Censorship of Online Encyclopedias: Implications for NLP Models

Eddie Yang and Margaret E. Robert

Presenter: Yong Xie

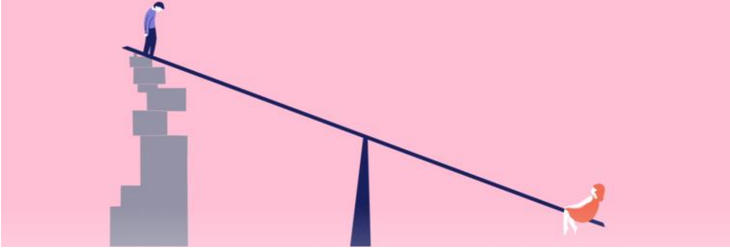
Outlines

- Motivation
- Background
- Word Embedding Comparison
- Effect on Downstream Applications

Introduction

Many biases origin from the data, and the biases are introduced collectively.

- Gender bias
- Racial bias
- Stereotypes
- …



What if the data are affected by specific institutions?

Background on Censorship

- Censorship on Wikipedia
  - May 2001: Chinese Wikipedia was launched
  - June 2015: Chinese Wikipedia was blocked in China
  - April 2019: All versions of Wikipedia were block in China
- Consequence
  - Decreased views
  - Less contributions
  - Strengthened Chinese equivalence Baidu Baike

繁简两种字体的网站标志

Background on Censorship

- Censored Topics
  - Democratic concepts
  - Propaganda
  - Historical events pertaining to Chinese Communist Party (CCP)
  - Political figures

What are the effect of the censorship on word embeddings and NLP applications?

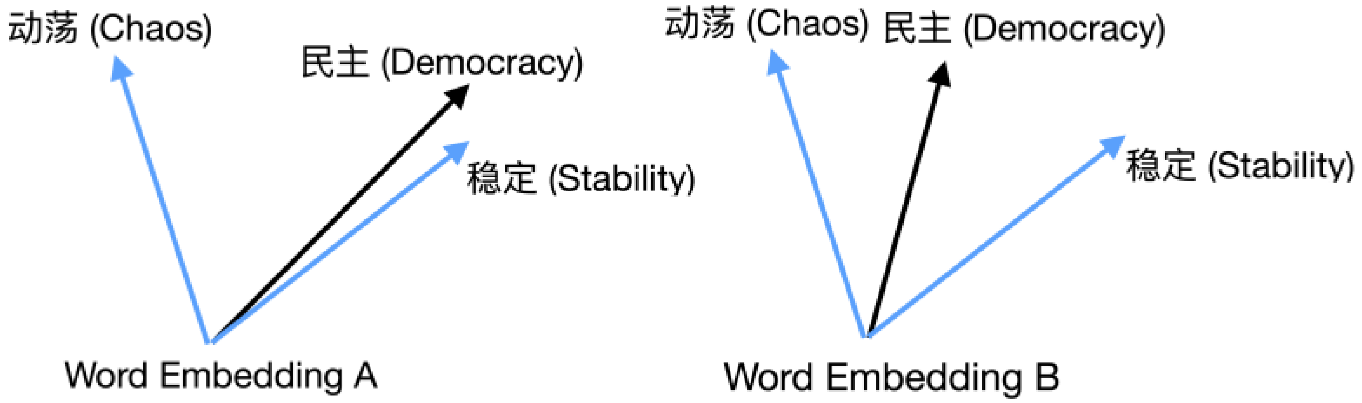
Word Embedding Comparison

- Method
  - Comparing the distance between a set of target words and attribute words in embedding space

Target words	Democratic concepts and ideas	Democratic values, procedures of democracy, channels for voicing
	Targets of propaganda	Social control, CCP, historical events, important figures
Attribute words	Propaganda attribute words	Stability/chaos, prosperity/decline
	Evaluative attribute words	General words of polarized sentiment

  - We can evaluate each target word (the concept we are interested in) relative to a positive/negative attribute word.

Word Embedding Comparison



- ‘Democracy’ in word embedding A has a more positive connotation than in word embedding B.

