Sentiment analysis with covariate-assisted word embeddings

Xu, Shirong; Dai, Ben; Wang, Junhui

Published in:
Electronic Journal of Statistics

Published: 01/01/2021

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1214/21-EJS1854

Publication details:
Xu, S., Dai, B., & Wang, J. (2021). Sentiment analysis with covariate-assisted word embeddings. *Electronic Journal of Statistics*, *15*(1), 3015-3039. https://doi.org/10.1214/21-EJS1854

Electronic Journal of Statistics
Vol. 15 (2021) 3015–3039
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1854

Sentiment analysis with covariate-assisted word embeddings*

Shirong Xu

School of Data Science, City University of Hong Kong, Kowloon Tong, Hong Kong
e-mail: shirongxu2-c@my.cityu.edu.hk

Ben Dai

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: bdai@umn.edu

Junhui Wang

School of Data Science, City University of Hong Kong, Kowloon Tong, Hong Kong
e-mail: j.h.wang@cityu.edu.hk

Abstract: Sentiment analysis measures inclination of textual documents, aiming to extract and quantify their subjective sentiment polarity. In literature, most sentiment analysis methods first numericalize textual documents through certain word embeddings framework, and then formulate sentiment analysis as an ordinal regression or classification task. Yet it is often ignored that different people may have different preference of wording, and thus a uniform word embeddings often leads to suboptimal performance. In this article, to accommodate the heterogeneity among individual persons, we propose a covariate-assisted word embeddings in a margin-based ordinal regression framework, where covariates are incorporated through scaling factors to adjust the word embeddings. Moreover, we employ a block-wise coordinate descent scheme to tackle the resultant large-scale optimization task, and establish theoretical results to quantify the asymptotic behavior of the proposed method, guaranteeing its fast convergence rate in terms of prediction accuracy. Finally, we demonstrate the advantages of the proposed method over its competitors in both the Yelp Challenge dataset and synthetic datasets.

MSC2020 subject classifications: Primary 62H30.
Keywords and phrases: Personalized prediction, sentiment analysis, word embeddings, unstructured data, ordinal regression.

Received October 2020.

Contents

*This work is supported in part by HK RGC grants GRF-11303918, GRF-11300919 and GRF-11304520.

1. Introduction

Unstructured text data has become increasingly important in recent years, due to the fast advancement of information technology and evolution of information storage. It typically arises from text-heavy documents, including customer reviews, news stories, or online twits. One of the central tasks of text data analysis is to extract subjective sentiment polarity of the textual documents, which has been an essential component in modern business analytics and political surveys [1, 2].

In literature, most sentiment analysis methods first convert textual documents into numerical vectors, and then formulate it as a classification task, where sentiment levels are treated as binary or ordinal responses [3, 4, 5]. The numericalization is often done by using the bag-of-words framework [6], where word presences or frequencies in the textual documents are extracted as the numerical features. The bag-of-words framework is interpretable and easy to implement, but it fails to capture the relationship among meaningful words. Recently, embedding technique [7] has drawn significant interests from both statistics and machine learning communities for its flexibility and interpretability in representing textual documents, including Word2Vec [8] and Global Vectors for Word Representation (GloVe) [9]. The key idea of word embeddings is to embed each word into a low-dimensional vector space so that the corresponding vectors of relevant words are close in the embedded space. A number of embeddings schemes have been proposed from various perspectives, in order to obtain a uniform word embeddings for all individual persons to facilitate the subsequent text data analysis.

A uniform word embeddings is simple in nature but also suffers from some intrinsic limitations, due to the fact that different people may have different preferences of wording. For example, the word "interesting" can be used to express neutral or even negative sentiment level by people who tend not to use negative words to show politeness. Also, it appears very common that people like to use sarcastic expressions on internet. One review in Yelp dataset says

"I feel so excited to check out" to express the extremely negative sentiment without using any negative words. Analyzing such textual statements can be easily misled by the presence of positive words with strong polarity [10, 11]. In literature, difference in wording across genders, ages, educational background and political background has been widely reported [1, 12, 13]. It is thus natural to consider adaptive word embeddings to capture the heterogeneity among individual persons so that their preferences of wording can be incorporated to improve the prediction accuracy. Yet, only a few attempts have been made in literature, including time-varying word embeddings [14] and topic-adaptive word embeddings [15, 16].

In this paper, we propose a sentiment analysis method based on a novel covariate-assisted word embeddings, which integrates covariates into a ordinal regression framework [17] to refine word embeddings for better prediction accuracy. Specifically, a sentiment lexicon and the corresponding word embeddings will be employed to construct covariate-adjusted representation of each textual document. For each covariate level, an adjusting factor is introduced to scale the original word embeddings, which quantifies the deviation of semantics from the pre-trained word embeddings. Furthermore, we also develop a scalable block-wise coordinate descent algorithm to tackle the resultant large-scale optimization task. Theoretically, the asymptotic convergence rate of the proposed method is established in terms of sample size, sentiment levels, covariate levels, lexicon size.

The rest of the paper is organized as follows. Section 2 presents the proposed covariate-assisted word embeddings in an ordinal regression framework for sentiment analysis, as well as the block-wise coordinate descent algorithm. Section 3 establishes the asymptotic results for the proposed method assuring its fast convergence rate under several situations. Section 4 conducts a simulation study to examine the numerical performance of the proposed method in various synthetic datasets and applies the proposed method to analyze the Yelp challenge dataset. A brief summary is given in Section 5, and the Appendix contains the technical proofs.

2. Proposed methodology

2.1. Preambles

In sentiment analysis, a training dataset consists of $\{(\boldsymbol{t}_{ij}, y_{ij}); i = 1, \cdots, u, j = 1, \cdots, N_i\}$, where $\boldsymbol{t}_{ij}$ is the j-th textual document made by the i-th person, and $y_{ij} \in \{1, \cdots, K\}$ indicates its sentiment level with ordering $1 \prec 2 \prec \cdots \prec K$, where $\prec$ denotes less positive in terms of sentiment level. The primary goal of sentiment analysis is to construct a decision function $\phi(\boldsymbol{t}_{ij})$ to accurately predict the sentiment level of $\boldsymbol{t}_{ij}$, so that the disagreement between $\phi(\boldsymbol{t}_{ij})$ and y_{ij} can be minimized.

Various disagreement metrics for ordinal regression have been considered in literature [18], including mean absolute error(MAE), mean zero-one error

(MZOE) and mean square error(MSE). However, neither MZOE nor MSE is originally designed for ordinal regression, and both metrics have their own limitations for analyzing ordinal data. Particularly, MZOE fails to take the ordinality into account, which is particularly undesirable in sentiment analysis, where mis-classifying a positive review as neural is less severe than as negative, whereas these two type of misclassifications are treated equally in MZOE. As for MSE, it regards ordinal response as continuous, leading to unnecessary bias, especially when ordinal responses are only encoded to reflect the ordering but imply no elaboration of the difference among the ordered values. By contrast, MAE appears to be a reasonable choice for ordinal regression, and widely used in literature [17]. It can be written as

$$\mathrm{MAE}(\phi) = E\left(\left|y - \phi(\boldsymbol{t})\right|\right) = \sum_{k=1}^{K-1} E\Big(I\big(\operatorname{sgn}(y-k)\operatorname{sgn}\left(\phi(\boldsymbol{t})-k\right) \le 0\big)\Big), \quad (2.1)$$

where $I(\cdot)$ is an indicator function and $\operatorname{sgn}(x) = 1$ when $x > 0$, and -1 otherwise. It follows immediately from (2.1) that minimizing MAE is equivalently transformed into solving $K-1$ binary classification problems, where $\operatorname{sgn}(y-k)$ can be treated as the binary class label and $\operatorname{sgn}\left(\phi(\boldsymbol{t})-k\right)$ denotes the corresponding classification decision function.

Instead of estimating $\phi(\boldsymbol{t})$ directly, it is common to introduce $K-1$ functions with $f_{K-1}(\boldsymbol{t}) \le \cdots \le f_1(\boldsymbol{t})$, and set $\phi(\boldsymbol{t}) = \min\{k : f_k(\boldsymbol{t}) \le 0\}$. Then the estimation of ϕ is converted to estimating $\boldsymbol{f} = (f_1, \ldots, f_{K-1})$, and

$$\mathrm{MAE}(\phi) = \mathrm{MAE}(\boldsymbol{f}) = \sum_{k=1}^{K-1} E\Big(I\big(\operatorname{sgn}(y-k) f_k(\boldsymbol{t}) \le 0\big)\Big). \quad (2.2)$$

The indicator function in (2.2) is computationally intractable for optimization, thus we replace it by some surrogate margin losses. Specifically, an empirical version of (2.2) with a surrogate loss and regularization term can then be constructed to estimate $\boldsymbol{f}$,

$$\min_{\boldsymbol{f}} \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{K-1} V(\operatorname{sgn}(y_{ij}-k) f_k(\boldsymbol{t}_{ij})) + \lambda J(\boldsymbol{f}) \quad (2.3)$$

$$\text{subject to } f_{K-1}(\boldsymbol{t}_{ij}) \le f_{K-2}(\boldsymbol{t}_{ij}) \le \ldots \le f_1(\boldsymbol{t}_{ij}); i = 1, \ldots, n, j = 1, \ldots, N_i,$$

where $N = \sum_{i=1}^{n} N_i$, $V(z)$ is a surrogate margin loss function non-decreasing in z, $J(\boldsymbol{f})$ is a regularization term, and λ is a tuning parameter. Here $V(z)$ can take various forms. For instance, $V(\cdot)$ can be the hinge loss $V(u) = (1-u)_+$ [19], the ψ-loss $V(u) = \min((1-u)_+, 1)$ [20], or the logistic loss $V(u) = 1/(1+\exp(-u))$ [21]. It is interesting to note that setting V as the hinge loss or the logistic loss in (2.3) resembles the ALL-threshold method [22] for ordinal regression.

To facilitate the modelling of $\boldsymbol{f}$, textual documents need to be pre-processed into numerical vectors. In literature, primitive approaches extract word presence or frequency in $\boldsymbol{t}$ as predictors under the bag-of-words framework [6, 23].

Recently, more informative word embeddings frameworks have been developed, such as Word2Vec and GloVe [8, 9]. In particular, Word2Vec learns the word representation via a three-layer neural network, which assumes that words with similar linguistic meaning should be close in textual documents, resulting in frequent co-occurrences within a fixed-sized context window.

Specifically, let $\mathcal{D} = \{\omega_1, \omega_2, \ldots, \omega_d\}$ be a lexicon of sentiment words and $\boldsymbol{E} \in \mathbb{R}^{p\times d}$ be an embedding matrix, where p is the dimension of the embedded space and each column denotes the embedding of the corresponding word in $\mathcal{D}$. We further define $\boldsymbol{B}(\boldsymbol{t}) = (b_1, \ldots, b_d)^T$ to be the frequency vector of $\boldsymbol{t}$ based on $\mathcal{D}$. Then $\boldsymbol{EB}(\boldsymbol{t})$ is the averaged embeddings of words appearing in $\boldsymbol{t}$, which can be viewed as the representation of $\boldsymbol{t}$ in the embedded space, and the sentiment function f_k can be formulated as

$$f_k(\boldsymbol{t}_{ij}) = \boldsymbol{\beta}^T \boldsymbol{EB}(\boldsymbol{t}_{ij}) + \beta_{0,k}, \tag{2.4}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\beta_{0,K-1} \leq \ldots \leq \beta_{0,1}$. Clearly, $f_1, \ldots, f_{K-1}$ are parallel, and their ordering is inherent in $\beta_{0,k}; k = 1, \ldots, K-1$. Such structures have been commonly used in literature to enforce ordering among multiple functions [24, 25, 26].

