Cross-structural Factor-topic Model: Document Analysis with Sophisticated Covariates

Chien Lu CHIEN.LU@TUNI.FI
Jaakko Peltonen JAAKKO.PELTONEN@TUNI.FI
Timo Nummenmaa TIMO.NUMMENMAA@TUNI.FI
Jyrki Nummenmaa JYRKI.NUMMENMAA@TUNI.FI
Kalervo Järvelin KALERVO.JARVELIN@TUNI.FI
Faculty of Information Technology and Communication Sciences, Tampere University, Finland

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Modern text data is increasingly gathered in situations where it is paired with a high-dimensional collection of covariates: then both the text, the covariates, and their relationships are of interest to analyze. Despite the growing amount of such data, current topic models are unable to take into account large amounts of covariates successfully: they fail to model structure among covariates and distort findings of both text and covariates. This paper presents a solution: a novel factor-topic model that enables researchers to analyze latent structure in both text and sophisticated document-level covariates collectively. The key innovation is that besides learning the underlying topical structure, the model also learns the underlying factorial structure from the covariates and the interactions between the two structures. A set of tailored variational inference algorithms for efficient computation are provided. Experiments on three different datasets show the model outperforms comparable topic models in the ability to predict held-out document content. Two case studies focusing on Finnish parliamentary election candidates and game players on Steam demonstrate the model discovers semantically meaningful topics, factors, and their interactions. The model both outperforms state-of-the-art models in predictive accuracy and offers new factor-topic insights beyond other topic models.

Keywords: Probabilistic Modeling, Natural Language Processing, Topic Modeling

1. Introduction

In multiple domains, textual data is paired with accompanying numerical covariates. Examples include questionnaires where free-choice text fields are paired with a set of numerical (continuous or discrete-choice) answers to different questions (often on a Likert scale); political discussion where statements of public figures are paired with their voting record; product reviews where review text is paired with covariates either describing the reviewer along different attributes or scoring the product by multiple criteria; and many others. Such datasets contain structure both within the text content, often described as underlying topics; structure within the set of covariates; and structure linking the text and the covariates. The structures of the covariates and how they interplay with the text content play crucial roles and offer valuable insights. However, current generative probabilistic models do not work well in this setting: the models have overemphasized the text structure only with little attention to modeling structure in covariates. Available topic models either ignore covariates or simplistically model only direct influence of individual covariates, which yields poor overfitted performance when covariates are high-dimensional. Besides poor predictive performance, such models are also unable to provide insight into the structure in covariates and its relationship to topics. In this paper, we present a solution.

We introduce the *Cross-structural Factor Topic Model* (CFTM), a novel generative probabilistic model which can model the structure of both the text and its high-dimensional numerical covariates. We describe the generative structure of the model, and a parallelizable inference algorithm based on variational approximation. We show in experiments on several data sets that the method yields good performance in modeling held-out document content and yields meaningful insights about structures of covariates and text content.

The rest of the paper is structured as follows. Section 2 discusses related work. Sections 3 and 4 present the proposed method: Section 3 describes the generative model and Section 4 presents the inference approach. Empirical analysis including quantitative and qualitative evaluation is presented in Section 5. Conclusions are given in Section 6.

2. Related Work

For modeling text content of documents alone, topic models of multiple kinds have been proposed. Among them, Latent Dirichlet Allocation (LDA, Blei et al. 2003) is the classical method, which models document content as a bag of words whose word counts arise out of a mixture of latent topics, each of which has its own multinomial word distribution. Nonparametric topic models have been proposed, including Hierarchical Dirichlet Processes (Teh et al. 2006) which aim to learn the number of topics from data. Nonparametric modeling is a direction of future extension for our work.

The Entity topic model (ETM, Kim et al. 2012) models the influence of entities on word content by generating entity mentions from topics and then words from entity-describing word distributions. However, entity mentions are part of text content, no covariates are considered. An Author Topic Model (Rosen-Zvi et al. 2004) was introduced to model relationships between authors, documents, topics and words; however, such models only consider author identity and do not consider author attributes as covariates.

Supervised LDA (sLDA, Mcauliffe and Blei 2008) was developed to model labeled documents. An extended approach called Dirichlet-multinomial regression (DMR, Mimno and McCallum 2008) introduces regression model on topic mixture over covariates. The Sparse additive generative text model (SAGE, Eisenstein et al. 2011) allows topic content to fluctuate by the covariates. However, none of these models allows covariates to affect both topic prevalence and content; our proposed model addresses this.

MetaLDA (Zhao et al. 2017) and Structural topic model (STM, Roberts et al. 2016) both allow covariates to influence topic prevalence and content. However, MetaLDA does not provide a generative model of covariates, and only takes into account simple binary label covariates in modeling topics. STM was recently developed based on SAGE. It is an integrated solution to model covariates (both categorical and continuous) and text. However, the covariates that affect topic content have a limitation, as they allow discrete values only. Thus STM cannot handle sophisticated covariates. We will show in our experiments that STM performance is drastically worsened when the dimension of covariates is high.

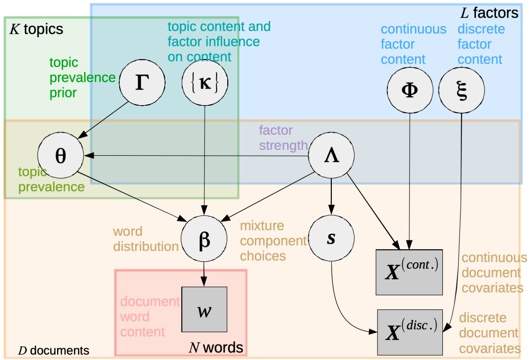


Figure 1: Proposed model. Plates denote topics, factors, documents, and words. The top row of latent variables (circles) describe topics, factors, and their interaction; the 2nd row ($\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$) describe prevalence of topics and factors in a document; the bottom part describes document content (gray boxes). Factor-loading prior parameter $\boldsymbol{\alpha}$ and noise variances σ, σ_γ, Σ_η, Σ_τ, σ_ϕ, σ_β omitted for clarity.

Distributed Multinomial Regression (Taddy 2015) is an alternative approach which directly models the relationship between the word occurrences and the covariates, but it does not model any structure among the covariates.

Another group of works focuses on combining neural models and topic modeling (Srivastava and Sutton 2017; Card et al. 2018; Gui et al. 2019; Wang and Yang 2020). Among them, SCHOLAR (Card et al. 2018) can be seen as a similar work to STM which incorporates covariates with a variational autoencoder (Kingma and Welling 2014). However, despite their flexibility these models do not generate structure within covariates, they only use covariate values as additional inputs in document content generation. Besides topic models, Non-negative Matrix Factorization (NMF) models are also used for text analysis. In general, a Poisson likelihood is employed to model the observed text whereas multinomial distributed likelihood are typically used by topic models. Many NMF-style works have been proposed (Hu et al. 2016; Acharya et al. 2015; da Silva et al. 2017; Zhao et al. 2018); among such works the most relevant is CTPF (Gopalan et al. 2014) which incorporates a multivariate user-rating matrix into account as covariates, and we will compare to it.

