on participants.

For the three validation studies, we similarly collected 198, 100, and 102 responses, respectively.

A.3.2 Argument Selection for Validation Studies

Following the curation of a set of *proto-arguments*, we generated additional arguments using GPT-4, as outlined in Section 3.5. This process resulted in the creation of 760 Synthesis Arguments and 60 Stronger Emphasis Arguments. For the first validation study, we selected 30 Synthesis Arguments and 15 Stronger Emphasis Arguments.

Two primary criteria guided the selection process. First, we filtered arguments that met the quantitative requirements: a higher predicted persuasiveness score compared to the original proto-arguments and the property that the two main latent topics of the proto-arguments are two topics with the highest loadings in the new arguments. From the pool of arguments that satisfied these criteria, we manually selected those with high predicted persuasiveness scores, ensuring diversity by excluding arguments that were overly similar. For instance, for each proto-argument, we generated three Stronger Emphasis Arguments; if two such arguments met the numeric thresholds but were highly similar, we only included one in the study. After re-evaluating the argument filtering, we discovered that we included one Synthesis Argument that did not have a higher predicted persuasion score than both of its proto-arguments. However, excluding this argument and its pairwise comparisons from Validation Study 1 and 2 does not meaningfully affect the results.

For the third validation study, we employed a different argument generation strategy. We randomly selected arguments and prompted GPT-4 to rewrite them to either increase or decrease the presence of topic (2), *inefficient use of resources*. We then filtered the revised arguments based on their inferred scores for topic (2), ensuring they reflected the intended changes. As with the first validation study, we ensured that the selected arguments were sufficiently distinct from one another, beyond their inferred topic loadings.

In all validation studies, the process of generating arguments by giving GPT-4 one or two initial inputs and specifying desired changes resulted in coherent arguments that were in line with the original argument collection. However, as expected, it was more challenging to increase the presence of topics that were already prominent in an argument, and it was similarly difficult to revise high-performing arguments to achieve even higher predicted persuasiveness scores. These relatively small margins in inferred topic loadings are less robust, and the filtering for Validation Study 1 is more affected by changing our topic inference method than the filtering for Validation Study 3, as discussed in Appendix B.

All the arguments utilized in validation studies are included in the Supplementary Materials.

A.3.3 Alternative Embedding Models

For our case study, we are utilizing using OpenAI's "text-embedding-ada-002". However, we also tested and found almost identical predictive performance with both the new 'small' and 'large' embedding model 3 of OpenAI and the open-source SBERT paraphrase-MiniLM-L6-v2 model (Reimers and Gurevych, 2019). When inspecting a well-performing 10-topic model fit based on SBERT embeddings, we found that the identified topics roughly map pairwise to the 10 topics reported in this paper. Specifically, the correlation of topic loadings between these topic pairs across the 1308 original arguments ranged from 0.4 to 0.8, indicating that we can discover similar topics on embeddings derived from different models.

A.3.4 Data Preparation and Processing

We standardize our embedding representation and response variables to make the variance across the two data types approximately equal. This step ensures that neither the embedding nor the response variable mechanically dominates the loss function merely because the variance in one is much larger than the variance in the other. In practice, we divided the 1536-dimensional embeddings used in the applications by 2 after we standardized them.

Further, note that for any solution to the optimization problem, we can scale up $\widehat{\mathbf{W}}$ without affecting the result, as long as we scale down $\widehat{\mathbf{H}}$ accordingly and vice versa. To deal with this scale invariance problem, common across every matrix factorization task, we standardize the results. We suggest dividing each column of $\widehat{\mathbf{W}}$ by its standard deviation, multiplying the rows of $\widehat{\mathbf{H}}$ with the corresponding standard deviations.

A.4 Additional Results - Persuasiveness of Arguments

A.4.1 Differentiation of Similar Topics

The topic labels introduced in section 3.4 encapsulate distinct themes, despite some apparent overlapping among Topics (4), (6), and (7). In particular, Topic (4), *morals, ethics, and justifications*, emphasizes historical justifications for meat consumption, the role of societal norms, and the moral implications of human choices in eating meat or abusing animals. In contrast, Topic (7), *animal rights and speciesism*, centers on the standing of animals as an oppressed group, discussing their rights, well-being, and the species-wide discrimination they

Dep. Variable:	Pers. Score	**R-squared:**	0.259
Model:	OLS	**Adj. R-squared:**	0.242
No. Observations:	505	**F-statistic:**	15.66
Covariance Type:	nonrobust	**Prob (F-statistic):**	2.31e-26

	Coefficient	**Std Err**	**t**	**P> |t|**
const	-0.0144	0.119	-0.121	0.904
(1)	-0.0164	0.028	-0.585	0.559
(2)	0.1319	0.028	4.717	0.000
(3)	0.0233	0.032	0.737	0.462
(4)	-0.1348	0.028	-4.836	0.000
(5)	-0.0198	0.026	-0.763	0.446
(6)	0.0569	0.028	2.044	0.042
(7)	-0.1765	0.029	-6.115	0.000
(8)	0.1041	0.029	3.642	0.000
(9)	-0.0918	0.028	-3.242	0.001
(10)	0.0393	0.029	1.341	0.180
Arg. Length	0.0052	0.001	7.092	0.000

Table 3: Summary statistics of the causal effect estimation of the different topics discovered in the analysis of arguments for veganism including argument length (characters).

