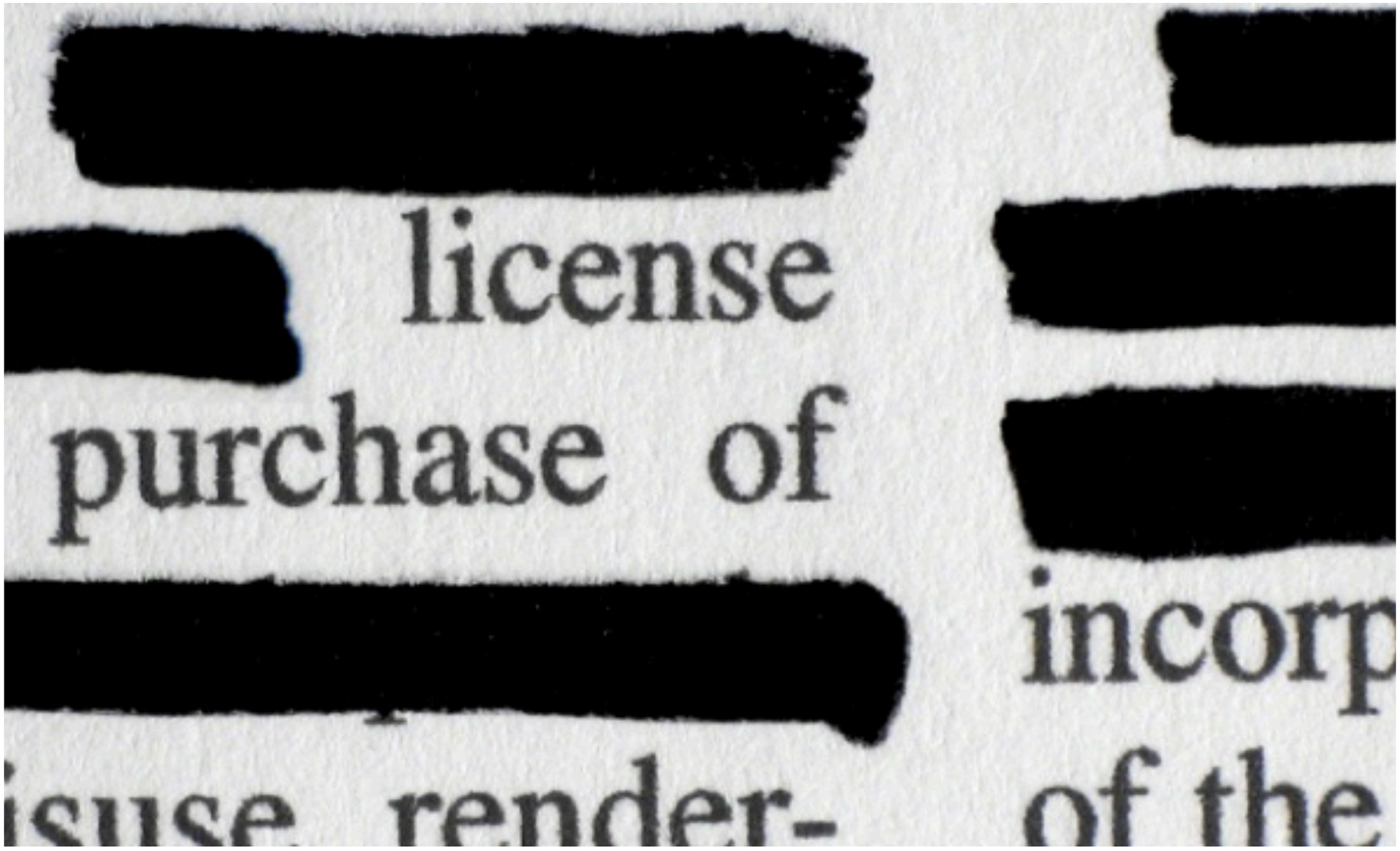

Chinese Language Under Censorship

Shifei Chen

Motivation & Target

- 64% of population’s opinion is suppressed
- China is the worst abuser
- What kind of language will be censored
- How to escape censorship