3. Proposed Method

We model a collection of documents indexed by $d \in \{1, \dots, D\}$ with text content and covariates jointly by a probabilistic model. Word content is distributed over a vocabulary of V unique words indexed by $v \in \{1, \dots, V\}$ and covariates are indexed by $p \in \{1, \dots, P\}$. Word content arises from K underlying latent topics indexed by $k \in \{1, \dots, K\}$, and covariates from $L < P$ underlying latent factors indexed by $l \in \{1, \dots, L\}$. Topics and factors interact: the strength of the latent factors affects the prevalence of topics and content (word distribution) in each topic. Figure 1 shows the plate model representation of the overall model. We next describe the generative model of the covariates and text content.

3.1. Document-level Latent Variables

Factor Loadings. Each document d is attached with a loading vector over L factors,

$$\boldsymbol{\Lambda}_d = [\lambda_{d,1}, \dots \lambda_{d,L}]^\top \sim Dir(\boldsymbol{\alpha}) \ . \tag{1}$$

Interaction Coefficients. For each topic $k \in \{1, ...K-1\}$ a L-length coefficient vector is generated as

$$\mathbf{\Gamma}_k \sim \boldsymbol{N}(0, \sigma_\gamma^2 \mathbf{I}_L) \tag{2}$$

to model the relation between the factors and the prevalence of the topic. Note that coefficient vectors for the first $K-1$ topics suffice since topic prevalences sum to 1.

Topic Prevalence. For each document d, the topic prevalence vector $\boldsymbol{\theta}_d = [\theta_{d,1}, \ldots, \theta_{d,K}]$ is generated as $\boldsymbol{\theta}_d = softmax(\boldsymbol{\eta}_d)$ where the auxiliary variables are generated as

$$\boldsymbol{\eta}_{d,1:(K-1)} \sim \boldsymbol{N}(\mathbf{\Gamma}^\top \mathbf{\Lambda}_d, \mathbf{\Sigma}_\eta) \tag{3}$$

and the $\eta_{d,K}$ is fixed to 0.

3.2. Structure of the Covariates

We assume the text content in each document is paired with a set of covariates. Different covariates in the set may require different model types to properly model their structure. Let $x_d^{(p)}$ denote the p:th covariate of document d. We model covariates with two kinds of structure: mixture model and factorization model. The former is suitable especially for discrete covariates, such as multiple-choice values, and the latter for continuous covariates. In both cases the covariate generation depends on a vector $\mathbf{\Lambda}_d$ of L latent parameters. We describe both types of covariate generation next.

Mixture Model. In this structure the p:th covariate is generated from a mixture. The mixture component membership label of the p:th covariate in document d is first generated from a categorical distribution

$$s_d^{(p)} \sim Cat(\mathbf{\Lambda}_d) \tag{4}$$

and the covariate $x_d^{(p)}$ corresponding to the label $s_d^{(p)}$ is then generated as $x_d^{(p)} \sim p(x_d^{(p)} | \boldsymbol{\xi}^{(p)}_{s_d^{(p)}})$ with parameter $\boldsymbol{\xi}^{(p)}_{s_d^{(p)}}$. We model the distribution in each mixture component as a Poisson distribution for covariates that are a count of rare events and as a multinomial distribution for categorical covariates.

Factorization Model. The covariate is directly generated from an exponential family distribution as

$$x_d^{(p)} | \mathbf{\Lambda}_d, \boldsymbol{\phi}^{(\boldsymbol{p})} \sim \mathbf{ExpFam}\left(\zeta\left(\mathbf{\Lambda}_d, \boldsymbol{\phi}^{(\boldsymbol{p})}\right), T\left(x_d^{(p)}\right)\right) \tag{5}$$

in which the natural parameter ζ is a weighted average of factor-wise parameters $\phi_l^{(p)} \sim N(0, \sigma_\phi^2)$ weighted by the document-specific factor loadings $\mathbf{\Lambda}_d$, so that

$$\zeta\left(\mathbf{\Lambda}_d, \boldsymbol{\phi}^{(\boldsymbol{p})}\right) = g^{(p)}\left(\sum_{l=1}^{L} \phi_l^{(p)} \lambda_{d,l}\right) \tag{6}$$

where g is the link function of the exponential-family model. For example, if a Gaussian with a known variance σ^2 is taken as the distribution, we have $x_d^{(p)} \sim N(\sum_{l=1}^L \phi_l^{(p)}, \sigma^2)$.

3.3. Structure of Text

Topic Content. We model the word generation process with a SAGE-inspired structure in which each document is attached with a latent vector $\boldsymbol{\beta}_d$ of length V. The v:th element of the latent vector is generated as

$$\beta_{d,v} = \kappa_v^{(w)} + \sum_k \theta_{d,k} \kappa_{v,k}^{(t)} + \sum_l \lambda_{d,l} \kappa_{v,l}^{(f)} + \sum_k \sum_l \theta_{d,k} \lambda_{d,l} \kappa_{v,l,k}^{(i)} + \epsilon_\beta \tag{7}$$

where $\epsilon_\beta \sim N(0, \sigma_\beta^2)$. The $\boldsymbol{\kappa}^{(w)}$ is a vector of length V controlling the overall word prevalence. The overall topic content $\boldsymbol{\kappa}^{(t)}$ is a $V \times K$ matrix, factor influence $\boldsymbol{\kappa}^{(f)}$ is a $V \times L$ matrix, and $\boldsymbol{\kappa}^{(i)}$ is a $V \times L \times K$ array which governs factor-topic interactions on the topic content level, that is, the value of $\kappa_{v,l,k}^{(i)}$ reflects the strength of how much the factor l alters the word probability of v in topic k.

To generate the observed words in the document, for the nth word in document d, the word $w_n^{(d)}$ is sampled from a multinomial distribution

$$w_n^{(d)} \sim MN\left(softmax\left(\boldsymbol{\beta}_d\right)\right) \ . \tag{8}$$

This model design allows the latent factors and topics to interact on both topic prevalence and topic content levels.

4. Variational Inference

We carry out variational inference for the model; variational inference aims to approximate the posterior distribution of model parameters by a factorized distribution q whose components are from known families. Unlike point estimate methods such as maximum a posteriori (MAP), variational inference is able to model a full distribution for parameters based on observations. The parameters of the factorized distribution are optimized by minimizing Kullback-Leibler divergence from the factorized distribution to the true parameter posterior, which becomes equivalent to maximizing the Evidence Lower Bound (ELBO). Iterative optimization optimizes each component distribution given the others; depending on the form of the observation probability and parameter priors, the optimum is obtained analytically for some parameters and by optimization techniques for others. In particular it turns out a crucial part, inference of the topic content, is nontrivial to do computationally efficiently–naive inference is slow; we solve this by a distributed multinomial regression approach with a kernel trick.

