



Intentional Control of Type I Error Over Unconscious Data Distortion: A Neyman–Pearson Approach to Text Classification

Lucy Xia^a, Richard Zhao^b, Yanhui Wu^{c,d}, and Xin Tong^e

^aDepartment of ISOM, School of Business and Management, Hong Kong University of Science and Technology, Kowloon, Hong Kong; ^bDepartment of Computer Science and Software Engineering, The Behrend College, The Pennsylvania State University, Erie, PA; ^cFaculty of Business and Economics, University of Hong Kong, Pokfulam, Hong Kong; ^dDepartment of Economics and Finance, University of Southern California, Los Angeles, CA; ^eDepartment of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA

ABSTRACT

This article addresses the challenges in classifying textual data obtained from open online platforms, which are vulnerable to distortion. Most existing classification methods minimize the overall classification error and may yield an undesirably large Type I error (relevant textual messages are classified as irrelevant), particularly when available data exhibit an asymmetry between relevant and irrelevant information. Data distortion exacerbates this situation and often leads to fallacious prediction. To deal with inestimable data distortion, we propose the use of the Neyman–Pearson (NP) classification paradigm, which minimizes Type II error under a user-specified Type I error constraint. Theoretically, we show that the NP oracle is unaffected by data distortion when the class conditional distributions remain the same. Empirically, we study a case of classifying posts about worker strikes obtained from a leading Chinese microblogging platform, which are frequently prone to extensive, unpredictable and inestimable censorship. We demonstrate that, even though the training and test data are susceptible to different distortion and therefore potentially follow different distributions, our proposed NP methods control the Type I error on test data at the targeted level. The methods and implementation pipeline proposed in our case study are applicable to many other problems involving data distortion. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY

Received March 2019
Accepted March 2020

KEYWORDS





Censorship; Data distortion; Neyman–Pearson classification paradigm; Social media; Text classification; Type I error


1. Introduction


The rise of social media platforms has spurred the extensive use of large-scale textual data for both academic and nonacademic purposes. However, textual data on open digital platforms are susceptible to manipulation, evident from the continuous debates about fake news, censorship, internet trolls, and social bots (Woolley and Howard 2016a, 2016b). Within an environment of data distortion, the utilization of textual data for information collection (e.g., gauging public opinion) and event discovery (e.g., monitoring social unrest) can be challenging. In the context of textual classification, this article shows the powerlessness of existing classification approaches to handling unknown or inestimable data distortion. We then propose and illustrate the use of the recently developed Neyman–Pearson (NP) classification approach that aims to asymmetrically control classification errors (Cannon et al. 2002; Scott 2005; Rigollet and Tong 2011; Li and Tong 2016; Tong, Feng, and Li 2018) in some common situations of data distortion, such as data obtained from censored Chinese social media.

Since 2009 when *Sina Weibo*—the Chinese equivalent to Twitter—was launched, social media have created an unprecedented informational shock to the Chinese society. Notably, *Sina Weibo* enables millions of citizens to generate and

communicate political information that is scarce in traditional media. Government agents, media outlets, NGOs and firms, and researchers have invested heavily in machine learning techniques to mine the wealth of textual information circulated on *Sina Weibo* (China Internet Network Information Center 2013, 2014; Economist 2013). However, due to the potential effect of widespread political information on social unrest and regime stability, the Chinese government extensively censors social media (Chen and Ang 2011; King, Pan, and Roberts 2013, 2014). Such censorship gives rise to two major challenges faced by data analysts in their endeavor of text mining. First, although the Chinese government allows for relatively free information flow on social media for the purposes of surveillance and monitoring officials (Qin, Strömberg, and Wu 2017), censorship substantially reduces the amount of information circulating on social media that can practically be used to classify data and predict hidden social events. The objective of minimizing the overall classification error, which is used by most existing machine learning algorithms, can cause an undesirably large error of missing important information. Second, social media censorship in China relies mostly on ad hoc human manipulation to fine-tune the extent of censorship in response to the changing local and temporal social conditions (Bamman, O'Connor, and Smith 2012; Zhu et al. 2013). This censorship strategy makes

CONTACT Yanhui Wu  yanhuiwu@marshall.usc.edu  Department of Economics and Finance, University of Southern California, Los Angeles, CA; Xin Tong  xintong@marshall.usc.edu  Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

 These materials were reviewed for reproducibility.

© 2020 American Statistical Association

it infeasible to infer the censorship rate. Thus, the traditional solution that corrects the potential bias due to data truncation through a parametric estimation of the censorship rate is hardly a practical choice. We propose the use of the NP classification approach to precisely overcome these two challenges.

To make our discussion more concrete, consider that a decision maker wishes to use social media posts about political issues and social events to discover and monitor grass-root political actions such as protests, petitions, or worker strikes. To this end, the decision maker must use algorithms trained on labeled data to classify a large number of posts, that is, to predict discrete outcomes (class labels) for upcoming posts. In a binary classification setting, a post is coded in $\{0, 1\}$, where class 0 means relevant to a specific topic, and class 1 means irrelevant. Two types of errors occur: Type I error (mislabel class 0 as class 1) and Type II error (mislabel class 1 as class 0).¹ The default classification objective in practice, which is referred to as *the classical classification paradigm* in this article, is the one that minimizes the overall classification error, which is a weighted sum of Type I and Type II errors, with weights being the proportions of classes. When controlling one type of error is dominantly important, a conflict occurs between the need for asymmetric control over classification errors and the neglect of such consideration in the overall classification error. Data distortion can exacerbate such a conflict. If a fraction of class 0 data is eliminated, then in the objective function of the classical paradigm, the weight of Type I error is reduced. Minimizing this objective naturally increases Type I error, which is undesirable when controlling Type I error to avoid overlooking relevant events is crucial to decision making.

In this article, we first derive the *classical oracle classifier* (theoretically optimal classifier under the classical paradigm) regarding the post-distortion population, and then demonstrate that, without precise knowledge about the data distortion rates, the pre-distortion classical oracle classifier cannot be recovered even if we have access to the entire post-distortion population. As a solution, we propose to use the Neyman–Pearson classification paradigm (NP paradigm) which minimizes Type II error under a user-specified Type I error constraint. The NP paradigm has the advantage that the NP oracle classifier (theoretically optimal classifier under the NP paradigm) is invariant to the class size proportion in the population. This property guarantees the invariance of the NP oracle under any distortion scheme as long as the class conditional distributions of the features remain the same.

To bring our theoretical discussion to live, we focus on an exemplary case in the general setting of Chinese social media, in which we classify a large number of posts about worker strikes published on Sina Weibo. Accurately identifying strike events in a timely manner is highly valuable for many decision makers, including governments, firms, and social scientists studying social movement. On the other hand, as a type of collective action, posts about strikes are prone to censorship, the extent of which varies across regions and over time. We show that applying existing classification methods leads to a considerable

Type I error, which can result in oversight or fallacious outcomes in decision making. We then use an NP umbrella algorithm (Tong, Feng, and Li 2018) in combination with state-of-the-art machine learning techniques to classify the posts. Consistent with the NP oracle’s invariance property to data distortion, we find that even though the training and test data are susceptible to different distortion rates and are thus differently distributed, the NP classifiers hold Type I errors well controlled at the targeted level on the test data. Furthermore, we demonstrate that for the purpose of controlling Type I errors, the NP classification methods allow decision makers to borrow data generated in an information-abundant environment to classify data generated in an information-scarce environment. This advantage is important when decision making is constrained by time and resources.

Our study of data distortion is essentially an inquiry into the validity of statistical prediction when the process of data generation is a primary concern. This concern is not dismissible even in the era of big data. Instead, it can be exacerbated when data sources are vulnerable to human intervention. One candidate solution to data distortion is to estimate and correct potential bias by assuming precise knowledge regarding data generation and distortion. This is analogous to the parametric estimation of censored or truncated data in classical statistical inference (Chung et al. 1991). Unfortunately, such a solution is infeasible when data are generated from diverse sources and are affected by complex interactions. Another potential solution, which is popular in the traditional statistics literature, is the development of sampling techniques that aim to obtain more representative samples from the population (Luborsky and Rubinstein 1995). However, sampling methods do not solve the data distortion problem in our study because even if the entire post-distortion population were available, knowledge about the pre-distortion population is still limited by unknown or inestimable distortion rates. In contrast, the NP classification approach we propose allows researchers to bypass one common kind of distortion which changes the class proportions but not the class conditional feature distributions.

The setting in this article might seem similar to domain adaptation (Ben-David et al. 2010; Chen, Weinberger, and Blitzer 2011), a type of transfer learning. However, the data distortion problem in our study differs fundamentally from the problems studied in domain adaptation. In domain adaptation, a key assumption is that the “source domain” and “target domain” share the same feature space, but have different feature distributions. A domain adaptation algorithm takes not only labeled data from the source domain, but also data (labeled or unlabeled) from the target domain. In contrast, the only available training data in our study are the labeled data from the post-distortion population (i.e., the source domain) without using any data (regardless of being labeled or unlabeled) from the pre-distortion population (i.e., the target domain). In this sense, the data-distortion problem we address is more challenging because data from the target domain is not available. To overcome such a challenge, the NP classification approach invokes the assumption that the features have the same conditional distributions in the source and target domains.

¹In the verbal discussion, Type I and Type II errors can also be thought of as the probability of making such errors.

