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

Yanhui Wu

Department of Finance and Business Economics
University of Southern California

Joint with Xin Tong (USC), Lucy Xia (Stanford), and Richard Zhao (Penn State)

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)

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)

Example: Social Media and Political Action in China

- Qin, Stromberg, and Wu (2017, JEP)
- Study how Chinese governments use social media for surveillance, monitoring, and propaganda
- 13.2 billion posts from Sina Weibo - the Chinese equivalent to Twitter during 2009-2013
- Use simple textual analysis techniques for data description and event prediction.

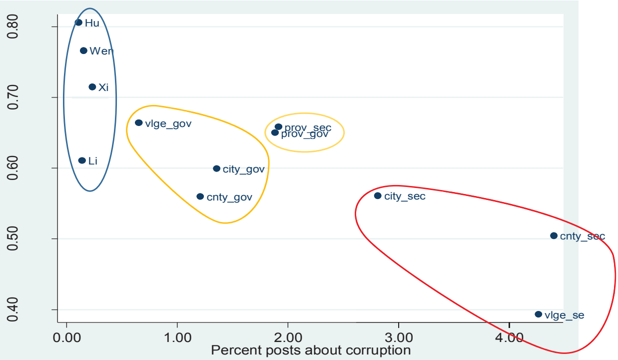
Step 1: Topic Modeling

- Use key words (e.g., strike, protest) to filter relevant posts
- Apply topic modeling (e.g., LDA)

Conflict			Protest			Strike		
Sensitivity: Very High #posts: 382,232			High 2,526,325			Medium 1,348,964		
Freq.	Word	Translation	Freq.	Word	Translation	Freq.	Word	Translation
322,797	镇压	Suppression	647,711	示威	Demonstration	1,361,854	罢工	Strike
32,117	冲突	Conflict	534,784	静坐	Sit-in	69,068	罢课	Student strike
19,124	警民	Police and People	430,112	自焚	Self-immolation	101,887	工人	Workers
17,460	催泪弹	Tear-gas bomb	260,574	讨薪	Ask for compensation	98,822	电脑	Computer
31,161	矛盾	Contradictory	346,836	游行	Parade	65,557	出租车	Taxi
40,286	警察	Police	164,367	请愿	Petition	164,549	泪	Tears
14,271	官民	Officials and people	113,936	示威者	Demonstrators	46,219	工会	Trade union
31,935	暴力	Violence	109,339	堵路	Stops up the road	91,051	抓狂	Driven nuts
130,036	被	By	166,600	抗议	Protest	55,687	司机	Drivers
74,391	政府	Government	101,845	集会	Assembly	48,845	集体	Collective
12,002	宽恕	Forgiveness	118,262	农民工	Migrant workers	52,066	员工	Staff
12,764	武力	Military force	103,975	思	Thinking	157,937	今天	Today
18,951	军队	Army	80,481	静静	Static	24,477	的士	Taxi
29,566	民众	Populace	60,237	闲谈	Chat	22,559	法国人	French
14,701	叙利亚	Syria	58,318	人非	Shortcomings of people	51,479	上班	Going to work
20,170	抗议	Protest	72,753	民工	Laborers	16,290	罢市	Merchant strike
60,068	人民	People	63,719	白宫	White House	40,827	抗议	Protest
21,521	村民	Villagers	130,198	坐	Sitting	86,612	手机	cellphone
10,264	起义	Revolt	60,957	己	Oneself	17,679	罢	Strike
10,150	开枪	Gunfire	37904	玩火自焚	Being made to pay for one's evil doings	41586	工资	Wages

Step 2: Sentiment Analysis

- Use a standard Chinese dictionary to count positive vs. negative words in a post



Step 3: Event Discovery

- Classify posts into two categories: event vs. others
- Training data: manually classify a sample of 6000 randomly drawn posts (after filtering)
- Machine learning (SVM): automatically classify all relevant posts after data testing
- Use the automatically-labelled posts to predict real events (location/time) based on certain statistical models

Step 3: Event Discovery

- Classify posts into two categories: event vs. others
- Training data: manually classify a sample of 6000 randomly drawn posts (after filtering)
- Machine learning (SVM): automatically classify all relevant posts after data testing
- Use the automatically-labelled posts to predict real events (location/time) based on certain statistical models