2.2. Covariate-assisted word embeddings

In many scenarios, some covariates $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijL})^T$ are also available for each observation $(\boldsymbol{t}_{ij}, y_{ij})$, where $x_{ijl} \in \{1, \ldots, m_l\}$ denotes the l-th covariate and m_l denotes its number of distinct levels. To incorporate available covariates in the word embeddings, the proposed covariate-assisted sentiment function f_k can be formulated as

$$f_k(\boldsymbol{t}_{ij}, \boldsymbol{x}_{ij}) = \boldsymbol{\beta}^T \boldsymbol{E}\big(\boldsymbol{B}(\boldsymbol{t}_{ij}) \circ \boldsymbol{w}^{(1)}_{x_{ij1}} \circ \cdots \circ \boldsymbol{w}^{(L)}_{x_{ijL}}\big) + \beta_{0,k}(\boldsymbol{x}_{ij}), \tag{2.5}$$

where $\circ$ denotes the entry-wise product, $\boldsymbol{w}^{(l)}_{x_{ijl}} \in \mathbb{R}^d$ denotes the parameter vector corresponding to x_{ijl}, and the intercept $\beta_{0,k}(\boldsymbol{x})$ is allowed to vary with $\boldsymbol{x}$. Note that both $\boldsymbol{E}$ and $\boldsymbol{B}$ are pre-trained or pre-specified, and only $\boldsymbol{\beta}$, $\beta_{0,k}$ and $\boldsymbol{w}^{(l)}_{x_{ijl}}$ are the unknown parameters in (2.5) that need to be estimated. Particularly, $\boldsymbol{w}^{(l)}_{x_{ijl}}$ serves the purpose of adjusting $\boldsymbol{B}(\boldsymbol{t}_{ij})$ in $f_k(\boldsymbol{t}_{ij}, \boldsymbol{x}_{ij})$, and the varying intercept can be viewed as baseline in predicting sentiment for each covariate level. Furthermore, although $\boldsymbol{\beta}$ and $\boldsymbol{w}^{(l)}_{x_{ijl}}$ may not be identifiable, they contribute to $f_k(\boldsymbol{t}_{ij}, \boldsymbol{x}_{ij})$ only through their product, and thus does not affect the predictability of the proposed method. If interpretability is also of interest, one may fix $\|\boldsymbol{\beta}\| = 1$ to avoid the non-identifiability, which only requires an additional normalization step of $\boldsymbol{\beta}$. Additionally, the proposed method in (2.5) is mainly designed for binary and categorical covariates, to which a direct extension to continuous covariates is to divide the domain of covariates into exclusive subsets, which are then treated as distinct categorical levels.

Let $\overline{\boldsymbol{w}}_{\boldsymbol{x}_{ij}} = \boldsymbol{w}^{(1)}_{x_{ij1}} \circ \cdots \circ \boldsymbol{w}^{(L)}_{x_{ijL}}$ be the overall adjusting effect, then $f_k(\boldsymbol{t}_{ij}, \boldsymbol{x}_{ij})$ can be rewritten as

$$f_k(\boldsymbol{t}_{ij}, \boldsymbol{x}_{ij}) = \boldsymbol{\beta}^T \big[\boldsymbol{E} \circ (\mathbf{1}_p \otimes \overline{\boldsymbol{w}}^T_{\boldsymbol{x}_{ij}})\big] \boldsymbol{B}(\boldsymbol{t}_{ij}) + \beta_{0,k}(\boldsymbol{x}_{ij}), \tag{2.6}$$

where $\otimes$ denotes the Kronecker product. This formulation leads to the proposed covariate-assisted word embeddings, where $\overline{\boldsymbol{w}}_{ij}$ calibrates the embedding matrix $\boldsymbol{E}$ by multiplying its embedding vector with a scaling factor. It allows for a refined word embedding by incorporating the available covariates, which is in sharp contrast to the uniform word embeddings in literature. For example, a positive word can be used to express negative sentiment when its embedding vector is multiplied by a negative scalar. This flexibility is particularly attractive when analyzing sarcastic statements on internet. Additionally, the overall adjusting effect $\overline{\boldsymbol{w}}_{ij}$ of documents issued by the same person will be close since they share common covariates, implying similarity among wordings of different textual documents by the same person. It also allows certain similarity among wording of different persons depending on the level of their common covariates.

With the modelling of f_k in (2.6), the proposed method can be organized as

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\beta}_0} \quad & \frac{1}{N} \sum_{i=1}^{u} \sum_{j=1}^{N_i} \sum_{k=1}^{K-1} V\big(\operatorname{sign}(y_{ij} - k)(\boldsymbol{\beta}^T \boldsymbol{E}\big(\boldsymbol{B}(\boldsymbol{t}_{ij}) \circ \bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}\big) + \beta_{0,k}(\boldsymbol{x}_{ij}))\big) \\ & \qquad + \lambda_1 J(\boldsymbol{\beta}) + \lambda_2 J(\boldsymbol{W}) \\ \text{subject to} \quad & \beta_{0,K-1}(\boldsymbol{x}_{ij}) \le \beta_{0,K-2}(\boldsymbol{x}_{ij}) \le \cdots \le \beta_{0,1}(\boldsymbol{x}_{ij}) \text{ for all } \boldsymbol{x}_{ij}, \end{aligned} \tag{2.7}$$

where $\boldsymbol{W} = [\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(L)}]$, and $\boldsymbol{W}^{(l)} = [\boldsymbol{w}^{(l)}_1, \ldots, \boldsymbol{w}^{(l)}_{m_l}]$ denotes the adjusting matrix of l-th categorical covariate. Here the number of parameters of $\boldsymbol{\beta}$ and $\boldsymbol{W}$ are p and $d\sum_{l=1}^{L} m_l$, respectively. To control the complexity of sentiment functions, $J(\boldsymbol{\beta})$ and $J(\boldsymbol{W})$ can be any regularization term. In the sequel, we illustrate the proposed method by setting $V(\cdot)$ is the hinge loss, $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$, and $J(\boldsymbol{W}) = \|\boldsymbol{W}\|_F^2$.

Note that the proposed method in (2.7) consists of two tuning parameters λ_1 and λ_2, and Lemma 1 shows they play a similar role in (2.7) and thus significantly simplifies the tuning process for λ_1 and λ_2.

Lemma 1. *The solution to (2.7) remains the same as long as $\lambda_1 \lambda_2^L$ stays the same.*

Lemma 1 implies that the optimization task in (2.7) with tuning parameters (λ_1, λ_2) has the same solution as that with (λ, λ) satisfying $\lambda^{L+1} = \lambda_1 \lambda_2^L$. Therefore, we simplify the cost function in (2.7) by setting $\lambda_1 = \lambda_2 = \lambda$ in the sequel. The proof of Lemma 1 and all other technical proofs are provided in a supplementary file [27].

2.3. Scalable computation

Note that optimization task in (2.7) is a bi-convex optimization problem with respect to $\boldsymbol{\beta}$ and $\bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$, and hence we employ a block-wise coordinate descent

algorithm to update $\boldsymbol{\beta}$ and $\bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ sequentially. By introducing a slack variable ξ_{ijk}, (2.7) can be reformulated as

$$\min_{\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\beta}_0,\xi_{ijk}} \quad \sum_{i,j,k} \xi_{ijk} + \lambda\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{W}\|_F^2 \tag{2.8}$$

$$\text{subject to} \quad \operatorname{sgn}(y_{ij}-k)\big(\boldsymbol{\beta}^T\boldsymbol{E}(\boldsymbol{B}(\boldsymbol{t}_{ij})\circ\bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}) + \beta_{0,k}(\boldsymbol{x}_{ij})\big) \geq 1-\xi_{ijk},\ \xi_{ijk}\geq 0,$$
$$\beta_{0,K-1}(\boldsymbol{x}_{ij}) \leq \beta_{0,K-2}(\boldsymbol{x}_{ij}) \leq \cdots \leq \beta_{0,1}(\boldsymbol{x}_{ij}).$$

We then break (2.8) into multiple sub-tasks, and update $\boldsymbol{\beta}, \boldsymbol{W}$ and $\boldsymbol{\beta}_0$ alternatively. Specifically, when $\boldsymbol{W}$ and $\boldsymbol{\beta}_0$ are fixed, $\boldsymbol{\beta}$ can be updated by solving

$$\min_{\boldsymbol{\beta},\xi_{ijk}} \quad \sum_{i,j,k} \xi_{ijk} + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{2.9}$$

$$\text{subject to} \quad \operatorname{sgn}(y_{ij}-k)\big(\boldsymbol{\beta}^T\boldsymbol{E}(\boldsymbol{B}(\boldsymbol{t}_{ij})\circ\bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}) + \beta_{0,k}(\boldsymbol{x}_{ij})\big) \geq 1-\xi_{ijk},\ \xi_{ijk}\geq 0.$$

Note that the optimization task in (2.9) resembles linear support vector machine (SVM) in nature except additional varying intercepts, which can be efficiently solved by `Liblinear` [28]. Therefore, we develop a similar optimization scheme based on dual coordinate descent method as in `Liblinear` to solve (2.9), which is named as driftSVM and available in Python package `VarSVM`.

When $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ are fixed, our strategy is to use the back-fitting scheme to update $\boldsymbol{W}^{(l)}; l = 1, \cdots, L$ sequentially. Particularly, $\boldsymbol{W}^{(l)}$ can be updated by solving

$$\min_{\boldsymbol{W}^{(l)},\xi_{ijk}} \quad \sum_{q=1}^{m_l} \sum_{\{\boldsymbol{x}_{ij}:\ x_{ijl}=q\}} \sum_k \xi_{ijk} + \lambda\sum_{q=1}^{m_l}\|\boldsymbol{w}_q^{(l)}\|_2^2$$

$$\text{subject to} \quad \operatorname{sgn}(y_{ij}-k)\big((\boldsymbol{w}_q^{(l)})^T\boldsymbol{B}_{ij}^{-\boldsymbol{w}_{x_{ijl}}}\boldsymbol{E}^T\boldsymbol{\beta} + \beta_{0,k}(\boldsymbol{x}_{ij})\big) \geq 1-\xi_{ijk},\ \xi_{ijk}\geq 0,$$

where $\boldsymbol{B}_{ij}^{-\boldsymbol{w}_{x_{ijl}}} = \operatorname{diag}\{\boldsymbol{B}(\boldsymbol{t}_{ij})\circ\boldsymbol{w}_{x_{ij1}}^{(1)}\circ\cdots\circ\boldsymbol{w}_{x_{ij,l-1}}^{(l-1)}\circ\boldsymbol{w}_{x_{ij,l+1}}^{(l+1)}\circ\cdots\boldsymbol{w}_{x_{ijL}}^{(L)}\}$ is a diagonal matrix. Furthermore, $\boldsymbol{w}_q^{(l)}$ can be optimized in a parallel fashion. That is, each $\boldsymbol{w}_q^{(l)}, q = 1, \cdots, m_l$ can be updated by solving

$$\min_{\boldsymbol{w}_q^{(l)},\xi_{ijk}} \quad \sum_{\{\boldsymbol{x}_{ij}:x_{ijl}=q\}} \sum_k \xi_{ijk} + \lambda\|\boldsymbol{w}_q^{(l)}\|_2^2 \tag{2.10}$$

$$\text{subject to} \quad \operatorname{sgn}(y_{ij}-k)\big((\boldsymbol{w}_q^{(l)})^T\boldsymbol{B}_{ij}^{-\boldsymbol{w}_q^{(l)}}\boldsymbol{E}^T\boldsymbol{\beta} + \beta_{0,k}(\boldsymbol{x}_{ij})\big) \geq 1-\xi_{ijk},\ \xi_{ijk}\geq 0.$$

The optimization task in (2.10) is exactly the same as (2.9), and hence can be solved by following identical optimization scheme.