2. Classification and Unknown Distortion Scheme

Binary classification is a supervised learning procedure frequently used in textual analysis. It aims to classify a piece of textual message into a category that is relevant to either a specific purpose or an irrelevant category. Formally, the aim of binary classification is to accurately predict class labels (i.e., $Y = 0$ or 1) for new observations (i.e., features $X \in \mathbb{R}^d$) on the basis of labeled training data. For the rest of the discussion, we treat the relevant information category as class 0 and the irrelevant one as class 1, so that missing a class 0 message is more consequential than missing a class 1 message. Concretely, let $h : \mathbb{R}^d \rightarrow \{0, 1\}$ be a binary classifier, $R_0(h) := \mathbb{P}(h(X) \neq Y | Y = 0)$ denote Type I error, and $R_1(h) := \mathbb{P}(h(X) \neq Y | Y = 1)$ denote Type II error. Then, the (population) classification error $R(h)$ can be decomposed as $R(h) = R_0(h) \cdot \mathbb{P}(Y = 0) + R_1(h) \cdot \mathbb{P}(Y = 1)$. We use the term *classical paradigm* to refer to the learning objective of minimizing $R(\cdot)$. The classical oracle classifier, that is, the classifier that minimizes $R(\cdot)$ among all functions, is $h^*(x) = \mathbb{I}(\eta(x) > 1/2)$, where $\eta(x) = \mathbb{E}(Y | X = x) = \mathbb{P}(Y = 1 | X = x)$. The classical oracle h^* is achievable only if the entire population is available. In practice, we have to train a classifier based on a finite sample.

2.1. Data Distortion Scheme

In this article, we restrict our attention primarily to the type of distortion that changes the class proportion of the population without changing the class conditional distributions of the features. In other words, we assume that distortion changes $\mathbb{P}(Y = 0)$ and $\mathbb{P}(Y = 1)$, but does not change the distributions of $X | (Y = 0)$ or $X | (Y = 1)$. The assumption that features have the same class-conditional distributions is justified if the distortion scheme in the dataset (e.g., deleting sensitive social media posts) is random. We will show that this assumption can be approximated by the data distortion situation in our case study and other real world applications. We discuss more general conditions in Appendix C of the supplementary materials.

2.2. Oracle Under Data Distortion

Denote the class 0 distortion rate by $\beta_0 = \beta_0^- - \beta_0^+$, where β_0^- is the class 0 downward-distortion rate and β_0^+ is the class 0 upward-distortion rate. These rates are the proportions of class 0 texts that are randomly deleted or injected, respectively. For example, $(\beta_0^-, \beta_0^+) = (0.2, 0.1)$ means 20% of class 0 texts are randomly deleted from the population, and 10% of class 0 texts are artificially injected, so the net effect is a $\beta_0 = 10\% = 20\% - 10\%$ decrease in class 0 texts. Since we cannot disentangle the upward and downward forces just from the post-distortion population, we will formulate the theory only on the net decrease effect β_0 . Similarly, β_1 is defined for class 1. Below, we derive the formula of the (classical) oracle classifier regarding the post-distortion population.

Theorem 1. Let f_0 and f_1 denote the pre-distortion probability density functions of $X | (Y = 0)$ and $X | (Y = 1)$, and $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$ be the class priors. Suppose the distortion scheme does not change the distributions for $X | (Y =$

$0)$ and $X | (Y = 1)$ but only the class proportions. Let β_0 and β_1 be the distortion rates of class 0 and class 1, respectively. Then, the classical oracle classifier regarding the pre-distortion population is

$$h^*(x) = \mathbb{I} \left(\frac{f_1(x)}{f_0(x)} > \frac{\pi_0}{\pi_1} \right),$$

and that regarding the post-distortion population is

$$h_{(\beta_0, \beta_1)}^*(x) = \mathbb{I} \left(\frac{f_1(x)}{f_0(x)} > \frac{1 - \beta_0}{1 - \beta_1} \cdot \frac{\pi_0}{\pi_1} \right).$$

In this theorem, the explicit analytic form of the classical pre-distortion oracle classifier h^* is a well-known result, while that of $h_{(\beta_0, \beta_1)}^*$ is new. See Appendix A in the supplementary materials for its proof. The thresholds of f_1/f_0 in oracle classifiers h^* (pre-distortion) and $h_{(\beta_0, \beta_1)}^*$ (post-distortion) differ by a multiplicative constant $(1 - \beta_0)/(1 - \beta_1)$. This difference in thresholds reflects a change in the class proportions in the population. If the entire post-distortion population is available, we can calculate the class conditional densities f_0 and f_1 as well as the post-distortion class proportions

$$\begin{aligned} \pi_0^{(\beta_0, \beta_1)} &= \frac{(1 - \beta_0)\pi_0}{(1 - \beta_0)\pi_0 + (1 - \beta_1)\pi_1}, \\ \pi_1^{(\beta_0, \beta_1)} &= \frac{(1 - \beta_1)\pi_1}{(1 - \beta_0)\pi_0 + (1 - \beta_1)\pi_1}. \end{aligned}$$

Then, $h_{(\beta_0, \beta_1)}^*$ can be recovered. However, there is no hope to recover or estimate h^* , unless β_0 and β_1 are known or estimable.

2.3. Impact of Censorship Rate Under the Gaussian Model

To visualize and quantify the result in Theorem 1, we study an example with $\beta_0 > 0$ and $\beta_1 = 0$ under a canonical linear discriminant analysis model. Let $f_0 \sim \mathcal{N}(\mu_0, \Sigma)$ and $f_1 \sim \mathcal{N}(\mu_1, \Sigma)$, where μ_0 and μ_1 represent mean vectors for classes 0 and 1, respectively, and Σ is the common covariance matrix. In this model, the decision boundary of the oracle h^* is

$$x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log \left(\frac{\pi_0}{\pi_1} \right) = 0. \quad (1)$$

When only $(1 - \beta_0)$ proportion of observations from class 0 remains, the post-distortion oracle classifier $h_{\beta_0}^* := h_{(\beta_0, 0)}^*$ has the following decision boundary:

$$\begin{aligned} x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) \\ + \log \left(\frac{(1 - \beta_0)\pi_0}{\pi_1} \right) = 0. \end{aligned} \quad (2)$$

Comparing (1) and (2), the shape of the decision frontier remains the same, but the left hand of the equations differs by a constant $\log(1 - \beta_0)$. To visualize the difference in decision boundaries, we plot an example in Figure 1. Proposition 1 further explores the relationship between Type I error and the censorship rate of class 0 for balanced classes.

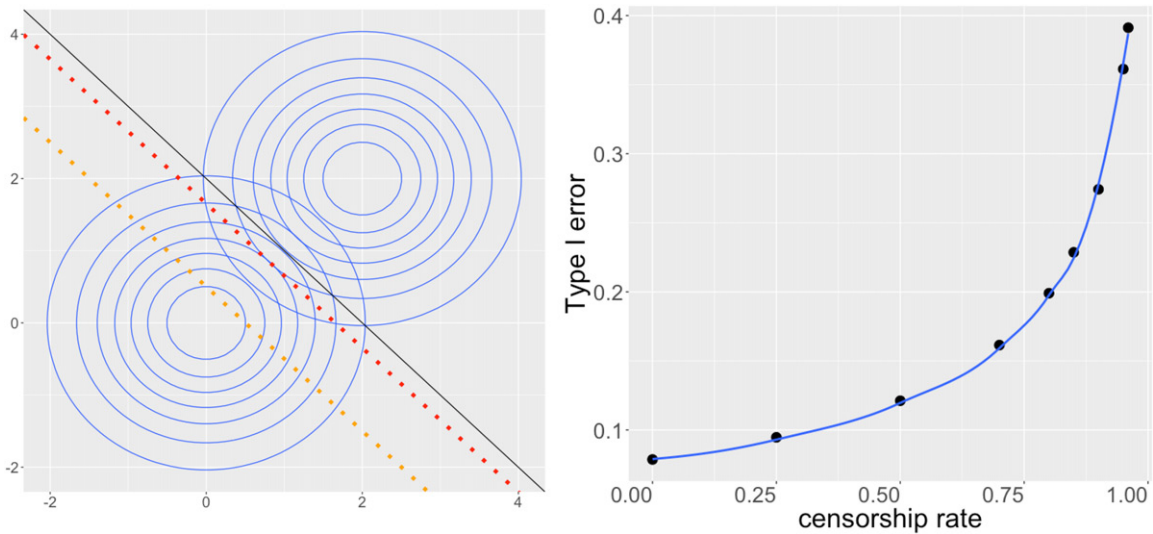


Figure 1. The left panel shows the shift of the oracle decision boundary due to distortion under a linear discriminant analysis model: $\mu_0 = (0, 0)^\top$, $\mu_1 = (2, 2)^\top$, $\Sigma = I$, $\pi_0 = 0.5$. The horizontal axis and vertical axis are the two feature measurements, and the contours represent different density levels of each class. The black line is the original oracle decision boundary; the red dashed line and the orange dashed line are the oracle decision boundaries after censorship on class 0 with $\beta_0 = 0.5$ and $\beta_0 = 0.95$, respectively. The right panel plots Type I error of $h_{\beta_0}^*$ as a function of β_0 .