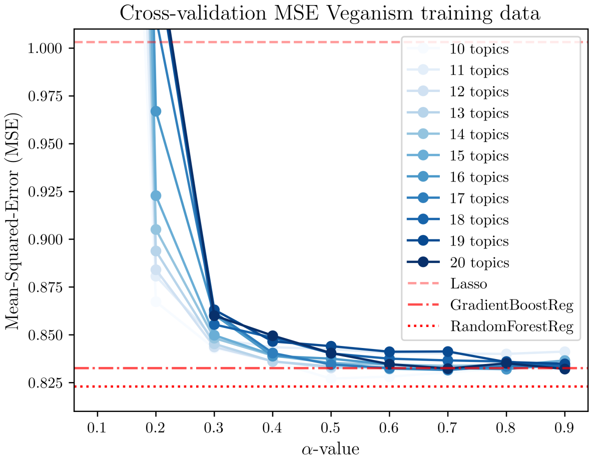


Figure 4: Out-of-sample predictive accuracy of SUN topic model for additional hyperparameter choices, as well as benchmark models on the training data. Results were calculated using 10-fold cross-validation.

face, often drawing parallels to other forms of historical oppression. Topic (6), *individual contributions and responsibilities*, shifts the focus to the direct impact of personal actions, highlighting the cumulative effects of individual choices on alleviating suffering through conscious consumption.

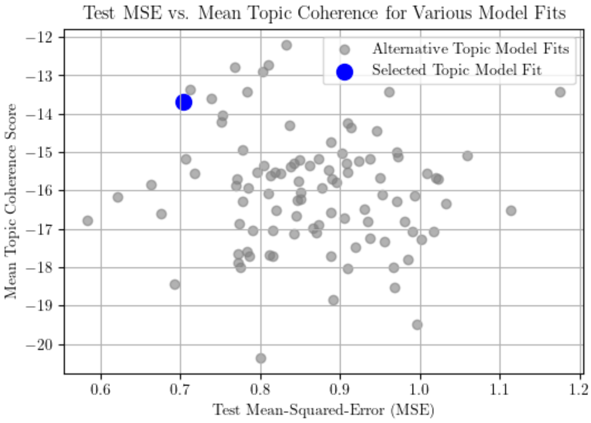


Figure 5: Topic coherence and out-of-sample predictive accuracy on 20% holdout of the training data for parameter choices $\alpha = 0.5$ and the number of topics $J = 10$.

B Converging Topic Inference

Following Ding et al. (2010) and evaluations of the predictive performance in our cross-validation step, we use 100 iterations of the semi-nonnegative matrix factorization updating steps when fitting our model. One step includes updating both $\mathbf{H}$ and $\mathbf{W}$. For our original results, reported above, we used the updating step for $\widehat{\mathbf{W}}$ while holding $\widehat{\mathbf{B}}$ fixed to minimize $\mathcal{L}_A$ to infer topic loadings for new documents based on a previously selected model fit. Matching our original topic fitting, we ran this updating step 100 times to infer topic loadings.

However, the convex sub-problem of inferring $\widehat{\mathbf{W}}$ given $\widehat{\mathbf{B}}$ did not consistently converge with only 100 steps. While this approach converges eventually, we also implemented a new extension to the SUN topic model where topic inference is done using the convex optimization solver CVXPY (Diamond and Boyd, 2016). This allows for faster convergence.

In practice, this means that originally, when convergence was not met, the inferred topic loadings of a new document were marginally affected by the random initialization of its topic loadings and the set of documents (other rows) in $\widehat{\mathbf{W}}$ for which we simultaneously inferred topic loadings. While this does not have a meaningful effect on our causal estimates or predictive performance, the new inference method yields more robust results when inferring topic loadings.

However, running $\widehat{\mathbf{W}}$ to convergence results in some documents with very high topic loadings across all latent topics, which reduces scarcity and complicates the interpretability of the inferred topic loadings. Identifying the ideal topic inference approach that balances robust results and the benefits of early stopping might be the subject of future research.