Target Word Selection

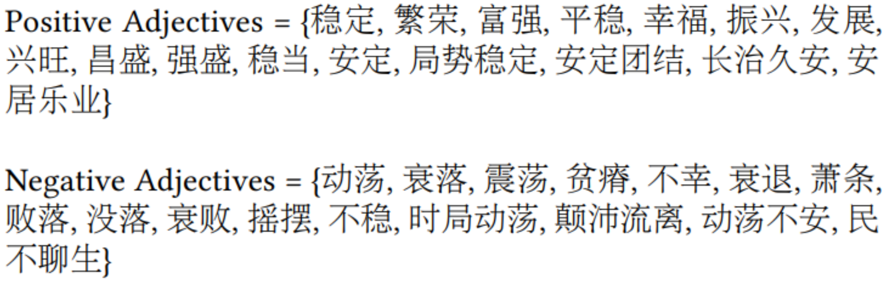
1. Select the most representative word by human for each category, e.g., the representative word for procedure of democracy is ‘election’.
2. Select 50 closest words in the embedding spaces in terms of cosine similarity.
3. Select words that are thought to be synonymous among the 100 closest words, after dropping some domain specific words.

- P.S. Use names of figures or events as the words.
- Step 1 and 3 are done by native speakers.

Attribute Word Selection

- Propaganda attribute words
  - Whether the word association is consistent with CCP propaganda
- Evaluative attribute words
  - Whether the target words are more generally associated between corpus.
  - A sentiment lexicon by Wang and Ku (2016), after dropping neutral words.

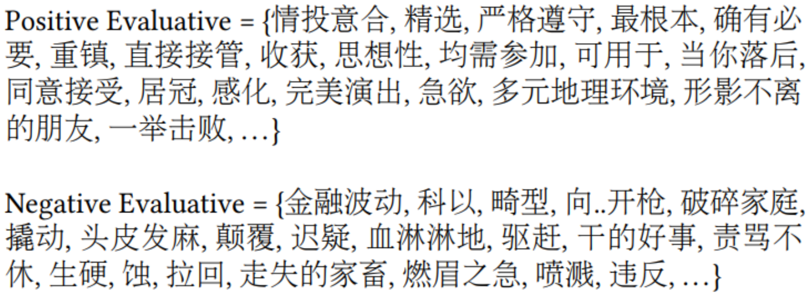
Propaganda Attribute Word



Positive Adjectives = {稳定, 繁荣, 富强, 平稳, 幸福, 振兴, 发展, 兴旺, 昌盛, 强盛, 稳当, 安定, 局势稳定, 安定团结, 长治久安, 安居乐业}

Negative Adjectives = {动荡, 衰落, 震荡, 贫瘠, 不幸, 衰退, 萧条, 败落, 没落, 衰败, 摇摆, 不稳, 时局动荡, 颠沛流离, 动荡不安, 民不聊生}

Examples of Evaluative Attribute Words



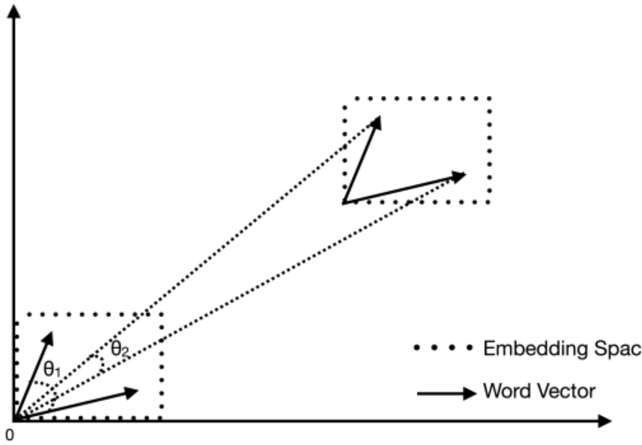
Positive Evaluative = {情投意合, 精选, 严格遵守, 最根本, 确有必要, 重镇, 直接接管, 收获, 思想性, 均需参加, 可用于, 当你落后, 同意接受, 居冠, 感化, 完美演出, 急欲, 多元地理环境, 形影不离的朋友, 一举击败, ...}

Negative Evaluative = {金融波动, 科以, 畸型, 向..开枪, 破碎家庭, 撬动, 头皮发麻, 颠覆, 迟疑, 血淋淋地, 驱赶, 干的好事, 责骂不休, 生硬, 蚀, 拉回, 走失的家畜, 燃眉之急, 喷溅, 违反, ...}

