A Bayesian Topic Model for Human-Evaluated Interpretability

Justin Wood, Corey Arnold, Wei Wang

Departments of Computer Science, University of California, Los Angeles, CA, 90095

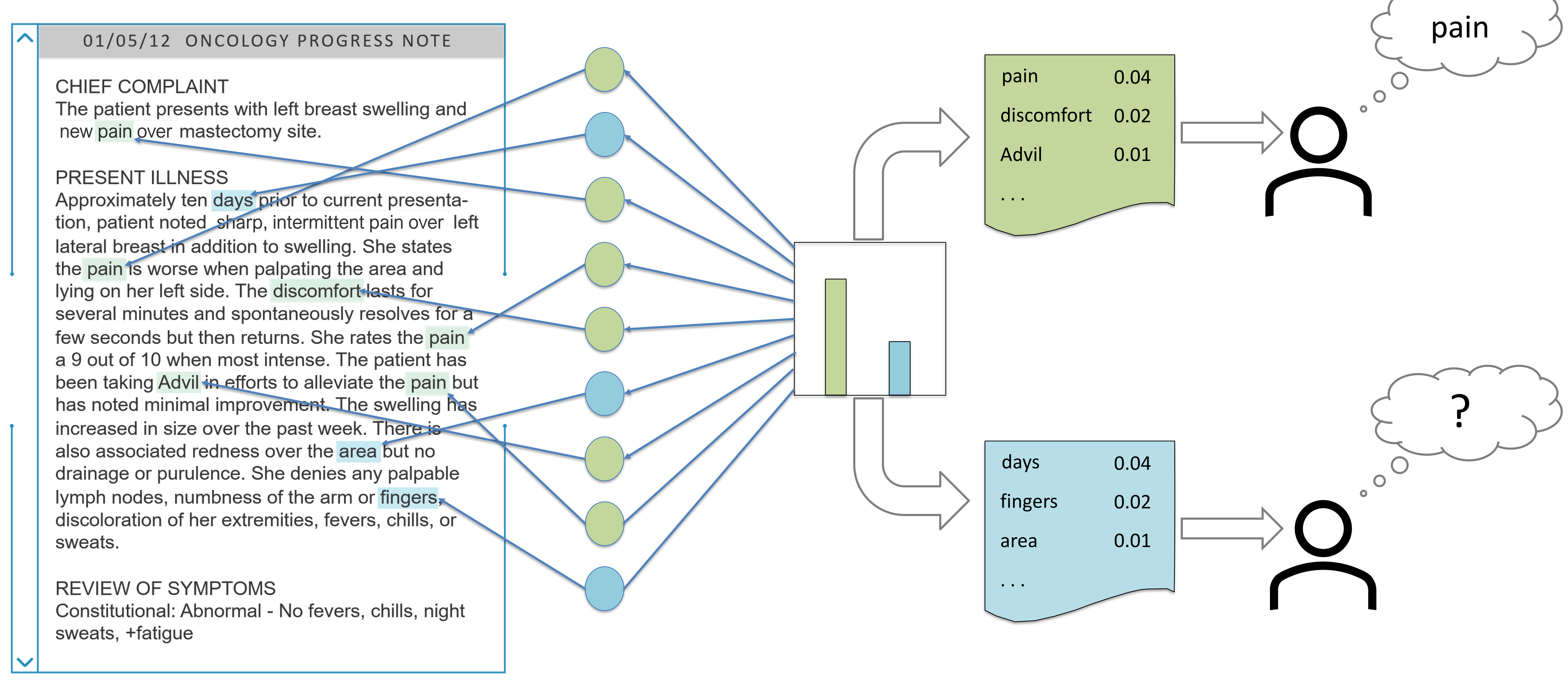
SUMMARY

One goal of topic modeling is to produce topics which are interpretable. However, existing methods often produce topics which are difficult for a human evaluator to accurately describe with a single label. This paper aims to improve interpretability in topic modeling by providing a novel, outperforming interpretable topic model. Our approach combines two previously established subdomains in topic modeling: nonparametric and weakly-supervised topic models. Given a nonparametric topic model, we can include weakly-supervised input using novel modifications to the nonparametric generative model. These modifications lay the groundwork for a compelling setting—one in which most corpora, without any previous supervised or weakly-supervised input, can discover interpretable topics. Combining nonparametric topic models with weakly-supervised topic models leads to an exciting discovery—a complete, self-contained and outperforming topic model for interpretability.