When $\boldsymbol{\beta}$ and $\boldsymbol{W}$ are fixed, $\boldsymbol{\beta}_0$ can be updated by solving

$$\min_{\boldsymbol{\beta}_0,\xi_{ijk}} \quad \sum_{i,j,k} \xi_{ijk} \tag{2.11}$$

$$\text{subject to} \quad \operatorname{sgn}(y_{ij}-k)\beta_{0,k}(\boldsymbol{x}_{ij}) + \xi_{ijk} \geq 1 - \operatorname{sgn}(y_{ij}-k)\big(\boldsymbol{\beta}^T\boldsymbol{E}(B(\boldsymbol{t})_{ij}\circ\bar{\boldsymbol{w}}_{\boldsymbol{x}_{ij}})\big),$$

$$\beta_{0,K-1}(\boldsymbol{x}_{ij}) \leq \beta_{0,K-2}(\boldsymbol{x}_{ij}) \leq \cdots \leq \beta_{0,1}(\boldsymbol{x}_{ij}),\ \xi_{ijk} \geq 0.$$

It is clear that (2.11) is a standard linear programming formulation with respect to $\boldsymbol{\beta}_{\boldsymbol{x}_{ij}}$ and ξ_{ijk}, and can be efficiently solved by the popular interior-point algorithm which is available in Python package `cvxopt` [29].

The parallel block-wise coordinate descent algorithm for the proposed method is summarized in Algorithm 1. In essence, this is an implementation of the block successive convex minimization, and hence it is guaranteed to converge to a stationary point [30].

Algorithm 1:

(Initialization): Set initial values $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}_0^{(0)}, \boldsymbol{W}^{(0)})$, tuning parameters λ, the tolerance error ϵ_{tol}, and iteration $t = 1$, $\epsilon_0 = 1$;
while $\epsilon_{t-1} > \epsilon_{tol}$ **do**

(Update for $\boldsymbol{\beta}$):
Estimate $\boldsymbol{\beta}^{(t)}$ by solving (2.9) with $\boldsymbol{\beta}_0^{(t-1)}, \boldsymbol{W}^{(t-1)}$;

(Update for $\boldsymbol{W}$): Set $\epsilon_W = 1, \boldsymbol{W}^{\text{new}} = \boldsymbol{W}^{(t-1)}$;
while $\epsilon_W < \epsilon_{tol}$ **do**

$\boldsymbol{W}^{\text{old}} \leftarrow \boldsymbol{W}^{\text{new}}$
for $(l = 1, \cdots, L)$ **do**
Estimate $(\boldsymbol{w}_m^{(l)})^{\text{new}}$ by solving (2.10) in a parallel fashion with $\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}_0^{(t-1)}$ and $\boldsymbol{W}^{(l')}$ as $(\boldsymbol{W}^{(l')})^{(t)}$ for $l' = 1, \cdots, l$;
end
$\epsilon_W = \sum_{l=1}^{L} \|(\boldsymbol{W}^{\text{new}})^{(l)} - (\boldsymbol{W}^{\text{old}})^{(l)}\|_F^2 / \sum_{l=1}^{L} \|(\boldsymbol{W}^{\text{old}})^{(l)}\|_F^2$;
end
$\boldsymbol{W}^{(t)} = \boldsymbol{W}^{\text{new}}$;

(Update for $\boldsymbol{\beta}_0$):
Estimate $\boldsymbol{\beta}_0^{(t)}$ by solving (2.11) with $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{W}^{(t)}$;

Set $\epsilon_t = \sum_{l=1}^{c} \|(\boldsymbol{W}^{(l)})^{(t)} - (\boldsymbol{W}^{(l)})^{(t-1)}\|_F^2 / \sum_{l=1}^{c} \|(\boldsymbol{W}^{(l)})^{(t-1)}\|_F^2$
$+ \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2^2 / \|\boldsymbol{\beta}^{(t-1)}\|_2^2 + \|\boldsymbol{\beta}_0^{(t)} - \boldsymbol{\beta}_0^{(t-1)}\|_2^2 / \|\boldsymbol{\beta}_0^{(t-1)}\|_2^2$
end

3. Theory

This section establishes the asymptotic convergence of the proposed method in estimating the ideal sentiment function $\boldsymbol{f}^0 = (f_1^0, \ldots, f_{K-1}^0)$, which is showed to satisfy that $\text{sgn}(f_k^0) = \text{sgn}(P(y \geq k|\boldsymbol{x}, \boldsymbol{t}) - \frac{1}{2})$ for $k = 1, \ldots, K-1$ [17]. Then the regret of $\boldsymbol{f}$ is defined as

$$e(\boldsymbol{f}, \boldsymbol{f}^0) = (K-1)^{-1} \sum_{k=1}^{K-1} E\big(I(\text{sgn}(f_k(\boldsymbol{x}, \boldsymbol{t})) \neq \text{sgn}(f_k^0(\boldsymbol{x}, \boldsymbol{t}))\big).$$

Furthermore, denote $\bar{V}(y, \boldsymbol{f}(\boldsymbol{x},\boldsymbol{t})) = (K-1)^{-1}\sum_{k=1}^{K-1} V(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t}))$, and $e_V(\boldsymbol{f},\boldsymbol{f}^0) = E\bar{V}(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t})) - E\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))$. We further denote $\mathcal{F} = \{\boldsymbol{f} = (f_1,\ldots,f_{K-1}) : f_k(\boldsymbol{x},\boldsymbol{t}) = \boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}} + \beta_{0,k}(\boldsymbol{x}), \beta_{0,K-1}(\boldsymbol{x}) \le \beta_{0,K-2}(\boldsymbol{x}) \le \cdots \le \beta_{0,1}(\boldsymbol{x})\}$, where $\boldsymbol{E_t} = \boldsymbol{E}\circ(\boldsymbol{1}_p \otimes \boldsymbol{B}(\boldsymbol{t}))$. Let $n = \sum_{i=1}^{u} N_i$ denote the total number of observations, and two technical assumptions are made to quantify the asymptotic behavior of the proposed method.

Assumption A. For any $\xi_n > 0$, there exists $\boldsymbol{f}^* \in \mathcal{F}$ such that $e_V(\boldsymbol{f}^*,\boldsymbol{f}^0) \le \xi_n$.

Assumption A assures that the approximation error of $\mathcal{F}$ in approximating $\boldsymbol{f}^0$ is governed by ξ_n, which eventually will impact the asymptotic behavior of the proposed sentiment function $\widehat{\boldsymbol{f}}$.

Next, let $V^T(u) = \min(V(u), T)$, where T is chosen so that $V(\mathrm{sgn}(y-k)f_k^0(\boldsymbol{x},\boldsymbol{t})) \le T$ almost surely. In addition, we let $\bar{V}^T(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t})) = (K-1)^{-1}\sum_{k=1}^{K-1} V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t}))$ and define $e_{V^T}(\boldsymbol{f},\boldsymbol{f}^0) = E\left(\bar{V}^T(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t}))\right) - E\left(\bar{V}^T(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))\right)$.

Assumption B. There exist constants $\alpha > 0, 1 \ge \gamma \ge 0$ and $a_1, a_2 > 0$ such that for any sufficiently small $\delta_n > 0$,

$$\begin{aligned} &\sup_{\boldsymbol{f}\in\mathcal{F}_{\delta_n}} e(\boldsymbol{f},\boldsymbol{f}^0) \le a_1\delta_n^{\alpha}, \\ &\sup_{\boldsymbol{f}\in\mathcal{F}_{\delta_n}} \mathrm{Var}(\bar{V}^T(\mathrm{sgn}(y-k), f_k(\boldsymbol{x},\boldsymbol{t})) - \bar{V}(\mathrm{sgn}(y-k), f_k^0(\boldsymbol{x},\boldsymbol{t}))) \le a_2\delta_n^{\gamma}, \end{aligned} \tag{3.1}$$

where $\mathcal{F}_{\delta_n} = \{\boldsymbol{f}\in\mathcal{F} : e_{V^T}(\boldsymbol{f},\boldsymbol{f}^0) \le \delta_n\}$.

Assumption B implies the local smoothness of $e(\boldsymbol{f},\boldsymbol{f}^0)$ and $\mathrm{Var}(\bar{V}^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t})) - \bar{V}(\mathrm{sgn}(y-k)f_k^0(\boldsymbol{x},\boldsymbol{t})))$ within a neighborhood of $\boldsymbol{f}^0$. Here α and γ are determined by the joint distribution of $(\boldsymbol{x},\boldsymbol{t})$ and the loss function V. Additionally, (3.1) provides a connection between the first and second moments of $\bar{V}^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t})) - \bar{V}^T(\mathrm{sgn}(y-k)f_k^0(\boldsymbol{x},\boldsymbol{t}))$, which is essential for establishing the subsequent large deviation inequalities. In fact, Assumption B is a mild assumption and has been verified for various losses and distributions in literature [31, 32]. For example, Assumption B holds true for the hinge loss and any distribution $P(\boldsymbol{x},\boldsymbol{t})$ with $\alpha = 1$ and $\gamma = 1$ [32].

Theorem 1. *Suppose that Assumptions A-B are met. Then for the proposed sentiment function $\widehat{\boldsymbol{f}}$ defined in (2.7), there exists a constant $a_3 > 0$ such that*

$$P(e(\widehat{\boldsymbol{f}},\boldsymbol{f}^0) \ge a_1\delta_n^{2\alpha}) \le 3.5\exp\left(-a_3 n(\lambda J^*)^{2-\min(\gamma,1)}\right),$$

provided that $\delta_n^2 \ge 4\lambda J^*$, where $J^* = \min\{\|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{W}\|_F^2 : \mathrm{sgn}(f_k) = \mathrm{sgn}(f_k^0); k = 1,\ldots,K-1, \boldsymbol{f}\in\mathcal{F}\}$, $\delta_n^2 = \min(\epsilon_n^2 + 2\xi_n, 1)$, and $\epsilon_n^2 = \left(D_1 n^{-1}\log(n/D_1)\right)^{1/(2-\gamma)}$, where $D_1 = \max\{p + d\sum_{l=1}^{c} m_l, \overline{M}K\}$, $\overline{M} = \prod_{l=1}^{L} m_l$, and a_1, γ, ξ_n are defined as in Assumptions A-B.