Word Association

- Embeddings (Li et al, 2018)
  - A 300-dimensional embedding trained on Baike with Word2Vec Algo.
  - A 300-dimensional embedding trained on Chinese Wikipedia with Word2Vec Algo.
  - Embeddings trained on *People's Daily*, a state-run newspaper.
- Metric: Cosine Similarity
  - Word embeddings produced by Word2Vec

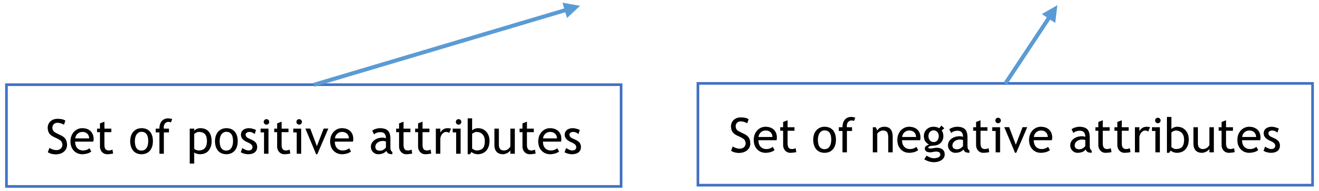
    Embed words in non-aligned space.
  - Standardization to make origin centroids



Word Association

1. Calculate positive connotation of target words in different embeddings

$$s(t, A, B) = \text{mean}_{p \in A} \cos(\vec{t}, \vec{p}) - \text{mean}_{q \in B} \cos(\vec{t}, \vec{q})$$



Set of positive attributes

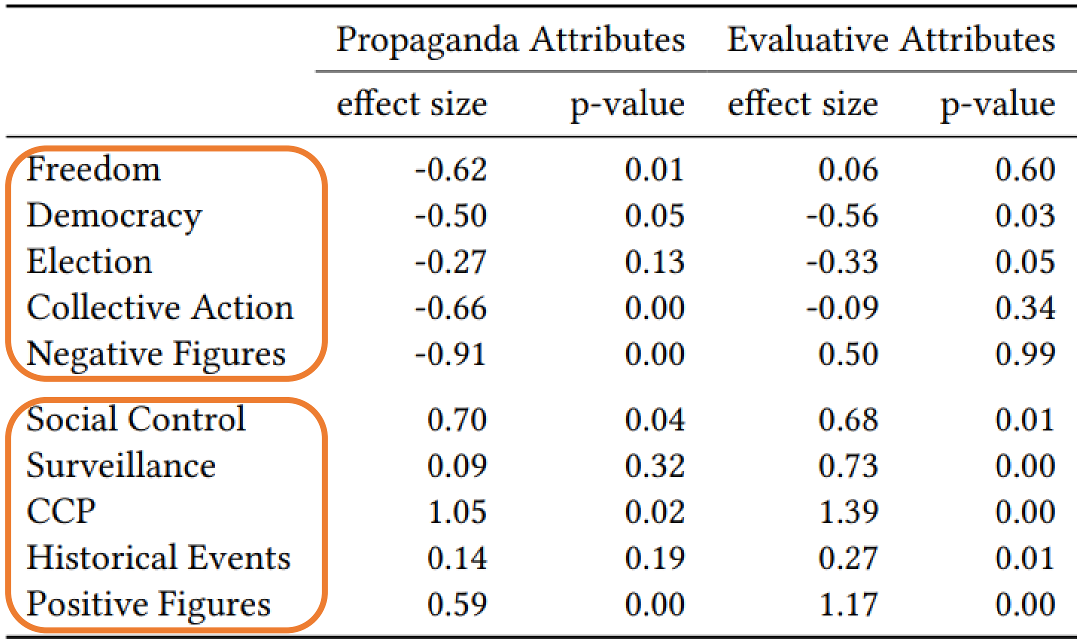
Set of negative attributes

2. Pair t-test compare the positive connotation of target words

$$\frac{\text{mean}_{i \in a} s(x_i, A_i, B_i) - \text{mean}_{i \in b} s(x_i, A_i, B_i)}{\text{std.dev}_i s(x_i, A_i, B_i)}$$