Topic Prevalence. Using Laplace Variational Inference (Braun and McAuliffe 2010; Wang and Blei 2013), the variational distribution of η_d is obtained as

$$q(\boldsymbol{\eta}_d) \approx \mathbf{N}(\hat{\boldsymbol{\eta}}_d, -\nabla^2 \mathcal{L}(\hat{\boldsymbol{\eta}}_d)^{-1}) \tag{9}$$

where the mean $\hat{\boldsymbol{\eta}}_d$ is the MAP solution, i.e., optimum of

$$\mathcal{L}(\boldsymbol{\eta}_d) \propto -\frac{1}{2} \boldsymbol{\eta}_d^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\eta}_d + \boldsymbol{\eta}_d^\top \boldsymbol{\Sigma}_\eta^{-1} \boldsymbol{\Gamma}^\top \boldsymbol{\Lambda}_d + \sum_v c_{d,v} \log \sum_k u_{d,k,v} \exp(\eta_{d,k}) - W_d \log \sum_k \exp(\eta_{d,k}) \tag{10}$$

and $u_{d,k,v}$ is an auxiliary variable

$$u_{d,k,v} = \frac{\exp\left(\kappa_v^{(w)} + \kappa_{v,k}^{(t)} + E[\mathbf{\Lambda}_d]^\top \boldsymbol{\kappa}_v^{(f)} + E[\mathbf{\Lambda}_d]^\top \boldsymbol{\kappa}_{v,k}^{(i)}\right)}{\sum_v \exp\left(\kappa_v^{(w)} + \kappa_{v,k}^{(t)} + E[\mathbf{\Lambda}_d]^\top \boldsymbol{\kappa}_v^{(f)} + E[\mathbf{\Lambda}_d]^\top \boldsymbol{\kappa}_{v,k}^{(i)}\right)} \,. \tag{11}$$

The $\nabla^2 \mathcal{L}(\hat{\boldsymbol{\eta}}_d)$ is a Hessian matrix of $\mathcal{L}(\boldsymbol{\eta}_d)$ at $\hat{\boldsymbol{\eta}}_d$. We find $\hat{\boldsymbol{\eta}}_d$ with the "L-BFGS" optimizer.

Mixture Covariates Model. We infer the component parameters for each membership label $l = 1, \ldots, L$. We present separately the cases for count data and for categorical data. We also infer the distribution of the membership labels.

When the p:th covariate is Count Data (with a Poisson model), we consider the Poisson parameter $\xi_l^{(p)}$ for each membership label $l = 1, \ldots, L$. The optimum of the variational distribution has an analytical form and becomes

$$q(\xi_l^{(p)}) = Gamma(a^{(p)} + E[\mathbf{\Lambda}_l]^\top \boldsymbol{X}_d^{(p)}, b^{(p)} + \sum_d (E[\mathbf{\Lambda}_l])) \tag{12}$$

When the p:th covariate is Categorical Data (with a Multinomial model), we consider for each membership label $l = 1, \ldots, L$ the multinomial parameter $\xi_l^{(p)}$, i.e., the vector of category probabilities. The solution has an analytical form $q(\boldsymbol{\xi}_l^{(p)}) = Dir(\boldsymbol{a}^{(p)} + \sum_d E[s_d^{(p)} = l]\boldsymbol{X}_d^{(p)})$ where $s_d^{(p)}$ is the current membership label for covariate p of document d.

Membership Labels. The variational distribution of the mixture membership label $s_d^{(p)}$ for covariate p of document d is multinomial and the optimum has an analytical form $\log q(s_d^{(p)} = l) \propto \log E[\lambda_{d,l}] + \sum_p \log E[p(x_d^{(p)} | \boldsymbol{\xi}_l^{(p)})]$.

Factorization Covariates Model. Taking advantage of conjugacy, the variational posterior of the factor-wise natural parameters $\boldsymbol{\phi}_p^{(A)}$ is $q(\boldsymbol{\phi}^{(p)}) = \mathbf{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\phi}^{(p)}}, \hat{\mathbf{\Sigma}}_{\boldsymbol{\phi}^{(p)}})$ where the covariance matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{\phi}}$ and the mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\phi}^{(p)}}$ are

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{\phi}} = \left({\Sigma_\sigma}^{-1} + \frac{1}{\sigma_\phi^2} \sum_d E_q\left[\mathbf{\Lambda}_d \mathbf{\Lambda}_d^\top\right]\right)^{-1} \;,\;\; \hat{\boldsymbol{\mu}}_{\boldsymbol{\phi}^{(p)}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{\phi}} \frac{\sum_d E_q\left[\mathbf{\Lambda}_d\right] x_{d,p}^{(A)}}{{\sigma_\phi}^2} \tag{13}$$

where $\Sigma_\sigma = Diag(\sigma^2, \ldots, \sigma^2)$.

Factor Loading. The variational posterior of $\mathbf{\Lambda}_d$ is a Dirichlet distribution parameterized by pseudocount vector $\boldsymbol{\alpha}_{\mathbf{\Lambda}_d}$. To derive the variational posterior (i.e. find the $\boldsymbol{\alpha}_{\mathbf{\Lambda}_d}$), we set up an objective function proportional to the ELBO; the objective function is

$$E_q[(\mathbf{a} + s_d - 1)^\top \log \mathbf{\Lambda}_d + \frac{1}{2}(2\mathbf{b}^\top \mathbf{\Lambda}_d - {\mathbf{\Lambda}_d}^\top \mathbf{A} \mathbf{\Lambda}_d)] - H(\mathbf{\Lambda}_d) \tag{14}$$

where we have

$$\mathbf{b} = E_q\left[{\boldsymbol{\eta}_d}^\top \mathbf{\Sigma}_\eta^{-1} \Gamma^\top + \sum_k \theta_{d,k} ({\mathbf{w}_d}^\top \boldsymbol{\kappa}_k^{(i)})\right] + E_q\left[{\mathbf{X}_d^{(A)}}^\top \mathbf{\Sigma}_{(A)}^{-1} \boldsymbol{\phi}^\top\right] \;\text{ and} \tag{15}$$

$$\mathbf{A} = E_q\left[\Gamma \mathbf{\Sigma}_\eta^{-1} \Gamma^\top + \boldsymbol{\phi} \mathbf{\Sigma}_{(A)}^{-1} \boldsymbol{\phi}^\top\right] \;, \tag{16}$$

and $H(\mathbf{\Lambda}_d)$ is the entropy of the Dirichlet distribution. Using a Taylor approximation to simplify the computation, denote

$$f(\mathbf{\Lambda}_d) = (\mathbf{a} + s_d - 1)^\top \log \mathbf{\Lambda}_d + \frac{1}{2}(2\mathbf{b}\mathbf{\Lambda}_d - \mathbf{\Lambda}_d \mathbf{A} \mathbf{\Lambda}_d) \ . \tag{17}$$

Then the objective function becomes

$$E_q[f(\mathbf{\Lambda}_d)] - H(\mathbf{\Lambda}_d) \approx f(\hat{\mathbf{\Lambda}}_d) + \frac{1}{2} tr(\nabla^2 f(\hat{\mathbf{\Lambda}}_d) Cov_q(\hat{\mathbf{\Lambda}}_d)) - H(\mathbf{\Lambda}_d) \tag{18}$$

where $\hat{\mathbf{\Lambda}}_d = \frac{\boldsymbol{\alpha}_{\mathbf{\Lambda}_d}}{\sum_l \alpha_{\mathbf{\Lambda}_d l}}$ is the mean of $\mathbf{\Lambda}_d$ and $\nabla^2 f(\hat{\mathbf{\Lambda}}_d) = Diag(\frac{(1-\mathbf{a}-s_d)}{\hat{\mathbf{\Lambda}}_d^{-2}}) - \mathbf{A}$ is the Hessian matrix. The L-BFGS optimizer is used to optimize (18) with respect to $\boldsymbol{\alpha}_{\mathbf{\Lambda}_d}$.