ACKNOLWEDGEMENT

This work was supported by the NIH-National Library of Medicine R21LM011937 to CA, and NIH U01HG008488, NIH R01GM115833, NIH U54GM114833, and NSF IIS-1313606 to WW.

INTRODUCTION



METHODS

- The generative model for non-parametric topic modeling (left)
- Weakly-supervised topic modeling is injected into ϕ_i
- A distribution is placed over each hyperparameter set

$$\theta_d = \sum_{i=1}^{\infty} q_{d,i} \cdot \prod_{l=1}^{i-1} (1 - q_{d,l})\, \delta_{\phi_{d,i}}$$

$$q_{d,i} \sim Beta(1, \gamma)$$

$$\phi_{d,i} \sim P$$

$$P = \sum_{i=1}^{\infty} r_i \cdot \prod_{l=1}^{i-1} (1 - r_l)\, \delta_{\phi_i}$$

$$r_i \sim Beta(1, \zeta)$$

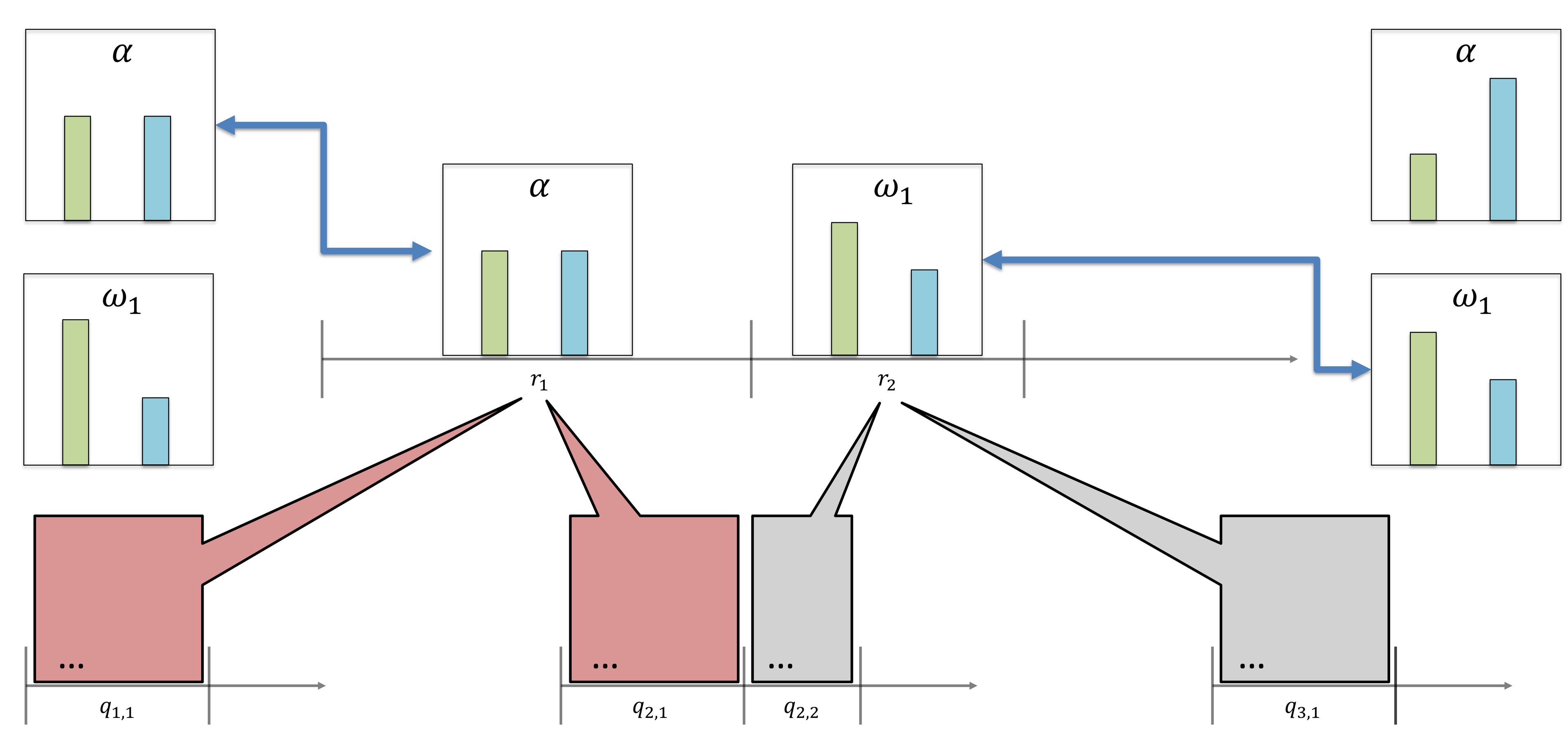
$$\phi_i \sim Dir(\alpha)$$

$$\phi_i \sim M$$

$$M = (1 - \xi) \cdot \delta_A + \frac{\xi}{B} \cdot \sum_{i=1}^{B} \delta_{\Omega_i}$$

$$A \sim \mathrm{Dir}(\alpha)$$

$$\Omega_i \sim \mathrm{Dir}(\omega_i)$$

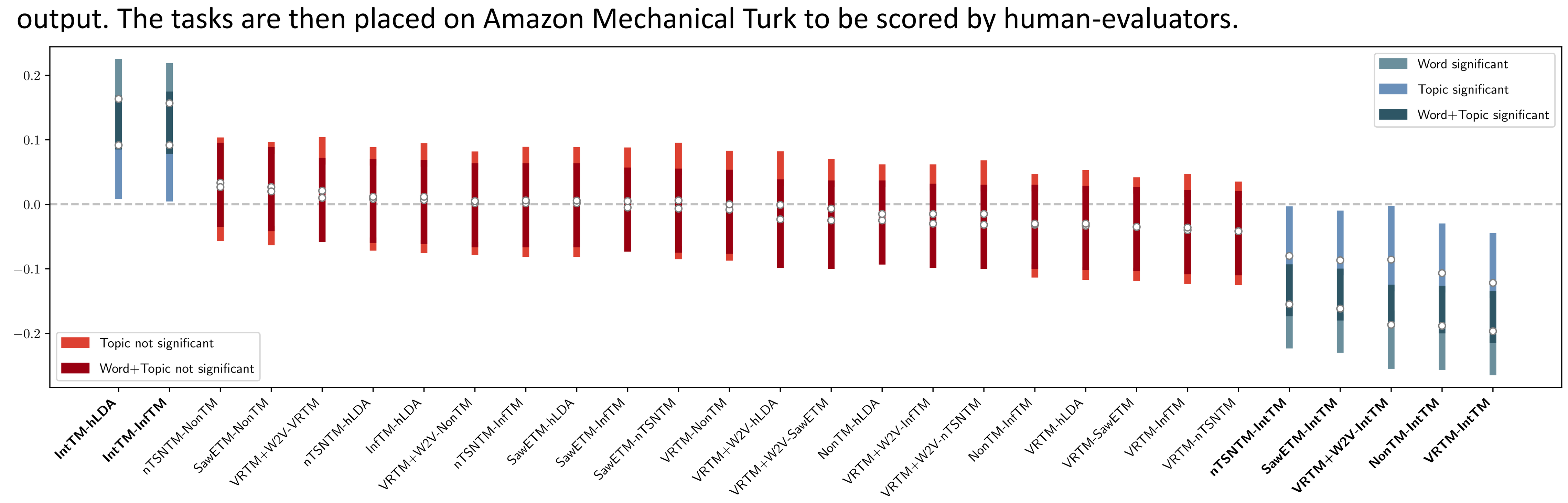


- A visual example of the generative model for three documents.
- The words of each document are portioned into bias from distributions in a stick break.
- Each stick break contains a distribution that was built from either α or ω_i hyperparameters.
- The hypermeters are drawn from a discrete distribution over the set $\{\alpha, \omega_i\}$.