Proposition 1. Suppose the probability densities of class 0 ($X|Y = 0$) and class 1 ($X|Y = 1$) follow distributions $\mathcal{N}(\mu_0, \Sigma)$ and $\mathcal{N}(\mu_1, \Sigma)$, respectively, and the two classes are balanced in the pre-distortion population (i.e., $\pi_0 = \pi_1 = 0.5$). Suppose that the censorship rate of class 0 is $\beta_0 \in (0, 1)$ and class 1 is not distorted ($\beta_1 = 0$). To keep notations simple, let $h_{\beta_0}^* = h_{(\beta_0, 0)}^*$ be the classical oracle classifier in the post-distortion population. Then, the Type I error of $h_{\beta_0}^*$ is

$$R_0(h_{\beta_0}^*) = \Phi\left(\frac{-\frac{1}{2}C - \log(1 - \beta_0)}{\sqrt{C}}\right), \quad (3)$$

where $C = (\mu_0 - \mu_1)^\top \Sigma^{-1} (\mu_0 - \mu_1)$. Clearly, $R_0(h_{\beta_0}^*)$ increases with $\beta_0 \in (0, 1)$.

Proposition 1 is proved in Appendix E of the supplementary materials. When censorship on class 0 texts intensifies, class 0 in the post-distortion population represents a smaller proportion, and the post-distortion oracle will favor class 1 more, leading to a rise in Type I error. Note that C captures the difficulty of the classification problem: the larger C , the better class separation, and the easier the classification problem.

3. Neyman–Pearson (NP) Classification Paradigm

One existing solution to the problem of data distortion is to collect information so as to better understand the data generation process. For example, one might spend efforts estimating the distortion rates β_0 and β_1 . However, such a solution is usually costly and practically infeasible. Another idea is to adjust the weight placed on each of the two types of errors in the objective function of the classical classifier. This is the cost-sensitive learning paradigm (Elkan 2001; Zadrozny, Langford, and Abe 2003), in which users impose different costs to the two types of errors to address the issue of asymmetric error importance. However, such a method does not solve the data distortion problem, as discussed in Appendix B of the supplementary materials. To

tackle the data distortion issue and Type I error control objective simultaneously, we propose to adopt the NP paradigm.

3.1. NP Oracle Invariant to Distortion

The NP oracle ϕ_α^* arises from the famous Neyman–Pearson lemma in statistical hypothesis testing (attached in Appendix F of the supplementary materials). Instead of minimizing $R(h) = R_0(h) \cdot \mathbb{P}(Y = 0) + R_1(h) \cdot \mathbb{P}(Y = 1)$ as in the classical paradigm, the NP classification paradigm aims to mimic the NP oracle ϕ_α^* , where

$$\phi_\alpha^* = \arg \min_{\phi: R_0(\phi) \leq \alpha} R_1(\phi), \quad (4)$$

in which α is a user-specified upper bound on Type I error. Under the NP classification paradigm, α reflects the level of a user’s conservativeness toward the Type I error. In some biomedical applications, there is clear choice of α , such as 0.01 and 0.05, due to either government regulation or common practice. In social sciences applications, the choice of α is more subjective. Some suggestions in choosing α can be found in Tong, Feng, and Li (2018).

The NP classification paradigm has three advantages: (i) bypass data distortion, (ii) address the class imbalance issue, and (iii) control the more severe error type (typically, Type I error) under a user-specified level. The third advantage is self-evident; the first two are illustrated as follows.

Theorem 2. Suppose that the distortion scheme does not change the distributions for $X|(Y = 0)$ and $X|(Y = 1)$. The NP oracle classifier ϕ_α^* defined in (4) is invariant under distortion at various rates β_0 (on class 0) and β_1 (on class 1), regardless of whether pre-distortion classes are balanced.

Theorem 2 (proof in Appendix A of the supplementary materials) implies that in an idealized situation when one has access to the entire post-distortion population, he/she can reconstruct

the NP oracle classifier as if the entire pre-distortion population is available. The rationale is that the NP oracle depends only on the conditional distributions of $X|Y = 0$ and $X|Y = 1$ but not on the marginal distribution of Y . This means that, as long as these conditional distributions do not change, the NP oracle will stay the same.

Figure 2 illustrates the difference between a classical oracle classifier and its NP counterpart in both balanced and imbalanced Gaussian settings. While the classical oracles are different, the NP oracle is the same in both settings. As the type of data distortion in our study amounts to a change in the class proportion, this figure also demonstrates a contrast between a shift in decision boundary of the classical oracle and the invariance of the NP oracle under data distortion.

The main theoretical results, Theorems 1 and 2, do not require any parametric assumptions. We only use parametric Gaussian examples as an illustration of these two theorems. Specifically, we use Proposition 1 and Figures 1 and 2 to illustrate (1) the impact of data distortion on classical oracles and (2) the invariance of the NP oracles to data distortion.

Theorem 2 suggests that, using samples from the post-distortion population, we can train classifiers to mimic the pre-distortion NP oracle classifier due to its invariance to distortion, and thus bypass the need to estimate the data distortion scheme. Another key implication is that one could train an NP classifier on data from a distorted population with distortion rates β'_0 and β'_1 and test the classifier on data from another distorted population with different distortion rates β''_0 and β''_1 . In other words, if the training and test data undergo different censorship schemes, the NP paradigm can still be applied. Regarding the

practical implementation of the NP paradigm, we will introduce the NP umbrella algorithm (Tong, Feng, and Li 2018), which is compatible with all the scoring-type classification methods (e.g., logistic regression, support vector machines, and random forest), parametric or nonparametric.

In Appendix C of the supplementary materials, we discuss a situation in which the class conditional densities of features are also changed by distortion. We derive the necessary and sufficient condition for the invariance property of the NP oracles in such a more general situation. Essentially, the general condition requires that the post-distortion class conditional density ratio is a multiple of the pre-distortion one, and that a good tail behavior is satisfied for the density ratios. We also construct concrete examples showing that these abstract generalization conditions could materialize in common model settings. Nevertheless, we choose to present the more-specific condition in Theorem 2 because it is transparent and easy to interpret.

3.2. NP Umbrella Algorithm

To construct a classifier under the NP paradigm, one can plug the class conditional feature densities and the threshold estimates into the NP oracle classifier suggested by the Neyman–Pearson lemma (Appendix F in the supplementary materials). Plug-in NP classifiers have been constructed in two settings: low-dimensional (Tong 2013) and high-dimensional with independent features (Zhao et al. 2016). However, plug-in procedures suffer from the curse of dimensionality in more general high-dimensional settings. To make the NP paradigm more

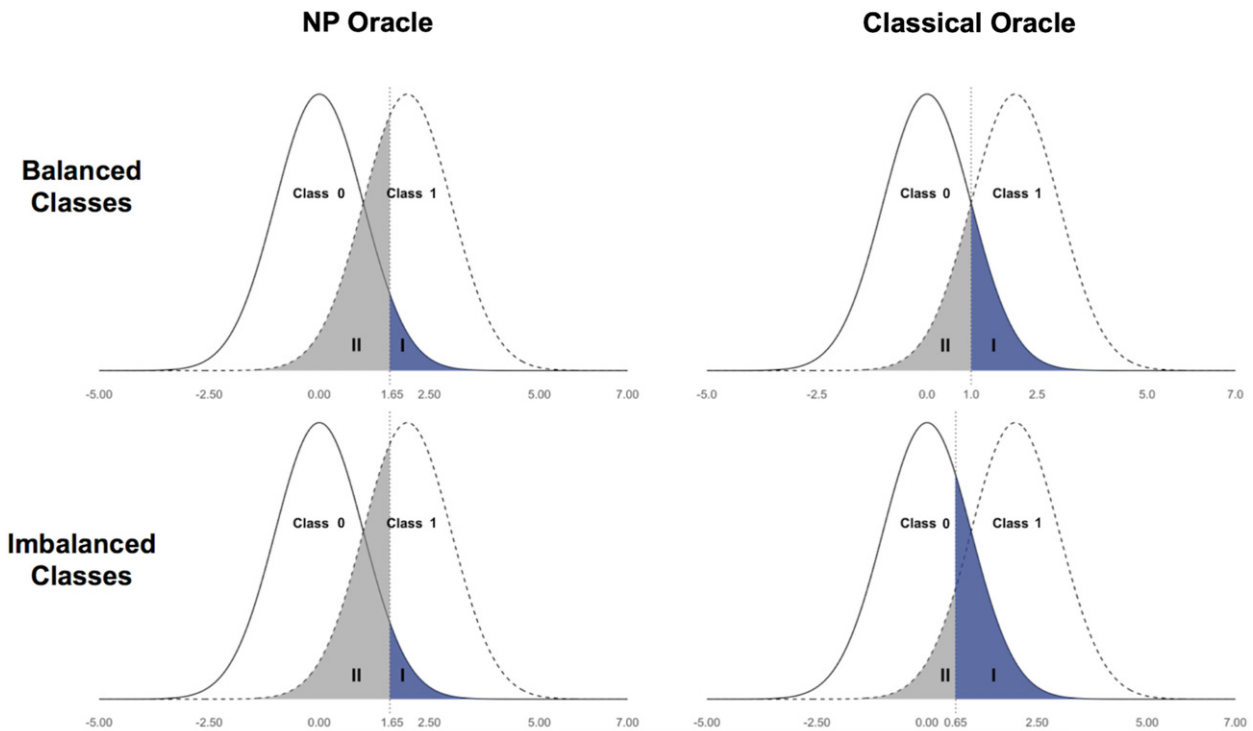


Figure 2. NP versus classical oracle classifiers in a Gaussian model example. The conditional distributions of X under the two classes are $\mathcal{N}(0, 1)$ and $\mathcal{N}(2, 1)$, respectively. Suppose that a user prefers a Type I error $\leq \alpha = 0.05$. When the two classes are balanced (i.e., $P(Y = 0) = P(Y = 1)$), the classical oracle $\mathbf{1}(X > 1)$ that minimizes the risk would result in a Type I error = 0.159. On the other hand, the NP oracle $\mathbf{1}(X > 1.65)$ that minimizes the Type II error under the Type I error constraint (≤ 0.05) delivers the desirable Type I error. In an imbalanced situation where $2P(Y = 0) = P(Y = 1)$, while the NP oracle does not change and retains the desirable Type I error, the decision boundary of the classical oracle shifts left to 0.6534 and results in a much larger Type I error = 0.257.

