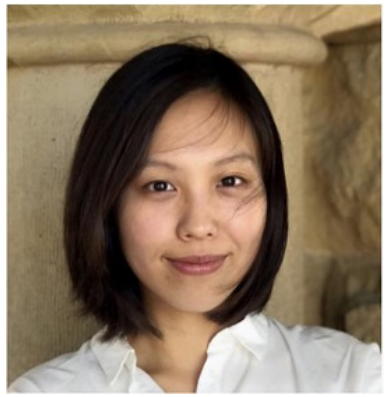
Intentional Control of Type I Error over Unconscious Data Distortion: A Neyman-Pearson Approach to Text Classification

Richard Zhao

Assistant Professor of Computer Science, Penn State University

Joint Work with Xin Tong (USC), Yanhui Wu (USC) and Lucy Xia (Stanford / Hong Kong University of Science and Technology)





Computational Textual Analysis in Social Sciences

- Political science (Grimmer and Stewart 2013; Lucas et al. 2015; Wilkerson and Casas 2017)
- Sociology (Evans and Aceves 2016; Lazer and Radford 2017)
- Economics (Gentzkow, Kelly, and Taddy 2018)
- Media bias (Groseclose and Milyo 2005; Gentzkow and Shapiro 2010; Qin, Stromberg, and Wu 2018)
- Economic uncertainty (Baker, Bloom, and David 2016; Bloom et al. 2018)
- Industrial organization (Hoberg and Phillips 2016)
- Financial markets (Tetlock 2007)

Problems in Text Classification

- Textual analysis for data description: fine
- Textual analysis to generate estimates of socially relevant phenomena (e.g., event discovery; nowcasting): maybe problematic
  - Training environment: feature engineering, labeling
  - Sampling: non-random sample
  - Generalization: too many but setting-specific data
  - **Data distortion: observed data misrepresent the true population**
- Textual data are vulnerable to manipulation.

Data Distortion

- Downward distortion: censorship
  - Chinese government extensively censors social media (e.g., King et al. 2013, 2014)
  - Censorship is ad hoc and unpredictable (e.g., Chen et al. 2011); hard to figure out the censorship scheme
- Upward distortion: information inflation
  - Manipulation behind closed doors: posts injected by bots, Internet trolls
  - "Yes Men": say what your boss wants you to say, e.g., propaganda
  - Herding: say what your peers say, e.g., Facebook "disinformation"

This Talk

- Studies problems with classical classification methods in the presence of data distortion
- Offers a solution based on the Neyman-Pearson classification paradigm

This Talk

- Studies problems with classical classification methods in the presence of data distortion
- Offers a solution based on the Neyman-Pearson classification paradigm
- Roadmap
  - NP-classification paradigm
  - Case study: use censored social media data to discover political events (strikes)

Binary Classification

- Features $X \in \mathcal{X} \subset \mathbb{R}^p$
- Class labels $Y \in \{0, 1\}$
- A classifier is a data dependent mapping $h : \mathcal{X} \to \{0, 1\}$
- Classification error ("risk")

$$\begin{aligned} R(h) &= \mathbb{P}(h(X) \neq Y) \\ &= \mathbb{P}(Y = 0) R_0(h) + \mathbb{P}(Y = 1) R_1(h), \end{aligned}$$

where

  - $R_0(h) = \mathbb{P}\left(h(X) \neq Y | Y = 0\right)$ denotes the type I error,
  - $R_1(h) = \mathbb{P}\left(h(X) \neq Y | Y = 1\right)$ denotes the type II error.

- Classical goal: find a classifier h to minimize $R(h)$

Oracle under Data Distortion

- Class priors: $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$
- Distortion rates: $\beta_0 = (\beta_0^-, \beta_0^+)^\top$ $\beta_1 = (\beta_1^-, \beta_1^+)^\top$
- Oracle classifier: $h^*(x) = \mathbb{I}(\eta(x) > 1/2)$, where $\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$
- Suppose that Class 0 $(X|Y = 0)$ and Class 1 $(X|Y = 1)$ have probability density functions f_0 and f_1. The oracle classifier under the classical paradigm regarding the after-distortion population is

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

- Impossible to recover the true oracle classifier (even with unlimited data) unless the distortion rates are known!

Intentional Control over Errors

- The after-distortion classical oracle classifier may have type-I error out of control.
- Tentative solution: reweigh the objective function
  - cost-sensitive learning (Elkan, 2001; Zadrozny et al, 2003)
  - ad hoc assignment of costs can be misleading
- What if we decouple type-I and type-II errors? (Neyman-Pearson Lemma)
- Construct $\hat{h}$ such that

$$\mathbb{P}(R_0(\hat{h}) \leq \alpha) > 1 - \delta,$$

for given α and δ, where δ is a user-specified violation rate.