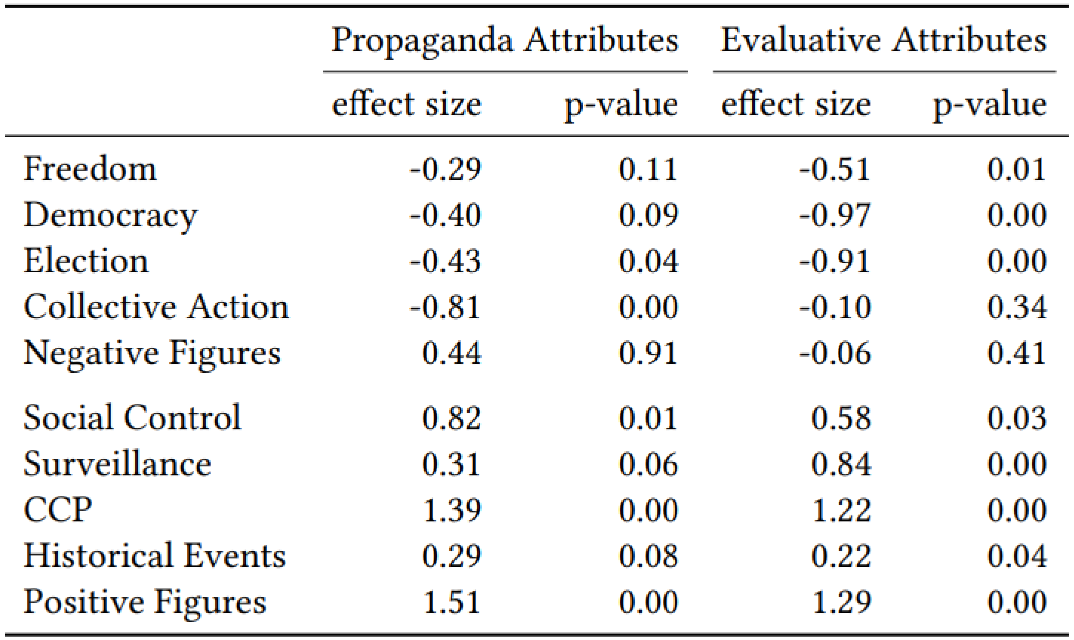
Results

Table 2: Wikipedia vs. Baidu Baike



	Propaganda Attributes		Evaluative Attributes	
	effect size	p-value	effect size	p-value
Freedom	-0.62	0.01	0.06	0.60
Democracy	-0.50	0.05	-0.56	0.03
Election	-0.27	0.13	-0.33	0.05
Collective Action	-0.66	0.00	-0.09	0.34
Negative Figures	-0.91	0.00	0.50	0.99
Social Control	0.70	0.04	0.68	0.01
Surveillance	0.09	0.32	0.73	0.00
CCP	1.05	0.02	1.39	0.00
Historical Events	0.14	0.19	0.27	0.01
Positive Figures	0.59	0.00	1.17	0.00

Table 3: Wikipedia vs. *People's Daily*



	Propaganda Attributes		Evaluative Attributes	
	effect size	p-value	effect size	p-value
Freedom	-0.29	0.11	-0.51	0.01
Democracy	-0.40	0.09	-0.97	0.00
Election	-0.43	0.04	-0.91	0.00
Collective Action	-0.81	0.00	-0.10	0.34
Negative Figures	0.44	0.91	-0.06	0.41
Social Control	0.82	0.01	0.58	0.03
Surveillance	0.31	0.06	0.84	0.00
CCP	1.39	0.00	1.22	0.00
Historical Events	0.29	0.08	0.22	0.04
Positive Figures	1.51	0.00	1.29	0.00

Effect on Downstream Applications

- Task: news headline sentiment classification
- Data
  - Train: 5000 headlines from TNEWS,
  - Test: (5291+, 3913=, 3424-) from Google News
    - 100 headlines containing each target word
  - Average of word embedding from three embeddings are used as features.
- Models:
  - Naïve Bayes, Support vector machine, and TextCNN.
  - Trained on three set of features from different embeddings.