practical, Tong, Feng, and Li (2018) proposed an NP umbrella algorithm, a wrapper method that allows users to apply their favorite scoring-type classification methods, such as logistic regression, support vector machines, and random forests, under the NP paradigm. Figure 3 illustrates the pseudocode of the NP umbrella algorithm. This umbrella algorithm uses part of class 0 data and all class 1 data to train a scoring-function and use the left-out class 0 data to determine a threshold for the scoring function. To use the algorithm, a user specifies a desired upper bound α for the (population) Type I error and an upper bound for the Type I error violation rate δ (i.e., the probability that Type I error exceeds α). Proposition 2 and Corollary 1 provide a theoretical warranty for the control of Type I error using the classifiers constructed based on samples.

Proposition 2 (Adapted from Tong, Feng, and Li (2018)). Suppose that we divide the training data into two parts, one with data from both classes 0 and 1 for training a base algorithm (e.g., svm,

random forest, etc.) to obtain f and the other as a left-out class 0 sample for choosing the threshold. Applying f to the left-out class 0 sample of size n , we denote the resulting classification scores as T_1, \dots, T_n , which are real-valued random variables. Then, we denote by $T_{(k)}$ the k th order statistic (i.e., $T_{(1)} \leq \dots \leq T_{(n)}$). For a new observation X , if we denote its classification score $f(X)$ as T , we can construct classifiers $\hat{\phi}_k(X) = \mathbb{I}(T > T_{(k)})$, $k \in \{1, \dots, n\}$. Then, the population Type I error of $\hat{\phi}_k$, denoted by $R_0(\hat{\phi}_k)$, is a function of $T_{(k)}$ and hence a random variable, and it holds that

$$\mathbb{P} [R_0(\hat{\phi}_k) > \alpha] \leq \sum_{j=k}^n \binom{n}{j} (1 - \alpha)^j \alpha^{n-j}. \quad (5)$$

That is, the probability that the Type I error of $\hat{\phi}_k$ exceeds α is under a constant that only depends on k, α , and n . We call this probability the violation rate of $\hat{\phi}_k$ and denote its upper bound by $v(k) = \sum_{j=k}^n \binom{n}{j} (1 - \alpha)^j \alpha^{n-j}$.

```

1: input:
   training data: a mixed i.i.d. sample  $\mathcal{S} = \mathcal{S}^0 \cup \mathcal{S}^1$ , where  $\mathcal{S}^0$  and  $\mathcal{S}^1$  are class 0 and
   class 1 samples respectively
    $\alpha$ : type I error upper bound,  $0 \leq \alpha \leq 1$ ; [default  $\alpha = 0.05$ ]
    $\delta$ : a small tolerance level,  $0 < \delta < 1$ ; [default  $\delta = 0.05$ ]
    $M$ : number of random splits on  $\mathcal{S}^0$ ; [default  $M = 1$ ]
2: function RANKTHRESHOLD( $n, \alpha, \delta$ )
3:   for  $k$  in  $\{1, \dots, n\}$  do                                     ▷ for each rank threshold candidate  $k$ 
4:      $v(k) \leftarrow \sum_{j=k}^n \binom{n}{j} (1 - \alpha)^j \alpha^{n-j}$            ▷ calculate the violation rate upper bound
5:    $k^* \leftarrow \min \{k \in \{1, \dots, n\} : v(k) \leq \delta\}$        ▷ pick the rank threshold
6:   return  $k^*$ 
7: procedure NPCLASSIFIER( $\mathcal{S}, \alpha, \delta, M$ )
8:    $n = \lceil |\mathcal{S}^0|/2 \rceil$                                            ▷ denote half of the size of  $|\mathcal{S}^0|$  as  $n$ 
9:    $k^* \leftarrow$  RANKTHRESHOLD( $n, \alpha, \delta$ )                       ▷ find the rank threshold
10:  for  $i$  in  $\{1, \dots, M\}$  do                                     ▷ randomly split  $\mathcal{S}^0$  for  $M$  times
11:     $\mathcal{S}_{i,1}^0, \mathcal{S}_{i,2}^0 \leftarrow$  random split on  $\mathcal{S}^0$              ▷ each time randomly split  $\mathcal{S}^0$  into two halves with
    equal sizes
12:     $\mathcal{S}_i \leftarrow \mathcal{S}_{i,1}^0 \cup \mathcal{S}^1$                                ▷ combine  $\mathcal{S}_{i,1}^0$  and  $\mathcal{S}^1$ 
13:     $\mathcal{S}_{i,2}^0 = \{x_1, \dots, x_n\}$                                  ▷ write  $\mathcal{S}_{i,2}^0$  as a set of  $n$  data points
14:     $f_i \leftarrow$  classification algorithm( $\mathcal{S}_i$ )                 ▷ train a scoring function  $f_i$  on  $\mathcal{S}_i$ 
15:     $\mathcal{T}_i = \{t_{i,1}, \dots, t_{i,n}\} \leftarrow \{f_i(x_1), \dots, f_i(x_n)\}$  ▷ apply the scoring function  $f_i$  to  $\mathcal{S}_{i,2}^0$  to
    obtain a set of score threshold candidates
16:     $\{t_{i,(1)}, \dots, t_{i,(n)}\} \leftarrow$  sort( $\mathcal{T}_i$ )              ▷ sort elements of  $\mathcal{T}_i$  in an increasing order
17:     $t_i^* \leftarrow t_{i,(k^*)}$  ▷ find the score threshold corresponding to the chosen rank threshold  $k^*$ 
18:     $\phi_i(X) = \mathbb{I}(f_i(X) > t_i^*)$  ▷ construct an NP classifier based on the scoring function  $f_i$ 
    and the threshold  $t_i^*$ 
19: output:
   an ensemble NP classifier  $\hat{\phi}_\alpha(X) = \mathbb{I} \left( \frac{1}{M} \sum_{i=1}^M \phi_i(X) \geq 1/2 \right)$  ▷ by majority vote

```

Figure 3. Pseudocode for the NP umbrella algorithm adapted from Tong, Feng, and Li (2018) with permission.

Corollary 1. Suppose that the distortion scheme does not change the distributions for $X|Y = 0$ and $X|Y = 1$. The NP umbrella algorithm (with $M = 1$) presented in Figure 3 yields a classifier $\hat{\phi}$ such that $\hat{\phi}$ has Type I error violation rate controlled, that is, $\mathbb{P}(R_0(\hat{\phi}) \leq \alpha) \geq 1 - \delta$, and attains the smallest Type II error given a user-specified method.

Corollary 1 follows from Proposition 2. The proof of Corollary 1 can be briefly described as following. It is obvious that $\nu(k)$ decreases as k increases. To choose from $\hat{\phi}_1, \dots, \hat{\phi}_n$ such that a classifier achieves minimal Type II error with Type I error violation rate less than or equal to a user's specified δ , the right order is

$$k^* = \min \{k \in \{1, \dots, n\} : \nu(k) \leq \delta\}. \quad (6)$$

Notice that the NP umbrella algorithm does not guarantee the Type II error to be close to the oracle level, because it does not rely on assumptions of the distribution of (X, Y) or the chosen classification method.

4. Case Study

We present a case study regarding how to classify posts about strike events in Chinese social media. This case empirically illustrates the problem of unknown data distortion in text classification and the relevance of the NP classification approach to real-world decision making. Moreover, we demonstrate how to implement and assess various NP classification methods so that researchers of interest can adopt them.

Information regarding collective action events such as worker strikes and protests is important for citizens' participation in politics, policy implementation by governments, the accountability of political leaders, and business decisions of firms. In authoritarian countries, however, this type of information has been scarce in the public sphere because of strict government control over the mass media. The emergence of social media enables citizens to circulate information about social events and voice their opinions on political issues. This has inspired local governments, nongovernment organizations, firms and investors, and particularly social scientists to gather, decode and analyze the information produced on social media in authoritarian countries. However, in their endeavor to utilize information found on social media, these decision makers face the challenge of data distortion caused by extensive censorship of social media information. The NP classification approach is intended to help make use of the limited set of useful information that remains on social media to better discover and predict hidden social events.

In this section, we first depict the Chinese government's social media censoring strategy and explain how it fits the theoretical setup outlined in Section 2. Second, we describe our research design and data collection. Third, we detail the pipeline of data analysis including data preprocessing, feature engineering, and the implementation of each NP classification method. As a preview, Figure 4 shows the entire chain of empirical analysis. Finally, we present the results in a baseline sample and then in four augmented samples to further illustrate the advantage of the NP classification approach.

