Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums
Supplementary Document

Minghui Qiu [*] Yanchuan Sim [†] Noah A. Smith [†] Jing Jiang [*]

1 Introduction

This document supplements the paper published at SDM 2015 with the same title.

2 Model

We propose an integrated model to jointly model user arguments, interactions and attributes for the task of stance prediction. The graphical representation of the model can be found in Figure 1.

The model is composed of four parts: *i.)* *user profiling*, which considers a regression-based latent factorization method to incorporate user attributes for profiling users; *ii.)* *user stance*, which contains a binomial matrix factorization method for modeling categorical stance data; *iii.)* *user arguments*, which incorporate textual cues in threaded posts; and *iv.)* *user interaction*, which integrates the positive and negative interaction attributes between users.

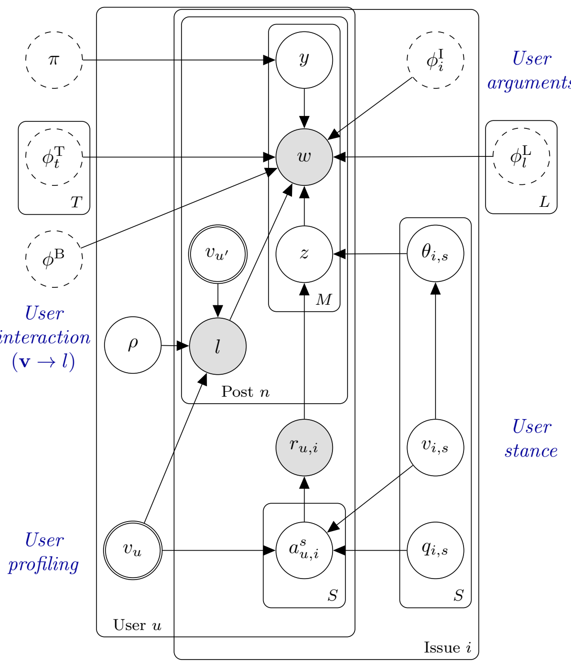

Figure 1: Plate notation for our model. The dashed variables will be collapsed out during Gibbs sampling. $\rho = \{c, b, q_u, q_{u'}\}$, representing two parameters used in user interaction modeling and two biases specific to a user and her recipient. v_u, $v_{u'}$, q_u and $q_{u'}$ are fixed by a regression based latent factorization method. Hyperparameters are omitted for clarity.

3 Inference and Learning

Our goal is to learn the hidden factor vectors and topics of the textual content to accurately model user stances and maximize the probability of generating the textual content. Hence our objective function is defined as:

$$J = -\sum_{u,i,n} \Bigg(\log p(r_{u,i} \mid \rho_{u,i,n}) + \log p(l_{u,i,n} \mid \rho_{u,i,n}) + \log p(\rho_{u,i,n} \mid \Upsilon) + \log p(w_{u,i,n} \mid l_{u,i,n}, v_{i,s}, \Omega) \Bigg),$$

where u, i, n are user, issue and argument index respectively. $\rho_{u,i,n} = \{v_i, q_i, G, g, \delta_u, \delta_{u'}, b_u, b_{u'}, c_1, c_2\}$ refers to the set of latent variables related to user u, recipient u' of the nth post of user u, and issue i, and Υ is the set of Gaussian priors for all the variables in $\rho_{u,i,n}$. Ω denotes all the Dirichlet prior hyperparameters for ϕ. The first three terms denote the probability of generating user stance and interaction given the priors Υ, where the variable in $\rho_{u,i,n}$ are to be optimized to minimize the objective function. The last term denotes the probability of observing the text conditioned on $\theta_{i,s}$ from learnt vector $v_{i,s}$, interaction $l_{u,i}$, and Dirichlet priors Ω.

Exact inference under the posterior distribution is intractable. We use Monte Carlo EM [1; 2], an inference method that alternates between collapsed Gibbs sampling [3] and gradient descent, to estimate parameters in the model. In the E-Step, we perform Gibbs sampling for variables $\{y, z\}$, fixing the values of

[*]School of Information System, Singapore Management University

[†]Language Technologies Institute, Carnegie Mellon University

ρ. In the M-step, we perform gradient descent to update latent variables in ρ, fixing the values of $\{y, z\}$.