Blacklists

- Scraped from Github
- The naive approach classification -> search “法轮”
- The machine-learning approach classification -> SVM

Top 20 Keywords

Top 1–10			Top 11–20		
n	Keyword	Translation	n	Keyword	Translation
703	鸡巴	dick	621	台独	Taiwanese independence
689	法轮	Falun	620	阴唇	labia
665	李洪志	Li Hongzhi	618	真善忍	truthfulness, tolerance
640	阴道	vagina	616	疆独	Xinjiang independence
638	阴茎	penis	616	做爱	making love
635	藏独	Tibetan independence	611	口交	blowjob
633	龟头	glans	604	法轮功	Falun Gong
629	淫水	kinky	597	性交	sex
626	肛交	anal sex	596	共匪	CCP bandit
622	小穴	small hole	593	江泽民	Jiang Zemin

Table 1: Top 20 keywords as ranked by n, the number of lists each keyword appears in.

Similarity

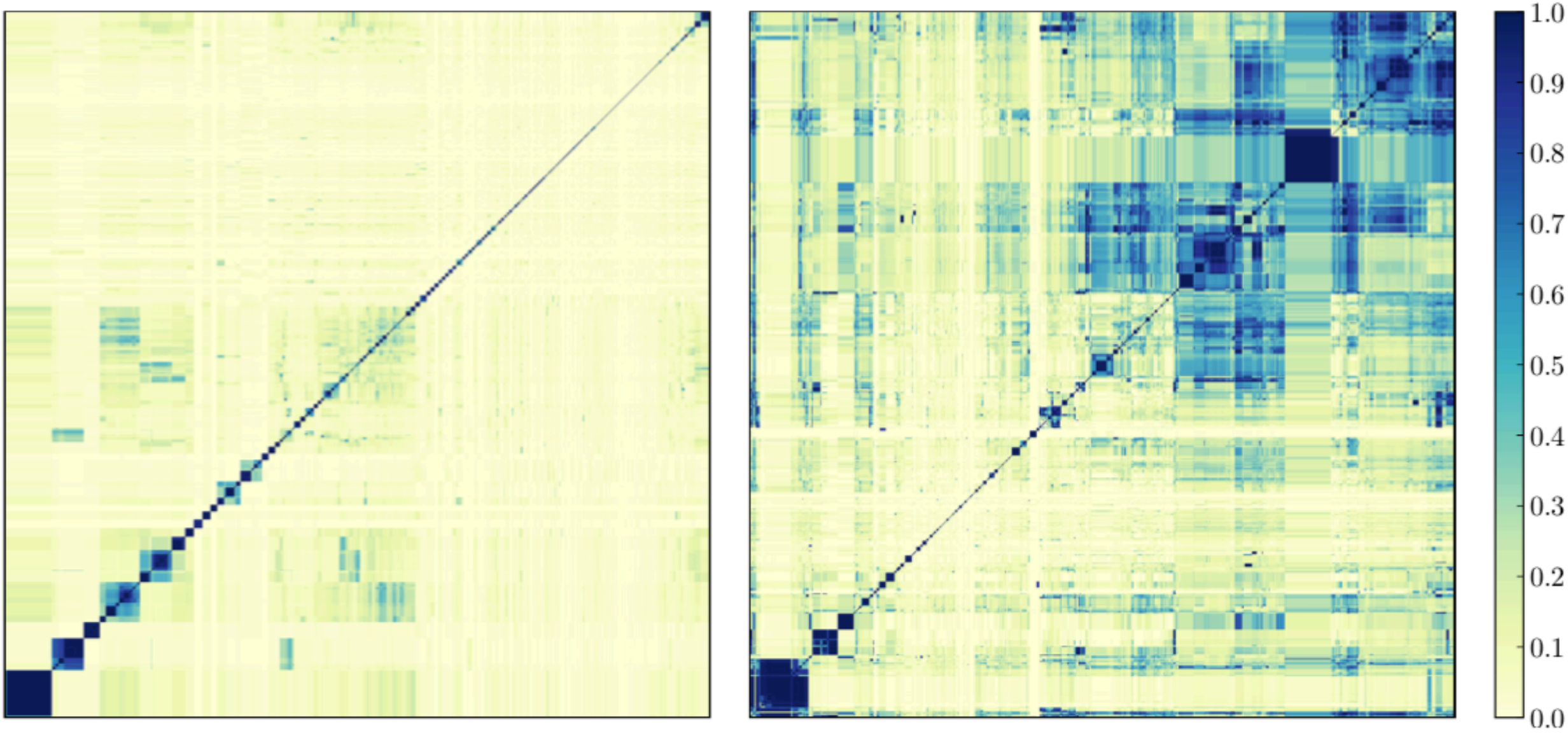

Figure 3: Left, lists clustered by Jaccard similarity; right, lists clustered by *similarity*$(x, y) = \max(\%$ of x in $y, \%$ of y in $x)$.

Beyond Blacklists

- What kind of languages are more censorable?
- Corpus from Weibo posts (divided into 4 themes)
- Extract linguistic features and analyse them

Linguistic Features

- Frequency of sensitive keywords
- Sentiment (positive or negative)
- Word frequency
- Character frequency
- Number of semantic classes
- Number of idioms
- Readability
- Word embedding and eigenfeatures

Classification

- Do you think this post was censored or not?
- Human baseline -> 63.51%, Cohen’s kappa at 0.07
- Machine-learning classification -> Naive Bayes and SVM

Results

Features		**Censored**			**Uncensored**		