B.1 Changes in Results

While we find that the main results of our work do not change meaningfully, we include results corresponding to this updated topic inference approach in this section of the Appendix. As detailed in the following sections, we observe two main effects of the new inference method. First, the confidence intervals of our causal estimates are smaller, indicating a more precise inference of topic loadings. Second, some arguments selected for our validation studies no longer satisfy the filtering cutoffs based on inferred topic loadings. In particular, for Validation Study 1 & 2, we selected and generated new arguments at the very tails of the topic loading distributions. The differences in topic loadings and predicted persuasiveness scores between original, proto-arguments, and newly generated arguments were often small. These small differences are affected by the new topic inference method, leading to different argument filtering results. However, our main results remain consistent when re-evaluating the validation studies, considering only arguments that passed the filtering using the newly inferred topic loadings.

B.2 Cross-Validation

Figure 6 shows the cross-validation results using the CVXPY-based topic inference for the out-of-sample predictions. The results remain virtually unchanged and the combination of $J = 10$ topics and $\alpha = 0.5$ remains the best-performing hyperparameter choice.

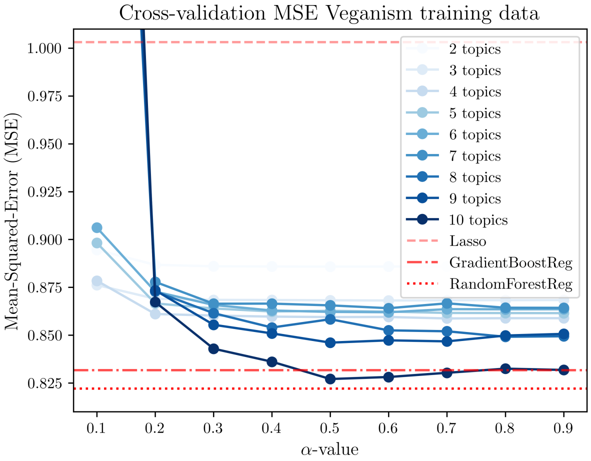


Figure 6: Out-of-sample predictive accuracy of SUN topic model for different parameter choices, as well as benchmark models on the training data. Results were calculated using 10-fold cross-validation and topic inference using CVXPY.

B.3 Causal Inference

Using the same topic model fit, we now infer the topic loadings on our estimation set using the new inference approach. As shown in Figure 7 and Table 4, our estimates remain mostly unchanged. However, we do observe smaller confidence intervals.

B.4 Validation Studies

While the previous results were only marginally affected by the new topic inference, the selection

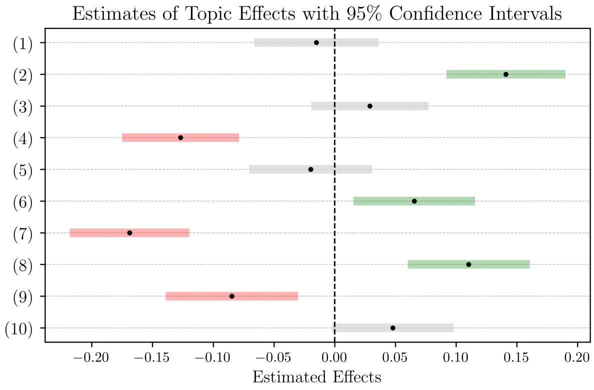


Figure 7: Estimated effects of discovered latent topics on the persuasiveness score of arguments for veganism, based on topic inferences using CVXPY. Refer to Table 4 for more details.

Dep. Variable:	Pers. Score	**R-squared:**	0.259
Model:	OLS	**Adj. R-squared:**	0.243
No. Observations:	505	**F-statistic:**	15.67
Covariance Type:	nonrobust	**Prob (F-statistic):**	2.25e-26

	Coefficient	**Std Err**	**t**	**P> |t|**
const	−0.0612	0.035	−1.765	0.078
(1)	−0.0151	0.026	−0.581	0.562
(2)	0.1410	0.025	5.648	0.000
(3)	0.0290	0.025	1.182	0.238
(4)	−0.1269	0.025	−5.179	0.000
(5)	−0.0197	0.026	−0.763	0.446
(6)	0.0655	0.025	2.570	0.010
(7)	−0.1689	0.025	−6.747	0.000
(8)	0.1104	0.026	4.321	0.000
(9)	−0.0847	0.028	−3.049	0.002
(10)	0.0479	0.026	1.877	0.061
Arg. Length	0.0052	0.001	7.094	0.000

Table 4: Summary statistics of the causal effect estimation of the different topics discovered in the analysis of arguments for veganism and inferred using CVXPY.

criteria for new arguments for our validation studies were more significantly impacted. As we are applying the relatively strict numerical filter for selecting arguments, we need to update our selected arguments for validation studies 1, 2, and 3.

In Validation Studies 1 and 2, we selected new arguments that had higher predicted persuasion scores and whose highest topic loadings corresponded to the targeted topics (two topics for synthesis arguments and one topic for stronger emphasis arguments). When we use the newly derived topic loadings, only 17/30 of the previously selected synthesis arguments and 10/15 of the stronger emphasis arguments meet these criteria.