4.1. Data Distortion in Chinese Social Media

In China, social media are typically owned by private service providers. For example, *Sina Weibo*—the microblogging platform in this study—is owned by Sina Corp., which is a company listed in NASDAQ. However, the Chinese central government controls the infrastructure based on which the social media platform operates and thus has the de facto right of censoring social media. Numerous studies have documented that the Chinese government extensively censors social media information, particularly political information that may undermine the leadership of the Chinese Communist Party, trigger large-scale collective action, and cause social unrest (Chen and Ang 2011; King, Pan, and Roberts 2013, 2014). Nevertheless, this does not mean that all politically sensitive information is censored. Using a dataset of 13.2 billion posts published in Sina Weibo from 2009 to 2013, Qin, Strömberg, and Wu (2017) documented millions of posts published in Sina Weibo that discussed protests, demonstrations, strikes, and corruption. Based on the posting activities of users who had published this politically sensitive information, they conclude that the Chinese government allows for the circulation of some political information on social media with an intention to encourage participation and collect information for surveillance and monitoring local officials. Other studies (Lorentzen 2014; Qin, Strömberg, and Wu 2019) suggested that the Chinese government's strategy of censoring social media revolves around a trade-off between utilizing bottom-up information and avoiding accumulation and spread of information that may scale up existing events (e.g., protests and strikes) or spur new action. Such a tradeoff leads to the following common censorship practice: information about small local social events is not censored until a scale shift of information is detected (Bamman, O'Connor, and Smith 2012; Zhu et al. 2013). In other words, when the quantity of sensitive information exceeds some threshold, censorship is triggered.

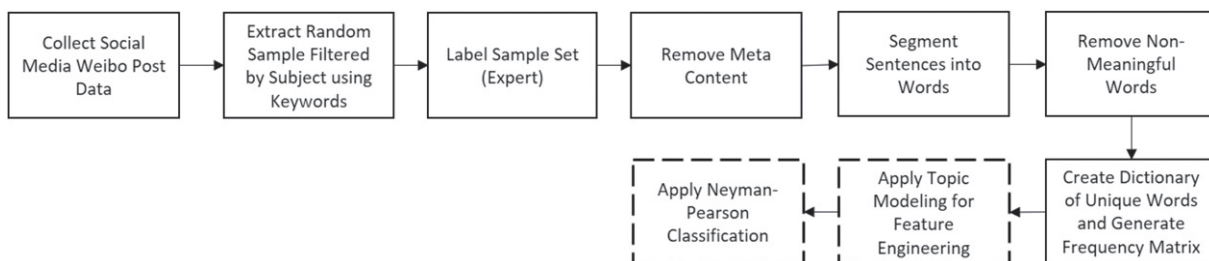


Figure 4. Illustration of the data processing pipeline with the preprocessing steps in the solid squares.

Unlike in Russia where the manipulation of online information is mostly through the deployment of bots to perform automated tasks, in China, censorship of social media is largely implemented in an ad hoc manner. The threshold of censorship depends on local social and political conditions (Chen and Ang 2011; Bamman, O'Connor, and Smith 2012). It is well known that during the period of Congressional meetings or national celebration and in regions where social conflicts are pronounced, the Chinese government tends to tighten censorship to contain potential social unrest. This ad hoc censorship policy provides an explanation for the wide range of censorship rates estimated in existing studies (Chen and Ang 2011; Bamman, O'Connor, and Smith 2012; Fu, Chan, and Chau 2013).

In practice, the censorship on Chinese social media involves three additional parties other than the central government: (1) social media providers, private IT companies which implement censorship, (2) government information officers who enforce the implementation of censorship, and (3) local governments who find ways to interfere with the operation of social media. These parties may have different objectives than the central government. For example, to maintain a high level of information traffic, social media providers do not completely comply with the government's censorship demands. Moreover, the enforcement of censorship by government information officers is based on ad hoc issuing of directives, depending on the involving officers' collection and interpretation of information (Chen and Ang 2011; Zhu et al. 2013). Finally, although local governments do not have the right to censor social media, they may bribe employees of social media providers to delete information that may reflect negatively on them.

The above characteristics regarding censorship make the Chinese social media an ideal setting to study the problem of classification in the presence of data distortion and the NP classification methods as a solution to the problem. A decision maker who wishes to extract useful information about certain issues or events from post-censorship social media posts faces the problem of data distortion as we formulate in the previous sections. The quantity-based censorship suggests that the features of information in the relevant class are likely to remain stable despite that censorship significantly reduces the quantity of this type of information. Therefore, the key assumption under which the invariance property of the NP oracle classifier is approximately true. Importantly, the ad hoc nature of censorship and the involvement of multiple parties in its implementation render the actual censorship scheme highly volatile and unpredictable. It is practically infeasible for a decision maker to infer the rate of data distortion due to censorship.

4.2. Data Collection and Research Design

For this study, we collected public user posts related to sensitive social issues from the microblogging site *Sina Weibo*. Through a third-party content crawling agency, we obtained a dataset of approximately 10 million raw posts about public issues and social events in 2012. We are interested in classifying posts about the subject "worker strikes." We focus on strikes for several reasons. First, the number of strikes in China has surged in the last decade, and strikes have become an important form of

worker movement (China Labour Bulletin 2012, 2018). Accurately identifying strike events in a timely manner is important for a wide range of decision makers, including governments, firms, and social scientists. Second, as an indicator of collective action, posts about strikes are prone to censorship. Third, the degree of censorship of posts about strikes varies across regions and over time. For instance, censorship tends to be more intense toward the end of the year when workers' yearly compensation is due and in regions where economic conditions are worse and unemployment rates have increased. As explained below, this variation provides a partial test of the assumptions that entail the application of our proposed NP classification methods as well as an opportunity to demonstrate the advantage of the NP classification approach.

We extract a subset of posts filtered according to a preselected list of keywords.² This filtering generates 221,229 posts linguistically relating to strikes. From this dataset, we extract a random sample of 2500 posts that were published in the first quarter of 2012 and were originated from Guangdong—a coastal province where strike incidence occurred most frequently among all provinces in China during the sample period. This sample serves as a baseline for our data analysis as well as an illustration of various NP classification methods. We then extract three random samples of the same size (2500 posts) in the 2nd, 3rd, and 4th quarters of 2012 from Guangdong, respectively. We will apply an NP classifier trained in the baseline sample to these three samples in other periods. Good performance (in terms of controlling Type I error) of this classifier across different samples provides suggestive evidence on the stable distribution of features in the presence of data distortion. Finally, we select a random sample of 2500 posts from all the posts originated from three inland provinces—Gansu, Qinghai, and Xinjiang—during the entire year of 2012. Evidence shows that these three provinces were among regions where censorship on social media was most intense (Bamman, O'Connor, and Smith 2012). Therefore, information that can be used to discover and predict strikes is expected to be scarce in these provinces. Again, we will apply the NP classifier trained in the baseline sample to this non-Guangdong sample. If this classifier performs well on the new sample, decision makers can use a classifier trained in an environment with relatively abundant information to overcome the challenge of classification in an information-scarce environment where labeling of posts is likely to be much more costly and cannot be done in a timely manner. This is a potential advantage of the NP classification approach in its ability to transfer knowledge from one domain to another.

4.3. Data Preprocessing

We now describe how we process the unstructured raw Sina Weibo posts so that they can be fed to learning algorithms. The first step is to generate post labels. A decision maker's interest is to learn strike events which are a form of workers'

²The filter for strike includes the following list of keywords, which commonly appear with the subject: "罢工(worker strike)," "工潮(worker strike)," "罢市(shopkeeper strike)," "罢课(class boycott)," "罢驶(stop driving)," "罢驾(stop driving)," "罢运(transportation worker strike)."

collective action and reflect ongoing social and economic problems. Labeling posts according to the decision makers' interest turns out to be nontrivial for two reasons. First, in terms of substance, some posts related to strikes are about events in history or in other countries without implications for current events. Second, linguistically, the word "strike" is widely used in many different contexts, literally and metaphorically. For example, in Chinese, in the sentence "my computer/my cell phone is on strike," "strike" means "has stopped working." In the sentence "A person's body/brain is on strike," "strike" means "is not functioning normally." This type of linguistic ambiguity exists in many languages. We specified a set of rules to capture these subtleties.

As a trial, we outsourced the labeling task to Amazon Mechanical Turk. Despite the active responses, the label quality was subpar, having many errors and inconsistencies. Realizing the difficulty of the task, we switched to expert labeling. We hired two Chinese-speaking experts to manually categorize the raw posts into "strike related" (class 0) and "strike unrelated" (class 1). Class 0 posts are about worker strikes, including student strikes, taxi driver strikes, and merchant strikes, whereas class 1 posts contain the keyword "strike" but are using the word metaphorically to describe the malfunctioning of computers, elevators or other objects.

After trial and error, the two experts achieved high quality and consistent labeling in several trial samples. They then labeled the five aforementioned random samples: the baseline sample (GD-Q1), the three other samples in Guangdong after the first quarter in 2012 (GD-Q2, GD-Q3, and GD-Q4), and the sample outside of Guangdong (NGD). Overall, among the 12,500 posts in these five samples, 3237 posts are labeled as "strike related" (Class 0) and 9263 as "strike unrelated" (Class 1).