Corollary 1. *Suppose the assumptions of Theorem 1 are met, and that* λ *satisfies* $n^{-1}(\lambda J^*)^{\min(\gamma,1)-2} = o(1)$. *Then it holds true that*

$$e(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0) = O_p(\delta_n^{2\alpha}) \text{ and } E|e(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0)| = O(\delta_n^{2\alpha}).$$

Clearly, the rate δ_n^2 is governed by both ϵ_n^2 and ξ_n, where ϵ_n^2 is determined by the complexity of $\mathcal{F}^V$ depending on the dimension of the embedded space p, the size of sentiment lexicon d, and the number of levels of all covariates $m_l; l = 1, \ldots, L$. Usually, there is a trade-off between the approximation error ξ_n and the complexity of $\mathcal{F}^V$ over the choice of $\boldsymbol{f}^0$, so as to attain the optimal convergence rate $\delta_n^{2\alpha}$.

4. Numerical experiments

In this section, we conduct a series of numerical experiments on simulated datasets and the Yelp challenge dataset to examine the performance of the proposed method. We compare it against various baseline word embedding methods in literature, including word embeddings based on Google news trained by word2vec technique [8], word embeddings based on Wikipedia trained by GloVe [9], and random word embeddings generated by multivariate normal distribution. Here the random word embeddings are used as baseline to verify the effectiveness of the other two pre-trained embeddings. Moreover, we let Google_p, Wiki_p, and Random_p denote the corresponding covariate-assisted word embeddings, whereas Google, Wiki and Random denote their corresponding baseline embeddings in (2.4), respectively.

For each method, the tuning parameters are selected via grid search over $[10^{-6}, 10^3]$, and their numerical performance is measured by MAE evaluated on a test set,

$$\text{TE}(\boldsymbol{f}) = \frac{1}{n_{\text{test}}(K-1)} \sum_{k=1}^{K-1} \sum_{i=1}^{n_{\text{test}}} I(\text{sgn}(y_i - k) \neq \text{sgn}(f_k(\boldsymbol{x}_i, \boldsymbol{t}_i)), \tag{4.1}$$

where n_{test} is the size of the test set.

4.1. Yelp challenge

The Yelp challenge dataset consists of four parts, including "business", "review", "user" and "check-in", and is publicly available at https://www.yelp.com/dataset/challenge. In "business", it contains location, latitude-longitude, averaged stars, opening hours, review counts and business categories. In "review", a specific review is composed of textual comment, stars, business, user and corresponding feedback given by other users. In "user", personal information associated with each user is given, including users' social network, starting time and elite-experience in the Yelp community. Additionally, users' behavior like votes and stars are also provided. In "checking", the counts of check-ins at each

business are provided. In fact, all capitalized words in the reviews are converted into lower case by using the `nltk` package in Python. Other pre-processing steps, including removing spaces, stop words and punctuation, are also conducted.

We implement the proposed method based on the "review" and "user" parts. The "review" part contains reviews with star ratings from 1 to 5. Due to the imbalance of classes in "stars" of reviews, we encode "1" and "2" as 1, "3" as 2, and "4" and "5" as 3. In the pre-processing step, all capitalized words are converted into lower case, stop words and punctuation are removed, and frequency vectors are constructed for each review under the bag-of-words framework against a sentiment lexicon consisting of about 6,800 positive and negative words [33] combined with 1,000 1-gram features extracted based on term frequency–inverse document frequency (TF-IDF). The "user" part provides a personal social network, number of fans, counts of "useful", counts of "cool", counts of "funny" and elite-experience. In particular, elite-experience indicates the years when the user was selected as elite for well-written reviews, high-quality tips, or a detailed personal profile. Furthermore, elite users are characterized by pertinent comments and useful tips, with which other users have resonated to cast "useful", "funny" and "cool" votes. About 3.25% of the users in the Yelp community have elite experience.

One salient difference between elite and non-elite users is their preference of wording, in particular the frequencies of sentiment words in reviews. For instances, as showed in Figure 1, "reputable", "diligence" and "abnormal" are used much more frequently by non-elite users, whereas "slut" and "catchy" are much more popular among elite users. Also, as surprising as it appears, non-elite users tend to use "reputable" in an ironic way to show that the service they receive does not live up to their expectations, such as "A reputable apartment would try to fix their errors", "I'm spending my money on an experience from

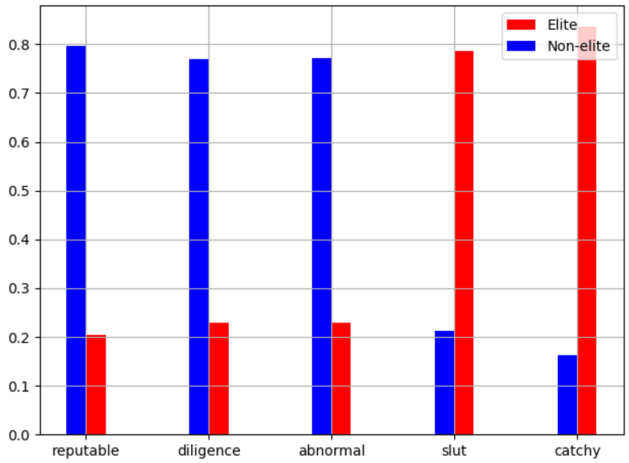


Fig 1: Relative frequency of words in average by elite and non-elite users.

a reputable Salon that has been nothing but rude and unhelpful", and so on. In sharp contrast, elite users use "reputable" as a positive comment, leading to a 3.34 stars in average among those reviews containing "reputable", while only 2.00 for by non-elite reviews.

Another interesting difference between elite and non-elite users is their preferences in giving "stars". As seen in Figure 2, the distribution of averaged stars given by elite users appears to be a normal distribution, whereas the stars given by non-elite users appear to be more disordered in that they tend to give 1-star or 5-star reviews.

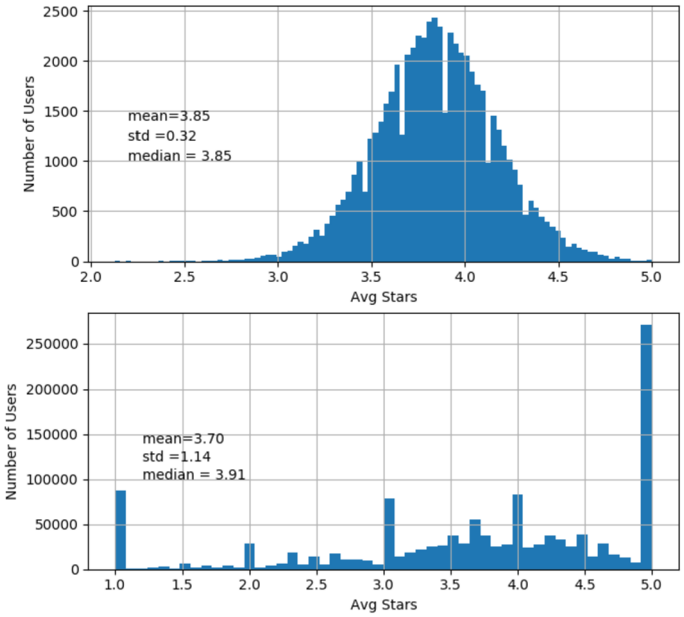


Fig 2: Histograms of averaged stars for elite and non-elite users, respectively.

Furthermore, number of feedbacks including "helpful", "cool", and "funny" tend to provides useful information about users. Users with a large number of feedbacks are popular for their objective comments, interesting expressions or humorous reviews. In fact, these three covariates are proportional to each other, and hence we only include "helpful" for application, which is converted to a binary covariate indicating whether "useful" count of the user is in the top 10 percent.

In this numerical experiment, 100,000 reviews are sampled from the Yelp challenge dataset, which are then split into three equal-sized sets used for training, validation and testing. The averaged test errors and their standard errors over 100 replications are reported in Table 1. To further evaluate the effectiveness of the proposed method in sentiment analysis, we also include the word embeddings trained by Embeddings from Language Models (ELMo) for comparison

[34], which generate word embeddings with information of context. Specifically, each review is converted to averaged word embeddings of length 1,024 obtained from ELMo based on 1 Billion Word Benchmark [35].

TABLE 1
Averaged test errors of various methods as well as their standard errors (in parentheses) over 100 replications in the Yelp dataset.

Google	Google$_p$	Wiki	Wiki$_p$	Random	Random$_p$	ELMo
0.1492	0.1407	0.1534	0.1410	0.1833	0.1443	0.1423
(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0003)

Table 1 shows that the proposed method is able to improve the performance of Google, Wiki and Random by incorporating covariates, with improvement ranging from 5.7% to 21.3%. Even though word embeddings by ELMo appears to be more accurate than those trained by Word2Vec and GloVe, the proposed method yields the best performance on the word embeddings on Google news, showing that the proposed method is capable of learning covariate-varying word embeddings to improve prediction accuracy. Interestingly, when using covariates to adjust random embeddings which does not make use of semantic information at all, the improvement is much more significant and the performance of Random$_p$ is almost comparable to Google$_p$ and Wiki$_p$, showing that the proposed method is capable of training a word embeddings adaptive to prediction task. To verify the significance of the improvement, we further conduct t-tests between the proposed method on three word embeddings with their corresponding baselines, and the best performer with ELMo.

TABLE 2
P-values for various pair-wise t-tests.

Google vs Google$_p$	Wiki vs Wiki$_p$	Random vs Random$_p$	Google$_p$ vs ELMo
0.0000	0.0000	0.0000	0.0003

As shown in Table 2, the improvements on three baseline word embeddings are statistically significant, showing that a uniform word embeddings in sentiment analysis may lead to sub-optimal performance, and hence entails adjusting effect from covariates. Additionally, Google$_p$ also outperforms ELMo with statistically significant improvement, suggesting the proposed framework is competitive in sentiment analysis.

4.2. Simulations

We further verify the effectiveness of our proposed method under the assumed model in 2.4 with various numbers of covariates and degrees of covariate effect. The simulated examples are generated as follows. We first choose 100 words from the sentiment lexicon [33] and obtain corresponding word embeddings $\boldsymbol{E}$ based on GoogleNews. Then we generate $\boldsymbol{w}_j = (\mathbf{1}_{100-r}, \widetilde{\boldsymbol{w}}_j); j = 1, \ldots, m$, where $\widetilde{\boldsymbol{w}}_j \in$

$\mathbb{R}^r$ with each component generated from $unif(-1,1)$ and r adjusts the degree of covariate effect. Then we generate $(x_i, \boldsymbol{b}_i); i = 1, \ldots, n$, where $\boldsymbol{b}_i \in \mathbb{R}^{100}$ is a sequence of integers denoting word frequencies with each element of $\boldsymbol{b}_i$ generated independently from $Pois(1)$, and x_i is uniformly chosen from $\{1, \ldots, m\}$. The sentiment level y_i is generated via $y_i = \max\{k : \boldsymbol{\beta}^T \boldsymbol{E}(\boldsymbol{b}_i \circ \boldsymbol{w}_{x_i}) + \beta_{0,k}(\boldsymbol{x}_i) \leq 0\}$, where $\boldsymbol{\beta}$ is generated from $N(\boldsymbol{0}, \boldsymbol{I}_{300})$ and $\boldsymbol{\beta}_0$ is set to generate K equal-sized classes.