Effect on Downstream Applications

- Method
  - Investigate mis-classification for each category of words
    - Models are pre-disposed to associate with more positive words will have more false-positives
    - Models are pre-disposed to associate with more negative words will have more false-negatives
  - Linear mixed effect model to mitigate noise

$$y_j = \alpha_{ij} + X_j\beta_j + \epsilon_j$$

    - where y is the difference between prediction and ground-truth, larger y implies that classification are more positive.
    - X is a dummy for models, X=1 if trained on Baidu Baike, X=0 if trained on Wikipedia

Results – Accuracy

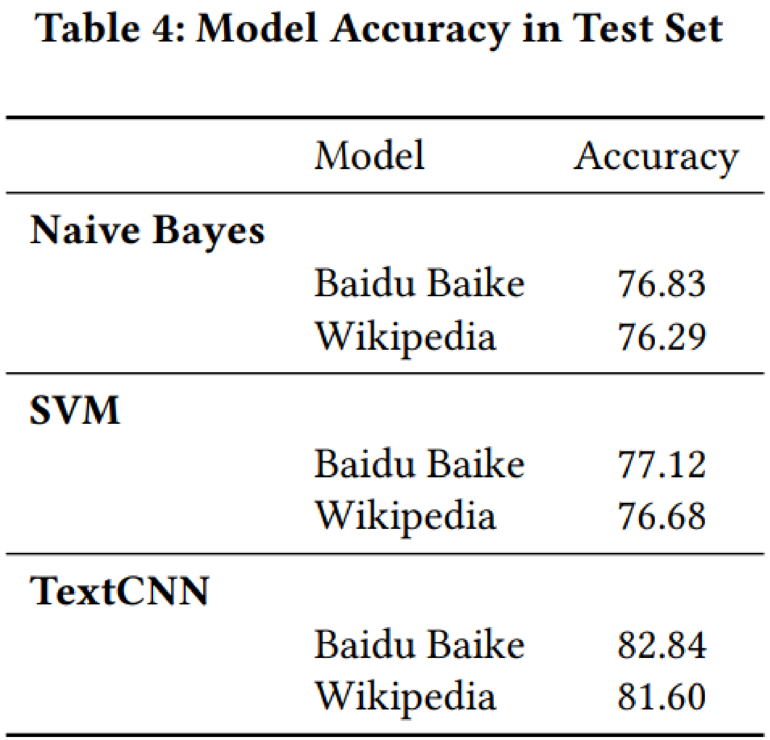


Table 4: Model Accuracy in Test Set

	Model	Accuracy
Naive Bayes		
	Baidu Baike	76.83
	Wikipedia	76.29
SVM		
	Baidu Baike	77.12
	Wikipedia	76.68
TextCNN		
	Baidu Baike	82.84
	Wikipedia	81.60