3.1 E-Step. We present the derived Gibbs sampling update rules and assume the reader is familiar with the approach. Interested readers are referred to [3] for more details.

For the term in the mth position of argument n from user u on issue i, we jointly sample its switching variable $y_{u,i,n,m}$ and topic $z_{u,i,n,m}$, conditioned on its Markov blanket. Let $w = w_{u,i,n,m}$, $s = s_{u,i}$ and $l = l_{u,i,n}$, let d denote the set of variables $\{u, i, n, m\}$.

$$
\begin{aligned}
& p(y_d = y, z_d = z | y_{\neg_d}, w, v_{i,s}, \Omega) \\
& \propto (C^y_{\neg_d} + \gamma) \cdot \left[\frac{C^w_{y,\neg_d} + \eta^{\mathsf{B}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d} + V\eta^{\mathsf{B}}_{(\cdot)}} \frac{1}{T} \right]^{I(y=\mathsf{B})} \\
& \cdot \left[\frac{C^w_{y,\neg_d,i} + \eta^{\mathsf{I}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,i} + V\eta^{\mathsf{I}}_{(\cdot)}} \frac{1}{T} \right]^{I(y=\mathsf{I})} \\
& \cdot \left[\frac{C^w_{y,\neg_d,z} + \eta^{\mathsf{T}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,z} + V\eta^{\mathsf{T}}_{(\cdot)}} \theta^z_{i,s} \right]^{I(y=\mathsf{T})} \\
& \cdot \left[\frac{C^w_{y,\neg_d,l} + \eta^{\mathsf{L}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,l} + V\eta^{\mathsf{L}}_{(\cdot)}} \frac{1}{T} \right]^{I(y=\mathsf{L})},
\end{aligned} \tag{3.1}
$$

where $C^w_{y=\mathsf{I},\neg_d,i}$ denotes the number of times that w is sampled as a issue-specific term in issue i excluding the current term assignment; all the other Cs are defined in the same way. $I(\cdot)$ is an indicator function. $\eta_{(\cdot)}$ is a summation over all the terms η_w. Note that, when $y = \mathsf{T}$, the term is topical term, we need to sample a topic label from $\theta_{i,s}$, which is a deterministic logit transformation of $v_{i,s}$, specifically, $\theta^z_{i,s} = \frac{\exp(v^z_{i,s})}{\sum_t \exp(v^t_{i,s})}$.

3.2 M-Step. In this step, we perform gradient descent to learn latent variables in ρ by fixing the values of y and z. We then reformulate the objective function. $J_{u,i,n} =$

$$
\begin{aligned}
& -\sum_{u,i,n} \Big(\log p(r_{u,i} \mid \rho_{u,i,n}) + \log p(l_{u,i,n} \mid \rho_{u,i,n}) + \\
& \log p(\rho_{u,i,n} \mid \Upsilon) + \log p(\{z\}_{y=\mathsf{T}} | \theta_{i,s}, \Omega) \Big) \\
= & -\sum_{u,i,n} \Big(\log p(r_{u,i} \mid \rho_{u,i,n}) + \log p(l_{u,i,n} \mid \rho_{u,i,n}) + \\
& \log p(\rho_{u,i,n} \mid \Upsilon) + \sum_z N^z_{u,i,n} \log \theta^z_{i,s} \Big) + O,
\end{aligned} \tag{3.2}
$$

where $N^z_{u,i,n}$ is the number of times topic z appears in user u's arguments in issue i. We used the expected counts obtained during the E-Step as we have assigned values to all the topics and switches. $\rho_{u,i,n}$ refers a set of latent variables $\{v_i, q_i, G, g, \delta_u, \delta_{u'}, b_u, b_{u'}, c_1, c_2\}$, and Υ is the set of Gaussian priors for all the variables in $\rho_{u,i,n}$. O is a constant that does not depend on the variables in $\rho_{u,i,n}$.

By computing first derivatives of J with respect to the variables in ρ, we can then update them using gradient descent.