Under this data generation scheme, we consider various cases with (n, m, K) = (2000,2,5), (2000,4,5), (4000,2,5) and (4000,4,5) and $r = 20, 40, 60, 80$, respectively. In each case, we split the dataset into training, validation and test sets with ratio 1:1:1. The averaged test errors of various methods over 100 replications are summarized in Table 3.

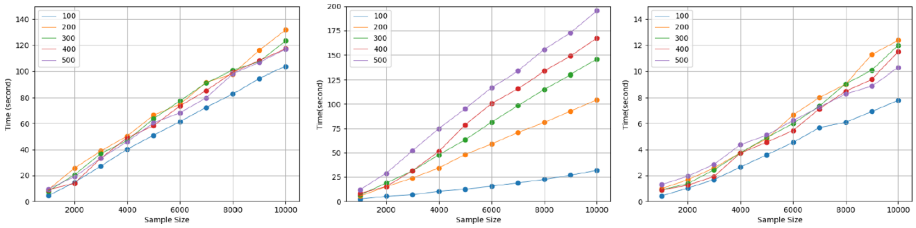
TABLE 3
Averaged test errors of various methods as well as their standard errors (in parentheses) over 100 replications.

	(n, m, K)	Google_p	Google
$r = 20$	(2000,2,5)	0.0679(0.0030)	0.1122(0.0030)
	(2000,4,5)	0.0794(0.0036)	0.1261(0.0027)
	(4000,2,5)	0.0659(0.0035)	0.1135(0.0034)
	(4000,4,5)	0.0664(0.0031)	0.1246(0.0024)
$r = 40$	(2000,2,5)	0.0632(0.0028)	0.1446(0.0024)
	(2000,4,5)	0.0697(0.0024)	0.1687(0.0021)
	(4000,2,5)	0.0503(0.0026)	0.1340(0.0026)
	(4000,4,5)	0.0583(0.0024)	0.1575(0.0018)
$r = 60$	(2000,2,5)	0.0497(0.0008)	0.1593(0.0024)
	(2000,4,5)	0.0612(0.0008)	0.1936(0.0024)
	(4000,2,5)	0.0443(0.0015)	0.1559(0.0023)
	(4000,4,5)	0.0495(0.0013)	0.1957(0.0015)
$r = 80$	(2000,2,5)	0.0420(0.0008)	0.1830(0.0024)
	(2000,4,5)	0.0539(0.0008)	0.2311(0.0013)
	(4000,2,5)	0.0380(0.0012)	0.1766(0.0020)
	(4000,4,5)	0.0427(0.0007)	0.2273(0.0015)

As shown in Table 3, the proposed method outperforms its baseline embedding method under all settings, showing that our proposed method is capable of enhancing the sentiment performance by incorporating covariate effect into word embeddings. The advantage becomes more substantial when r gets larger and $\boldsymbol{w}_j$'s become more different, showing that employing a homogeneous word embeddings may yield poor performance when embeddings varies with some covariates.

To verify the efficiency of the proposed method, we examine the computing time of three sub-optimization tasks, where the sample size increases from 1,000 to 10,000, or the dictionary size increases from 100 to 500. The averaged computing time over 50 replications of three sub-optimization tasks under all settings are reported in Figure 3.

As shown in Figure 3, the averaged computing time of three optimizations tasks are all linearly proportional to sample size. This is due to the fact that the



(a) Updating $\boldsymbol{\beta}$ (b) Updating $\boldsymbol{W}$ (c) Updating $\boldsymbol{\beta}_0$

Fig 3: The averaged computing time of 50 replications of three sub-optimization tasks.

optimization tasks for $\boldsymbol{\beta}$ and $\boldsymbol{W}$ resemble linear SVM in nature except varying intercepts, and hence can be solved efficiently by a dual coordinate descent algorithm as in `Liblinear` [28]. Moreover, the computational time for updating $\boldsymbol{W}$ also depends on the size of dictionary, whereas those for updating $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ appear to be less affected.

5. Summary

This article proposes a flexible framework for covariate-assisted sentiment analysis by incorporating covariates into word embeddings to improve prediction accuracy. Specifically, the proposed method admits varying document representations over covariate information, such as gender, education level and so on. This is also equivalent to admitting varying sentiment functions over levels of covariates, and then endows the proposed method with the ability to capture distinctions in wording and sentiment derived from covariate. Additionally, we propose an scalable block-wise coordinate descent algorithm to solve the resultant optimization task. We also establish the asymptotic properties of the proposed method, which provides a theoretical guarantee of its convergence to the ideal sentiment function. Note that even though our proposed method is formulated under the ordinal regression framework, the key idea of integrating covariates can be employed in other models.

Acknowledgments

The authors are also grateful to the editor, associate editor, and anonymous reviewers for their constructive comments and suggestions, which have significantly improved the manuscript.

Appendix

Lemma 2. *The solution to* (7) *remains the same as long as* $\lambda_1\lambda_2^L$ *stays the same.*

Proof of Lemma 2. Let

$$G(\boldsymbol{\beta},\boldsymbol{W},\boldsymbol{\beta}_0;\lambda_1,\lambda_2)=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{K-1}V\big(\operatorname{sgn}(y_{ij}-k)\boldsymbol{f}_k(\boldsymbol{x}_{ij},\boldsymbol{t}_{ij})\big)+\lambda_1\|\boldsymbol{\beta}\|^2+\lambda_2\|\boldsymbol{W}\|_F^2,$$

where $\|\boldsymbol{W}\|_F=\sum_{l=1}^{L}\|\boldsymbol{W}^{(l)}\|_F^2$, and then

$$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{\beta}}_0)=\operatorname*{argmin}_{\boldsymbol{\beta},\boldsymbol{W},\boldsymbol{\beta}_0}G(\boldsymbol{\beta},\boldsymbol{W},\boldsymbol{\beta}_0,\lambda_1,\lambda_2).$$

Consider another tuning parameter pair $(\tilde{\lambda}_1,\tilde{\lambda}_2)$ satisfying $\lambda_1\lambda_2^L=\tilde{\lambda}_1\tilde{\lambda}_2^L$, and denote $(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{W}},\tilde{\boldsymbol{\beta}}_0)=\operatorname*{argmin}_{\boldsymbol{\beta},\boldsymbol{W},\boldsymbol{\beta}_0}G(\boldsymbol{\beta},\boldsymbol{W},\boldsymbol{\beta}_0;\tilde{\lambda}_1,\tilde{\lambda}_2)$. Then we have

$$G(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{\beta}}_0;\lambda_1,\lambda_2)=G(\sqrt{\lambda_1/\tilde{\lambda}_1}\widehat{\boldsymbol{\beta}},\sqrt{\lambda_2/\tilde{\lambda}_2}\widehat{\boldsymbol{W}},\widehat{\boldsymbol{\beta}}_0,\tilde{\lambda}_1,\tilde{\lambda}_2).$$

It hence follows that $(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{W}},\tilde{\boldsymbol{\beta}}_0)=(\sqrt{\lambda_1/\tilde{\lambda}_1}\widehat{\boldsymbol{\beta}},\sqrt{\lambda_2/\tilde{\lambda}_2}\widehat{\boldsymbol{W}},\widehat{\boldsymbol{\beta}}_0)$, which leads to

$$\begin{aligned}&\tilde{\boldsymbol{\beta}}^T\boldsymbol{E}(\boldsymbol{B}(\boldsymbol{t})\circ\overline{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}})\\&=\sqrt{\lambda_1/\tilde{\lambda}_1}\widehat{\boldsymbol{\beta}}^T\boldsymbol{E}(\boldsymbol{B}(\boldsymbol{t})\circ\sqrt{\lambda_2/\tilde{\lambda}_2}\widehat{\boldsymbol{w}}_{x_1}^{(1)}\circ\sqrt{\lambda_2/\tilde{\lambda}_2}\circ\cdots\circ\sqrt{\lambda_2/\tilde{\lambda}_2}\widehat{\boldsymbol{w}}_{x_L}^{(L)})\\&=\sqrt{\lambda_1\lambda_2^L/(\tilde{\lambda}_1\tilde{\lambda}_2^L)}\widehat{\boldsymbol{\beta}}^T\boldsymbol{E}(\boldsymbol{B}(\boldsymbol{t})\circ\overline{\widehat{\boldsymbol{w}}}_{\boldsymbol{x}}).\end{aligned}$$

The desired result then follows immediately after the fact that $\lambda_1\lambda_2^L=\tilde{\lambda}_1\tilde{\lambda}_2^L$. □

Lemma 3. *Let* $C_1(\tau)=d^{\frac{1}{2}}(\tau J^*)^{\frac{L}{2}+1}/L^{\frac{L}{2}}\|\boldsymbol{E}_{\boldsymbol{t}}\|_F$, *then for any* $\beta_{0,k}(\boldsymbol{x})$ *with* $|\beta_{0,k}(\boldsymbol{x})|\geq C_1(\tau)+T+1$, *we have*

$$\begin{aligned}&V^T\big(\operatorname{sgn}(y-k)(\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t}}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\beta_{0,k}(\boldsymbol{x}))\big)\\&=\begin{cases}TI(\operatorname{sgn}(y-k)\neq 1), & \text{if } \beta_{0,k}(\boldsymbol{x})\geq C_1(\tau)+T+1;\\ TI(\operatorname{sgn}(y-k)=1), & \text{if } \beta_{0,k}(\boldsymbol{x})\leq -C_1(\tau)-T-1,\end{cases}\end{aligned}\tag{5.1}$$

for any $(\boldsymbol{\beta},\boldsymbol{W})\in\{(\boldsymbol{\beta},\boldsymbol{W}):\|\boldsymbol{\beta}\|^2+\|\boldsymbol{W}\|_F^2\leq J^*\tau\}$.