To decipher which Chinese characters form meaningful words, we apply *The Stanford Segmenter* (Tseng et al. 2005), which uses a Chinese treebank (CTB) segmentation model and breaks down input messages into disjointed words. After removing nonmeaningful stop words, we create a dictionary of unique words and generate a frequency matrix that counts the number of times each word appears in each post, based on the dictionary. The *strikes* matrix, containing 12,500 rows (posts) and 34,968 columns (features), is used to engineer features in topic modeling.

4.4. Feature Engineering

In the preprocessed *strikes* dataset, the size of vocabulary dictionaries is much larger than the number of posts. This high-dimensional problem can be handled with various techniques. For example, one can use marginal screening methods such as sure independence screening (Fan and Lv 2008), nonparametric independence screening (Fan, Feng, and Song 2011), and the Kolmogorov-Smirnov (KS) test, interaction screening methods (Hao and Zhang 2014; Fan et al. 2015), the forward stepwise selection, shrinkage methods such as LASSO (Tibshirani 1996) and SCAD (Fan and Li 2001), or dimension reduction methods such as principal component analysis.

These methods, however, all overlook the semantic structures possessed by corpora datasets. Thus, we adopt latent

Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003; Teh, Newman, and Welling 2007; Grimmer and Stewart 2013), which is a popular generative probabilistic model designed for large corpora. In this model, documents (posts) are represented as random mixtures over latent topics and each topic is represented as a distribution over words. We train the LDA model using the R package `topicmodels` and select "Gibbs sampling" as the fitting method. With a predetermined K , we extract K topics that serve as new features. The posterior distribution over these K topics in each document will be the feature values.

4.5. Results

In this subsection, we present the main results of the analysis which follows the pipeline depicted in Figure 4. Alongside the results, we discuss their real-world implications. We also address several nuanced technical issues that are important for the implementation of classification methods, hoping to provide quantitative social scientists some implementation guidelines to analyze their classification problems in various empirical settings.

4.5.1. Topic Modeling

In the use of LDA for feature engineering, specifying the number of topics K is essential. We use a *stability* criterion to select K . Concretely, for a candidate K , we randomly select half of the posts to apply LDA. This process is repeated 50 times. Every time, LDA outputs K topics. Each document is represented by posterior probabilities over these K topics, and each topic is represented by posterior probabilities over the vocabulary dictionary. We look at the top 20 keywords that have the largest posterior probabilities in each of the K topics. Based on these words, we decide whether a topic is truly related to the subject. We consider the number of topics K to be suitable if over 50 repetitions, the proportions of relevant topics have low variance. For illustrative purpose, we compare $K = 5$ and $K = 10$.

Table 1 lists the top 20 keywords for each topic in one repetition when $K = 10$, using the entire dataset of 12,500 strike posts. Even a casual reader (of the Chinese characters or their corresponding English translation) will recognize the 4th, 6th, and 10th topics as about actual worker strikes. In particular, the 4th topic is mostly about students boycotting classes, evident from the keywords "school," "student," "teacher," "student strike," and "demonstration." The 6th topic is about worker strikes in firms, evident from the keywords "company," "employee," "wage," "protest," "collective," "factory," and "staff." The 10th topic is about strikes in the transportation sector, evident from the keywords "strike," "driver," "vehicle," "taxi," "public transportation," "collective," "road," "bus," and "traffic." The remaining seven topics are irrelevant. Thus, in this repetition, the proportion of relevant topics is 3/10. Over the 50 repetitions, we calculated the variances of these proportions. In this regard, $K = 5$ and 10 output variances 0.0037 and 0.0018, respectively. By the stability criterion, we prefer $K = 10$.

One interesting observation is that, for a different choice of K , the feature words in a specific topic may contain different

Table 1. Top 20 keywords for ten topics from one repetition on the entire *strikes* dataset.

Topic 1	去 go 下午 afternoon	今天 today 然后 afterwards	吃 eat 鼻屎 mucus	明天 tomorrow 挖 pick	做 do 回来 come back	上班 work 回家 go home	爱 love 睡 sleep	睡觉 sleep 买 buy	今晚 tonight 晚上 night	偷笑 smirk 累 tired
Topic 2	罢工 strike 换 change	电脑 computer 最近 recently	手机 cellphone 直接 directly	打 beat 结果 consequence	发现 realize 系统 system	居然 surprisingly 部 part	电话 telephone 问题 problem	突然 sudden 彻底 thoroughly	开 open 博 broad	发 send 家里 home
Topic 3	罢工 strike 前 before	天 day 真的 really	说 speak 早上 morning	现在 now 知道 know	小 small 去 go	点 bit 分钟 minute	今天 today 走 walk	后 after 思考 think	下 down 一直 always	三 three 真是 indeed
Topic 4	年 year 学生 student	工人 worker 学校 school	罢课 student strike 美国 United States	中国 China 国家 country	上 previous 生活 life	月 month 组织 organization	工会 union 人民 people	游行 demonstration 中 in	政府 government 领导 leader	老师 teacher 举行 hold
Topic 5	能 can 希望 hope	让 let 出来 come out	可以 may 里 inside	时候 time 工作 work	没有 none 身体 body	还是 still 感觉 feel	这个 this 太阳 Sun	种 plant 还有 still	上 up 出 out	觉得 feel 一定 definitely
Topic 6	公司 company 新闻 news	员工 employee 集体 collective	事件 event 问题 problem	工资 wage 三 three	工作 work 分享 share	抗议 protest 月日 month-date	对 right 工厂 factory	罢市 shopkeeper strike 人员 staff	中 in 新 new	发生 happen 政府 government
Topic 7	罢工 strike 抓狂 go crazy	抓 clutch 眼泪 tears	狂 crazy 天气 weather	泪 tear 委屈 be wronged	系 be 话 word	今日 today 鄙视 despise	可怜 pity 住 reside	地 ground 空调 air-conditioner	生病 sick 搞到 get	衰 unfortunate 甘 this
Topic 8	罢工 strike 汗 sweat	想 think 看到 see	人 human 伤心 sad	玩 play 事情 thing	哈哈 haha 很多 many	事 thing 闹钟 alarm clock	找 find 放假 holiday	为什么 why 二 two	心情 mood 个人 individual	回复 reply 双 pair
Topic 9	罢工 strike 第一 first	次 time 对 right	开始 begin 周 week	时间 time 所有 all	过 pass 下 down	已经 already 第二 second	终于 finally 机场 airport	最后 at last 地方 place	继续 continue 草草 hasty	嘻嘻 LOL 完全 completely
Topic 10	罢工 strike 钱 money	司机 driver 没有 none	车 car 回 back	出租车 taxi 公交车 bus	的士 taxi 江门 Jiangmen	公交 public transportation 半 half	集体 collective 交通 traffic	小时 hours 事 thing	汕头 Shantou 全部 all	路 road 广州 Guangzhou

NOTE: The English translation of some keywords in Topic 7 are based their Cantonese meaning.

information. For example, the topic regarding “strikes in the transportation sector (topic 10 in Table 1)” appears both when $K = 5$ and when $K = 10$. In addition to relating to the subject, the topic keywords also contain information about the location of the events, which is valuable for decision making. However, when $K = 5$, only one location “汕头” (Shantou) appears as a feature in the selected topic; whereas when $K = 10$, two locations, “汕头” (Shantou) and “江门” (Jiangmen), appear. To investigate the cause of this difference, we manually read through the 2500 posts we selected from GD-Q1. Of them, 230 posts are about strikes in “Shantou” and 161 posts are about strikes in “Jiangmen.” We suspect that it is the relatively low frequency of “Jiangmen” that makes it vanish as a feature in the selected topic when $K = 5$. Thus, choosing a larger K may have the advantage of capturing a greater amount of valuable information. In the remaining part of the article, we set $K = 10$ unless otherwise specified.

It should be noted that, in this study, we choose $K = 5$ or $K = 10$ simply for an illustrative purpose. Practitioners can

select a set of desirable K 's based on their domain knowledge, time constraint, and financial budget.

4.5.2. NP Classification in the Baseline Sample

Fixing $K = 10$ in LDA, we apply both the classical and NP classification algorithms to the baseline dataset GD-Q1. The NP algorithms are implemented through the R package `nproc` (also available in Python). To better demonstrate the performance of NP classifiers, we implement three settings.

- Setting 1: We randomly split GD-Q1 into training and test sets of equal sizes (half of class 0 and half of class 1 data in training) 100 times. Hence, the class 0 proportion in a training set is the same as that in a test set. We set NP parameters: $\alpha = 0.2$ and $\delta = 0.3$.
- Setting 2: We randomly split class 0 data into three folds of equal sizes, and split class 1 data into two halves. We take 1/3 (one fold) class 0 data and 1/2 class 1 data as the training set

and use the other 2/3 class 0 data and the other half class 1 data as the test set. Thus, the class 0 proportion in the training set is half as much as in the test set. We again repeat the experiment 100 times. We set NP parameters: $\alpha = 0.2$ and $\delta = 0.3$.

- Setting 3: The same as in Setting 1, except that we now set NP parameters: $\alpha = 0.1$ and $\delta = 0.3$.

In each training set, we run LDA ($K = 10$) and construct a transformed training set which utilizes the learned topics as new features and the posterior probabilities over these topics as feature values. We then train classifiers based on the transformed training datasets. Type I and Type II errors are calculated using the corresponding transformed test set. The classification methods implemented include the classical versions of penalized logistic regression (PLR), naive Bayes (NB), support vector machines (SVM), random forest (RF), and sparse linear discriminant analysis (sLDA), together with their NP counterparts with corresponding parameters (e.g., $\alpha = 0.2$ and $\delta = 0.3$ for Setting 1 and Setting 2; $\alpha = 0.1$ and $\delta = 0.3$ for Setting 3).