3.3 Fast Inference. Generally, the E-Step takes more time than the M-Step, since in the E-Step, we need to update topic and switch assignments for all the terms (bigrams). For each term, we jointly sample its corresponding topic z and switch y, and for $y = \mathsf{T}$, we sample topic from $\frac{C^w_{y,\neg_d,z} + \eta^{\mathsf{T}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,z} + V\eta^{\mathsf{T}}_{(\cdot)}} \theta^z_{i,s}$, otherwise we don't need to sample a topic label. Hence, each term takes $O(Y) + O(T)$ time to update, where Y is switch size, T is topic size.

To speed up the inference step, we consider the inference method used in SparseLDA [4]. In SparseLDA, it takes only $O(K_w + K_d)$ instead of $O(T)$ time to sample a topic for a word w in document d, K_w and K_d denote the number of topics associated with w and d respectively. However, unlike SparseLDA, $\theta_{i,s}$ is a T-dimension dense term that cannot be further decomposed. Thus we resolve to use the following treatment.

$$
\begin{aligned}
\frac{C^w_{y,\neg_d,z} + \eta^{\mathsf{T}}_w}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,z} + V\eta^{\mathsf{T}}_{(\cdot)}} \theta^z_{i,s} &= A(z) + B(z). \\
\text{where} \quad A(z) &= \frac{C^w_{y,\neg_d,z} \theta^z_{i,s}}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,z} + V\eta^{\mathsf{T}}_{(\cdot)}}. \\
B(z) &= \frac{\eta^{\mathsf{T}}_w \theta^z_{i,s}}{\sum_{w'=0}^V C^{w'}_{y,\neg_d,z} + V\eta^{\mathsf{T}}_{(\cdot)}}.
\end{aligned}
$$

Here $A(z)$ contains K_w elements, corresponding to the number of topics co-occurring with the term w, and $B(z)$ has T elements. To sample a topic, we first compute $\bar{A} = \sum_z A(z)$ and $\bar{B} = \sum_z B(z)$. We then choose $\bar{A}$ or $\bar{B}$ to proceed based on their proportions. With the data structure used in SparseLDA, and by storing encoded values of $(z, C^w_{y,\neg_d,z})$ in reverse-sorted arrays, we can calculate$\bar{A}$ and sample topic from $\bar{A}$ in $O(K_w)$ time. Note that $\bar{B}$ is the same for all the terms from issue i and stance s, that means to update $\bar{B}$ is cheap. As a result, with an initial cost for computing $\bar{B}$, it takes only $O(1)$ time to update $\bar{B}$ for a term. But to sample a topic from $\bar{B}$ takes $O(T)$ time. This means

"Religion"	"Healthcare"	"Politics"	"Same-sex marriage"	"Death Penalty"	"Bin Laden"
god exists	health care	united states	gay marriage	death penalty	bin laden
no god	american people	barack obama	gay people	morally correct	al qaeda
prove god	federal government	white house	sexual orientation	life begins	al queda
christian god	tax cuts	bin laden	same-sex marriage	intense suffering	osama bin
richard dawkins	health insurance	foreign policy	equal rights	kill people	united states
atheists believe	wall street	democratic party	straight people	gay marriage	death penalty
agnostic atheist	social security	8 years	gay couple	chemical energy	no evidence
belief system	private sector	fox news	civil rights	moral agency	true true
against god	bush administration	republican party	gay couples	past tense	middle east
god told	small businesses	president obama	opposite sex	against israel	civilian casualties
no bearing	create jobs	president bush	sex marriage	earn money	human cost
lack belief	raise taxes	john mccain	against gay	no mind	iraq war
harry potter	economic crisis	sarah palin	consenting adults	equally bad	saudi arabia
evidence against	al qaeda	george bush	homosexual parents	evil equally	foreign policy
no faith	middle east	black people	gay rights	good number	military bases
modern science	higher taxes	osama bin	civil unions	thousand horsemen	vietnam war
jesus christ	voted against	ron paul	gay man	twelve thousand	armed forces
physical evidence	financial crisis	mitt romney	federal government	thousand stalls	political gain
god exist	track record	iraq war	gay sex	electrical energy	openly admit
believe god	billion dollars	bill clinton	born gay	muslim belief	million people

Table 1: Top topic terms from ϕ_t^{T}.