Proof of Lemma 3. We first show that $|\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t}}\bar{\boldsymbol{w}}_{\boldsymbol{x}}|\leq C_1(\tau)$, for any $(\boldsymbol{\beta},\boldsymbol{W})\in\{(\boldsymbol{\beta},\boldsymbol{W}):\|\boldsymbol{\beta}\|^2+\|\boldsymbol{W}\|_F^2\leq J^*\tau\}$. Particularly,

$$\begin{aligned}&|\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t}}\bar{\boldsymbol{w}}_{\boldsymbol{x}}|=|(\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t},1},\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t},2},\ldots,\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t},d})\bar{\boldsymbol{w}}_{\boldsymbol{x}}|\\&\leq\|\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t}}\|_2\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}\|_2=\left(\sum_{j=1}^{d}(\boldsymbol{\beta}^T\boldsymbol{E}_{\boldsymbol{t},j})^2\right)^{1/2}\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}\|_2\\&\leq\left(\sum_{j=1}^{d}\|\boldsymbol{\beta}\|_2^2\|\boldsymbol{E}_{\boldsymbol{t},j}\|_2^2\right)^{1/2}\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}\|_2=\|\boldsymbol{\beta}\|_2\|\boldsymbol{E}_{\boldsymbol{t}}\|_F\|\boldsymbol{w}_{\boldsymbol{x}_1}^{(1)}\circ\boldsymbol{w}_{\boldsymbol{x}_2}^{(2)}\cdots\circ\boldsymbol{w}_{\boldsymbol{x}_L}^{(L)}\|_2\end{aligned}$$

$$
\begin{aligned}
&=\|\boldsymbol{\beta}\|_2\|\boldsymbol{E_t}\|_F\left(\sum_{j=1}^{d}(\prod_{i=1}^{L}\boldsymbol{w}_{x_i,j}^{(i)})^2\right)^{\frac{1}{2}} \le \|\boldsymbol{\beta}\|_2\|\boldsymbol{E_t}\|_F\left(\sum_{j=1}^{d}(\sum_{i=1}^{L}\frac{(\boldsymbol{w}_{x_i,j}^{(i)})^2}{L})^{L}\right)^{\frac{1}{2}}\\
&\le \left(\|\boldsymbol{\beta}\|_2^2+\|\boldsymbol{W}\|_F^2\right)^{1/2}\|\boldsymbol{E_t}\|_F d^{\frac{1}{2}}\left(\frac{\|\boldsymbol{W}\|_F^2}{L}\right)^{\frac{L}{2}}\\
&\le \left(\|\boldsymbol{\beta}\|_2^2+\|\boldsymbol{W}\|_F^2\right)^{1/2}\|\boldsymbol{E_t}\|_F d^{\frac{1}{2}}\left(\frac{\tau J^*}{L}\right)^{\frac{L}{2}} \le d^{\frac{1}{2}}\frac{(\tau J^*)^{\frac{L}{2}+1}}{L^{\frac{L}{2}}}\|\boldsymbol{E_t}\|_F = C_1(\tau),
\end{aligned}
$$

where $\boldsymbol{E}_{\boldsymbol{t},j}$ is the j-th column of $\boldsymbol{E_t}$, the first inequality follows from the Cauchy-Schwarz inequality, the third inequality follows from the inequality of arithmetic and geometric means.

Next we verify only (5.1) for $\beta_{0,k}(\boldsymbol{x}) \ge C_1(\tau)+T+1$, and it can be verified similarly for $\beta_{0,k}(\boldsymbol{x}) \le -C_1(\tau)-T-1$. For any $\beta_{0,k}(\boldsymbol{x}) \ge C_1(\tau)+T+1$, when $\operatorname{sgn}(y-k)=1$ we have

$$
0 \le V^T\big(\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\beta_{0,k}(\boldsymbol{x})\big) \le V^T\big(-|\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}|+C_1(\tau)+T+1\big) \le 0,
$$

where both inequalities follow from the non-increasing property of $V^T(\cdot)$. When $\operatorname{sgn}(y-k)=-1$, we have

$$
T \ge V^T\big(-\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}-\beta_{0,k}(\boldsymbol{x})\big) \ge V^T(|\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}|-C_1(\tau)-T-1) \ge T.
$$

The desirable result then follows. □

Lemma 4. *Let* $C_2(\tau)=\|\boldsymbol{E_t}\|_F\sum_{k=0}^{L}M(\sqrt{J^*\tau})^k$. *For any* $f_k,\tilde{f}_k \in \mathcal{F}_k(\tau)$, *we have*

$$
|f_k(\boldsymbol{x},\boldsymbol{t})-\tilde{f}_k(\boldsymbol{x},\boldsymbol{t})| \le C_2(\tau)\left(\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_2+\|\boldsymbol{W}-\tilde{\boldsymbol{W}}\|_F\right)+|\beta_{0,k}(\boldsymbol{x})-\tilde{\beta}_{0,k}(\boldsymbol{x})|.
$$

Proof of Lemma 4. Let $f_k,\tilde{f}_k \in \mathcal{F}_k(\tau)$, then we have

$$
\begin{aligned}
&|f_k(\boldsymbol{x},\boldsymbol{t})-\tilde{f}_k(\boldsymbol{x},\boldsymbol{t})|\\
=&|\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\beta_{0,k}(\boldsymbol{x})-\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}-\tilde{\beta}_{0,k}(\boldsymbol{x})|\\
=&|\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\beta_{0,k}(\boldsymbol{x})-\tilde{\boldsymbol{\beta}}^T\boldsymbol{E_t}\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}+\tilde{\boldsymbol{\beta}}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}-\tilde{\boldsymbol{\beta}}^T\boldsymbol{E_t}\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}-\tilde{\beta}_{0,k}(\boldsymbol{x})|\\
\le&|(\boldsymbol{\beta}-\boldsymbol{\beta}')^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\tilde{\boldsymbol{\beta}}\boldsymbol{E_t}(\bar{\boldsymbol{w}}_{\boldsymbol{x}}-\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}})|+|\beta_{0,k}(\boldsymbol{x})-\tilde{\beta}_{0,k}(\boldsymbol{x})|\\
\le&\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_2\|\boldsymbol{E_t}\|_F\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}\|_2+\|\tilde{\boldsymbol{\beta}}\|_2\|\boldsymbol{E_t}\|_F\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}-\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}\|_2+|\beta_{0,k}(\boldsymbol{x})-\tilde{\beta}_{0,k}(\boldsymbol{x})|
\end{aligned}
$$

By the definition of $\mathcal{F}_k(\tau)$, $\boldsymbol{\beta}$ is bounded by $J^*\tau$. It suffices to bound $\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}\|_2$ and $\|\bar{\boldsymbol{w}}_{\boldsymbol{x}}-\bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}\|_2$ respectively. For $\bar{\boldsymbol{w}}_{\boldsymbol{x}}$, by inequality of arithmetic and geometric means, we have

$$
\|\boldsymbol{w}_{\boldsymbol{x}_1}^{(1)}\circ\boldsymbol{w}_{\boldsymbol{x}_2}^{(2)}\cdots\circ\boldsymbol{w}_{\boldsymbol{x}_L}^{(L)}\|_2^2=\sum_{j=1}^{d}(\prod_{i=1}^{L}w_{x_i,j}^{(i)})^2 \le \sum_{j=1}^{d}\left(\sum_{i=1}^{L}\frac{(w_{x_i,j}^{(i)})^2}{L}\right)^{L}
$$

$$\leq d\left(\frac{J(\boldsymbol{W})}{L}\right)^L \leq d\left(\frac{J^*\tau}{L}\right)^L \equiv M$$

Similarly for $\|\bar{\boldsymbol{w}}_{\boldsymbol{x}} - \bar{\tilde{\boldsymbol{w}}}_{\boldsymbol{x}}\|_2$, for any integer $1 \leq m \leq L$, we have

$$\begin{aligned}
&\|\boldsymbol{w}_{x_1}^{(1)} \circ \boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{\boldsymbol{x}_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_1}^{(1)} \circ \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{\boldsymbol{x}_m}^{(m)}\|_2 \\
\leq &\|\boldsymbol{w}_{x_1}^{(1)} \circ \boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{\boldsymbol{x}_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_1}^{(1)} \circ \boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{\boldsymbol{x}_m}^{(m)} + \\
&\quad \tilde{\boldsymbol{w}}_{x_1}^{(1)} \circ \boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{\boldsymbol{x}_m}^{(m)} - \tilde{\boldsymbol{w}_{x_1}}^{(1)} \circ \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{x_m}^{(m)}\| \\
= &\|(\boldsymbol{w}_{x_1}^{(1)} - \tilde{\boldsymbol{w}}_{x_1}^{(1)}) \circ \boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{x_m}^{(m)} + \\
&\quad \tilde{\boldsymbol{w}}_{x_1}^{(1)} \circ (\boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{x_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{x_m}^{(m)})\| \\
\leq &\|\boldsymbol{w}_{x_1}^{(1)} - \tilde{\boldsymbol{w}}_{x_1}^{(1)}\|_2 \|\boldsymbol{w}_{x_2}^{(2)} \cdots \circ w_{x_m}^{(m)}\|_2 + \\
&\quad \|\tilde{\boldsymbol{w}}_{x_1}^{(1)} \circ (\boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{x_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{x_m}^{(m)})\| \\
\leq &d\|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F \left(\frac{J^*\tau}{L}\right)^L + \|\tilde{\boldsymbol{w}}_{x_1}^{(1)}\|_2 \|\boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{x_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{x_m}^{(m)})\|_2 \\
\leq &d\|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F \left(\frac{J^*\tau}{L}\right)^L + \sqrt{J^*\tau}\|\boldsymbol{w}_{x_2}^{(2)} \cdots \circ \boldsymbol{w}_{x_m}^{(m)} - \tilde{\boldsymbol{w}}_{x_2}^{(2)} \cdots \circ \tilde{\boldsymbol{w}}_{x_m}^{(m)})\|_2 \\
\leq &\sum_{k=1}^{L} M(\sqrt{J^*\tau})^k \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F
\end{aligned}$$

where the last inequality follows by applying similar steps iteratively. This completes the proof. □

Proof of Theorem 1. By Assumption B, we have $\{e(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0) \geq a_1\delta_n^{2\alpha}\} \subset \{e_{V^T}(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0) \geq \delta_n^2\}$, and thus it suffices to bound $P(e_{V^T}(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0) \geq \delta_n^2)$. Since $\widehat{f}$ is a global minimizer of (7) of the manuscript, it yields that

$$\begin{aligned}
&P(e(\widehat{\boldsymbol{f}}, \boldsymbol{f}^0) \geq a_1\delta_n^{2\alpha}) \\
\leq &P\left(\sup_{e_{V^T}(\boldsymbol{f}, \boldsymbol{f}^0) \geq \delta_n^2} \sum_{i,j} \tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) - \sum_{i,j} \tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) \geq 0\right) \equiv I,
\end{aligned} \tag{5.2}$$

where $\tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) = \bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) + \lambda J(\boldsymbol{\beta}) + \lambda J(\boldsymbol{W})$. Next we define a scaled empirical process as

$$\begin{aligned}
&E_n(\tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))) - \tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}))) \\
= &\frac{1}{n}\sum_{i,j}(\tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))) - \tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))) \\
&- E(\tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))) - \tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}))).
\end{aligned}$$