Results - Betas

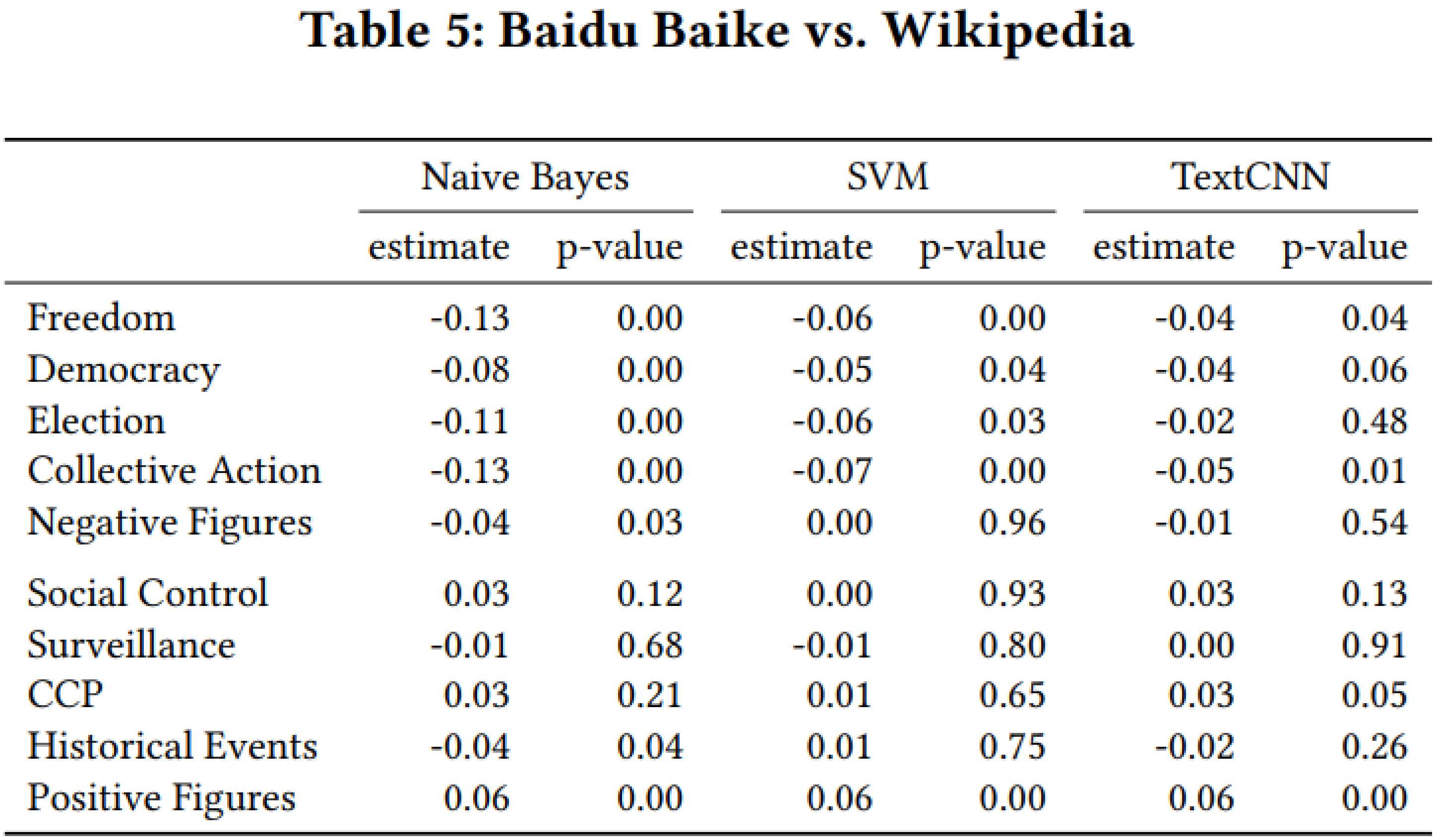
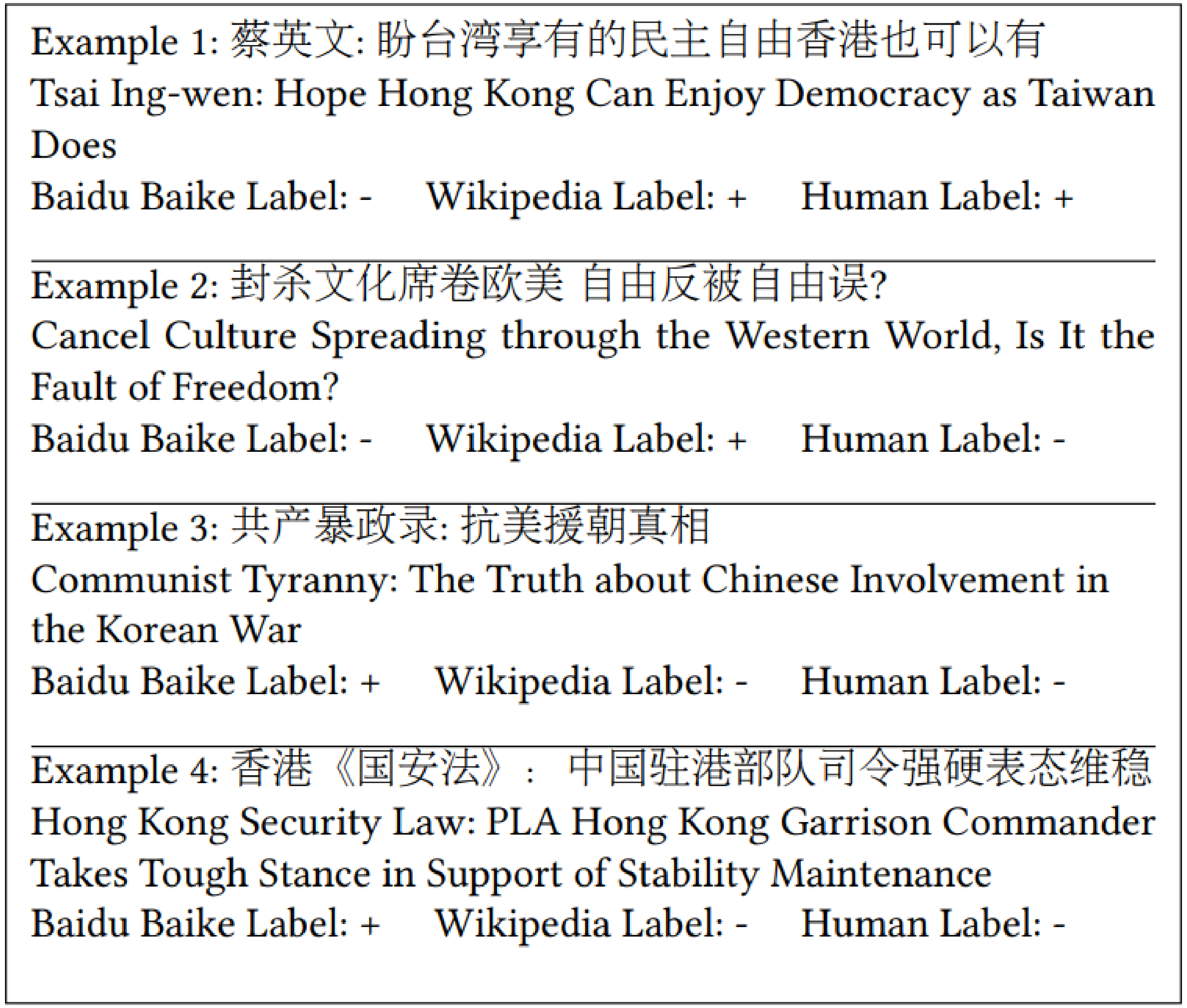