Topic-Factor Interaction. For $k \in \{1, \ldots, K-1\}$, we derive the variational posterior of the interaction coefficient vector $\mathbf{\Gamma}_k$ which defines the effect of factor loadings on the topic prevalence. As the prior of the coefficients and the likelihood are both normal, taking the advantage of the conjugacy we have the analytical posterior $q(\mathbf{\Gamma}_k) = \mathbf{N}(\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}_k}, \hat{\mathbf{\Sigma}}_{\mathbf{\Gamma}_k})$ where the covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{\Gamma}_{\boldsymbol{k}}}$ and mean $\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}_k}$ are

$$\hat{\mathbf{\Sigma}}_{\mathbf{\Gamma}_{\boldsymbol{k}}} = \left({\mathbf{\Sigma}_\eta}^{-1} \sum_d E_q \left[\mathbf{\Lambda}_d \mathbf{\Lambda}_d^\top \right] + {\Sigma_{\boldsymbol{\gamma}}}^{-1} \right)^{-1} \quad \text{and} \tag{19}$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}_k} = \left(E_q \left[\mathbf{\Lambda} \right] E_q \left[\mathbf{\Lambda} \right]^\top + {\mathbf{\Sigma}_\gamma}^{-1} \right)^{-1} E_q \left[\mathbf{\Lambda}_d \right]^\top E_q \left[\boldsymbol{\eta} \right] \ . \tag{20}$$

Topic Content. The complexity of topic content $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ leads to challenges of efficiency and accuracy. A naive derivation of a variational posterior would yield computationally inefficient and non-scalable equations involving inverses of huge matrices and other expensive computations. Instead, we develop a set of tailored inference algorithms based on distributed multinomial regression (Taddy 2015) and a kernel trick (Agrawal et al. 2019), as described next in Proposition 1, Proposition 2, and Theorem 1. The propositions and theorem show how the text structure inference algorithm can be implemented with parallel computation (each vocabulary term can be run in parallel) to enhance efficiency.

Proposition 1 (Distributed Multinomial Regression) *The inference of the $\boldsymbol{\beta}$ in (8) can be performed through conducting inference on independent Poisson models for each word, where each word v has the following generative model:*

$$\boldsymbol{\kappa}_v \sim N(0, \Sigma_\tau) \ , \quad \beta_{d,v} = {\boldsymbol{\kappa}_v}^\top \mathbf{\Psi}_d + \epsilon_\beta \ , \quad w_{d,v} \sim Poisson \left(e^{\beta_{d,v} + \kappa_v^{(w)} + \log m_d} \right) \tag{21}$$

where $\epsilon_\beta \sim N(0, \sigma_\beta^2)$ is the random noise. The notation $\boldsymbol{\kappa}_v = [\kappa_{v,1}^{(t)}, \ldots \kappa_{v,K}^{(t)}, \kappa_{v,1,1}^{(i)}, \ldots, \kappa_{v,L,K}^{(i)}]$ joins together the topic and topic-factor interaction coefficients affecting word v. Correspondingly, $\mathbf{\Psi}_d$ is a mapping function that represents the combined influence terms of both topic prevalences and factor loadings and is defined as

$$\mathbf{\Psi}_d \triangleq \mathbf{\Psi} \left(\boldsymbol{\theta}_d, \mathbf{\Lambda}_d \right) := \left[\theta_{d,1}, \ldots, \theta_{d,K}, \Lambda_{d,1}, \ldots, \Lambda_{d,L}, \Lambda_{d,1}\theta_{d,1}, \ldots, \Lambda_{d,L}\theta_{d,K} \right] \tag{22}$$

where the first K elements in the vector are the topic prevalence, the positions are corresponding to the $\boldsymbol{\kappa}_v^{(t)}$ and the rest are topic-factor interactions, their position are corresponding to $\text{vec}(\boldsymbol{\kappa}_v^{(i)})$. In the following, for simplicity we abuse the notation, using $\mathbf{z}_d$ to denote the collection of $\{\boldsymbol{\theta}_d, \boldsymbol{\Lambda}_d\}$. The logarithm of the document length $\log m_d$ is plugged in to serve as the fixed effect (exposure) in the Poisson model.

This framework was proposed by Taddy (2015) to transform the multinomial logistic model into a collection of independent Poisson models to circumvent expensive computations resulting from softmax transformation. Moreover, since the Poisson models are independent of each other, one can easily introduce parallel computation techniques (e.g. map-reduce, Dean and Ghemawat 2008) to speed up the computation. STM also adopted this approach; we apply the framework in a novel factor-topic modeling context. By adapting the framework, the likelihood model (8) is factorized into V independent term-wise Poisson models with a plug-in fixed effect (exposure) shared across terms.

Proposition 2 (Gaussian Process Reparametrization) *The generative model in Proposition 1 can be reparameterized as*

$$g_v \sim GP(0, k_\tau)\ , \quad \beta_{d,v} = g_v(\boldsymbol{z}_d)\ , \quad w_{d,v} \sim Poisson\left(e^{\beta_{d,v} + \kappa_v^{(w)} + \log m_d}\right) \tag{23}$$

where the equation of $\beta_{d,v}$ in (21) is seen as a function with inputs θ_d and λ_d and is then presented as the equation of $\beta_{d,v}$ in (23), and with a Gaussian process prior .

Combining the propositions 1 and 2, taking the weight-space view (see Rasmussen and Williams 2006), the prior of $\boldsymbol{\beta_v}$ becomes

$$\boldsymbol{\beta_v} \sim N(0, K_\tau + \sigma_\beta^2 \mathbf{I}_D) \tag{24}$$

where K_τ is a $D \times D$ matrix with $k_\tau\left(\mathbf{z}_d, \mathbf{z}_{d'}\right) \triangleq \boldsymbol{\Psi}_d^\top \Sigma_\tau \boldsymbol{\Psi}_{d'}$. We first find the point estimate $\boldsymbol{\beta_v}^* \triangleq \underset{\boldsymbol{\beta}}{\text{argmax}} f(\boldsymbol{\beta})$ with the objective function

$$f(\boldsymbol{\beta}) = \sum_d \log p(w_{d,v} | \beta_d, m_d) - \log \boldsymbol{\beta}^\top R_\tau \boldsymbol{\beta} \tag{25}$$

where $R_\tau = \left(K_\tau + \sigma_\beta^2 \mathbf{I}_D\right)^{-1}$. We use "L-BFGS" to get the fixed $\kappa_v^{(w)}$ value by $\kappa_v^{(w)} = \frac{1}{D}\sum_d \beta_{d,v}^*$, the margin $\boldsymbol{\beta}^{(m)} = \left[\boldsymbol{\beta}_1^* - \kappa_1^{(w)} \ldots, \boldsymbol{\beta}_V^* - \kappa_V^{(w)}\right]$ is then the posterior mode of $\boldsymbol{\beta}$. These equations infer the posterior of the word distribution parameters $\boldsymbol{\beta}$ which are combinations of topic and factor influences. Next we infer the influence variables $\boldsymbol{\kappa}_v^{(t)}$ of topics to each word and $\boldsymbol{\kappa}_{v,,k}^{(i)}$ of factors to each word and topic, with the following theorem.