In Validation Study 3, we intervened to either increase or decrease the presence of topic (2), and then selected arguments that reflected the targeted change in their topic (2) loading. Of the previously selected increased presence argument, we retain 67/80, and of the previously selected decreased presence arguments, we retain 20/20.

	Win Rate (%)	95% CI
Validation Study 1		
Argument Synthesis (SY)	**54.5**	[48.5, 60.2]
Original (OG)	45.2	[40.5, 50.0]
GPT-best (GPT)	53.2	[46.4, 60.4]
Stronger Emphasis (SE)	51.6	[44.4, 58.9]
Validation Study 2		
Argument Synthesis (SY)	46.7	[40.5, 53.3]
Original (OG)	**54.2**	[48.8, 59.3]
Stronger Emphasis (SE)	41.8	[29.1, 54.4]
Validation Study 3		
Increased Topic (2) Args	70.0	[64.6, 74.8]
Decreased Topic (2) Args	30.6	[22.3, 40.3]

Table 5: Results of Validation Study 1, 2, and 3 when only considering arguments that pass the filtering using the CVXPY topic inference approach. For each study, win rates are calculated as the share of comparisons with arguments of other origins that are won. Confidence intervals are calculated based on 500 bootstraps of the individual comparison outcomes. The highest score for each study is bolded, the second-highest score is underlined.

Only considering the outcomes of pairwise comparisons of arguments that we retain based on these new filtering results, we are left with 594/990, 295/510, and 434/500 pairwise comparisons for Validation Study 1, 2, and 3 respectively.

We recalculate the win rates per argument group for Validation Studies 1, 2, and 3 as summarized in Table 5. While the lower number of comparisons leads to wider confidence intervals, there are no fundamental changes to the results of the validation studies. The main change is that Stronger Emphasis arguments are no longer the second best performing in terms of win rate in Validation Study 1. Yet, all other relative performance rankings are preserved, and the main findings persist.

Argument	Topic Loadings									
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Many believe that most meat is sourced from humane, small farms, but the reality is factory farms are the major source. Their production processes are highly secretive, including the methods of slaughter. Even terms like ’free range’ do not guarantee the absence of animal suffering during slaughtering.	**4.87**	1.64	1.20	2.70	1.21	1.00	0.03	1.07	0.56	1.40
Slaughterhouses generally operate under a veil of secrecy and often deny external access. This lack of transparency raises questions about whether or not they adhere to ethically sound practices in the treatment of animals.	**4.71**	0.33	0.84	2.52	0.90	0.71	0.57	1.57	0.61	1.63
Free range eggs may have a reputation for coming from chickens that live in idyllic settings. However, many hens remain living in confined, overcrowded sheds with limited access to daylight. Moreover, they often undergo distressing beak trimming measures.	**4.52**	0.53	0.00	1.04	4.14	3.58	2.78	2.89	0.84	1.70
While it might not be universally accepted, it’s a fact that battery farming practices can lead to hens being kept in smaller spaces that might seem inhumane. Regrettably, despite some regulatory efforts, these practices continue in many regions across the globe, including more developed regions like the USA and EU.	**4.43**	0.94	0.00	0.50	3.20	3.20	3.02	2.16	1.77	2.93
One perspective is that animal farming might exploit laborers, possibly contributing to their physical and mental health stress. Critics argue that the sector evades workers’ compensation and could potentially involve vulnerable individuals.	**4.39**	1.13	1.04	0.13	0.87	0.27	1.87	1.92	2.50	2.80
...										
Just as we humans value and deserve bodily autonomy, so do animals. Exploiting them for products like honey denies them their rights, erodes their freedom, and imposes our will on their natural existence. Embrace veganism to respect and uphold these rights.	**0.04**	1.23	2.83	0.57	2.20	0.59	3.69	1.22	0.83	1.53
Choose veganism, contribute to water conservation. Producing cow’s milk and beef necessitates more than triple the water used in making soya milk & vegan burgers. Given escalating global water scarcity, adopting a vegan diet is a practical and impactful solution!	**0.04**	3.83	2.09	0.72	2.22	1.32	0.65	0.93	3.34	1.44
Every single life, including those of animals, is precious and should be respected. We shouldn’t sacrifice their existence to fulfill our dietary preferences. Upholding their right to life by adopting a vegan lifestyle is a compassionate choice that respects all beings.	**0.03**	1.07	3.21	1.35	2.00	2.21	2.43	1.16	0.91	0.57

Table 6: Overview of arguments with very high and very low loadings for Topic (1).
This presents a preview of the documents that an analyst might inspect when developing the label for Topic (1). We place a strong emphasis on this manual inspection step when it comes to evaluating a topic model fit and deriving topic labels.