Table 5: Baidu Baike vs. Wikipedia

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.13	0.00	-0.06	0.00	-0.04	0.04
Democracy	-0.08	0.00	-0.05	0.04	-0.04	0.06
Election	-0.11	0.00	-0.06	0.03	-0.02	0.48
Collective Action	-0.13	0.00	-0.07	0.00	-0.05	0.01
Negative Figures	-0.04	0.03	0.00	0.96	-0.01	0.54
Social Control	0.03	0.12	0.00	0.93	0.03	0.13
Surveillance	-0.01	0.68	-0.01	0.80	0.00	0.91
CCP	0.03	0.21	0.01	0.65	0.03	0.05
Historical Events	-0.04	0.04	0.01	0.75	-0.02	0.26
Positive Figures	0.06	0.00	0.06	0.00	0.06	0.00

Results - Examples



Example 1: 蔡英文: 盼台湾享有的民主自由香港也可以有
Tsai Ing-wen: Hope Hong Kong Can Enjoy Democracy as Taiwan Does
Baidu Baike Label: - Wikipedia Label: + Human Label: +

Example 2: 封杀文化席卷欧美 自由反被自由误?
Cancel Culture Spreading through the Western World, Is It the Fault of Freedom?
Baidu Baike Label: - Wikipedia Label: + Human Label: -