Theorem 1 (Kappa Recovery) *Let $\boldsymbol{\theta}_k$ be a k-th unit vector with length K, $\boldsymbol{\Lambda}_l$ be a l-th unit vector with length L, $\boldsymbol{z}_k$ denote the collection $\{\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_0\}$, $\boldsymbol{z}_l$ denote the collection $\{\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_l\}$, and $\boldsymbol{z}_{k,l}$ denote the collection $\{\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_l\}$. Then the posterior of $\kappa_{v,k}^{(t)}$ is $N(\mu_{\kappa_{v,k}^{(t)}}, \sigma^2_{\kappa_{v,k}^{(t)}})$, and the posterior of $\kappa_{v,l}^{(f)}$ is $N(\mu_{\kappa_{v,l}^{(f)}}, \sigma^2_{\kappa_{v,l}^{(f)}})$ where*

$$\mu_{\kappa_{v,k}^{(t)}} = K_\tau\left(\boldsymbol{z}_k, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau \boldsymbol{\beta}_v^{(m)}\ , \quad \mu_{\kappa_{v,l}^{(f)}} = K_\tau\left(\boldsymbol{z}_l, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau \boldsymbol{\beta}_v^{(m)}\ , \tag{26}$$

$$\sigma^2_{\kappa^{(t)}_{v,k}} = k_\tau\left(\boldsymbol{z}_k, \boldsymbol{z}_k\right) + K_\tau\left(\boldsymbol{z}_k, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\boldsymbol{z}_d\}_{d=1}^D, \boldsymbol{z}_k\right) \ , \tag{27}$$

$$\sigma^2_{\kappa^{(f)}_{v,l}} = k_\tau\left(\boldsymbol{z}_l, \boldsymbol{z}_l\right) + K_\tau\left(\boldsymbol{z}_l, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\boldsymbol{z}_d\}_{d=1}^D, \boldsymbol{z}_l\right) \tag{28}$$

and K_τ and R_τ are defined in Proposition 2. The posterior of $\kappa^{(i)}_{v,k,l}$ is $N(\mu_{\kappa^{(i)}_{v,k,l}}, \sigma_{\kappa^{(i)}_{v,k,l}})$ with

$$\mu_{\kappa^{(i)}_{v,k,l}} = [-1,-1,1]\, K_\tau\left(\{\boldsymbol{z}_k, \boldsymbol{z}_l, \boldsymbol{z}_{k,l}\}, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau \boldsymbol{\beta}_v^{(m)} \tag{29}$$

where $[-1,1,1]$ is simply the 1×3 matrix with elements 1 and -1, and the variance is

$$\begin{aligned}\sigma^2_{\kappa^{(i)}_{v,k,l}} &= k_\tau\left(\boldsymbol{z}_k, \boldsymbol{z}_k\right) + k_\tau\left(\boldsymbol{z}_l, \boldsymbol{z}_l\right) + k_\tau\left(\boldsymbol{z}_{k,l}, \boldsymbol{z}_{k,l}\right) + K_\tau\left(\boldsymbol{z}_k, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\boldsymbol{z}_d\}_{d=1}^D, \boldsymbol{z}_k\right) + \\ &K_\tau\left(\boldsymbol{z}_l, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\boldsymbol{z}_d\}_{d=1}^D, \boldsymbol{z}_l\right) + K_\tau\left(\boldsymbol{z}_{k,l}, \{\boldsymbol{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\boldsymbol{z}_d\}_{d=1}^D, \boldsymbol{z}_{k,l}\right) \ .\end{aligned} \tag{30}$$

Proof By Proposition 2, $g(\mathbf{z}_k) = \kappa^{(t)}_{v,k}$, thus, given the multivariate normal distribution

$$\begin{bmatrix} \boldsymbol{\beta}_v^{(m)} \\ g(\mathbf{z}_k) \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K_\tau + \sigma_\beta^2 \mathbf{I}_D & K_\tau\left(\{\mathbf{z}_d\}_{d=1}^D, \mathbf{z}_k\right) \\ K_\tau\left(\mathbf{z}_k, \{\mathbf{z}_d\}_{d=1}^D\right) & k_\tau\left(\mathbf{z}_k, \mathbf{z}_k\right) \end{bmatrix}\right) \tag{31}$$

the posterior mean and variance of $\kappa^{(t)}_{v,k}$ can be obtained as

$$\mu_{\kappa^{(t)}_{v,k}} = E\left[g(\mathbf{z}_k) | \{\mathbf{z}_d\}_{d=1}^D, \boldsymbol{\beta}_v^*\right] = K_\tau\left(\mathbf{z}_k, \{\mathbf{z}_d\}_{d=1}^D\right) R_\tau \boldsymbol{\beta}_v^{(m)} \ \text{and} \tag{32}$$

$$\sigma^2_{\kappa^{(t)}_{v,k}} = Var\left(g(\mathbf{z}_k) | \{\mathbf{z}_d\}_{d=1}^D, \boldsymbol{\beta}_v^*\right) = k_\tau\left(\mathbf{z}_k, \mathbf{z}_k\right) + K_\tau\left(\mathbf{z}_k, \{\mathbf{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\mathbf{z}_d\}_{d=1}^D, \mathbf{z}_k\right) \ . \tag{33}$$

Similarly, the posterior mean and variance of $\kappa^{(f)}_{v,l}$ are

$$\mu_{\kappa^{(t)}_{v,k}} = K_\tau\left(\mathbf{z}_k, \{\mathbf{z}_d\}_{d=1}^D\right) R_\tau \boldsymbol{\beta}_v^{(m)}, \ \sigma^2_{\kappa^{(f)}_{l,k}} = k_\tau\left(\mathbf{z}_l, \mathbf{z}_l\right) + K_\tau\left(\mathbf{z}_l, \{\mathbf{z}_d\}_{d=1}^D\right) R_\tau K_\tau\left(\{\mathbf{z}_d\}_{d=1}^D, \mathbf{z}_l\right) \ . \tag{34}$$

Since we have $g(\mathbf{z}_{k,l}) = \kappa^{(t)}_{v,k} + \kappa^{(f)}_{v,l} + \kappa^{(i)}_{v,k,l}$, given the multivariate normal distribution

$$\begin{bmatrix} \boldsymbol{\beta}_v^{(m)} \\ g(\mathbf{z}_{k,l}) \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K_\tau + \sigma_\beta^2 \mathbf{I}_D & K_\tau\left(\{\mathbf{z}_d\}_{d=1}^D, \mathbf{z}_{k,l}\right) \\ K_\tau\left(\{\mathbf{z}_{k,l}, \mathbf{z}_d\}_{d=1}^D\right) & k_\tau\left(\mathbf{z}_{k,l}, \mathbf{z}_{k,l}\right) \end{bmatrix}\right) \tag{35}$$

the posterior mean and variance of $\kappa^{(i)}_{v,k,l}$ can be obtained accordingly via

$$\mu_{\kappa^{(i)}_{v,k,l}} = E[g(\mathbf{z}_{k,l}) - g(\mathbf{z}_k) - g(\mathbf{z}_l) | \{\mathbf{z}_d\}_{d=1}^D, \boldsymbol{\beta}_v^{(m)}] \ \text{and} \tag{36}$$

$$\sigma^2_{\kappa^{(t)}_{v,k}} = Var\left(g(\mathbf{z}_{k,l}) - g(\mathbf{z}_k) - g(\mathbf{z}_l) | \{\mathbf{z}_d\}_{d=1}^D, \boldsymbol{\beta}_v^{(m)}\right) \ . \tag{37}$$