Comparison of Two Classification Paradigms

Binary classification

Paradigm	Oracle classifier
Classical	$h^* = \arg\min R(h)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\leq\alpha} R_1(\phi)$

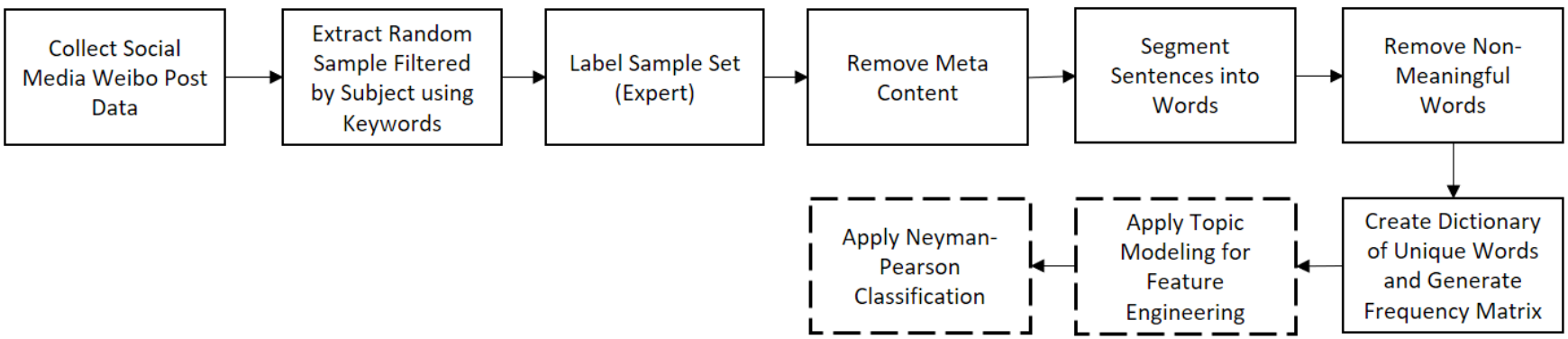
where α reflects users' conservative attitude towards the type I error.

Political Information on Social Media in China

- Public information on political issues and social events is scarce in authoritarian governments
- Sina Weibo - the Chinese equivalent to Twitter
- Fine-tuned censorship: ad hoc deletion of posts instead of closing user accounts
- Effectiveness of after-censorship information: quite useful (Qin, Stromberg, and Wu 2017; 2018)
- Facing a problem of text classification in the presence of unpredictable censorship

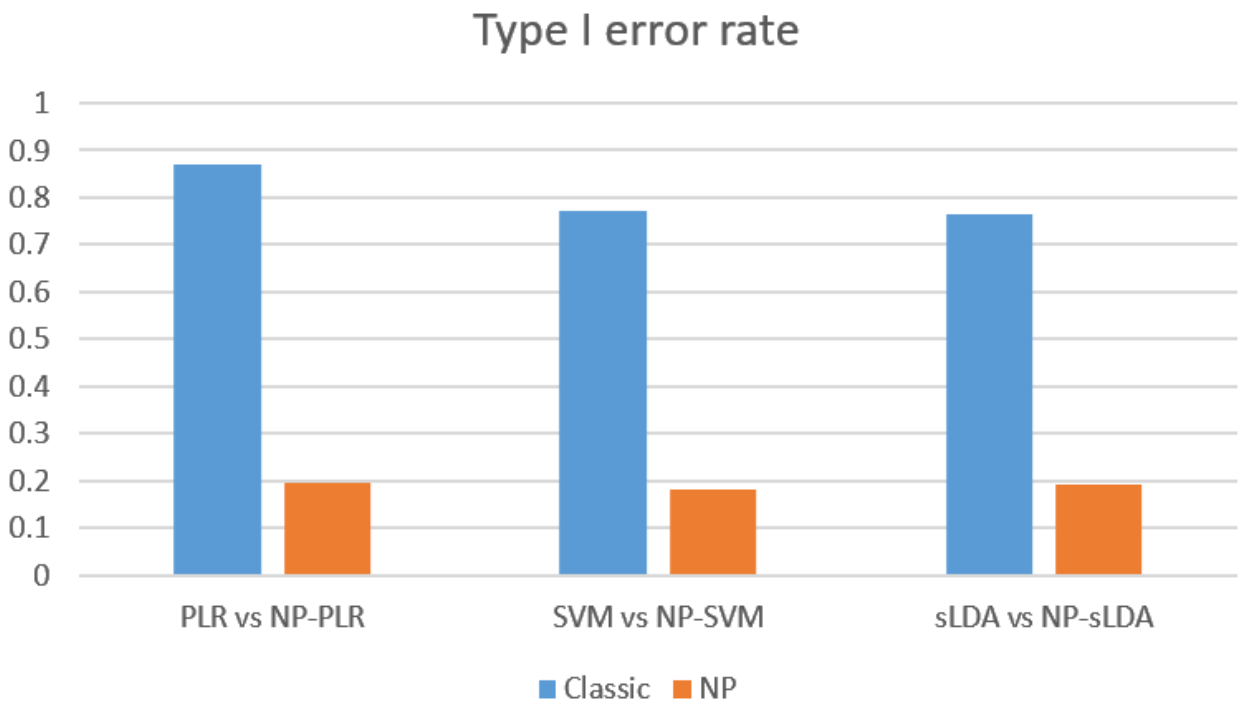
Data Processing

- Crawl 10 million posts about political issues from Sina Weibo in 2012
- Filter by subjects to obtain 221k posts about strikes.
- Sample selection: a sample of 4579 strike posts in two randomly selected months from a province (Guangdong)



Results: Strikes

- Detect posts: strikes (class 0) or not (class 1)
- Methods implemented:
  - penalized logistic regression (PLR)
  - support vector machine (SVM)
  - sparse linear discriminant analysis (sLDA)



Conclusion

- Problems with classification of large-scale textual data
  - A conflict between data distortion and the classical classification objective
  - The conflict is exacerbated when the cost of type-I error is large.
- Proposed solution
  - We propose a NP-classification method to bypass a class of data distortion problems and develop an algorithm that is flexible and adaptive to popular machine learning classification techniques.
  - We illustrate the proposed method by case studies using Chinese social media data to identify political events.

Thank you!