Detecting Censorable Content on Sina Weibo: A Pilot Study

Kei Yin Ng
Montclair State University
Montclair, NJ, USA
ngk2@montclair.edu

Anna Feldman
Montclair State University
Montclair, NJ, USA
feldmana@montclair.edu

Chris Leberknight
Montclair State University
Montclair, NJ, USA
leberknightc@montclair.edu

ABSTRACT

This study provides preliminary insights into the linguistic features that contribute to Internet censorship in mainland China. We collected a corpus of 344 censored and uncensored microblog posts that were published on Sina Weibo and built a Naive Bayes classifier based on the linguistic, topic-independent, features. The classifier achieves a 79.34% accuracy in predicting whether a blog post would be censored on Sina Weibo.

KEYWORDS

Chinese social media, censorship detection

1 INTRODUCTION

As the Internet continues to be an integral part of people's lives, more than half the world's Internet users are still having restricted access to information on the World Wide Web due to censorship. Our study takes a closer look at the censorship activities in mainland China, with a particular focus on one of its mainstream social media platforms – Sina Weibo. Internet censorship in mainland China consists of several layers: restricted access to certain websites, restricted access to certain search results, and removal of some information published by Internet users. Since censorship on social media typically happens after a user has successfully published on the platform, what gets censored or not is largely a "real-time" decision due to the unpredictable nature of published content. Discussions on sensitive topics do not always get censored, as evidenced by their accessibility on the platform. This study investigates the factors that contribute to censorship on Sina Weibo from a linguistic perspective. We locate sources that provide censored and uncensored Weibo texts, extract linguistic features from the corpus we collected and build a classifier that predicts censorship independent of discussion topics. It is hoped that by gaining insight into the linguistic features that contributes to censorship, social media users can better maneuver their text content to circumvent censorship.

2 RELATED WORK

Internet censorship has dramatically increased over the past five years and reports suggest that more than half of the world's Internet users live in a country where the Internet is censored or restricted [1]. While there have been significant efforts to develop strategies and technologies for evading censorship ([8]) , results from existing research fall short in delivering a number of essential aspects of online censorship. Many measurement and circumvention studies focus more on exploiting technological limitations with existing routing protocols ([5, 10, 11, 17]). While there have been many significant contributions in this area, little attention has focused on linguistically-inspired techniques to study online censorship. One notable exception applies linguistic steganography to obfuscate censored content [16]. Their results focus purely on circumvention while this research takes a linguistic approach to detect censorable content and thereby provides insight into evading censorship "linguistically".

In recent years, several detection mechanisms have been proposed to observe and categorize the type of content and keywords that are censored ([7, 18]). [6] analyze the content of censored and uncensored texts from various Chinese social media sources to study the relationship between criticism of the state and chance of censorship. Their main findings suggest that negative comments about the state did not always lead to censorship. Rather, the presence of Collective Action Potential (the potential to cause collective action in real life) in a text is what rendered a post susceptible to censorship. [9] explored the effectiveness of linguistic tactics in circumventing online censorship in China and argued that using parodic satire could most likely survive censorship because it relies heavily on users' and censors' ability to detect sensitive topics based on context. Similar to [9], the empirical study conducted by [2] discovered linguistically-informed ways to better survive the Chinese censorship. Their findings showed that the use of homophones of censored keywords on Sina Weibo could help extend the time a Weibo post could remain available online. The findings of [2, 6, 9] are all based on a significant amount of human effort to interpret and annotate texts to evaluate the likeliness of censorship. However, in case of change in censorship approach or direction imposed by the authority, time- and labor-intensive methods might not be eficient enough to inform Internet users the latest censorship environment.

Our work is similar to [15] on automatic fake news detection, we are interested in devising ways to automatically predict censorship by examining what a machine can learn from the linguistic signals embedded in existing data and how well the learning can generalize to out-of-sample data. We see fake news and censorable content detection as two related tasks: to inform Internet users ways to strive for a free, open, and reliable online information exchange platform.