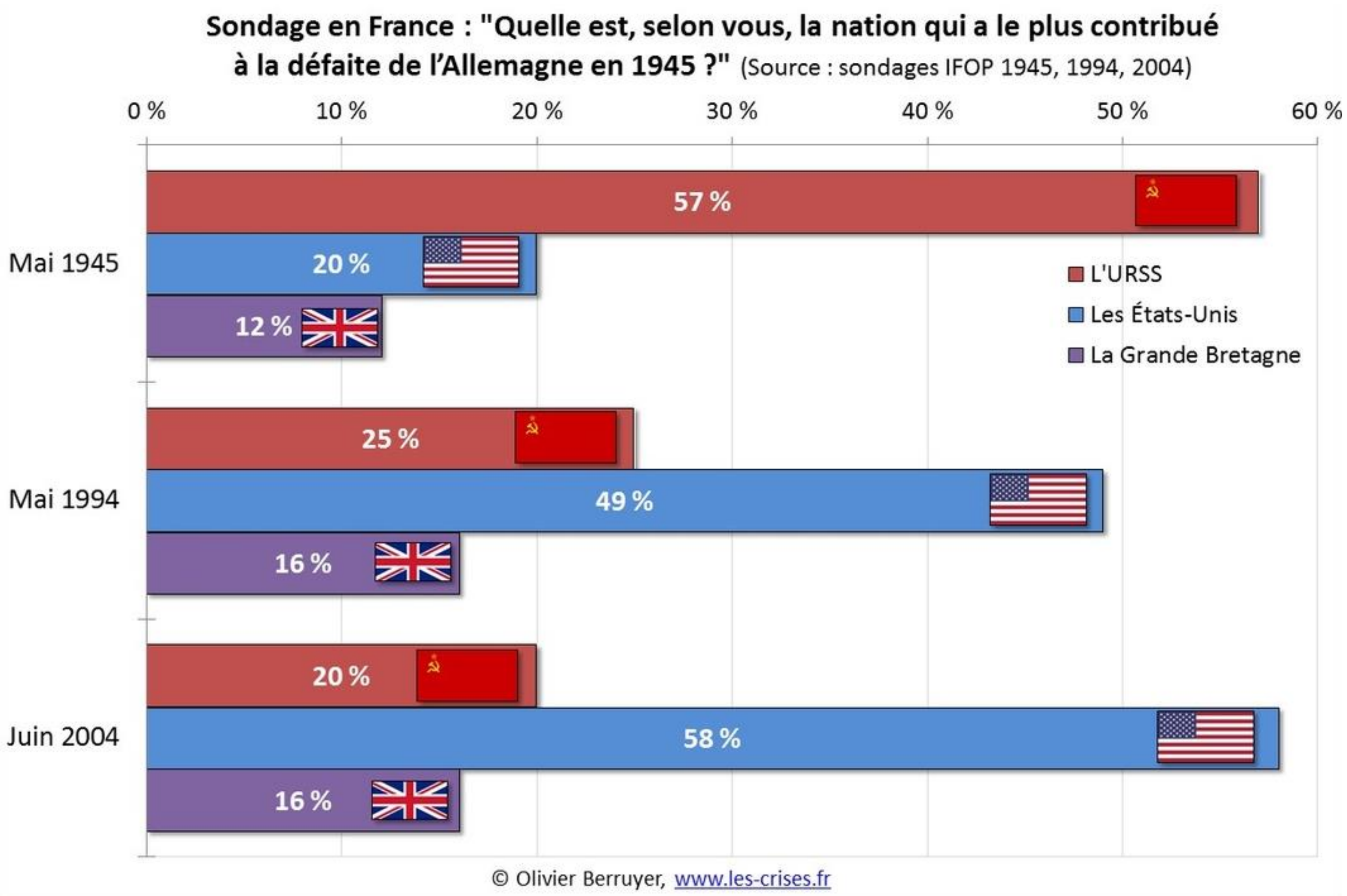
Example 3: 共产暴政录: 抗美援朝真相
Communist Tyranny: The Truth about Chinese Involvement in the Korean War
Baidu Baike Label: + Wikipedia Label: - Human Label: -

Example 4: 香港《国安法》： 中国驻港部队司令强硬表态维稳
Hong Kong Security Law: PLA Hong Kong Garrison Commander Takes Tough Stance in Support of Stability Maintenance
Baidu Baike Label: + Wikipedia Label: - Human Label: -

Remarks

- Censorship on certain topics can alter the word embeddings and further influence the downstream tasks.
- Hard to isolate the effect of censorship, though the results are suggestive that the difference is caused by censorship.
- Chinese Wikipedia used as close approximation to the counterfactual is problematic, as it also lacks contributors from mainland China.
- Some news are controversial, so demographics of annotators matters.
- Are propaganda and different ideology another type of bias to be de-biased?

Remarks



Sondage en France : "Quelle est, selon vous, la nation qui a le plus contribué à la défaite de l'Allemagne en 1945 ?" (Source : sondages IFOP 1945, 1994, 2004)

© Olivier Berruyer, www.les-crises.fr