predicted_probability	province	prefecture	date	event	_description	event_location
0.011485248	四川	成都	11/6/2009	私立学校教师罢课	Teachers in private schools strike	四川成都
0.013135893	陕西	西安	11/7/2009	私立学校教师罢课	Teachers in private schools strike	四川成都
0.011441577	湖北	武汉	1/1/2013	的士司机罢工	Taxi drivers strike	湖北武汉
0.012353521	广东	深圳	1/13/2013	深圳富士康罢工	Shenzhen Foxconn workers strike	广东深圳
0.011263852	江西	南昌	1/13/2013	南昌富士康工人罢工	Nanchang Foxconn workers strike	江西南昌
0.011537749	广东	广州	4/1/2013	香港码头工人罢工	Dockers in Hongkong srike	Hongkong
0.011536272	广东	广州	4/2/2013	香港码头工人罢工	Dockers in Hongkong srike	Hongkong
0.011378806	广东	广州	4/3/2013	香港码头工人罢工	Dockers in Hongkong srike	Hongkong
0.015047119	广东	广州	4/11/2013	凤凰古城罢市	Shopkeepers in Fenghuang strike	湖南湘西
0.012744553	广东	深圳	4/11/2013	凤凰古城罢市	Shopkeepers in Fenghuang strike	湖南湘西
0.01147429	湖北	武汉	4/11/2013	凤凰古城罢市	Shopkeepers in Fenghuang strike	湖南湘西
0.012634203	湖南	长沙	4/11/2013	凤凰古城罢市	Shopkeepers in Fenghuang strike	湖南湘西
0.012158257	四川	成都	4/11/2013	凤凰古城罢市	Shopkeepers in Fenghuang strike	湖南湘西
0.013271377	广东	广州	5/1/2013		Various strikes in other areas	
0.012999576	广东	深圳	5/1/2013		Various strikes in other areas	
0.012382323	广东	广州	4/22/2013		noisy information	
0.013629925	广东	广州	4/23/2013		noisy information	

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 mis-present 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

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 robots, internet trolls
  - "Yes Men": say what your boss wants you to say, e.g., propaganda
  - Herding: say what your peers say, e.g., Facebook information

This Paper

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

This Paper

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

Binary Classification

- Features $X \in \mathcal{X} \subset \mathbb{R}^p$
- Class labels $Y \in \{0, 1\}$

Binary Classification

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

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.

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 $\pi_1 = \mathbb{P}(Y = 1)$

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 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 $\eta(x) = \mathbb{E}(Y|X = x)$

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: $h^*(x) = \mathbb{1}(\eta(x) > 1/2)$, where $\eta(x) = \mathbb{E}(Y|X = x)$
- **Theorem 1** 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

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

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 $\eta(x) = \mathbb{E}(Y|X = x)$
- **Theorem 1** 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

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

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

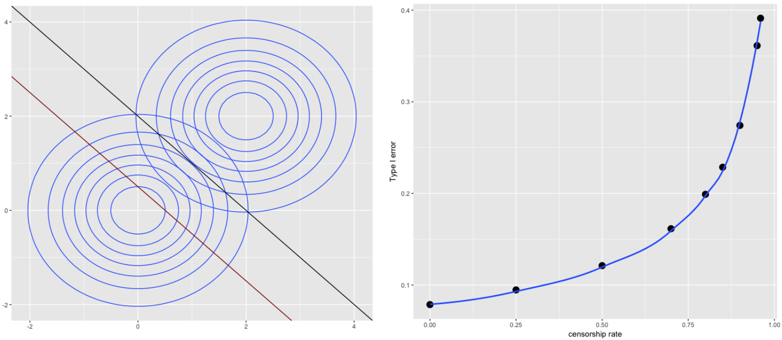
Classification Errors under Data Distortion

- **Corollary 1** Following Theorem 1, $R_0(h^*_{(\beta_0,\beta_1)})$, type I error of $h^*_{(\beta_0,\beta_1)}$, increases in β_0^- and decreases in β_1^-.

Classification Errors under Data Distortion

- **Corollary 1** Following Theorem 1, $R_0(h^*_{(\beta_0,\beta_1)})$, type I error of $h^*_{(\beta_0,\beta_1)}$, increases in β_0^- and decreases in β_1^-.
- Illustrative example
  - Only keep β_0^- active: censorship on Class 0
  - Gaussian distribution: $f_0 \sim \mathcal{N}(\mu_0, \Sigma)$ and $f_1 \sim \mathcal{N}(\mu_1, \Sigma)$
  - Parameters: $\mu_0 = (0,0)^\top$, $\mu_1 = (2,2)^\top$, $\Sigma = I$, $\pi_0 = 0.5$ and $\beta_0^- = 0.95$

Type-I Error and Censorship



Intentional Control over Errors

- The after-distortion classical oracle classifier may have type-I error out of control.

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

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)

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

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

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

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.
- Cost-sensitive learning does not deliver such an $\hat{h}$

Neyman-Pearson (NP) Classification Paradigm

The NP paradigm seeks a classifier that satisfies:

$$\min_{R_0(h)\leq\alpha} R_1(h)\,,$$

where α is a user-specified level (e.g., 5%).

Neyman-Pearson (NP) Classification Paradigm

The NP paradigm seeks a classifier that satisfies:

$$\min_{R_0(h) \leq \alpha} R_1(h) \,,$$

where α is a user-specified level (e.g., 5%).

- Early work in the engineering community: Cannon et.al. (2002); Scott (2005)
- Recent research on NP classification:
  - methodology: Rigollet and Tong (2011); Tong (2013); Zhao et al. (2016)
  - applications in bio/medicine: Li and Tong (2016); Tong et al. (2018)

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.