SETN '18, July 9–15, 2018, Rio Patras, Greece

ACM ISBN 978-1-4503-6433-1/18/07…$15.00
https://doi.org/10.1145/3200947.3201037

3 DATASET

To investigate linguistic features that might contribute to censorship, we need a corpus that consists of both censored and uncensored texts in order to analyze and compare the linguistic signals embedded in each category. Since our goal is to build a classifier that predicts censorship regardless of topics, a set of censored and uncensored texts needs to be relevant to the same topic in order to offset the effects of topics on censorship.

Our corpus contains censored and uncensored posts about Bo Xilai, a former Communist Party chief in Chongqing, China. In 2013, Bo Xilai was a likely candidate for promotion to the elite Politburo Standing Committee. However, he was found guilty of corruption and was expelled from the Communist Party, parliament and faced prosecution. He was sentenced to life imprisonment. His wife was given a suspended death sentence for the murder of a British businessman, Neil Heywood, in 2012. The Bo Xilai incident has spurred a lot of discussion on misconduct of Chinese political leaders and the Party. Only posts published between January 1, 2015 to January 1, 2018 were collected.

3.1 Uncensored Data

Sina Weibo is regarded as one of the most popular social media platforms in mainland China. It functions similarly as Twitter where users can publish, reblog and repost opinions and news on any topics. Although users can publish freely on Weibo, the published content is subject to scrutiny and would possibly be censored or deleted if it is considered to be violating Weibo's policies. Content that can be found on Sina Weibo has already passed the censorship mechanisms and is regarded as uncensored. We collect 152 uncensored posts from Weibo using the search keyword 薄熙来 (Bo Xilai).

3.2 Censored Data

FreeWeibo (https://freeweibo.com/) is a website dedicated to making censored Weibo posts available to the public. We collect 192 censored posts from FreeWeibo on the same topic published between January 1, 2015 to March 1, 2016. This date range is shorter than that of the uncensored data since the number of available uncensored posts is significantly less than that of censored. Therefore, a longer date range of uncensored posts is used in order to create a more balanced corpus. Below are some examples of censored and uncensored posts with their English translations:

Censored: 薄熙来是今天的高岗，周永康是今天的康生，两位都是为 党和国家立下汗马功劳的人，也都是被政治斗争构陷的人。

Bo Xilai is today's Gao Gang. Zhou Yongkang is today's Kang Sheng. They both made great contributions to the Party and the country. They both got framed by political infighting.

Censored: 薄熙来案件是审查过的司法机关人员可以旁听，不能电 视直播，微博截取直播。还是留下痕迹。 The observers of Bo Xilai scandal are judicial authorities who passed investigations. There is no live broadcast on TV or Weibo. This is still flawed.

Censored: 刚有朋友微信上问大伙，当年薄熙来在重庆办公室打了公 安局长王立军一个耳光，这算不算袭警？如果当时王立军开枪，算不 算是合理用枪？你我该如何回答？

My friend just asked some questions on Weibo: When Bo Xilai slapped Wang Lijun in the face in the Chongqing office, should he be accused of assaulting a police officer? If Wang Lijun shot him, would that be a legitimate use of gun? How should we answer these questions?

Uncensored: 薄熙来始终不认罪，戴械具，老周认罪，不戴，好看 点。

Bo Xilai still isn't pleading guilty. And he's cuffed. Old Zhao pleaded guilty, and he's not cuffed. That looked better.

Uncensored: 能不再挣扎心平气和的接受指控承认罪行，可是悔悟不 等人。从2014年7月到现在，短短一年的时间，头发白的这么快。想 起薄熙来在最后的宣判面前，安静和蔼的样子特别慈祥。可是，做不 好自己本分的事、管理不好自己的行为、自控能力差、不知足，就应 当为之付出相应的代价。所以，学做人是一辈子的事。

[He] calmly accepted the charges and pleaded guilty. But regrets wait for no one. Since July 2014 till now, within only a year, his hair has turned gray already. Bo Xilai looked quiet, gentle and peaceful when he was waiting for the verdict. However, failing to live up to one's responsibility, failing to manage one's behavior, poor self-discipline and not contented, are what make him liable to the price. Therefore, learning how to live wisely is a lifetime journey.