	Acc	Pre	Rec	F1	Pre	Rec	F1
NB all (147)	0.65	0.76	0.65	0.70	0.53	0.65	0.58
NB eigenvalues (40)	0.57	0.69	0.57	0.62	0.43	0.57	0.50
NB ling. features (107)	0.64	0.76	0.61	0.68	0.51	0.68	0.58
NB best features(17)	0.67	0.74	0.74	0.74	0.56	0.56	0.56
SMO all (147)	0.70	0.75	0.78	0.77	0.61	0.56	0.58
SMO eigenvalues (40)	0.62	0.63	0.92	0.75	0.44	0.11	0.17
SMO ling. features (107)	0.68	0.74	0.74	0.74	0.57	0.57	0.57
SMO best features (17)	0.72	0.71	0.91	0.80	0.73	0.39	0.50
majority class	0.63						

Table 2: Classification results for the Bo Xilai subcorpus.

Results

- Bad readability -> censored
- Longer -> censored
- “They”, “fuck”, “hate”, “never”, etc. -> censored

Language Encoding

“前天起，北京正在举办一场别开生面的**老年拔河比赛**，本次比赛共有**九名选手**参加。根据现场传回的报导，目前**第一轮**的比赛还在紧张进行中，**天线宝宝**队明显略占上风，相比之下**康师傅**队明显有些不从心。”

“A few days ago, Beijing was hosting an innovative **tug-of-war** for the **elderly**; this game has **nine contestants** in all. **The first round** of the contest is still intense ...The **teletubby** team noticeably has the advantage and, relatively, the **Master Kang** team is obviously falling short. ”

System Encoding

- Cipher for encoding -> Leet/Martian Script(火星文)
- Portmanteau neologism creation -> fuse existing words to create novel ones
- Dynamic phrasebook -> Good Morning = 古德莫宁
- Poetry passwords

Summary

- NLP studies confirms previous discoveries
- System encoding could be a potential field
- Language is evolving, so does censorship