$\blacksquare$

The process for the inference of text structure is summarized in Algorithm 1 and the entire inference process is shown in Algorithm 2, where the update steps correspond to the equations described in this section for each parameter.

```
Algorithm 1: Text Structure Inference
Data: Term-document Matrix
Hyper-parameters: $\sigma_\beta$, $\Sigma_\tau$
Result: $\{\boldsymbol{\beta}^{(m)}, \boldsymbol{\kappa}^{(w)}, \boldsymbol{\kappa}^{(t)}, \boldsymbol{\kappa}^{(i)}\}$
for $v$ in $1, \ldots, V$ do
    Obtain $\boldsymbol{\beta}_v^*$ with (25)
    Obtain $\kappa_v^{(w)}$ with (4)
    Obtain $\boldsymbol{\beta}_v^{(m)} = \boldsymbol{\beta}_v^* - \kappa_v^{(w)}$
    Recover $\boldsymbol{\kappa}_v$ with Theorem 1
end
```

```
Algorithm 2: Variational Inference
Data: Term-document Matrix $\mathbf{W}$,
      Covariates $\mathbf{X}$
Model Setting: $K$, $L$
Hyper-parameters: $\sigma_\beta$, $\Sigma_\tau$, $\Sigma_\eta$, $\sigma_\gamma$, $\sigma_\phi$
Result: $\{\boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathrm{s}, \boldsymbol{\phi}, \boldsymbol{\xi}\}$
for $t$ in $1, \ldots, maxit$ do
    Update $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$ (Text Structure)
    Update $\boldsymbol{\eta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, s (Local Variables)
    Update $\boldsymbol{\phi}$, $\boldsymbol{\xi}$ (Covariate Structure)
end
```

5. Empirical Study

The empirical study comprises two parts. In the first part we compare our model quantitatively with other state-of-the-art approaches. We will show that it outperforms the other methods with regard to predictive performance on held-out data. The second part contains qualitative evaluations on case studies which demonstrate the usability of CFTM for gaining insight into text data and their covariates. The fitted CFTM model is used to extract underlying topics, structure among the covariates, and their interactions.

5.1. Datasets

We perform the empirical study using three real-world datasets.

Yle Election Compass 2019 is a survey directed to candidates for Finnish parliamentary elections with results open to the public.[1] It collects each candidate's basic information and agreement with different statements about ideological viewpoints, societal issues and policies, measured by 29 Likert scale questions (score 1-5). Candidates can elaborate their answers to advertise or communicate to voters; We take the written content of each candidate as the text document, the Likert scale questions as continuous variables, and gender and native languages as categorical variables. Text is lemmatized, numbers, punctuation and stop-words are removed. Texts with more than 40 words are taken for analysis, the final dataset contains 1937 documents and 1764 vocabulary terms. The original text is in Finnish, in the case study shown in Section 5.3 we provide an English translation.

Doom Eternal Game Reviews were collected from Steam[2], a popular gaming platform with an abundance of player-written game reviews. We focused on a first-person shooter game "Doom Eternal". Review texts and corresponding metadata were collected via SteamAPI and profile data was crawled from public profile pages linked with collected Steam IDs. The positivity/negativity (if the reviewer recommends the game or not) is taken

[1] https://vaalikone.yle.fi/eduskuntavaali2019 ; https://yle.fi/uutiset/3-1072538.

[2] https://store.steampowered.com/

as a categorical variable. The number of submissions and guides of users are rare events so they are considered count variables. We identified 22 continuous variables such as number of achievements, and played time, etc. For text content, numbers, punctuations and stop words were removed and the text was lemmatized. Reviews with more than 40 words after processing were kept. Finally, a collection of 1144 reviews and 2377 terms remained.

Airport Lounge User Reviews were collected from Skytrax[3], in which customers can give numerical ratings (score 1-5, including aspects such as comfort and staff) to the airport lounges together with written reviews. Again, we keep reviews with more than 40 words, numbers, punctuations and stop words were removed, and texts are lemmatized. The processed data set contains 1311 reviews with 2799 vocabulary terms, paired with 8 numerical ratings, and 2 categorical ratings (recommend or not).

5.2. Quantitative Evaluation

We compare our model with four state-of-the-art models: LDA, STM, MetaLDA, and SCHOLAR. The performance comparison focuses on held-out prediction using the above-mentioned datasets. Details are described as follows.

Evaluation Metric. The held-out likelihood is used to evaluate model performance. The text content is randomly divided into training and held-out sets, each containing 50% of the original content [4]. The training set is used to fit the models. The fitted model is then used to predict the held-out text content and the held-out likelihood values are computed. Note that another typical metric, perplexity on the test set, is an exponential transformation of the held-out likelihood: higher held-out likelihood means lower perplexity.

Experimental Settings. CFTM is run with simple unoptimized prior settings $\alpha = 10 \cdot \mathbf{1}$, $\Sigma_\gamma = \Sigma_\tau = 10 \cdot \mathbf{I}$, $\sigma_\eta = \sigma_\phi = 0.1$, $\sigma_\beta = 0.01$. Other methods are run with their default values. We evaluate the model performance on different settings (combinations of the number of topics $K \in \{5, 10, 15, 20\}$ and number of factors $L \in \{5, 10\}$). To assess robustness of the methods to limited data, we run experiments both on the full data sets and on a random draw of 500 documents. The document subset sampling (in the limited-data case), train-test division, and model fitting are repeated 10 times for each setting.

Running time. We implemented our algorithms in R [5]. Using the parallel implementation, on average our model takes around 8 minutes and 13 minutes to converge using 8 and 4 cores respectively. In contrast, the R implementation of STM (a method also having covariates) takes 18 minutes to converge, clearly longer than our model.

Results. The result is shown in Figure 2. In most settings (Yle Compass with 500 samples, Doom Eternal full data set and 500-samples, Lounge reviews full) CFTM clearly and statistically significantly outperforms all other methods. In two settings results were closer: for Lounge reviews with 500 samples, CFTM with $L = 5$ is statistically significantly better than the closest competitor SCHOLAR for 5 and 10 topics and not significantly different for 15 and 20; for the Yle Compass full data set the difference to the closest competitor LDA is not statistically significant. Overall, CFTM has consistently good performance.

[3] www.airlinequality.com, we use the collection https://github.com/quankiquanki/skytrax-reviews-dataset

[4] Note that the 50%-50% division is chosen according the practice used in STM(Roberts et al. 2016).

[5] source code and data sets used in this work can be found in supplementary material

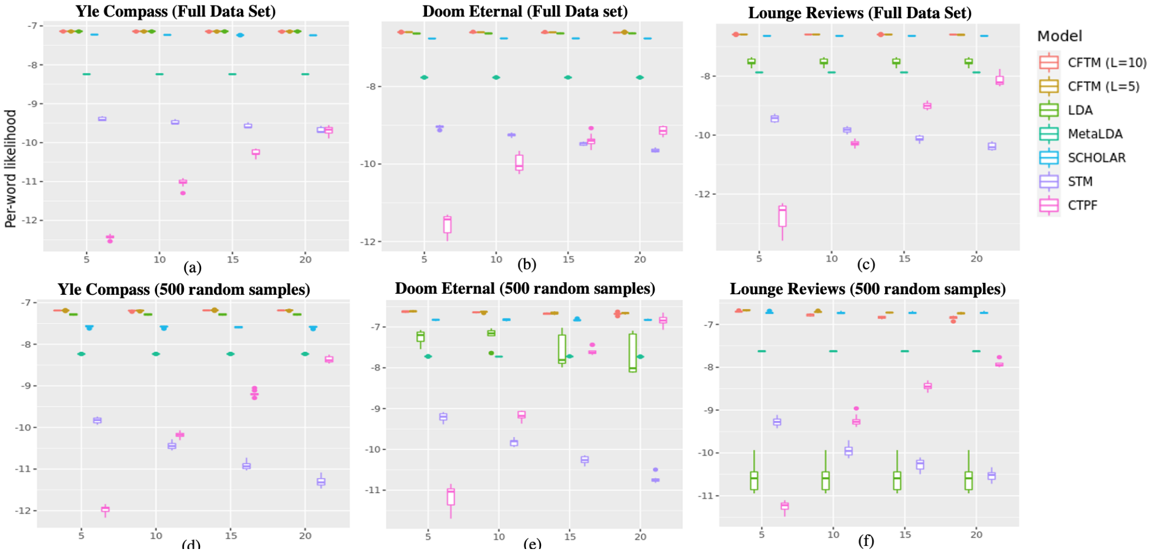


Figure 2: Quantitative evaluation, performance comparison with held-out likelihood (per word), higher is better. CFTM is compared to LDALDA, MetaLDA, SCHOLAR, STM, MetaLDA, and CTPF. (a)-(c): Comparison on the full dataset. (d)-(f): Comparison on 500 random samples. The box plots show variation of performance over random simulations or random divisions of data.