Uncensored: 怪不得第一次听薄熙来名字的时候感觉熟悉又陌生原来 是希伯来人作怪

No wonder why I felt both familiar and unfamiliar when I first heard the name "Bo Xilai" – the Hebrews were behind it.

3.3 Data Preprocessing

Name of author, friend tags and reblogged content were removed from all data. Hashtags were preserved as they might provide useful information. All non-textual information such as images and videos were also discarded. Since the Chinese language does not have word boundaries, word segmentation has to be carried out before certain linguistic features can be extracted. The word segmenter by Aihanyu Corpus (http://www.aihanyu.org/cncorpus/index.aspx) was used to segment all the data.

4 LINGUISTIC FEATURES

4.1 LIWC

LIWC ([13, 14]) is a program built on dominant theories in psychology, business, and medicine. It analyzes text on a word-by-word basis, calculating percentage of words that match certain language dimensions such as psychological processes (affect, drives etc.) and linguistic processes (adverbs, prepositions etc.). The idea is that certain words are strong indicators of people's emotional and cognitive worlds. Each word can belong to more than one category.

We used the Chinese LIWC to extract the frequency of word categories. The Chinese LIWC dictionary was developed by [3]. It was built by first translating from the English LIWC, and then further developed and modified to accommodate the linguistic differences between English and Chinese. Therefore, the Chinese Dictionary contains some categories that are not included in the English Dictionary, such as prepEnd (words that are appended to the end of other words), quanUnit (classifier/quantifier) and so on. The frequency score of each category is used as one feature.

4.2 Sensitive Keywords

We collected a list of keywords that are regarded as sensitive in mainland China and counted the frequency of keywords in each piece of data. The first source is a list of blacklisted keywords provided by Wikipedia (https://en.wikipedia.org/wiki/List_

of_blacklisted_keywords_in_China) and the second source is a list of sensitive Sina Weibo search terms provided by China Digital Times (https://chinadigitaltimes.net/china/sensitive-words-series/). Since accessibility of searches change from time to time, China Digital Times tested the "searchability" of each keyword and recorded the date of testing for reference. We collected keywords that were tested between January 1, 2015 to March 1, 2016, a period that overlapped with our data. A total of 598 keywords were collected from the two sources. Sensitive keywords were found in 31 out of 152 uncensored data and 60 out of 192 censored data. We implemented this feature as a relative frequency (normalized by the total number of word tokens in a post).

4.3 Sentiment

We used two Chinese sentiment analyzers to investigate sentiment presented in the data.

4.4 BosonNLP

For each piece of data, BosonNLP (https://bosonnlp.com/) provides a percentage score of non-negative sentiment, and another percentage score of negative sentiment. The two scores have a sum of 1. We selected two different training models for analysis – the General model and the Weibo model. Each model has been trained with different corpus. The General model applies to general Chinese texts and the Weibo model is trained with texts commonly found on Weibo. The average negative sentiment percentage obtained for censored texts are 53.9% for General model and 46.0% for Weibo model. The same percentage scores obtained for uncensored texts are 49.3% and 36.6% respectively.

4.5 BaiduAI

BaiduAI https://ai.baidu.com/ was used to obtain another set of sentiment scores. It provides a positive sentiment percentage score and a negative sentiment percentage score for each post, which also sum up to 1. The average negative percentage are 64.9% and 56.3% for censored and uncensored texts respectively.

4.6 Ngrams

We extract unigrams, bigrams and trigrams of words for each blog post. Before extracting ngrams, we apply a list of stop words that contains some single-character function words and punctuation. We included all the ngrams whose raw frequency was greater than 3. To account for the differences in blog post lengths, the ngrams are normalized.