Negative Interaction	Positive Interaction	"Does God Exist?"	"Renewable Energy Sources"	"Marines Urinating on Taliban"	"For or Against Gun Control?"
no evidence	good point	no god	wave energy	war crime	kill people
no god	health care	scientific method	energy technology	illegal invaded	balsthis sucks
no reason	no matter	natural sciences	offshore wave	war crimes	sucks balsthis
no matter	minimum length	no proof	total efficiency	official policy	ban guns
bin laden	totally agree	infinite religions	coal/nat gas	u.s. military	people kill
no longer	completely agree	jesus christ	sustainable energy	war logs	black market
al qaeda	god exists	no deity	onshore wave	accurate picture	gun control
no proof	years ago	natural laws	energy harnessing	invaded kuwait	banning guns
united states	looks like	religious law	good book	like torture	2nd amendment
people like	no reason	blame god	less feasilble	military operates	gun related
long time	good argument	morally perfect	viable comparison	dead bodies	gun laws
christian god	manhood academy	natural science	early 80s	u.s. armed	nuclear weapons
absolutely no	live debate	real science	due primarily	military code	related deaths
side supporting	dumb bitches	credible source	solar pv	civilian casualties	gun deaths
supporting mitt	common sense	vast majority	fuel mix	iraq invasion	guns illegal
al queda	look like	herd instinct	late 70s	like pissing	keep guns
sound like	human nature	infinite things	david mckay	little doubt	death rate
no idea	sounds like	immoral acts	cambridge david	dead human	less people
makes no	pretty good	great light	natural philosophy	u.s. army	save lives
sounds like	great point	infinite number	achieved efficiency	talk specifics	killing thousands

Table 2: Top interaction-specific terms from ϕ_l^{L}, and top issue-specific terms from ϕ_i^{I} for popular issues.

we only have a speed gain when we choose $\bar{A}$ to proceed.

In our experiment, we find $\frac{\bar{A}}{\bar{A}+\bar{B}} > 0.8$, which means, in most cases, we need only $O(K_w)$ to sample a topic. In all, to jointly sample a switch and a topic, for more than 80% of cases, we only need $O(Y) + O(K_w)$ time. We find this to be around three times as fast as the original method.

4 Qualitative Analysis

We present six popular topics based on $\theta_{i,s}$ across issues in Table 1. Topic labels are manually assigned.

We find "existence of God" and "same-sex marriage" are popular topics in our data. All these topics are readily identified based on their top topical words. Topical terms are similar to high-level issues of the existence of God, healthcare, and same-sex marriage. Since these topics are in the same space as the hidden factors in matrix factorization, they can serve as interpretable labels for the corresponding dimensions in matrix factorization.

We present top interaction words for both positive and negative interactions from ϕ_l^{L}; see Table 2. These words are automatically learned by our model, making use of interaction polarity of user arguments. The results show these interaction words are quite intuitive. We also present top issue-specific terms from ϕ_i^{I} for popular issues in Table 2. These issues are hand picked by the authors from popular issues to cover a wider variety of issues as some issues are conceptually similar. Labels are assigned manually. Overall, these issue-specific terms the model discovers are easy to interpret. For example, on the issue "Does God Exist?", top terms are “no God”, “scientific method,” and “no proof.” This shows that some users talk about the issue from a “science” perspective.

In all, these types of words discovered by our model provide a human-interpretable representation of topics, issues, and interactions present in the debate data, all through an unsupervised manner. Subjectively, we find these words to be meaningful.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA), by an A*STAR fellowship to Y. Sim, and by NSF CAREER IIS-IIS-1054319. We thank the reviewers for their valuable comments.

References

[1] H. M. Wallach, “Topic Modeling: Beyond Bag-of-words,” in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML ’06, 2006, pp. 977–984. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143967

[2] G. C. Wei and M. A. Tanner, “A monte carlo implementation of the em algorithm and the poor man’s data augmentation algorithms,” *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[3] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[4] L. Yao, D. M. Mimno, and A. McCallum, “Efficient methods for topic model inference on streaming document collections,” in *KDD*, 2009, pp. 937–946.