5.3. Case Studies

We conduct empirical analyses using CFTM on *Yle Election Compass 2019* and *Doom Eternal Game Reviews* datasets. The hyper-parameter setting is the same as above but the model is trained with the full datasets. Among multiple choices of the number of topics and factors, we use semantic coherence (Mimno et al. 2011) as the model selection criterion to choose the best CFTM model for inspection.

Spectrum of political positions. When fitting the *Yle Election Compass 2019* dataset, the CFTM model with 9 topics and 5 factors was selected. Figure 3(a) shows the top words for the 9 extracted topics. Topic names are assigned by authors by analysis of the topic words. CFTM has found clear topical content appropriate in the domain: each topic uncovers different aspects of political interests ranging from local politics (*Local Politics of Pirkanmaa*) to climate issues (*Climate Change and Costs*).

Figure 3 (b) displays the factor structures of three factors: *Eurosceptic*, *Green*, and *Pro-global* [6] and Figure 3 (c) presents their influences on topic content. Similarly to the topics, factor labels can be assigned by analyzing their feature weights (posterior mean of ϕ). For example, the factor *Green* supports environmental protection, having high agreement with statements such as “Climate is worth the cost”, “Discourage eating meat”, and “Reduce tree cutting”. The factor *Eurosceptic* agrees with statements “Leave eurozone” and “Immigrants

[6] The feature weights of all the 5 factors are provided in supplementary material

Topic	Top 10 Words
Bans and Actions	ban, have to, want, Swedish, choose, action, such, hundred, accept, positive
Climate Change and Costs	climate, get, suffice, Euro, price, company, discrimination, climate change, event, reality
Local Politics of Pirkanmaa	Pirkanmaa, Finland, such, must, nearby year, alcohol, God, habit, rich, decision maker
Support the Youth	end stage, young, use, business activity, rule-out, livelihood, external, support, society, both
Tourism Business	company, problem, tourism, money, strawberry, working life, solution, Russia, correct, Swedish
Young Immigrants	young, prerequisite, energy drink, decision, arrangement, crime, secure, immigration, job, can
Offer Support and Alternatives	increase, offer, support, act, alternative, reject, quite, sport, guide, grounds
Life Attitudes	working life, level, speak, decide, God, mind, suffice, act, possible, wrong
Problem Resolution	manner, part, situation, develop, strive, try, act, secure, exploit, problem

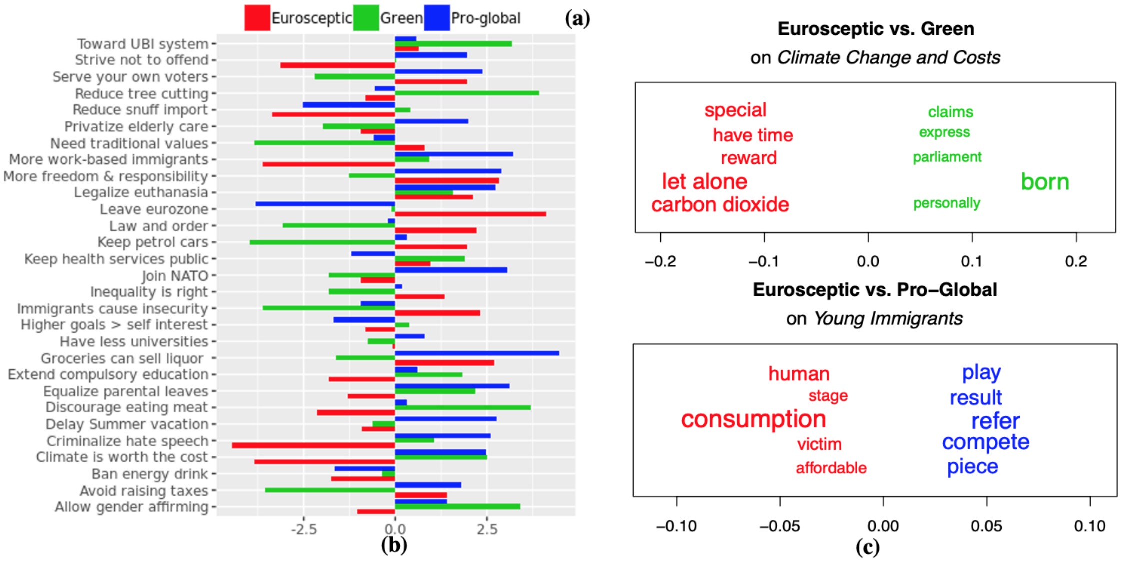


Figure 3: CFTM results for Yle Election Compass. (a) Extracted topics. Top words are shown in order of frequency in the topic. (b) Feature weights of factors *Eurosceptic*, *Green* and *Pro-global*. (c) Wording difference of *Eurosceptic* vs. *Green* on the topic *Climate Change and Costs*, and wording difference of *Eurosceptic* vs. *Pro-global* on topics *Young Immigrants*. Horizontal position of a term v shows the difference $\kappa^{(i)}_{v,l,k} - \kappa^{(i)}_{v,l',k}$ in topic k of factors l and l'.

cause insecurities", and disagrees with "Join NATO"; and the factor *Pro-global* holds an opposite position on the above statements and supports "More work-based Immigrants".

The impact of factors on wordings of a topic can be explored with the posterior of $\kappa^{(i)}$. Figure 3 (c) examines factor influence on wordings, showing the comparison of *Eurosceptic* vs. *Green* on the topic *Climate Change and Costs* and the comparison of *Eurosceptic* vs. *Pro-global* on the topic *Young Immigrants*. The horizontal axis reveals the difference of influence between two factors on prominence of words. Candidates with high loading along the *Eurosceptic* factor use more words 'let alone' and 'make time' when discussing the topic

Climate Change and Costs whereas candidates aligned along the factor *Green* emphasize 'claims', and 'personally'. On the other hand, when it comes to the topic emphasize *Young Immigrants*, candidates aligned along the factor *Eurosceptic* use more words such as 'victim', and 'consumption', whereas candidates aligned along the factor *Pro-global* use more words such as 'compete' and 'result'. The differing wording preferences among the factors corresponding to competing political orientations shows how the same issues (topics) are approached from very different perspectives by candidates aligned along those factors.