5 AUTOMATIC EXPERIMENTS

We extract a total of 408 features as described above. We build Naive Bayes classifiers [4] with various feature combinations and evaluate each performance with 10-fold cross-validation. Table 1 summarizes the results. All Punctuations and OtherP are both LIWC features. All Punctuations is the overall count of 10 common punctuations such as question mark, quotes, comma etc. plus a group of less common punctuations (OtherP) such as ellipsis and percent sign. The LIWC Summary includes WC (word count), WPS (average sentence length in words), Dic (percent of target words captured by the LIWC dictionary) and Other Grammar (verb, adjectives, quantifiers etc). The best 17 features were selected with the standard Information Gain feature selection algorithm [12] and were also evaluated with 10-fold cross-validation.

Table 1: Classification results for the Bo Xilai dataset

Feature Combination (# of features)		Censored			Uncensored		
	Acc	Pre	Rec	F1	Pre	Rec	F1
OtherP (1)	0.55	0.57	0.92	0.70	0.32	0.05	0.09
All Punctuation (11)	0.66	0.66	0.85	0.74	0.66	0.40	0.50
Sensitive keywords (1)	0.53	0.71	0.31	0.43	0.47	0.82	0.60
AllPunc+Sensitive keywords (12)	0.65	0.67	0.76	0.71	0.60	0.50	0.55
LIWC Summary (4)	0.69	0.65	1.00	0.79	0.98	0.28	0.43
LIWC Psychological Processes (49)	0.62	0.77	0.50	0.60	0.54	0.79	0.64
LIWC Linguistic Processes (30)	0.60	0.72	0.49	0.58	0.52	0.75	0.61
LIWC all (95)	0.68	0.76	0.64	0.70	0.60	0.72	0.65
LIWC all+keywords (97)	0.67	0.76	0.64	0.69	0.59	0.72	0.65
Baidu Sentiment (3)	0.57	0.61	0.72	0.66	0.50	0.38	0.43
Boson Sentiment (4)	0.57	0.63	0.60	0.62	0.49	0.51	0.50
Baidu+Boson (7)	0.57	0.63	0.63	0.63	0.49	0.49	0.49
ngrams (299)	0.51	0.60	0.42	0.50	0.44	0.62	0.52
All features (408)	0.69	0.76	0.69	0.72	0.62	0.70	0.66
LIWC+Baidu+Boson+keywords (102)	0.72	0.79	0.69	0.74	0.65	0.75	0.70
Best features (17)	0.79	0.80	0.85	0.83	0.78	0.71	0.75

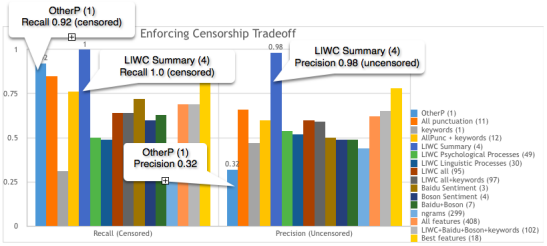
The best 17 features selected by the IG algorithm are listed below along with the Domain each of them belong to.

(1) feel (words related to physical feeling and touch) Perceptual
(2) informal (informal language markers (assents, fillers, swear words, netspeak) Informal Language
(3) prepEnd (postpositions, e.g., words expressing time and space e.g. 中, 为止, 之外) Function Words
(4) (dictionary words, i.e., percent of words captured by the LIWC dictionary) General
(5) nonflu (nonfluencies) Informal Language
(6) particle (function word associated with another word or phrase to impart meaning.) Function Words
(7) assent (words like agree, OK, yes) Informal Language
(8) see (words like view, see, seen) Perceptual
(9) body (words describing human body, e.g., cheek, hands, spit) Biological
(10) modal-pa (Modal particles) used at the end of sentences to indicate mood, or attitude. e.g. 了, 吗, 吧 Function Words
(11) general-pa (general particle) particle that is not modal e.g. 地, 得, 来着 Function Words
(12) AllPunc (all punctuation) General
(13) ppron (personal pronouns) Function Words
(14) quanunit (Chinese-specific quantity unit, e.g., quantifier 位, 元, 名) Function Words
(15) swear (swear words) Informal Language
(16) WC (Word count) General
(17) WPS (average sentence length) General