Let $A_{ij} = \{\boldsymbol{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^T}(\boldsymbol{f}, \boldsymbol{f}^0) \leq 2^i\delta_n^2, 2^{j-1}J^* \leq J(\boldsymbol{\beta}) + J(\boldsymbol{W}) \leq 2^j J^*\}$, and $A_{i0} = \{\boldsymbol{f} \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq e_{V^T}(\boldsymbol{f}, \boldsymbol{f}^0) \leq 2^i\delta_n^2, J(\boldsymbol{\beta}) + J(\boldsymbol{W}) \leq J^*\}$

for $i, j \geq 1$. Then we have

$$
\begin{aligned}
I \leq & \sum_{i,j\geq 1} P\Big(\sup_{\boldsymbol{f}\in A_{ij}} \sum_{i,j} \big(\tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) - \tilde{\bar{V}}(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))\big) \geq 0\Big) \\
& + \sum_{i=1}^{\infty} P\Big(\sup_{\boldsymbol{f}\in A_{i0}} \sum_{i,j} \big(\tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) - \tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))\big) \geq 0\Big) \\
= & \sum_{i,j\geq 1} P\Big(\sup_{\boldsymbol{f}\in A_{ij}} \frac{1}{n} \sum_{i,j} \big(\tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) - \tilde{\bar{V}}(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))\big) + \\
& \qquad M(i,j) \geq M(i,j)\Big) \\
& + \sum_{i=1}^{\infty} P\Big(\sup_{\boldsymbol{f}\in A_{i0}} \frac{1}{n} \sum_{i,j} \big(\tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}^*(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})) - \tilde{\bar{V}}^T(y_{ij}, \boldsymbol{f}(\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}))\big) + \\
& \qquad M(i,0) \geq M(i,0)\Big) \\
\leq & \sum_{i,j\geq 1} P\Big(\sup_{\boldsymbol{f}\in A_{ij}} E_n\big(\bar{V}^T(y, \boldsymbol{f}^*(\boldsymbol{x})) - \bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}))\big) \geq M(i,j)\Big) \\
& + \sum_{i=1}^{\infty} P\Big(\sup_{\boldsymbol{f}\in A_{i0}} E_n\big(\bar{V}^T(y, \boldsymbol{f}^*(\boldsymbol{x})) - \bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}))\big) \geq M(i,0)\Big) \equiv I_1 + I_2.
\end{aligned}
$$

To bound I, we proceed to bound I_1 and I_2 respectively. By Assumption A, we have

$$
\begin{aligned}
& E\left(\bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - E\bar{V}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))\right) \\
= & E\left(\bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - E\bar{V}^T(y, \boldsymbol{f}^0(\boldsymbol{x}, \boldsymbol{t}))\right) \\
& + E\left(\bar{V}^T(y, \boldsymbol{f}^0(\boldsymbol{x}, \boldsymbol{t})) - E\bar{V}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))\right) \\
\geq & e_{V^T}(\boldsymbol{f}, \boldsymbol{f}^0) - \delta_n^2/2,
\end{aligned}
$$

where the last inequality follows from Assumption A. Then for any $\boldsymbol{f} \in A_{ij}; i \geq 1, j \geq 1$,

$$
E(\tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - \tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))\big) \geq (2^{i-1} - \frac{1}{2})\delta_n^2 + \lambda(2^{j-1} - 1)J^* \equiv M(i,j),
$$

and for any $\boldsymbol{f} \in A_{i0}; i \geq 1$,

$$
E(\tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - \tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))\big) \geq (2^{i-1} - \frac{1}{2})\delta_n^2 - \lambda J^* \geq 2^{i-3}\delta_n^2 \equiv M(i,0).
$$

For the variance, by Assumptions A and B,

$$
\begin{aligned}
& \sup_{\boldsymbol{f}\in A_{ij}} \mathrm{Var}\Big(\tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - \tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{t}))\Big) \\
= & \sup_{\boldsymbol{f}\in A_{ij}} \mathrm{Var}\Big(\bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t})) - \bar{V}(y, \boldsymbol{f}^0(\boldsymbol{x}, \boldsymbol{t}))
\end{aligned}
$$

$$
\begin{aligned}
&+\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))-\bar{V}^T(y,\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{t}))\Big)\\
=&\sup_{\boldsymbol{f}\in A_{ij}}\mathrm{Var}\Big(\bar{V}^T(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t}))-\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))\Big)\\
&+\mathrm{Var}\Big(\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))-\bar{V}^T(y,\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{t}))\Big)\\
&+2\mathrm{Cov}(\bar{V}^T(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t}))-\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t})),\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))-\bar{V}^T(y,\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{t})))\\
\le&2\sup_{\boldsymbol{f}\in A_{ij}}\mathrm{Var}\Big(\bar{V}^T(y,\boldsymbol{f}(\boldsymbol{x},\boldsymbol{t}))-\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))\Big)+\\
&2\mathrm{Var}\Big(\bar{V}(y,\boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))-\bar{V}^T(y,\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{t}))\Big)\\
\le&2a_2(M(i,j))^\gamma+2a_2\xi_n^\gamma\le 4a_2(M(i,j))^\gamma\equiv v(i,j).
\end{aligned}
$$

To bound I_1 and I_2, we first denote $J^*=J(\boldsymbol{\beta}^*)+J(\boldsymbol{W}^*)$, and $\mathcal{F}(\tau)=\{\boldsymbol{f}=(f_1,f_2,\ldots,f_{K-1}):f_k\in\mathcal{F}_k(\tau)\}$ with

$$
\mathcal{F}_k(\tau)=\{f_k(\boldsymbol{x},\boldsymbol{t})=\boldsymbol{\beta}^T\boldsymbol{E_t}\bar{\boldsymbol{w}}_{\boldsymbol{x}}+\beta_{0,k}(\boldsymbol{x}):J(\boldsymbol{\beta})+J(\boldsymbol{W})\le\tau J^*\}.
$$

Then the associated spaces of the loss function are

$$
\begin{aligned}
\mathcal{F}^V(\tau)&=\{(K-1)^{-1}\sum_{k=1}^{K-1}V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t})):\boldsymbol{f}\in\mathcal{F}(\tau)\},\\
\mathcal{F}_k^V(\tau)&=\{V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{\tau})):f_k\in\mathcal{F}_k(\tau)\}.
\end{aligned}
$$

Next, we employ the L_2-metric entropy to measure the cardinality of $\mathcal{F}^V(\tau)$, denoted as $H(\epsilon,\mathcal{F}^V(\tau))$. For $\bar{V}(y,f(\boldsymbol{x},\boldsymbol{t})),\tilde{\bar{V}}(y,f(\boldsymbol{x},\boldsymbol{t}))\in\mathcal{F}^V(\tau)$, we have

$$
\begin{aligned}
&\|\bar{V}(y,f(\boldsymbol{x},\boldsymbol{t}))-\tilde{\bar{V}}(y,f(\boldsymbol{x},\boldsymbol{t}))\|_2\\
\le&\frac{1}{K-1}\sum_{k=1}^{K-1}\|V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t}))-V^T(\mathrm{sgn}(y-k)\tilde{f}_k(\boldsymbol{x},\boldsymbol{t}))\|_2. \qquad (5.3)
\end{aligned}
$$

Therefore, it suffices to compute the metric entropy of $\mathcal{F}_k^V(\tau)$.

Let $\overline{M}=\prod_{k=1}^L m_k$, and then $\mathcal{F}_k^V(\tau)$ can be expressed as the union of $4^{\overline{M}}$ subspaces

$$
\mathcal{F}_k^V(\tau)=\bigcup_{l,r\le 2^{\overline{M}}}\mathcal{F}_k^V(\tau,\boldsymbol{p}_l,\boldsymbol{q}_r), \qquad (5.4)
$$

where $\mathcal{F}_k^V(\tau,\boldsymbol{p}_l,\boldsymbol{q}_r)$ is defined as

$$
\begin{aligned}
\mathcal{F}_k^V(\tau,\boldsymbol{p}_l,\boldsymbol{q}_r)=\{&V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t})),\ f_k\in\mathcal{F}_k(\tau),\\
&p_{lu}|\beta_{0,k}(\boldsymbol{x}^{(u)})|\le p_{lu}(C_1(\tau)+T+1),q_{ru}\beta_{0,k}(\boldsymbol{x}^{(u)})\le 0;\ u=1,\ldots,\overline{M}\},
\end{aligned}
$$

where $\boldsymbol{p}_l,\boldsymbol{q}_r;\ l,r=1,\ldots,2^{\overline{M}}$ taking all possible values in $\{-1,1\}^{\overline{M}}$. Next, we proceed to compute the entropy of $\mathcal{F}_k^V(\tau,\boldsymbol{p}_l,\boldsymbol{q}_r)$. For any two functions $V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t})),V^T(\mathrm{sgn}(y-k)\tilde{f}_k(\boldsymbol{x},\boldsymbol{t}))\in\mathcal{F}_k^V(\tau,\boldsymbol{p}_l,\boldsymbol{q}_r)$, we have

$$
|V^T(\mathrm{sgn}(y-k)f_k(\boldsymbol{x},\boldsymbol{t}))-V^T(\mathrm{sgn}(y-k)\tilde{f}_k(\boldsymbol{x},\boldsymbol{t}))|
$$

$$
\begin{aligned}
&\le \sum_{u=1}^{\overline{M}} I(\boldsymbol{x}=\boldsymbol{x}^{(u)})(I(p_{lu}=-1)\cdot(I(q_{ru}=1)\cdot 0 + I(q_{ru}=-1)\cdot 0)\\
&\quad + I(p_{lu}=1)|f_k(\boldsymbol{x},\boldsymbol{t}) - \tilde{f}_k(\boldsymbol{x},\boldsymbol{t})|)\\
&\le \sum_{u=1}^{\overline{M}} I(\boldsymbol{x}=\boldsymbol{x}^{(u)}) I(p_{lu}=1)\Big(C_2(\tau)\big(\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_2 + \|\boldsymbol{W}-\tilde{\boldsymbol{W}}\|_F\big)\\
&\quad + |\beta_{0,k}(\boldsymbol{x}) - \tilde{\beta}_{0,k}(\boldsymbol{x})|\Big)\\
&\le C_2(\tau)\sqrt{2\big(\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_2^2 + \|\boldsymbol{W}-\tilde{\boldsymbol{W}}\|_F^2\big)}\\
&\quad + \sum_{u=1}^{\overline{M}} I(\boldsymbol{x}=\boldsymbol{x}^{(u)}) I(p_{lu}=1)|\beta_{0,k}(\boldsymbol{x}) - \tilde{\beta}_{0,k}(\boldsymbol{x})|,
\end{aligned}
$$

where the first and second inequalities follow from Lemma 3 and Lemma 4 respectively. Combined with (5.3), we have

$$
\begin{aligned}
&\|\bar{V}(y,f(\boldsymbol{x},\boldsymbol{t})) - \tilde{\bar{V}}(y,f(\boldsymbol{x},\boldsymbol{t}))\|_2\\
\le& C_2(\tau)\sqrt{2\big(\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_2^2 + \|\boldsymbol{W}-\tilde{\boldsymbol{W}}\|_F^2\big)}\\
&+ \frac{1}{K-1}\sum_{k=1}^{K-1}\sum_{u=1}^{\overline{M}} I_{\{\boldsymbol{x}=\boldsymbol{x}^{(u)}\}} I_{\{p_{lu}=1\}} |\beta_{0,k}(\boldsymbol{x}) - \tilde{\beta}_{0,k}(\boldsymbol{x})|.
\end{aligned}
$$

Let $\mathcal{B}(\tau J^*) = \{(\boldsymbol{\beta},\boldsymbol{W}) : \|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{W}\|_F^2 \le \tau J^*\}$, and the corresponding $\epsilon/(2\sqrt{2}C_2(\tau))$-covering number with L_2-norm is $O\big((\frac{2C_2(\tau))\sqrt{2\tau J^*}}{\epsilon})^{(p+d\sum_{l=1}^{L} m_l)}\big)$. For $u = 1,\ldots,\overline{M}$ we define the set

$$
\begin{aligned}
&\{(\beta_{0,1}(\boldsymbol{x}^{(u)}), \beta_{0,2}(\boldsymbol{x}^{(u)}), \ldots, \beta_{0,K-1}(\boldsymbol{x}^{u})) :\\
&\quad -(C_1(\tau)+T+1) \le \beta_{0,1}(\boldsymbol{x}^{(u)}) \le \ldots \le \beta_{0,K-1}(\boldsymbol{x}^{u}) \le C_1(\tau)+T+1\},
\end{aligned}
$$

and its corresponding number of $\epsilon/2$-covering balls with L_1-norm is $C_{\tilde{N}+K-2}^{K-1}$, where $\tilde{N} = \lceil \frac{4(C_1(\tau)+T+1)}{\epsilon} \rceil$ with $\lceil\cdot\rceil$ being the ceiling function.