Exploring player experiences. The CFTM model of 6 topics and 7 factors was selected when fitting the *Doom Eternal Game Reviews* dataset. Figure 4 (a) displays the topic words of the extracted topics. The topics cover game mechanics (e.g. *Fighting*, *Damage and Survival*) and more general views on the game (*Feelings and Experiences*) and issues external to the play experience (*Support and Services*). Figure 4 (b) [7] and (c) further present the feature weights of factors *Doom-focused Player*, *Game Collector*, and their influences on topics *Support and Services* and *Feelings and Experiences*. Players with high loading of the factor *Doom-focused Player* are more likely to use words like 'doom' and 'account' in the topic *Support and Services* and 'feel', 'weapon' in topic *Feelings and Experiences*, whereas players with a high loading of the factor *Game Collector* prefer words 'rip', 'tear' in both topics *Support and Services* and *Feelings and Experiences*. The topics, factors and interactions are well-suited for the domain.

6. Conclusions

We presented the Cross-structural Factor-Topic Model (CFTM), a novel generative probabilistic model for text documents occurring with sophisticated covariates. It represents latent topical structure in text, factor structure in covariates, and influence of the factors on both topic prevalence and content. The model is flexible, allowing both discrete covariates with a mixture structure and continuous covariates with a factorized structure in the same model. We proposed an efficient inference scheme coupling variational inference to efficient distributed inference. In experiments the model outperformed LDA, STM, MetaLDA, and SCHOLAR; moreover, CFTM discovered meaningful topics, factors, and factor influences in case studies investigating a political survey and reviews of a computer game.

Acknowledgments

This work is supported by the Academy of Finland decisions 312395, 313748, and 327352.

References

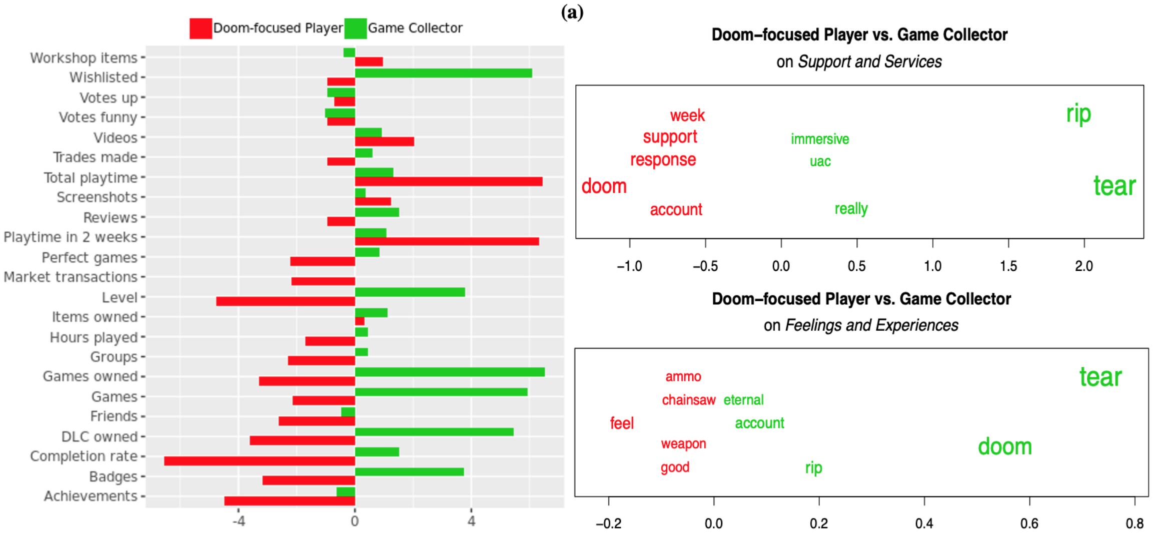
A. Acharya, D. Teffer, J. Henderson, M. Tyler, M. Zhou, and J. Ghosh. Gamma process Poisson factorization for joint modeling of network and documents. In *Proc. ECML PKDD*, pages 283–299. Springer, 2015.

R. Agrawal, B. Trippe, J. Huggins, and T. Broderick. The kernel interaction trick: Fast bayesian discovery of pairwise interactions in high dimensions. In *Proc. ICML*, pages 141–150, 2019.

[7] The feature weights of all the 7 factors are provided in supplementary material

Topic	Top 10 Words
Fighting	rip, tear, dream, frankly, potato, neon, success, kar, smoothly, hugo
Support and Services	support, response, week, account, anyway, offline, everything, paste, team, approve
Visuals and Features	doom, dream, neon, march, hdr, doot, mayhem, replayability, market, kickass
Damage and Survival	damage, thing, though, run, challenge, player, bad, lot, combat, people
Movement and Weapons	dash, weapon, contain, minute, maykr, grenade, teleport, switch, pad, ammo
Feelings and Experiences	really, game, good, recommend, software, feel, level, learn, start, whole

(a)



(b) **(c)**

Figure 4: CFTM results for Doom Eternal. (a) Extracted topics. (b) Feature weights of factors *Doom-focused Player* and *Game Collector*. (c) Wording difference of factors *Doom-focused Player* vs. *Game Collector* on the topic *Support and Services* and *Feelings and Experiences*.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

D. Card, C. Tan, and N. A. Smith. Neural models for documents with metadata. In *Proc. ACL*, pages 2031–2040. ACL, 2018.

E. de Souza da Silva, H. Langseth, and H. Ramampiaro. Content-based social recommendation with Poisson matrix factorization. In *Proc. ECML PKDD*. Springer, 2017.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

J. Eisenstein, A. Ahmed, and E. Xing. Sparse additive generative models of text. In *Proc. ICML*, pages 1041–1048. ACM, 2011.

P. K. Gopalan, L. Charlin, and D. Blei. Content-based recommendations with Poisson factorization. In *Proc. NIPS*, pages 3176–3184, 2014.

L. Gui, J. Leng, G. Pergola, R. Xu, and Y. He. Neural topic model with reinforcement learning. In *Proc. EMNLP-IJCNLP*, pages 3469–3474. ACL, 2019.

C. Hu, P. Rai, and L. Carin. Non-negative matrix factorization for discrete data with hierarchical side-information. In *Proc. AISTATS*, pages 1124–1132. PMLR, 2016.

H. Kim, Y. Sun, J. Hockenmaier, and J. Han. Etm: entity topic models for mining documents associated with entities. In *Proc. ICDM*, pages 349–358. IEEE, 2012.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014.

J. Mcauliffe and D. Blei. Supervised topic models. In *Proc. NIPS*, pages 121–128, 2008.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proc. UAI*, pages 411–418. AUAI Press, 2008.

D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. EMNLP*, pages 262–272. ACL, 2011.

C. Rasmussen and C. Williams. *Gaussian processes in machine learning.* MIT Press, 2006.

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

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. UAI*, pages 487–494. AUAI Press, 2004.

A. Srivastava and C. A. Sutton. Autoencoding variational inference for topic models. In *Proc. ICLR*. OpenReview.net, 2017.

M. Taddy. Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3): 1394–1414, 2015.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

C. Wang and D. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.

X. Wang and Y. Yang. Neural topic model with attention for supervised learning. In *Proc. AISTATS*, pages 1147–1156. PMLR, PMLR, 2020.

H. Zhao, L. Du, W. Buntine, and G. Liu. MetaLDA: A topic model that efficiently incorporates meta information. In *Proc. ICDM*, pages 635–644. IEEE, 2017.

H. Zhao, P. Rai, L. Du, and W. Buntine. Bayesian multi-label learning with sparse features and labels, and label co-occurrences. In *Proc. AISTATS*. PMLR, 2018.