6 DISCUSSION

We found it interesting that neither sensitive keywords nor sentiment analysis turned out to be strong indicators of censorship. To get a better understanding of the language differences between censored and uncensored text (which feature is more associated

Figure 1: Precision and recall tradeoff (best features)



with which class), for each of the 17 best feature, we subtract the average value of all censored posts from their corresponding uncensored values. A positive results indicates an association between a feature and uncensored class, whereas a negative results indicate an association between a feature and the censored class. The results are shown in Figure 2. All the differences shown in the graphs are statistically significant (two tailed unpaired t-test, $p<0.5$).

Features that are categorized by LIWC as Function Words, Informal Language, and Feel (Perceptual) seem to be the best indicators of censored content. While we cannot make any strong claims at this point, it seems that censored language contains more words that are richer in semantic variety and are more informal in nature. This is evidenced by its strong association with the Dic feature which represents the various semantic classes and dimensions of the LIWC dictionary. The use of informal language (informal, nonflu, swear) is also more associated with censored texts. The strong association with particles (which are frequently used to indicate mood (modal-pa) and relationship (general-pa) in Chinese) might entail some characteristics of censored language. Since Chinese general particles are typically used to specify relationship, it suggests that censored language might tend to mark relationships among people, matter etc. Modal particles are usually used to signify speaker's mood, attitude and tone. This means that censored language might tend to be more subjective. We hypothesize that sentiment does not contribute to the classification because it indicates only positive and negative, but not the intensity of opinions. We plan to address this issue in our future work. We also notice that words that are related to the sense see are indicators of the uncensored content, while words related to feel are associated with the censored content. Upon a closer look at the data and the words that fall under each category, we can see that for feel, many words are used as metaphors to express psychological feelings despite the fact that their literal sense refers to physical feelings. As for see, many words are used in their literal sense in the uncensored data e.g. showcase, depict, discover etc. They tend to be more objective in reporting the Bo Xilai incident, instead of being very opinionated on it.

Our findings go along with [6]'s Collective Action Potential (CAP) theory, which states that it is not controversial content that gets censored, but content that has CAP. In our case, sensitive keywords (the indicators of controversy) were not strong indicators of censorship.

We realize that we are working with a relatively small dataset and therefore, our results are preliminary. To get an idea of whether our

Figure 2: Language differences of censored and uncensored posts.

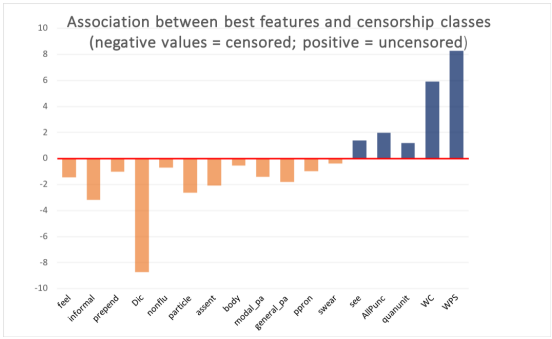
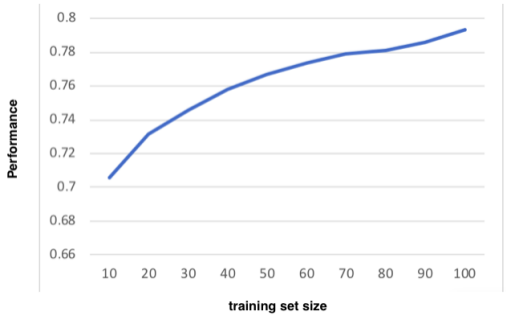


Figure 3: The relationship between training set size and best classifier performance (with 10-fold cross-validation)



classifier can improve with the increase of the training examples, we partitioned our entire dataset into 10%, 20% etc. and evaluate the performance of the best classifier on each partition. The learning curve depicted in Figure 3 suggests that if we collect a larger corpus, the classifier's performance will likely increase. We are in the process of experimenting with other classifiers, expanding the corpus to include more topics, and also extracting more linguistic features.