RESULTS

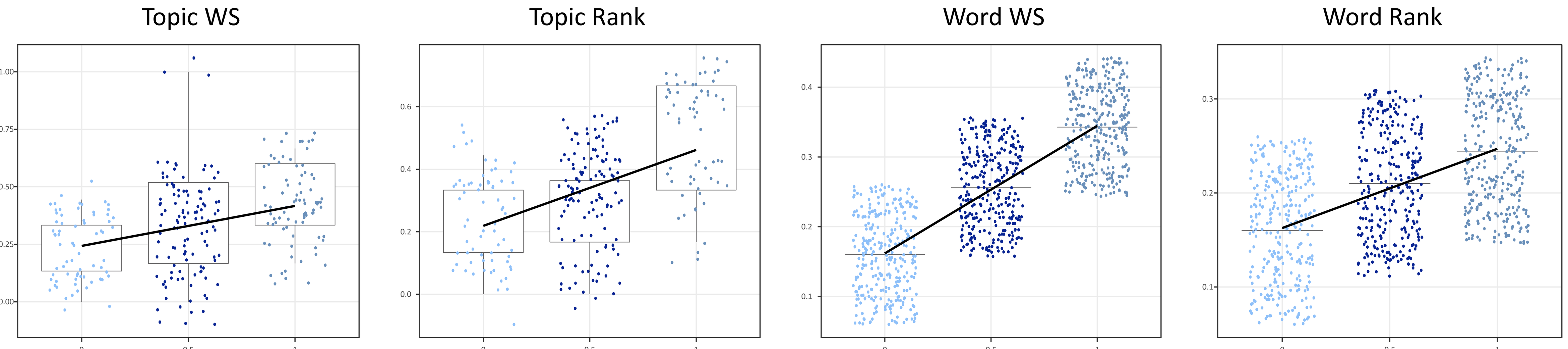
✧ Interpretability

We examine the interpretability of our method against state-of-the-art neural topic models and competing Bayesian models. For a robust set of data, we run each topic model and create word intrusion and topic intrusion tasks from the output. The tasks are then placed on Amazon Mechanical Turk to be scored by human-evaluators.



✧ Effect of ξ

To determine the effect that the ξ parameter has on interpretability we design an experiment that asks human evaluators to determine word and topic intrusions under different values of ξ. We also seek to determine the interpretability effect when discovering a knowledge source (Rank) and using a provided knowledge source (WS). All models show that as ξ increases, so does interpretability.



RESULTS

✧ Human-evaluated task analysis

We show the statistical analysis of our interpretability experiment which demonstrates for all datasets our interpretable topic modeling approach results in better interpretability than baseline methods.

Word Intrusion

	N	μ_1	MD	p-value
hLDA	600	0.15 ± 0.03	-0.02 ± 0.04	0.83
InfTM	600	0.15 ± 0.03	-0.01 ± 0.04	0.736
IntTM	600	**0.31 ± 0.04**	**0.14 ± 0.05**	**2.20e-09**
NonTM	600	0.12 ± 0.03	-0.04 ± 0.04	0.987
nTSNTM	600	0.15 ± 0.03	-0.01 ± 0.04	0.709
SawETM	600	0.15 ± 0.03	-0.02 ± 0.04	0.808
VRTM	600	0.11 ± 0.03	-0.05 ± 0.04	0.996
VRTM+W2V	600	0.12 ± 0.03	-0.04 ± 0.04	0.984

Topic Intrusion

	N	μ_1	MD	p-value
hLDA	500	0.27 ± 0.04	0.02 ± 0.05	0.236
InfTM	600	0.27 ± 0.04	0.02 ± 0.05	0.215
IntTM	600	**0.36 ± 0.04**	**0.11 ± 0.05**	**1.30e-05**
NonTM	600	0.26 ± 0.03	0.01 ± 0.05	0.421
nTSNTM	500	0.28 ± 0.04	0.03 ± 0.05	0.175
SawETM	600	0.28 ± 0.04	0.03 ± 0.05	0.107
VRTM	600	0.28 ± 0.04	0.03 ± 0.05	0.163
VRTM+W2V	600	0.24 ± 0.03	-0.01 ± 0.05	0.656

All models show that as ξ increases, so does interpretability. This demonstrates that the ξ acts as a parameter that controls the amount of interpretability into the topic model.