Finally, the metric entropy of $\mathcal{F}^V(\tau)$ can be upper-bounded as follows:

$$
\begin{aligned}
H(\epsilon,\mathcal{F}^V(\tau)) \le& \max\Big\{O\Big((p+d\sum_{l=1}^{L} m_l)\log\frac{2C_2(\tau)(2\tau J^*)^{1/2}}{\epsilon}\\
&+ \overline{M}K\log(\frac{4(C_1(\tau)+T+1)}{\epsilon}+K) - \overline{M}\log K!\Big), 1\Big\}\\
\le& \max\Big\{O\Big((p+d\sum_{l=1}^{L} m_l)\log\frac{2C_2(\tau)(2\tau J^*)^{1/2}}{\epsilon}\\
&+ \overline{M}K\log(\frac{4(C_1(\tau)+T+1)}{K\epsilon}+1)\Big), 1\Big\}
\end{aligned}
$$

$$\leq \max\{O(D_1 \log \frac{D_2(\tau)}{\epsilon} + 1), 1\},$$

where $D_1 = \max\{p+d\sum_{l=1}^{L} m_l, \overline{M}K\}$, $D_2(\tau) = \max\{2C_2(\tau)(2\tau J^*)^{1/2}, 4(C_1(\tau)+T+1)K^{-1}\}$, and the second inequality follows from the fact that $\log K! \geq K\log K - K - \log K$.

In the following, we proceed to verify (4.5) – (4.7) in [36]. We first notice that $\int_{\frac{\epsilon}{32}M(i,j)}^{v^{\frac{1}{2}}(i,j)} H^{\frac{1}{2}}(u, \mathcal{F}^V(\tau))du/M(i,j)$ is non-increasing in i and $M(i,j)$, then we have

$$\begin{aligned}&\int_{\frac{\epsilon}{32}M(i,j)}^{v^{\frac{1}{2}}(i,j)} H^{\frac{1}{2}}(u, \mathcal{F}^V(2^j))du/M(i,j)\\ &\leq \int_{\frac{\epsilon}{32}M(1,j)}^{v^{\frac{1}{2}}(1,j)} H^{\frac{1}{2}}(u, \mathcal{F}^V(2^j))du/M(1,j). \end{aligned} \tag{5.5}$$

Rearranging the right-hand side of (5.5) and let $\delta_n^2 = \big(D_1 n^{-1}\log(n/D_1)\big)^{1/(2-\gamma)}$, yielding

$$\int_{\frac{\epsilon}{32}M(1,j)}^{v^{\frac{1}{2}}(1,j)} H^{\frac{1}{2}}(u, \mathcal{F}^V(2^j))du/M(1,j) \leq n^{\frac{1}{2}}.$$

Then, it is easy to see that $\inf_{\boldsymbol{f}\in A_{ij}} E(\tilde{\bar{V}}^T(y, \boldsymbol{f}(\boldsymbol{x},\boldsymbol{t})) - \tilde{\bar{V}}^T(y, \boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{t}))\big)$ is lower bounded by

$$\begin{aligned}&\inf_{\boldsymbol{f}\in A_{ij}} E(\bar{V}^T(y, \boldsymbol{f}(\boldsymbol{x},\boldsymbol{t})) - \bar{V}^T(y, \boldsymbol{f}^0(\boldsymbol{x},\boldsymbol{t}))\big) + \lambda(J(\boldsymbol{f}) - J(\boldsymbol{f}^*))\\ &\qquad - e_{V^T}(\boldsymbol{f}^0 - \boldsymbol{f}^*) \geq M(i,j),\end{aligned}$$

it then follows that $M(i,j)/v(i,j) \leq 1/(8\max\{T,1\})$ and (4.7) directly implies (4.5).

According to Theorem 3 in [36] with $M = n^{1/2}M(i,j)$ and $v = v(i,j)$ yields that

$$\begin{aligned}I_1 &\leq \sum_{i,j:M(i,j)\leq T} 3\exp(-\frac{(1-\epsilon)nM(i,j)^2}{2(4v(i,j) + M(1,j)T/3)})\\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp(-C_5 n M^{2-\min(\gamma,1)}(i,j))\\ &= \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\Big((-C_5 n\big((2^{i-1}-\frac{1}{2})\delta_n^2 + \lambda(2^{j-1}-1)J^*\big)^{2-\min(\gamma,1)}\Big)\\ &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\Big(-C_5 n\big((i-\frac{1}{2})\delta_n^{4-2\min(\gamma,1)} + (j-1)(\lambda J^*)^{2-\min(\gamma,1)}\big)\Big)\\ &\leq \frac{3\exp(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)})}{\big(1-\exp(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)})\big)^2}.\end{aligned} \tag{5.6}$$

Similarly, I_2 can be bounded by

$$\begin{aligned} I_2 &\le \sum_{i=1}^{\infty} 3\exp(-\frac{(1-\epsilon)n(M(i,0))^2}{2(4v(i,0))+M(i,0)T/3}) \\ &\le \sum_{i=1}^{\infty} 3\exp\left(-C_5 n(M(i,0))^{2-\min(\gamma,1)}\right) \\ &\le \sum_{i=1}^{\infty} 3\exp\left(-C_5 n(2^{i-3}\delta_n^2)^{2-\min(\gamma,1)}\right) \\ &\le \frac{3}{1-\exp\left(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)}\right)}. \end{aligned} \tag{5.7}$$

Combining (5.6) and (5.7), we have $I \le \frac{6\exp(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)})}{\left(1-\exp(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)})\right)^2}$. To simplify, let $Q = \exp(-4C_5 n(\lambda J^*)^{2-\min(\gamma,1)})$, by the fact that $I \le I^{1/2} \le 1$ and $I^{1/2}(1-\sqrt{Q}) \le I^{1/2}(1-Q) \le \sqrt{6Q} \le 2.5\sqrt{Q}$, we have $I \le (I^{1/2}+2.5)\sqrt{Q} \le 3.5\sqrt{Q}$. The desired result then follows immediately. □

Supplementary Material

Supplement to "Sentiment analysis with covariate-assisted word embeddings"
(doi: 10.1214/21-EJS1854SUPP; .zip).

References

[1] Gentzkow, M., and Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, 78(1), 35-71. MR2642856

[2] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. *In Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 4, No. 1).

[3] Yan, X., and Bien, J. (2020). Rare feature selection in high dimensions. *Journal of the American Statistical Association*, 1-14.

[4] Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291-304. MR2408634

[5] Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503), 755-770. MR3174658

[6] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint* cs/0205070.

[7] Dai, B., Shen, X., and Wang, J. (2020). Embedding learning. *Journal of the American Statistical Association*, 1-13.

[8] Le, Q., and Mikolov, T. (2014, June). Distributed representations of sentences and documents. *In International conference on machine learning* (pp. 1188-1196). PMLR.

[9] Pennington, J., Socher, R., and Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[10] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. *In Proceedings of the 2013 conference on empirical methods in natural language processing*, (pp. 704-714).

[11] Chen, T., Xu, R., He, Y., and Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72, 221-230.

[12] Thelwall, M., Wilkinson, D., and Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, 61(1), 190-199.

[13] Kucuktunc, O., Cambazoglu, B. B., Weber, I., and Ferhatosmanoglu, H. (2012, February). A large-scale sentiment analysis for Yahoo! answers. *In Proceedings of the fifth ACM international conference on Web search and data mining, (pp. 633-642).*

[14] Bamler, R., and Mandt, S. (2017, July). Dynamic word embeddings. *In International conference on Machine learning*, (pp. 380-389). PMLR.

[15] Das, R., Zaheer, M., and Dyer, C. (2015, July). Gaussian lda for topic models with word embeddings. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 795-804).*

[16] Ren, Y., Wang, R., and Ji, D. (2016). A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences*, 369, 188-198.

[17] Wang, J., Shen, X., Sun, Y., and Qu, A. (2016). Classification with unstructured predictors and an application to sentiment analysis. *Journal of the American Statistical Association*, 111(515), 1242-1253. MR3561946

[18] Baccianella, S., Esuli, A., and Sebastiani, F. (2009, November). *Evaluation measures for ordinal regression. In 2009 Ninth international conference on intelligent systems design and applications (pp. 283-287).* IEEE.

[19] Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

[20] Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003). On ψ-learning. *Journal of the American Statistical Association*, 98(463), 724-734. MR2011686

[21] Zhu, J., and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1),

185-205. MR2137897

[22] Pedregosa, F., Bach, F., and Gramfort, A. (2017). On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18, 1-35. MR3687598

[23] Wang, S. I., and Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 90-94)*.

[24] Zou, H., and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3), 1108-1126. MR2418651

[25] Wu, Y., and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 801-817. MR2514189

[26] Liu, Y., and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*, 23(2), 415-437. MR2801302

[27] Xu, S., Dai, B., and Wang, J. (2021). Supplement to "Sentiment analysis with covariate-assisted word embeddings". DOI: 10.1214/21-EJS1854SUPP

[28] Hsieh, C. J., Chang, K. W., Lin, C. J., Keerthi, S. S., and Sundararajan, S. (2008, July). A dual coordinate descent method for large-scale linear SVM. *In Proceedings of the 25th international conference on Machine learning* (pp. 408-415).

[29] Andersen, M., Dahl, J., Liu, Z., Vandenberghe, L., Sra, S., Nowozin, S., and Wright, S. J. (2011). Interior-point methods for large-scale cone programming. *Optimization for Machine Learning 5583*.

[30] Razaviyayn, M., Hong, M., Luo, Z. Q., and Pang, J. S. (2014). Parallel successive convex approximation for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1406.3665*.

[31] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 138-156. MR2268032

[32] Shen, X., and Wang, L. (2007). Generalization error for multi-class margin classification. *Electronic Journal of Statistics*, 1, 307-330. MR2336036

[33] Hu, M., and Liu, B. (2004, August). Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[34] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[35] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... and Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. *In 12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265-283).

[36] Shen, X., and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 580-615. MR1292531