We also notice the importance of the trade-off between precision and recall for our task. Figure 1 illustrates which feature provides the highest recall and precision for censored and uncensored posts. Eventually we want to make it more dificult for the censors to detect content that may be "deemed inappropriate". We therefore want our model to have good coverage, which suggests improving the recall. We will explore this matter in future work.

7 CONCLUSION

In this paper we described a pilot study in which we built a model to classify censored and uncensored social media posts from mainland China. Our corpus deliberately contains only one topic, the Bo Xilai scandal. Our goal was to explore whether linguistic features could be effective in distinguishing censored and uncensored content. Our study suggests that subjective information, such as expressions of mood and feeling, and informal language can likely be indicators of censorable content.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1704113.

REFERENCES

[1] Sam Burnett and Nick Feamster. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. ACM SIGCOMM Computer Communication Review (2015).
[2] Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. 2015. Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions. In Ninth International AAAI Conference on Web and Social Media.
[3] Chin-Lan Huang, Cindy Chung, Natalie K. Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben C.P. Lam, Wei-Chuan Chen, Michael Bond, and James H. Pennebaker. 2012. The development of the Chinese linguistic inquiry and word count dictionary. Chinese Journal of Psychology 54, 2 (2012), 185–201.
[4] George H. John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, 338–345.
[5] S. Katti, D. Katabi, and K. Puchala. 2005. Slicing the onion: Anonymous routing without pki. Technical Report. MIT CSAIL Technical Report 1000.
[6] Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review 107, 2 (May 2013), 1–18.
[7] J. Knockel, M. Crete-Nishihata, J.Q. Ng, A. Senft, and J.R. Crandall. 2015. Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China. In Proceedings of the 5th USENIX Workshop on Free and Open Communications on the Internet.
[8] Christopher S. Leberknight, Mung Chiang, and Felix Ming Fai Wong. 2012. A Taxonomy of Censors and Anti-Censors: Part I-Impacts of Internet Censorship. International Journal of E-Politics (IJEP) 3, 2 (2012).
[9] S. Lee. 2016. Surviving Online Censorship in China: Three Satirical Tactics and their Impact. China Quarterly (2016).
[10] D. Levin, Y. Lee, L.Valenta, Z. Li amd V. Lai, C. Lumezanu, N. Spring, and B. Bhattacharjee. 2015. Alibi Routing. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication.
[11] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. [n. d.]. Defeating Image Obfuscation with Deep Learning. arXiv preprint arXiv:1609.00408. ([n. d.]).
[12] H. Peng, F. Long, and C. Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. EEE Transactions Pattern Analysis and Machine Intelligence 27, 8 (2005), 1226–1238.
[13] James W. Pennebaker, Roger Booth, and M.E. Francis. [n. d.]. Linguistic Inquiry and Word Count (LIWC2007).
[14] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric The Development of Psychometric Properties of LIWC. Technical Report. University of Texas at Austin.
[15] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. arXiv preprint arXiv:1708.07104. (2017).
[16] Iris Safaka, Christina Fragouli, , and Katerina Argyraki. 2016. Matryoshka: Hiding secret communication in plain sight. In 6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16). USENIX Association.
[17] Zachary Weinberg, Jeffrey Wang, Vinod Yegneswaran, Linda Briesemeister, Steven Cheung, Frank Wang, and Dan Boneh. 2012. StegoTorus: A Camouflage Proxy for the Tor Anonymity System. Proceedings of the 19th ACM conference on Computer and Communications Security (2012).
[18] T. Zhu, D. Phipps, A. Pridgen, JR Crandall, and DS Wallach. 2013. The velocity of censorship: high-fidelity detection of microblog post deletions. arXiv:1303.0597 [cs.CY]. (2013).