Table 2 summarizes the average Type I and Type II errors in Setting 1 using the above classification methods under the classical approach (odd columns) and the NP approach (even columns, named with a prefix NP) over all 100 repetitions. Notably, all the classical methods produce a large Type I error and a small Type II error, with naive Bayes being the most extreme one, where the Type I error is 1. This is in part caused by the relatively large size of Class 1 in the training dataset. By contrast, all NP methods successfully control the Type I error within the target level, while producing a larger Type II error than that of the classical methods. This means that a decision maker using the NP methods can more accurately discover true information about strike events at the cost of screening some extra irrelevant information. In the current study, missing a strike-related post (class 0) may lead to delayed government responses, oversight in business decisions, and under-estimates of the strike incidence frequency in social studies. It is particularly costly when a hidden event may compound into a large scale issue and spread to other regions. Generally, a decision maker cares more about Type I error than Type II error. The larger Type II error associated with the NP classifiers implies

that an excessive amount of irrelevant information has been collected and another round of screening may be needed. The cost of such further screening appears insignificant (Qin, Strömberg, and Wu 2017). Overall, the NP classifier is preferable in many real-world applications.

Table 3 summarizes the average Type I and Type II errors in Setting 2, in which the class 0 proportion in the training set is half as much as its proportion in the test set. This mimics the real life scenario when censorship of the more-sensitive information is tightened, resulting in more scarce relevant information (a smaller class 0) in the observed data. According to our previous theoretical discussion, such more stringent censorship would shift the decision boundary of the classical oracle classifier more drastically, worsening Type I error of the classical classification methods. For example, PLR produces a Type I error of 0.965, which is larger than 0.914—its counterpart in Setting 1. By contrast, the NP oracle is unaffected by data distortion, and NP-PLR has a Type I error controlled below the prespecified $\alpha = 0.2$ in both Setting 1 and Setting 2. This phenomenon is consistent across all the five methods we implemented.

Table 4 summarizes the average Type I and Type II errors of these methods in Setting 3, which is the same as in Setting 1 except that we now use a new set of parameters $\alpha = 0.1$, $\delta = 0.3$. This second set of parameters is chosen to represent a scenario when decision makers face a higher cost of missing a strike event and wish to impose more stringent control over Type I error. Tables 2 and 4 demonstrate that, across different NP classifiers, Type I errors are uniformly controlled under the target level. In particular, when the upper bound of Type I error is reduced from ($\alpha = 0.2$) to ($\alpha = 0.1$), Type I errors of the NP classifiers are reduced below the new target level 0.1. These observations suggest that the NP methods provide an instrument for decision makers to fine-tune the target level of Type I errors according to circumstances.

In summary, the parameters α and δ in NP classification methods govern the trade-off between Type I and Type II errors, and the balance of this trade-off depends on the decision maker's objective and resources available. In the *strikes* example, the consequence of making Type I errors is severe—it could threaten government stability, jeopardize a politician's career, or mislead business decisions, whereas the cost of dealing with Type II

Table 2. Average error rates with $\alpha = 0.2$, $\delta = 0.3$ for the strike dataset over 100 repetitions, under Setting 1.

Error rates	PLR	NP-PLR	NB	NP-NB	SVM	NP-SVM	RF	NP-RF	sLDA	NP-sLDA
Type I	0.914	0.196	1	0.193	0.816	0.179	0.684	0.184	0.825	0.194
Type II	0.005	0.427	0	0.482	0.014	0.598	0.047	0.502	0.014	0.423

Table 3. Average error rates with $\alpha = 0.2$, $\delta = 0.3$ for the strike dataset over 100 repetitions, under Setting 2.

Error rates	PLR	NP-PLR	NB	NP-NB	SVM	NP-SVM	RF	NP-RF	sLDA	NP-sLDA
Type I	0.965	0.184	1	0.183	0.918	0.166	0.822	0.169	0.872	0.185
Type II	0.002	0.498	0	0.571	0.005	0.732	0.023	0.588	0.010	0.494

Table 4. Average error rates with $\alpha = 0.1$, $\delta = 0.3$ for the strike dataset over 100 repetitions, under Setting 3.

Error rates	PLR	NP-PLR	NB	NP-NB	SVM	NP-SVM	RF	NP-RF	sLDA	NP-sLDA
Type I	0.912	0.095	1	0.089	0.824	0.084	0.690	0.083	0.826	0.097
Type II	0.006	0.659	0	0.733	0.014	0.806	0.047	0.740	0.014	0.651

errors is small. Considering this preference for controlling Type I error, together with the data distortion problem, it is highly valuable to use classification methods under the NP paradigm rather than under the classical paradigm. In Appendix D of the supplementary materials, we also demonstrate how sparsity-inducing methods, such as NP-sLDA, help select meaningful topics, so that our approach achieves both good prediction performance and good interpretability.

4.5.3. Knowledge Transfer: NP Classifiers Across Datasets

In practice, decision makers often need to make decisions quickly. This time constraint sometimes restricts the amount of information available for developing predictive algorithms. For instance, a decision maker wants to assess the work conditions of a region (e.g., province) in April using social media posts. However, the number of relevant posts may be too small to train an effective classifier, or there might not be enough time and resources to hire experts to label posts. If this decision maker could use a classifier trained with data collected in the first quarter of the year, his or her learning would be more efficient and timely. Similarly, in a region where information related to worker strikes is scarce because of extensive censorship or limited supply, data analysis based on machine learning will benefit substantially from information collected in other regions with less censorship or more information supply.

The above discussion conveys the notion of *knowledge transfer*, which is implied by the invariance property of the NP classification paradigm. We now examine its validity empirically. We use classifiers trained on all posts in the baseline sample (GD-Q1) to classify posts in other datasets (GD-Q2, GD-Q3, GD-Q4, and NGD). From the previous section (recall Tables 2–4), we find that NP-sLDA performs the best among all methods we compared in terms of Type II errors. Thus, in this section, we focus on NP-sLDA only. In Table 5, we first present the results with parameters $\alpha = 0.1$ and $\delta = 0.3$. Note that the Type I errors are slightly larger than the target control level $\alpha = 0.1$. Nevertheless, this does not mean the failure of applying the classifiers trained in GD-Q1 because some regional and time-varying features are specific to a dataset and cannot be used for learning in other datasets. For example, “汕头” (Shantou) is a prefecture in Guangdong province where taxi-driver strikes occurred multiple times in the first quarter of 2012, and thus this locality appeared as a pronounced feature in topics selected from the baseline dataset. Unless the strike events in this locality lasted for a long period and became national, we would not expect it to appear as an important feature for data from samples

in other periods or from non-Guangdong provinces. In other words, we expect that the underlying populations over time or in different regions are not identical.

Being aware of the above learning barrier caused by features that are specific to a particular sample, we propose to have a smaller tolerance level to control the desirable Type I error. In particular, we trained the classifier using ($\alpha = 0.1, \delta = 0.05$). Table 6 presents the results under this new criterion. In contrast to the results in Table 5, the Type I error is now well controlled under the target level 0.1.

The above results demonstrate that, armed with NP classifiers, a decision maker, who is constrained by available information and time, can leverage information collected from previous periods or in circumstances where useful information was not severely censored. Of course, this knowledge transfer is feasible only if the post-censorship feature distributions remain sufficiently stable across datasets. Therefore, the results in Tables 5 and 6 provide suggestive evidence that censorship does not distort feature distributions that are important for the algorithm’s learning process. In other words, the assumption that warrants the invariance property of the NP oracle classifier is partially justifiable in the current empirical setting, although this assumption is not directly testable because uncensored data are not available. As mentioned in Section 1, our practice of using the NP algorithm to handle the data distortion problems differs from any existing practice in domain adaptation in that we do not use any data (labeled or unlabeled) on the target domain in the algorithm training process.

4.5.4. Knowledge Accumulation: NP Classifiers With Enlarged Training Data

In reality, a decision maker often accumulates information from the past. In view of the invariance property of the NP methods, this accumulated information can be used to facilitate learning if the feature distributions remain stable. We illustrate this point in the current case study. We first repeat Setting 3 using all posts from the Guangdong province, and compare the results with those in Table 4. In the comparison (presented in Table 7), GD-Q1 recollects the results in Table 4, and GD-ALL reports the results obtained from information in Guangdong over the entire four quarters. Clearly, when we use the larger dataset, the Type II error decreases, while the Type I error remains under control at 0.1. Furthermore, we include all posts from Guangdong as the training data, and test on the NGD dataset. We keep the parameters ($\alpha = 0.1, \delta = 0.05$) the same for comparison with Table 6. Table 8 shows that, with Type I

Table 5. Average error rates with $\alpha = 0.1, \delta = 0.3$ for posts from GD-Q2, GD-Q3, GD-Q4, and NGD.

Error rates	Guangdong-Q2	Guangdong-Q3	Guangdong-Q4	Non-Guangdong
Type I	0.133	0.141	0.109	0.106
Type II	0.558	0.563	0.516	0.533

Table 6. Average error rates with $\alpha = 0.1, \delta = 0.05$ for posts from GD-Q2, GD-Q3, GD-Q4, and NGD.