NP Oracle Invariant to Class Priors

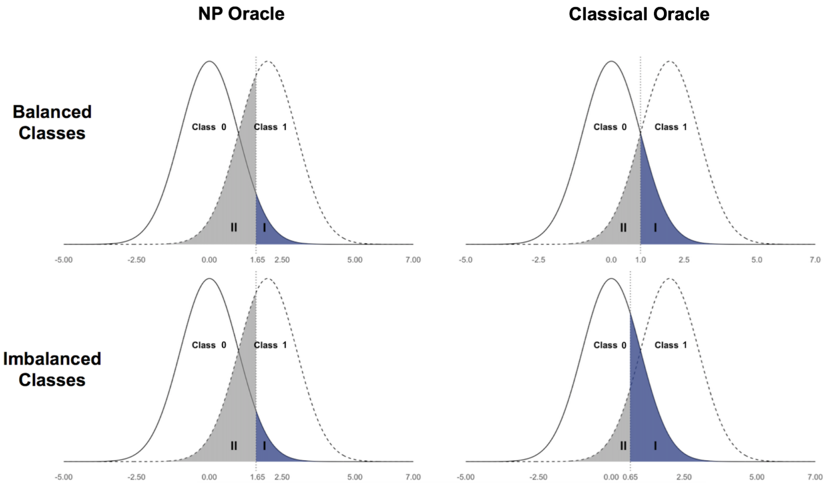
- **Theorem 2** Given any distributions for $(X|Y=0)$ and $(X|Y=1)$, the NP oracle classifier ϕ^*_α is invariant under distortion at various rates β_0 (on class 0) and β_1 (on class 1), regardless of whether before-distortion classes are balanced.

NP Oracle Invariant to Class Priors

- **Theorem 2** Given any distributions for $(X|Y=0)$ and $(X|Y=1)$, the NP oracle classifier ϕ^*_α is invariant under distortion at various rates β_0 (on class 0) and β_1 (on class 1), regardless of whether before-distortion classes are balanced.

- Proof: The constrained optimization that defines ϕ^*_α does not involve the class priors $\pi_0 = \mathbb{P}(Y=0)$ and $\pi_1 = \mathbb{P}(Y=1)$, so any change in class proportions (distortion) does not change the NP oracle.

Graphical Illustration



Political Information on Social Media in China

- Public information on political issues and social events is scarce in authoritarian regimes.
- Social media generate double-edge political information
  - facilitate political communication and improve surveillance
  - threaten regime stability

Political Information on Social Media in China

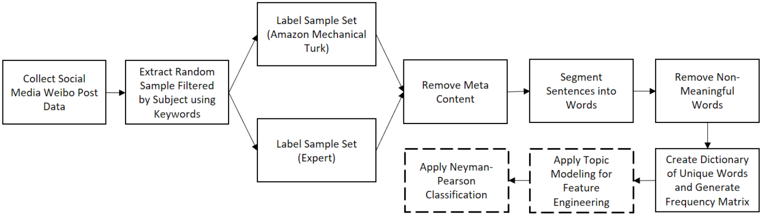
- Public information on political issues and social events is scarce in authoritarian regimes.
- Social media generate double-edge political information
  - facilitate political communication and improve surveillance
  - threaten regime stability
- Fine-tuned censorship: ad hoc deletion of posts instead of closing user accounts

Political Information on Social Media in China

- Public information on political issues and social events is scarce in authoritarian regimes.
- Social media generate double-edge political information
  - facilitate political communication and improve surveillance
  - threaten regime stability
- 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)
- Surveillance and data collection: how to use social media posts to predict and discover political events?
- We are 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)
- $n_0 = 774$, $n_1 = 3,805$ and $p = 16,895$
- Topic modeling to engineer new features and apply NP classifiers
- Methods implemented:

Results: Strikes

- Detect posts: strikes (class 0) or not (class 1)
- $n_0 = 774$, $n_1 = 3,805$ and $p = 16,895$
- Topic modeling to engineer new features and apply NP classifiers
- Methods implemented:
  - penalized logistic regression (PLR)
  - support vector machine (SVM)
  - sparse linear discriminant analysis (sLDA)

Results: Strikes

- Detect posts: strikes (class 0) or not (class 1)
- $n_0 = 774$, $n_1 = 3,805$ and $p = 16,895$
- Topic modeling to engineer new features and apply NP classifiers
- Methods implemented:
  - penalized logistic regression (PLR)
  - support vector machine (SVM)
  - sparse linear discriminant analysis (sLDA)
- Set $\alpha = 0.2$ and $\delta = 0.3$

Results: Strikes

- Detect posts: strikes (class 0) or not (class 1)
- $n_0 = 774$, $n_1 = 3,805$ and $p = 16,895$
- Topic modeling to engineer new features and apply NP classifiers
- Methods implemented:
  - penalized logistic regression (PLR)
  - support vector machine (SVM)
  - sparse linear discriminant analysis (sLDA)
- Set $\alpha = 0.2$ and $\delta = 0.3$
- Results from classic and NP methods:

Error rates	PLR	NP-PLR	SVM	NP-SVM	sLDA	NP-sLDA
type I	.869	.195	.772	.181	.763	.192
type II	.006	.352	.014	.683	.017	.358

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.

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!