Error rates	Guangdong-Q2	Guangdong-Q3	Guangdong-Q4	Non-Guangdong
Type I	0.094	0.100	0.087	0.078
Type II	0.642	0.654	0.622	0.611

Table 7. Average error rates using NP-methods with $\alpha = 0.1$, $\delta = 0.3$ over 100 repetitions.

Error rates	NP-PLR		NP-NB		NP-SVM		NP-RF		NP-sLDA	
	GD-Q1	GD-ALL	GD-Q1	GD-ALL	GD-Q1	GD-ALL	GD-Q1	GD-ALL	GD-Q1	GD-ALL
Type I	0.095	0.094	0.089	0.093	0.084	0.090	0.083	0.091	0.097	0.093
Type II	0.659	0.420	0.733	0.491	0.806	0.587	0.740	0.476	0.651	0.425

NOTE: A comparison between using GD-Q1 and all data from Guangdong.

Table 8. Average error rates with $\alpha = 0.1$, $\delta = 0.05$ for posts from NGD, using classifier NP-sLDA trained on GD-Q1 only and trained on all data from GD (including GD-Q1, GD-Q2, GD-Q3, and GD-Q4), respectively.

Error rates	Trained over GD-Q1 only	Trained over all data from GD
Type I	0.078	0.059
Type II	0.611	0.407

error under control using NP-sLDA, the larger size of the training data decreases the Type II error one would achieve on the posts from non-Guangdong, even if the underlying population distributions in the GD and NGD datasets can be different.

5. Conclusion

Digital texts have become an important source of data for social scientists. With increasing sophistication in text mining to discover social events and to predict social behaviors, accurate classification of textual data for specific purposes is key to successful empirical analysis. However, while a wide range of textual analysis and machine learning techniques have been introduced into the social sciences (Grimmer and Stewart 2013; Gentzkow, Kelly, and Taddy 2017; Wilkerson and Casas 2017), the problem of data distortion has received relatively little attention. Being a fundamental data generation issue in statistical analysis, data distortion can cause serious problems in sampling, inference, and prediction. The current article is among the first efforts to study data distortion problems in the context of classifying large-scale textual data. Theoretically, we show that in the presence of unknown data distortion, the classical oracle classifier cannot be recovered even when the entire post-distortion population is available. By contrast, the NP oracle classifier is unaffected by data distortion. Practically, we study a case in which a decision maker classifies posts about worker strikes obtained from Sina Weibo—a leading Chinese microblogging platform that is subject to government censorship. We demonstrate that when one type of classification error (e.g., Type I error) is dominantly important, the NP classification algorithms allow users to control that type of error below a prespecified level. Although our problem setup involves the distortion parameters, our objective is not to estimate them, but to bypass the estimation needs for prediction purpose. In other words, we target a prediction problem rather than an inference problem. Our approach is to construct classifiers under the NP paradigm, and the theoretical underpinning behind this construction is the invariance property of the NP oracle classifier. It is important to note that the NP classification approach we propose is not specific to text classification. Instead, it can be used to handle more-general classification problems in the big data era when classification errors are asymmetric in importance. Plausible applications include control of epidemic diseases, crime

detection, social surveillance, and monitoring risky financial decisions, among many others.

Supplementary Materials

Supplementary materials include proofs, lemmas and some detailed implementation of algorithms.

Acknowledgments

The authors would like to thank the editor, associate editor, two statistical content referees, and the referee for reproducibility, for many constructive comments which have greatly improved the article. We would also like to thank Professor Jingyi Jessica Li for rounds of thoughtful discussions and suggestions, and the seminar participants at UCLA.

Funding

This work was partially supported by National Science Foundation grant NSF DMS 1613338.

References

- Bamman, D., O'Connor, B., and Smith, N. (2012), "Censorship and Deletion Practices in Chinese Social Media," *First Monday*, 17(3), <https://doi.org/10.5210/fm.v17i3.3943>. [68,74,75]
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010), "A Theory of Learning From Different Domains," *Machine Learning*, 79, 151–175. [69]
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [76]
- Cannon, A., Howse, J., Hush, D., and Scovel, C. (2002), "Learning With the Neyman-Pearson and Min-Max Criteria," Technical Report LA-UR-02-2951. [68]
- Chen, M., Weinberger, K. Q., and Blitzer, J. (2011), "Co-Training for Domain Adaptation," in *Advances in Neural Information Processing Systems*, pp. 2456–2464. [69]
- Chen, X., and Ang, P. H. (2011), "Internet Police in China: Regulation, Scope and Myths," in *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, pp. 40–52, edited by David Kurt Herold, Peter Marolt, New York: Routledge. [68,74,75]
- China Internet Network Information Center (2013), "The 32nd Statistical Report on Internet Development in China." [68]
- (2014), "The 33rd Statistical Report on Internet Development in China." [68]
- China Labour Bulletin (2012), *A Decade of Change: The Workers' Movement in China 2000–2010*, Hong Kong: China Labour Bulletin. [75]
- (2018), *The Workers' Movement in China: 2015–2017*, Hong Kong: China Labour Bulletin. [75]
- Chung, C.-F., Schmidt, P., Witte, A. D., and Witte, A. D. (1991), "Survival Analysis: A Survey," *Journal of Quantitative Criminology*, 7, 59–98. [69]
- Economist (2013), "China's Internet: A Giant Cage," *The Economist*. [68]
- Elkan, C. (2001), "The Foundations of Cost-Sensitive Learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978. [71]
- Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557, DOI: 10.1198/jasa.2011.tm09779. [76]

- Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [76]
- Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [76]
- Fan, Y., Kong, Y., Li, D., and Zheng, Z. (2015), "Innovated Interaction Screening for High-Dimensional Nonlinear Classification," *The Annals of Statistics*, 43, 1243–1272, DOI: 10.1214/14-AOS1308. [76]
- Fu, K., Chan, C., and Chau, M. (2013), "Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy," *IEEE Internet Computing*, 17, 42–50. [75]
- Gentzkow, M., Kelly, B. T., and Taddy, M. (2017), "Text as Data," Technical Report, National Bureau of Economic Research. [80]
- Grimmer, J., and Stewart, B. M. (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, 21, 267–297. [76,80]
- Hao, N., and Zhang, H. H. (2014), "Interaction Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 109, 1285–1301, DOI: 10.1080/01621459.2014.881741. [76]
- King, G., Pan, J., and Roberts, M. E. (2013), "How Censorship in China Allows Government Criticism But Silences Collective Expression," *American Political Science Review*, 107, 326–343. [68,74]
- (2014), "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation," *Science*, 345, 1251722. [68,74]
- Li, J. J., and Tong, X. (2016), "Genomic Applications of the Neyman-Pearson Classification Paradigm," in *Big Data Analytics in Genomics*, ed. K. C. Wong, Cham: Springer, pp. 145–167. [68]
- Lorentzen, P. (2014), "China's Strategic Censorship," *American Journal of Political Science*, 58, 402–414. [74]
- Luborsky, M. R., and Rubinstein, R. L. (1995), "Sampling in Qualitative Research: Rationale, Issues, and Methods," *Research on Aging*, 17, 89–113. [69]
- Qin, B., Strömberg, D., and Wu, Y. (2017), "Why Does China Allow Freer Social Media? Protests Versus Surveillance and Propaganda," *The Journal of Economic Perspectives*, 31, 117–140. [68,74,78]
- (2019), "Social Media and Protests in China," Working Paper. [74]
- Rigollet, P., and Tong, X. (2011), "Neyman-Pearson Classification, Convexity and Stochastic Constraints," *Journal of Machine Learning Research*, 12, 2831–2855. [68]
- Scott, C. (2005), "Comparison and Design of Neyman-Pearson Classifiers," Unpublished. [68]
- Teh, Y. W., Newman, D., and Welling, M. (2007), "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems*, pp. 1353–1360. [76]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [76]
- Tong, X. (2013), "A Plug-In Approach to Neyman-Pearson Classification," *Journal of Machine Learning Research*, 14, 3011–3040. [72]
- Tong, X., Feng, Y., and Li, J. (2018), "Neyman-Pearson (NP) Classification Algorithms and NP Receiver Operating Characteristic (NP-ROC) Curves," *Science Advances*, 4, eaao1659. [68,69,71,72,73]
- Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005), "A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005," in *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. [76]
- Wilkerson, J., and Casas, A. (2017), "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges," *Annual Review of Political Science*, 20, 529–544. [80]
- Woolley, S. C., and Howard, P. N. (2016a), "Automation, Algorithms, and Politics," *International Journal of Communication*, 10, 4882–4890. [68]
- (2016b), "Social Media, Revolution, and the Rise of the Political Bot," *Handbook of Media, Conflict, and Security* edited by Romy Frolich and Piers Robinson. London, UK: Routledge, 4882–4890. [68]
- Zadrozny, B., Langford, J., and Abe, N. (2003), "Cost-Sensitive Learning by Cost-Proportionate Example Weighting," in *IEEE International Conference on Data Mining*, p. 435. [71]
- Zhao, A., Feng, Y., Wang, L., and Tong, X. (2016), "Neyman-Pearson Classification Under High Dimensional Settings," *Journal of Machine Learning Research*, 17, 7469–7507. [72]
- Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., and Wallach, D. S. (2013), "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions," in *USENIX Security Symposium*, pp. 227–240. [68,74,75]